STATISTICAL MODELS OF INCOMPLETE DATA AND THEIR USE IN THE SOCIAL SCIENCES

Ulrich Pötter

Habilitationsschrift Bochum 2008
Preface

The present document has been submitted as the author’s Habilitationsschrift to the faculty of Social Science at the Ruhr University Bochum. It is concerned with the problem of incomplete data in the social sciences.

While traditionally the focus of statistical treatments of the problem was confined to completely missing data, the extensions to partially missing values were pursued systematically only since about 15 years. This approach has opened the way to unify concepts and models from otherwise quite divergent fields. Some problems of model building and estimation within the general approach are documented in the following pages with particular emphasis on social science applications.

The document is based on six articles that appeared either in refereed journals or as refereed contributions to conference proceedings. The six articles are:


- A non-parametric mean residual life estimator, Metodoloski Zvezki, 19, 2003, 97–113 (with K. Kopperschmidt)

- Covariate effects in periodic hazard rate models, Metodoloski Zvezki, 17, 2002, 137–145 (with K. Kopperschmidt)


These articles are slightly revised to provide for a more unified notation. However, I have tried not to change the style or the arguments originally presented. All articles are followed by a new postscriptum highlighting later developments and putting them in the context of the general discussion of incomplete data.

The articles are preceded by two new chapters. The first chapter systematically introduces the theory of incomplete data. It surveys the probabilistic theory and highlights those features that are of critical importance for social science applications. The main advantages of a general formulation of the problem of incomplete data are presented and some alternative formulations are discussed.

A second new chapter presents a case study that serves both to illustrate the basic concepts and to emphasise those aspects of the theory that are of particular importance for social science applications. Also added is an appendix that treats some implementation problems of one of the central building blocks of the treatment of incomplete data, the estimation of distribution functions from incomplete data.
Contents

Preface i

1. Introduction 1

2. Probabilistic Models of Incomplete Data 17
   2.1. The Structure of Incomplete Data ............... 18
       2.1.1. A Typology of Missing Data Structures ... 22
       2.1.2. Graphical Representations of Coarsened Data . 25
       2.1.3. Consistent Distributions .................. 28
       2.1.4. The Geometry of Consistent Distributions ... 33
       2.1.5. Constructing Consistent Distributions ...... 38
       2.1.6. Selectors and Allocations .................. 43
   2.2. Coarsening at Random ........................... 50
       2.2.1. Does CAR Apply to Survey Non-response? ... 54
       2.2.2. CAR in Hierarchically Structured Models ... 62
       2.2.3. Non-Hierarchical CAR ....................... 66
       2.2.4. CAR is Everything ........................... 71
       2.2.5. Uniqueness of CAR ........................... 73
       2.2.6. Further CAR Formulations ................... 81
       2.2.7. Strong CAR? ................................ 91
   2.3. Coarsening Variables and Coarsening Completely at Random ................. 97
       2.3.1. Coarsening Variables ......................... 98
       2.3.2. Completely Coarsening at Random ............ 113
       2.3.3. Additional Data ............................. 115
   2.4. Conclusions: CAR Modelling ....................... 121

3. A Case Study: Parent's Length of Life 123
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1. The Structure of Descent</td>
<td>124</td>
</tr>
<tr>
<td>3.2. Sampling from the Children Generation</td>
<td>128</td>
</tr>
<tr>
<td>3.3. Induced Sampling from the Parent Population</td>
<td>129</td>
</tr>
<tr>
<td>3.4. Assumptions about Independence</td>
<td>136</td>
</tr>
<tr>
<td>3.5. The “Simple Device”</td>
<td>148</td>
</tr>
<tr>
<td>3.6. Left Truncated Data</td>
<td>154</td>
</tr>
<tr>
<td>3.7. Selection by Survival</td>
<td>162</td>
</tr>
<tr>
<td>3.7.1. The Simulation Model</td>
<td>162</td>
</tr>
<tr>
<td>3.7.2. Considering Left Truncation</td>
<td>170</td>
</tr>
<tr>
<td>3.7.3. Using Information from Children</td>
<td>176</td>
</tr>
<tr>
<td>3.7.4. Retrospective Surveys</td>
<td>181</td>
</tr>
<tr>
<td>3.8. Inferences from the GLHS and SOEP Data</td>
<td>183</td>
</tr>
<tr>
<td>3.8.1. Description of the Data</td>
<td>185</td>
</tr>
<tr>
<td>3.8.2. Survivor Functions of Parents</td>
<td>187</td>
</tr>
<tr>
<td>3.8.3. Visualisation of Death Rates</td>
<td>193</td>
</tr>
<tr>
<td>3.9. Conclusion</td>
<td>196</td>
</tr>
<tr>
<td>4. Probabilistische Selektionsmodelle</td>
<td>199</td>
</tr>
<tr>
<td>4.1. Summary</td>
<td>199</td>
</tr>
<tr>
<td>4.2. Zusammenfassung</td>
<td>200</td>
</tr>
<tr>
<td>4.3. Einleitung</td>
<td>201</td>
</tr>
<tr>
<td>4.4. Selektionsmodelle in den Sozialwissenschaften</td>
<td>204</td>
</tr>
<tr>
<td>4.5. Beispiel: Einkommensangaben im ALLBUS</td>
<td>210</td>
</tr>
<tr>
<td>4.6. Stichproben und probabilistische Auswahlmodelle</td>
<td>215</td>
</tr>
<tr>
<td>4.7. Ignorierbare Ausfälle: Parametrische und nichtparametrische Modelle</td>
<td>221</td>
</tr>
<tr>
<td>4.8. Ignorierbare Ausfälle: MAR, CAR and all das</td>
<td>224</td>
</tr>
<tr>
<td>4.9. Ignorierbare Ausfälle: Konditionale CAR-Modelle</td>
<td>230</td>
</tr>
<tr>
<td>4.10. Nicht ignorierbare Ausfälle</td>
<td>235</td>
</tr>
<tr>
<td>4.11. Sensitivität</td>
<td>238</td>
</tr>
<tr>
<td>4.12. Diskussion</td>
<td>240</td>
</tr>
<tr>
<td>4.13. Postscriptum</td>
<td>241</td>
</tr>
<tr>
<td>5. Causal Inference from Series of Events</td>
<td>245</td>
</tr>
<tr>
<td>5.1. Summary</td>
<td>245</td>
</tr>
<tr>
<td>5.2. The Tradition of Causal Analysis in Sociology</td>
<td>246</td>
</tr>
<tr>
<td>5.3. Temporal and Probabilistic Criteria of Causation</td>
<td>247</td>
</tr>
</tbody>
</table>
5.4. Mathematical Models ...................................... 251
5.5. Statistical Methods .................................... 256
5.6. Events and their Descriptions ......................... 256
5.7. Agents, Actions, and Events ......................... 259
5.8. Conceptual Problems ....................................
   5.8.1. Local Independence ............................... 262
   5.8.2. Autonomy ........................................ 264
5.9. Conclusions ............................................. 266
5.10. Postscriptum ............................................ 266

6. A Non-parametric Mean Residual Life Estimator ........ 275
   6.1. Summary ............................................. 275
   6.2. Introduction ....................................... 276
   6.3. An Example from Market Research .................. 276
   6.4. Mean Residual Life Function ....................... 280
   6.5. Kaplan–Meier Integrals ............................. 282
   6.6. Variance Formulae .................................. 283
   6.7. A Small Simulation Study ............................ 285
   6.8. Appendix ............................................. 293
   6.9. Postscriptum ........................................ 298

7. Covariate Effects in Periodic Hazard Rate Models ...... 301
   7.1. Summary ............................................. 301
   7.2. Introduction ....................................... 302
   7.3. Marginal and Conditional Hazard Rates ............. 304
   7.4. Hazard Envelopes ................................... 306
   7.5. Consequences for Model Choice .................... 309
   7.6. Postscriptum ........................................ 311

8. A Multivariate Buckley-James Estimator ................. 313
   8.1. Summary ............................................. 313
   8.2. Introduction ....................................... 313
   8.3. Buckley-James Estimators ......................... 314
   8.4. Score Functions and Censoring ..................... 318
   8.5. Multivariate Extensions ............................ 320
   8.6. The Missing Information Principle .................. 322
   8.7. The Multivariate Buckley-James Estimator ......... 329
Since 30 years or more, the literature on incomplete data is dominated by approaches that make essential use of statistical models and their underlying probabilistic desiderata: random variables, distributions, stochastic independence etc. In their famous book “Statistical Analysis with Missing Data”, Roderick Little and Donald Rubin state:

Missing data mechanisms are crucial since the properties of missing-data methods depend very strongly on the nature of the dependencies in these mechanisms. The crucial role of the mechanism in the analysis of data with missing values was largely ignored until the concept was formalized in the theory of Rubin (1976), through the simple device of treating the missing-data indicators as random variables and assigning them a distribution. (Little, Rubin 2002: 11)

This “simple device” is now nearly universally adopted both in theoretical research and in applied work, and particularly in empirical social research. This document is no exception and in effect most chapters that follow use the “simple device” without further justification.

However, this “simple device” is neither the most natural nor the only one available to the empirically working social scientist. Nor was the “simple device” accepted without resistance. In fact, it has often been doubted that a statistical treatment of the problem is at all possible, at least in the social sciences. Oskar Morgenstern, in his early study “On
the Accuracy of Economic Observations”, said that “the errors in these
data cannot always be formulated according to strict statistical theory
for the simple reason that no such exhaustive theory is available for
many social phenomena” (1963: 7). Some 20 years later, Tore Dalenius
commented on the statistical treatment of survey non-response:

I take a dim view of the usefulness of these endeavors on
two grounds. (1) First, it appears utterly unrealistic to pos-
tulate “response probabilities” which are independent of
the varying circumstances under which an effort is made to
elicit a response. ...(2) …it seems unavoidable to introduce
assumptions of unknown validity about probabilities. In
summary, I am inclined to reject approaches to the non-
response problem which involve “response probabilities”

While the resistance against statistical treatments of incomplete data has
dwindled away, this may only partly be due to the successes of the the
“simple device” of Little and Rubin. It may as well be due to the rather
naive hope for a “method” that could somehow amend the deficiencies
of data sets. The “simple device” plus a few “plausible assumptions” will
indeed allow for estimates nearly as precise as the ones one would expect
when problems of incomplete data were completely ignored.

In contrast, alternative approaches are much more cautious and will in
many cases even indicate the inadequacy of a given data set to provide
useful answers at all. Manski (2003: 18) illustrates the effect by looking
at the percentage of employed persons estimated from the National
Longitudinal Survey of Youth, a sample of 6812 young Americans. The
naive estimate using only complete data is 78% with a 95% confidence
interval of roughly ±1%. Both the estimate and the confidence interval
can be justified within the Little-Rubin setup by simply declaring that
the “missing data mechanism” is “missing at random” (Little, Rubin
2002: 12). But taking into account that people who did not answer to this
question (some 18%) might be either employed or not, the percentage of
employed may vary between 64% and 82%, an interval not only much
larger than the one suggested by the sampling distribution but also too
large for many practical purposes.
The computation is easy: For 82% of the interviewed the employment status is known, but for the other 18% it is not. Thus if all those not answering that question were unemployed or out of labour force, the proportion of employed in the sample would be \(0.82 \cdot 0.78 + 0.18 \cdot 0 = 0.6396\) or roughly 64%. On the other hand, all of those not answering might as well have been employed. In that case, the share of 18% not answering must be added to the 64%, resulting in 82%. Note that these two extreme possibilities are not only logically consistent with the data but that they cannot be ruled out by an appeal to the general statistical approach to missing data nor by the more special notion of “missing at random”. The only information available from the data alone is that the percentage of employed people in the sample is in the range of 64% to 82%. In consequence, any “inference” to the population at large must at least contain this interval, plus any uncertainty resulting from the sampling procedure.

It is obvious why most practitioners would prefer to report 78% plus minus 1% as the result of the survey instead of being forced to admit that the sought for percentage is somewhere between 64% and 82%. Conventional sampling theory suggests that in order to achieve a 95% (symmetric) confidence interval for the percentage of ±9% (the equivalent to the interval of 64% to 82% resulting from Manski’s bounds), a sample of size 125 would be sufficient. Why go through the trouble to interview 6812 people if the result was at most as precise as the answers of just 125 persons? Moreover, since of the 6812 interviewed persons 5556 actually gave an answer, should the survey not be counted as being as informative as a pure random sample based on 5556 persons without any missing values?

In summary, there are many reasons to prefer the statistical approach to incomplete data, especially in the social sciences, where non-response rates of 18% must be regarded as moderate. But neither convenience, nor cost, nor appeal to conventional wisdom in sampling theory can be a valid argument to the effect that 78% ±1% is a reasonable estimate of the employment rate. How can such an optimistically accurate estimation be justified? The statistical approach to incomplete data proceeds in several steps:
The most crucial step is to construct a special mathematical model to replace the simple question originally posed. To make this step more transparent and to prepare for some generalisations, some notation is needed: Let \( \mathcal{U} = \{u_1, \ldots, u_N\} \) denote (a list of) the population of interest, in the example Americans aged 25 to 35 in 1991. Let \( Y : \mathcal{U} \rightarrow \{0, 1\} =: \mathcal{Y} \) be a function that assigns the employment status (1 stays for ‘employed’, 0 for ‘not employed’) to all members of the population. Further, denote by \( S \subseteq \mathcal{U} \) the given survey sample. The sought for percentage of employed people in the sample and in the population, respectively, are then

\[
m(Y, S) := \frac{1}{|S|} \sum_{u \in S} Y(u) \quad m(Y, \mathcal{U}) := \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} Y(u)
\]

But some of the people did not answer to the question on their employment status and thus the range of \( Y \) must be enlarged to allow for this possibility. The function \( Y^* : \mathcal{U} \rightarrow \{\{0\}, \{1\}, \ast\} \cup \{\mathcal{Y}\} = \{\{0\}, \{1\}, \{0, 1\}, \ast\} =: \mathcal{Y}^* \) will describe the answer to the employment question, where a ‘\( \ast \)’ is used if people were not asked (were never in the survey) and a non-response is indicated by the set \( \{0, 1\} \): a non-respondent might be either employed or not and the union of the two possibilities is the set \( \{0, 1\} \).

The function \( Y^* \) is intended to capture the responses of sample members, while \( Y \) captures the (socially constructed) fact of employment. The latter is defined for essentially all members of the population. But an answer to a survey questions is not, unless one was willing to posit that all members of the population posses an “answer characteristic”, assumed fixed for the purposes of survey sampling, that they “activate” just in case that they are included in a survey. The symbol ‘\( \ast \)’ is introduced so that \( Y^* \) can still be defined for all members of the population without presupposing anything about “answer characteristics”. This construction may be needed even when no comparison with some population is contemplate. E.g., the National Longitudinal Survey of Youth was started
in 1979 as a panel study. Of the 1256 non-responses on the employment question in 1991, only 555 were approached in 1991 and either declined to be interviewed or did not answer to the question. The other 701 could not be located or were otherwise lost to follow up before 1991. To presuppose a certain “answer characteristic” and thus a particular value of \{0\}, \{1\}, or \{0, 1\} for them would be preposterous and an extra symbol like ‘∗’ should be introduced.

Allowing for non-response and thus using only the values of \(Y^\ast\) (restricted to the survey members), the value of \(m(Y, S)\) must be between

\[
\frac{1}{|S|} \left( \sum_{u \in S} \mathbb{I}[Y^\ast = \{1\}](u) \right) \quad \text{and} \quad \frac{1}{|S|} \left( \sum_{u \in S} \mathbb{I}[Y^\ast \in \{\{1\}, \{0, 1\}\}](u) \right)
\]

(1.1)

where \(\mathbb{I}[]\) denotes the indicator function.¹ These are, of course, the bounds derived earlier. Note that there is nothing in the data that would allow to narrow the bounds. Both extreme cases are logically and materially possible, since the only connection between the functions is that from \(Y^\ast(u) = \{1\}\) \(Y(u) = 1\) should follow, and similarly for \(Y^\ast(u) = \{0\}\).² The latter conditions stipulate that respondents tell the truth in that generally \(Y(u) \in Y^\ast(u)\). This will be assumed throughout the following pages, since otherwise there would be no necessary connection between the observations and the percentage of employed in the sample. In the general situation, the requirement \(Y(u) \in Y^\ast(u)\) will be termed consistency condition in the following.

The “simple device”, then, consists in adding some further structure to the setup. It does so by introducing a probability model: Let \((\Omega, \mathcal{B}, \lambda)\) be a probability space. As is usual in the construction of probability models, the nature of \((\Omega, \mathcal{B})\) is irrelevant if only the space \(\Omega\) and its \(\sigma\)–algebra \(\mathcal{B}\)

¹ As usual, the indicator function \(\mathbb{I}[A](x)\) takes the value 1 iff \(x \in A\), 0 otherwise. For a function \(f\), \(\mathbb{I}[f \in B](x)\) is short for the somewhat cumbersome \(\mathbb{I}[f^{-1}(B)](x)\) where \(f^{-1}(B)\) is the pre-image of \(B\).

² Note that the problem of “inference to the population” does not play any role in this setup. It will re-surface when sampling properties of the “simple device” are discussed.
is chosen large enough to allow for the following constructions. Now the functions \( Y \) and \( Y^* \) can be transformed into random variables by redefining them to have joint domain \( \mathcal{U} \times \Omega \):³

\[
(Y, Y^*): \mathcal{U} \times \Omega \rightarrow \mathcal{Y} \times \mathcal{Y}^*
\]

These new functions should be carefully distinguished from the ones previously defined. The latter are a simple bookkeeping device, referring to a listing of people together with their employment status and their answers to survey questions respectively. In contrast, the new versions are only defined within a probability model. They are a mathematical construction and as such do not refer to anything in the world except in the very limited sense that it may be assumed that there is an \( \omega \) such that the values of \( Y(., \omega) \) are equal to the values in the list of the population. Values of both \( Y \) and \( Y^* \) are assumed to be generated from the list \( \mathcal{U} \) together with draws from a random number generator that produces both values for each \( u \in \mathcal{U} \). This is certainly not the way in which either the employment status or answers in interviews come about in this world.

Nevertheless it is exactly the formulation of a mathematical model that is the main achievement of the statistical approach. Within the model it is possible to precisely state conditions that justify a point estimate of a 78% employment rate. In terms of the probability model itself, the condition uniquely identifies the distribution of \( Y \) from the (known) distribution of \( Y^* \). This condition is termed **missing at random** (MAR) or, in a more general context, **coarsening at random** (CAR). It states that the conditional probability of any set \( \{Y^* = y^*\} \) given \( \{Y = y\} \) for a particular value of \( y \) is the same for all \( y \in y^* \). Put another way, the conditional probability of \( \{Y = y\} \) given the set \( \{Y^* = y^*\} \) is the same as that of the conditional probability of \( \{Y = y\} \) given that \( \{Y \in y^*\} \). In other words, having “observed” \( \{Y^* = y^*\} \) this tells us that \( \{Y \in y^*\} \) and nothing more. In Bayesian terms, the “information” \( \{Y^* = y^*\} \) allows

³ In Chapter 4 I have chosen to make the sample \( S \) dependent on \( \omega \in \Omega \) as well so that one can treat questions of sampling design within the same model. However, for the purpose of this introduction such a generalisation may distract from the main argument. Thus, the sample \( S \) is taken to be a known and fixed set here.
to update the a priori distribution of $Y$ to the conditional probability $\Pr(Y = y \mid Y \in y^*)$.

In the case of employment status the latter formulation states that the probability of being employed among the non-responders must be the same as the probability of being employed. Thus the probability of being employed, say $p_1$, must be equal to the probability of a positive answer to the employment question (0.64) plus the fraction of employed among the non-respondents (0.18 $p_1$). This gives the probability of 78% employment.

This looks like a rather complicated reformulation of a simple solution that moreover could have been formulated without the help of the “simple device” and without any probability theory. Namely: the proportion of employed among the respondents is the same as that among the non-respondents. But that is just a reformulation of the original unfounded claim and obviously only begs the question. The somewhat complicated formulation of the MAR/CAR condition is not a pretentious though vacuous reformulation of this unjustified claim. The main point of the “simple device” lies in the use of conditional probabilities that provide the connection between the distribution of $Y$ and that of $Y^*$. In particular, the first version of the MAR (or CAR) condition is a statement about the probability of the type of response that is chosen given the underlying state of the respondent. This is what Little and Rubin call a “missing data mechanism”. While it can only be formulated within a model for the joined distribution of $(Y, Y^*)$, it certainly provides a prescription to simulate incomplete data from any assumption on the distribution of the variable of interest, $Y$. This model construction permits to argue about the form of the “mechanism”, about the validity of the MAR or CAR condition, and about consequences of deviations from it. And it is precisely this feature that makes the “simple device” potentially useful.

Still, the special formulation used here might look unfamiliar. Most texts, including the well known textbook by Little and Rubin, formulate the MAR/CAR condition using a further random variable indicating whether data are missing or not. However, using subsets of $\mathcal{Y}$ to present incomplete data opens the way to discuss many types of incomplete data including truncated, censored, grouped, or heaped data in a unified way. Compared to pure missing data problems these general types of
incomplete data provide both more information and more challenging problems. The latter aspect enriches the discussion of the appropriateness of the “simple device” in general and of certain formulations in particular.

The missing at random condition can be formulated only within the probability model. It has no real world counterpart. While the random variable \( Y^* \) of the model may be identified with the observed values of \( Y^*(u), u \in U \), the empirical counterpart \( Y(u), u \in U \) of the random variable \( Y \) is only partially known. Therefore, the missing at random conditions cannot be translated into conditions on empirical distributions. No reformulation of the missing at random conditions will help in this respect.\(^4\)

Moreover, there is no way to formulate a “missing data mechanism” within the sampling theory framework. Using its bookkeeping variables allows to express assumptions pertaining to the real world such as: The percentage of employed among the non-respondents is 80%. It is plain that it might have been otherwise. But there is no way to argue for a particular value. In the sampling approach values of \( Y \) and \( Y^* \) are taken to be fixed for each interviewee. There is no formal way to express the idea of a tendency to answer to an interview request. On the contrary, the approach explicitly rules out any speculations about other possible responses and employment statuses than those that actually obtain. But such speculations are the essence of the “simple device”.

It is certainly not enough simply to invoke the missing at random “assumption” as an assumption about the social world and then proceed using an estimate that simply ignores missing or incomplete observations. It is apparent that neither the “simple device” nor the MAR condition as such justify the use of a point estimate of 78% for the employment rate. Nor does it justify the general tendency to ignore missing values and incomplete data and to use only complete data.

The original formulation of the missing data problem concerned the percentage of employed among a certain group and the problem encountered was that not all people answered to the respective question. In this

\(^4\) A more formal statement including necessary regularity conditions are discussed in the next Chapter.
setting, giving a particular percentage as the result of the survey can only be justified by appeal to convention and convenience or to experiences gained in more favourable circumstances. The “simple device” redefines this simple problem by first introducing a lot of mathematical structure and then solving a mathematical problem, not the one the survey statistician or the sociologist is interested in. The simple bookkeeping quantities \((Y, Y^*)\) are transformed into random variables. Then the problem of missing values is translated into one of formulating a stochastic relation between \(Y\) and \(Y^*\), a relation that has no direct empirical counterpart. And finally, sets of “assumptions” are produced that would guarantee the identification of the distribution of \(\hat{Y}\) from the distribution of \(Y^*\).

Before one may enjoy the merits or otherwise of the approach, one should therefore answer to the objections of its critics. The main objection, at least from the perspective of the survey statistician, is that the probability model and in particular the stochastic relation between \(Y\) and \(Y^*\) introduces additional mathematical structure far beyond the usual practises. But introducing additional structure may simply be a form of begging the question presupposing answers that otherwise would not exist. Moreover, in so far as some of the additional concepts are open to empirical scrutiny they may fail to be useful or applicable. Dalenius’ and Morgenstern’s remarks, if correct, would indicate that the prime motivating areas of applications of the statistical approach are exactly those where it lacks applicability.

The second objection pertains to the missing at random condition. Taken as an assumption about the real world, the condition is generally said to be untestable. In fact, in the example \(Y(u) \in Y^*(u)\) by construction and the same relation is used in the probability model. Since only \(Y^*(u)\) is empirically accessible any joint distribution of \(Y(u, .)\) and \(Y^*(u, .)\) that obeys the restriction is compatible with the data. And there is nothing else in the social world that corresponds to the MAR/CAR condition. It would be strange or at best elliptical to say that the missing at random assumption would hold for the interviewees of the National Longitudinal Survey of Youth. After all, it is certainly not a property or characteristic of the respondents. Nor does it pertain to the sample drawn. Neither is there a direct connection between the social world and probability.
statements as used in the “simple device”. Tryfos formulated:

To our knowledge, there is no randomizer consulting a table of random numbers to determine, say, the value of an unknown determining factor. We know of no gambling “Nature”, no “invisible hand”, divine or otherwise, rolling dice or drawing from an urn. (Tryfos 2004: 69)

Nor is it possible to “apply” the model to the social world. Matheron clearly stated that

… the notion of probability is of a mathematical, not empirical nature, and thus the notion of objectivity as it is accepted in the positive sciences is not relevant to it. … [N]obody has ever applied either the theory of probability or for that matter any other mathematical theory, to reality. One can only “apply” to reality real (physical, technical, etc.) operations, not mathematical operations. … In other words, it is always to probabilistic models, and only to them, that we apply the theory of probability. (Matheron 1989: 27)

Nor does the probability model “describe” the world. Or at least, models are nearly never used to describe the world. Rather, descriptions in science document observations and facts. Statements in descriptions relate to facts. Statements made within a probabilistic framework—in particular the MAR and CAR conditions—do not.

Even though some of the proponents of the “simple device” and many practitioners deliberately ignore the distinction, the critics were well aware of it. In fact, Dalenius was speaking of “assumptions of unknown validity about probabilities”, thus referring to the probability model and not to the data or social reality. Within a given probability model, what then is the validity of assumptions of the MAR/CAR variety? It might seem to be an arbitrary mathematical prerequisite similar to the introduction of the probability axioms. But then there would be no point in discussing it, not even a point for calling it an assumption. One does not say “$1 + 2 = 2 + 1$ because I assume addition of natural numbers to be commutative”. Perhaps it is best seen as a modelling decision, a constitutive decision that “defines the general framework
within which we shall operate and determines the choice of the tools we use” (Matheron 1989: 52). But such an interpretation is reasonable only for the introduction of the basic probabilistic framework for \((Y, Y^*)\). That choice determines the tools we use. The MAR/CAR conditions do not determine the general framework but define a very special case of the general model.

On the other hand, the MAR/CAR conditions are not open to empirical assessment. In this respect, they are similar to the assumption of stochastic independence that is invoked in many social science models. While the latter assumption is not empirically accessible either, it does not prejudice a particular solution to a problem but sets the frame of reference in which solutions are formulated. The MAR/CAR assumptions are different in that they do lead to a particular solution. And this solution may turn out to be unreasonable, even though a decision can not be based on the available data. For such conditions, criteria for their appropriateness and usefulness are required beyond the consistency within a given model. Except for other data sources or general experience, the only possible base for a critical assessment of the MAR/CAR condition must rely on an operational model, a prescription on how to simulate the incomplete data within the model that avoids to exploit information on the complete data model above that revealed by the incomplete data. Procedural models that can be simulated without presupposing a knowledge of the “truth” have been developed for certain non-MAR/CAR models as well. An assessment of the merits of MAR/CAR or alternative conditions may then be based on a comparison of the outcomes of procedurally specified sub-models. Criteria for such comparisons, and for the appropriateness of conditions that are neither open to empirical scrutiny nor are constitutive for the model construction, have only recently been proposed.

I will discuss these problems in some detail in the first, newly written chapter. It presents a general probabilistic formulation of the incomplete data problem and the corresponding solution sets. I then proceed to study several alternative mathematical representations of the solution sets which will help to decide whether the probabilistic framework adds possibly too much structure, thus answering the first concern of
the critics. In a second step, several formulations of the MAR/CAR conditions are introduced. Differences between these formulations have given rise to many confusions. A closer look at the differences, however, reveals that they are closely connected to the structure of incomplete data as represented in probabilistic models. One particular formulation, the one that relies on the introduction of additional random variables (e.g. indicators of missing values) to express MAR/CAR conditions, will in general introduce further restrictive elements into the analysis that go far beyond the introduction of a probability model per se. Such constructions are very useful since they allow to formulate MAR/CAR conditions even in cases that can not be couched in terms of a joined distributions of some \((Y, Y^*)\). And they provide a simple language to state MAR/CAR conditions in terms of these variables. On the other hand, it turns out that such formulations are ambiguous: There are in general several versions, some of which will declare \((Y, Y^*)\) to be MAR/CAR while others will contradict. I will conclude that such additional random elements will obfuscate the analysis as long as there is a reasonable mathematical structure of \((\mathcal{Y}, \mathcal{Y}^*)\). In a final step, various procedural models for the MAR and CAR conditions are introduced. These serve to study the special status of the MAR/CAR conditions in models of incomplete data. The chapter closes with some remarks on criteria of the appropriateness of the MAR/CAR conditions.

The next chapter, Chapter 3, is also written for the present document. It complements the previous chapter, discussing a practical problem using the tools developed so far. The problem is that of inferring the distribution of life lengths of a parent generation from the information provided by their children in a survey. The survey participants were asked for the birth dates of their parents and, if they died before the survey date, the date of death. Such a data structure combines truncated and censored data with a specific selection of the observations. Information on life length is obviously only available for those who had children so that deaths in infancy are never reported. Parents who are still alive at the time of the survey contribute censored observations. And parents with several children are possibly over-represented in a survey. The interplay of several forms of incompleteness makes this case study particularly valuable in demonstrating several versions of the “simple device” and
their consequences.

Chapter 4 takes up the general discussion, this time concentrating on the selectivity of observations. In the simplest case, selection models are just a variant of missing data models in that information on a certain group of people is unavailable. But selectivity becomes a much more interesting problem when partially complete data are considered along with the completely missing case. A further extension of the basic model discussed in Chapter 2 is the introduction of conditional incomplete data models. Much more challenging variants of incomplete data models are models of self-selection. They are prevalent in many parts of micro-economics and sociology. The Roy model and the particular case of Heckman’s model are but two examples that fit in the general framework. Such models clearly violate the CAR condition but still permit to deduce a unique answer from observational data, at least when some additional “assumptions” are invoked. One of the questions that prompted the article was whether arguments within the probability model would justify such non-CAR models. Another question that is investigated is the stability of answers from CAR or non-CAR models to minor variations in assumptions on the selection “mechanism”. The newly written appendix provides some hints especially on the latter question. Techniques of sensitivity analyses have seen some important developments since the article was published. While the article surveys the main statistical methods suggested to deal with incomplete and self-selected data, the appendix updates hints to this rapidly growing literature. It also contains a short discussion on the appropriate standards by which to judge statistical methods via asymptotic considerations.

The next chapter, Chapter 5, discusses statistical approaches to causal inference. The predominant approach in statistics nowadays is based directly on the missing data paradigm. It defines a cause, or, perhaps better, the causal effect of the value of the “variable” \( X(u) = x \) on \( Y(u) \) to be the difference between \( Y(u) \mid X(u) = x \) and \( Y(u) \mid X(u) \neq x \). This is a missing data problem since for any given unit \( u \in U \), only either \( X(u) = x \) or \( X(u) \neq x \) but not both are observed. In consequence, most statistical approaches rely on some form of the CAR condition to identify the “causal” effect of \( X \). The article provides an alternative relying on
1. Introduction

the timing of events. The approach identifies events as the relata of a causal relation and is thus closer to the classical philosophical discussion than that based on variables. Also, since humans can bring about events and since sequences of such events constitute an important topic in the social sciences it is better suited for social science applications than a counterfactual account. And finally the approach avoids to mimic (statistical) experiments that inspire the counterfactual approach. The analogy with experiments that is sought by other approaches may in many cases of interest to the social scientists turn out to be counterproductive.

The new appendix to the chapter starts with a short review of the counterfactual approach as it is discussed in the recent statistical and social science literature. It then illuminates the direct connection between the approaches to incomplete data problems and statistical versions of counterfactual accounts of causality. Recent progress in the two main problem areas of the event based approach, the notion of autonomy and the definition of local independence, are also reviewed.

Chapters 6 and 7 deal with the special case of censored data. The first proposes a potentially very useful extension of non-parametric techniques for censored data using the mean residual life function. Its main technical contribution is a thorough investigation of variance estimators based on a representation of Kaplan-Meier integrals in terms of independent summands. The main contribution in the present context of incomplete data, however, is the reminder that some statistical methods within the incomplete data paradigm can be re-expressed in a way that both fit the standard statistical methods geared to independent observations and the requirements of incomplete data methods. Chapter 7, on the other hand, provides a discussion of some constraints on the form of hazard functions when a particular constraint, that of periodic effects, is invoked. Such constraints are often present in labour market studies as well as in marketing, and these examples provided the motivation of the article. The constraints imply restrictions on the form of the influence of covariates that can be seen as a restriction on the possible incomplete data models compatible with the constraints when covariates are present.

Chapter 8 discusses a regression model for multivariate censored observations. Historically, the non-parametric multivariate censored data
model has been one of the inspiring problems for a general account of incomplete data. It is one of the examples where the non-parametric maximum likelihood estimator is inconsistent. One of the consistent alternatives deliberately coarsens the data even further by grouping both event and censoring times. Perhaps unexpectedly, incomplete data may help in stabilising estimation problems. The article exploits this fact for the construction of a locally efficient regression estimator. Its construction is based on the missing information principle, a reformulation and extension of the CAR condition that shows how scores and other likelihood quantities transform when some data are partially incomplete. The appendix reviews new developments in the estimation theory with incomplete data. Recent variations and extensions of the Buckley-James estimator suggested in the paper are also discussed.

The last Chapter 9 investigates the effect of incorrect model choice in regression models with linear predictors. It turns out that even when the model is mis-specified the relative size of regression coefficients is correctly estimated. This result, it is noted, is directly connected with a situation where the data for the dependent variable are only incompletely observed. Thus, a form of duality between incomplete data models and the analysis of mis-specified regression models is established. This type of partial duality, though probably known in special cases and for a long time, is certainly not yet explored to the same extent as other areas of incomplete data analysis. The new appendix lists a few more recent hints to a duality between incomplete data and wrong statistical models.

Finally, the work is supplemented by an appendix documenting an implementation of CAR compatible non-parametric distribution estimators within TDA (Transition Data Analysis, www.stat.ruhr-uni-bochum.de/tda.html), a computer package designed specifically to deal with censored and otherwise incomplete data. It is the base for Buckley-James type estimators discussed in Chapter 8 and can easily be used as a building block for regression problems with more general types of incomplete data.
While the treatment of missing data has a long tradition in statistics, it was only 30 years ago that Rubin proposed to study such problems systematically from a probabilistic point of view. In retrospect, Rubin and Little dubbed the introduction of random variables, probability distributions and expectations to the study of missing data a “simple device”. This chapter studies the ramifications of the “simple device” in the context of general incomplete data. It outlines a theory of probabilistic incomplete data models that is general enough to cover all data structures discussed in later chapters. And it provides the basic framework used in the statistical approach to incomplete data. Within this framework, some of the problems mentioned in the introduction can be answered. Indeed, the very introduction of a probability model may at first sight seem extraneous to the problem of analysing incomplete data. The use of probability theory may simply presuppose solutions that otherwise do not exist. It is not clear at the outset whether the introduction of the “simple device” will prejudice its solutions, thus just begging the original question.

I will show that this is not the case in the following section. There I also explore the mathematical structure of incomplete data models and relate them to other well known structures in mathematics. The second section introduces several versions of the coarsening at random (CAR) model. It turns out that the differences between CAR formulations are
closely connected to the structure of incomplete data which therefore is an essential part of the “simple device” of Rubin and Little. The discussion will clarify several misunderstandings about the role of CAR models that pervade the current literature. Among these are the empirical content of CAR models and the uniqueness of CAR models. The next sections considers the role of missing indicators that play such a prominent role in the now classical version of the “simple device”. It is shown that the introduction of the missing indicator (or of any other further random variables like censoring variables) extends the field of applications considerably. But it does so at the cost of making the definition of CAR dependent on modelling decisions that can not be justified from considerations of the structure of incomplete data alone. In this sense, the introduction of random variables indicating the extent of incompleteness goes beyond a probabilistic re-expression of the incomplete data problem. A last section concludes by taking up the discussion of criteria of applicability of probabilistic incomplete data models.

2.1. The Structure of Incomplete Data

Statistical data as well as random variables are commonly defined as functions with values in a previously defined space $\mathcal{Y}$. Statistical variables take as their domains of definition a list representing the target population, say $\mathcal{U}$. On the other hand, random variables are defined on a probability space $\Omega$ equipped both with a probability measure $\Pr(\cdot)$ and a set of subsets of $\Omega$ to which probabilities can be assigned. The device of random variables as functions on $\Omega$ serves to induce further probability distributions and to define operations between random variables. In contrast, statistical variables are a device to keep track of properties of the members of the population. In both cases, however, the range space is chosen by the researcher. It is an a priori decision in Matheron's sense, constituting the questions that are contemplated.

More often than not, the detailed information necessary for the assessment of a statistical variable is not available. Employment status may be
2.1. The Structure of Incomplete Data

thoroughly defined, say following the ILO standard. Thus the range of a corresponding variable is well defined. But if people are asked about their employment status using that standard, it may not fit too well with the knowledge of the respondents. In consequence, respondents may only provide a rough characterisation of their employment status or even do not answer at all.

This situation can be represented by a further set $\mathcal{Y}^*$, a set of subsets of $\mathcal{Y}$. As an example, suppose that employment status is differentiated into the three categories ‘employed’, ‘unemployed’ and ‘out of labour force’ coded consecutively as $\mathcal{Y} = \{1, 2, 3\}$. Suppose that respondents are reluctant or unable to differentiate between unemployment and out of labour force. Further they may refuse to answer at all. Then the observed data can be represented by

$$\mathcal{Y}^* = \{\{1\}, \{2\}, \{3\}, \{2, 3\}, \{1, 2, 3\}\} \subseteq \mathcal{P}(\mathcal{Y}) \setminus \{\emptyset\}$$

where $\mathcal{P}(\mathcal{Y})$ is the power set of $\mathcal{Y}$. Here a non-response is represented by $\{1, 2, 3\}$ while someone who is unwilling to differentiate between unemployment and being out of labour force is represented by $\{2, 3\}$. Exact answers are represented by $\{1\}, \{2\},$ or $\{3\}$.

In general, let

$$Y: \mathcal{U} \longrightarrow \mathcal{Y}$$

be a statistical variable. Then coarse data can be represented by a statistical variable $Y^*$

$$Y^*: \mathcal{U} \longrightarrow \mathcal{Y}^* \subseteq \mathcal{P}(\mathcal{Y}) \setminus \{\emptyset\}$$

such that for all $u \in \mathcal{U}$: $Y(u) \in Y^*(u)$. The last condition excludes the possibility that respondents are misrepresenting their situation. I will term this requirement the consistency condition. In a purely set based model of incomplete data such a restriction is necessary since otherwise there would be no connection between $Y$ and $Y^*$ at all.

The consistency condition rules out some situations that sometimes are considered as incomplete data problems. If all respondents possibly
mis-classified their situation by indicating another employment than the one they are in, then the consistency condition would force the choice of \{1, 2, 3\} as a representation of all answers. But that would be completely uninformative.

Fortunately, most incomplete data problems encountered in social science applications, in particular grouped, truncated, and censored data, do have a representation in terms of subsets. Consider grouped data: If $\mathcal{Y}$ is an ordered set represented by \{0, 1, \ldots, \tau\}, and if $[y_i, y_{i+1}]$ are intervals into which some observations are grouped, then $\mathcal{Y}^* = \{\{0\}, \{1\}, \ldots, \{\tau\}, [y_0, y_1], \ldots, [y_{k-1}, y_k]\}$ may be used to represent the situation. Such a grouping together with the possibility to provide “exact” values is often employed in survey questions. An example from the ALLBUS is discussed in Chapter 4. A similar case of the grouping together of observations that is not restricted to intervals is given by a general function $g: \mathcal{Y} \rightarrow \mathcal{Z}$ that can be represented as $\mathcal{Y}^* = \{g^{-1}(z) \mid z \in \mathcal{Z}\}$.

Another important case of incompleteness is provided by censored or truncated data. If $\mathcal{Y}$ is again the ordered set \{0, 1, \ldots, \tau\}, then censored observations, i.e. observations for which it is only known that an event happened after a certain period, can be represented by \{y, \ldots, \tau\} such that $\mathcal{Y}^*$ becomes \{\{0\}, \{1\}, \ldots, \{\tau\}, \{0, \ldots, \tau\}, \{1, \ldots, \tau\}, \ldots\}. This structure is studied in most of the later chapters.

A parallel construction can be used within a probability model without any change, resulting in

$$(Y, Y^*): \Omega \longrightarrow \mathcal{Y} \times \mathcal{Y}^*$$

such that

$$Y(\omega) \in Y^*(\omega) \text{ for all } \omega \in \Omega \quad (2.2)$$

This will also be called consistency condition.$^{1}$

$^{1}$The condition (2.2) can be weakened to $\Pr(Y \in Y^*) = 1$. This might be necessary if $\Omega$ is chosen uncountable, e.g. to model continuous covariates. But with finite sets $\mathcal{Y}$ as used here, and without the necessity to model additional information, nothing essential is gained since one can always choose a finite $\Omega$ with $\Pr(\{\omega\}) > 0$ for all $\omega$. 

20
One aspect of incomplete data can be studied by just considering the structure of $Y^*$. To facilitate the discussion, I introduce three simple examples that will be considered throughout this Chapter.

Example 1
Suppose again that employment status is differentiated into the three categories ‘employed’, ‘unemployed’ and ‘out of labour force’ coded consecutively as $Y = \{1, 2, 3\}$. Suppose further that respondents are reluctant or unable to differentiate between unemployment and out of labour force. Further they may of course refuse to answer at all. Then the observed data can be represented by

$$Y_1^* = \{\{1\}, \{2\}, \{3\}, \{2, 3\}, \{1, 2, 3\}\}$$

Example 2
Respondents in addition to the previous reporting conditions would sometimes only report whether they are employed or unemployed vs. being out of labour force. Then the support of $Y^*$ becomes

$$Y_2^* = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$$

Example 3
The second example can be changed slightly by deleting $\{3\}$. Thus nobody would report to be out of labour force. The resulting $Y_3^*$ becomes

$$Y_3^* = \{\{1\}, \{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}$$

The sets of coarsened data are partially ordered by set inclusion. Thus, in the first example, one has $\{2\} \prec \{2, 3\} \prec \{1, 2, 3\}$. And in the second example $\{2\} \prec \{1, 2\}, \{2, 3\} \prec \{1, 2, 3\}$. The ordering relations of the elements of the three examples can be read from the corresponding Hasse diagrams presented in Figure 2.1
2. Probabilistic Models of Incomplete Data

Figure 2.1.: Hasse diagrams for the sets $\mathcal{Y}_1^*$, $\mathcal{Y}_2^*$ and $\mathcal{Y}_3^*$. The top nodes correspond to the set $\{1, 2, 3\}$, the bottom nodes to the sets $\{1\}$, $\{2\}$, $\{3\}$.

2.1.1. A Typology of Missing Data Structures

Example 1 is special in that the elements of $\mathcal{Y}_1^*$ are hierarchically ordered. By this I mean that for all elements $y_1^*, y_2^* \in \mathcal{Y}^*$ we have either $y_1^* \subseteq y_2^*$ or $y_2^* \subseteq y_1^*$ or $y_1^* \cap y_2^* = \emptyset$. Another way to describe this special situation that will be of importance later on is that $\mathcal{Y}^*$ is the union of a sequence of refinements of a partition of $\mathcal{Y}$. Here, the sequence of partitions is $\{\{1, 2, 3\}\}, \{\{1\}, \{2, 3\}\}, \{\{1\}, \{2\}, \{3\}\}$.

The case of hierarchically ordered incomplete data structures is encountered in many common models of incomplete data situations. The most obvious one is when only completely missing or exact observations of a single variable are considered. Then the resulting $\mathcal{Y}^*$ is trivially hierarchically ordered.

If observations are discrete and possibly right censored durations, censored observations are presented by sets of the form $\{y, \ldots, \tau\}$. Such observations just tell that an event happened after period $y - 1$, but they don’t reveal the exact time of the event. Rearranging terms, the set $\mathcal{Y}^*$ becomes $\{\{0, \ldots, \tau\}\} \cup \{\{0\}, \{1, \ldots, \tau\}\} \cup \{\{0\}, \{1\}, \{2, \ldots, \tau\}\} \cup \ldots$ so that it is a union of refinements of a partition and thus hierarchically ordered.

Grouped data provide another important example. If $\mathcal{Y} = \{0, \ldots, \tau\}$ and if there is just one level of grouping $\mathcal{Y}$ into disjoint intervals $[y_i, y_{i+1}]$ whose union is $\mathcal{Y}$, then the $\mathcal{Y}^*$ resulting from taking the union of the intervals with the set of all singletons and possibly the whole set (completely missing data) will result in an hierarchically ordered $\mathcal{Y}^*$. 
Adding further levels of grouping still will lead to hierarchically ordered models as long as the further levels are either refinements or coarsenings of the original grouping.

Perhaps the most important example is provided by monotone missing data patterns. Consider several variables $Y_1, \ldots, Y_k$ with values in $\mathcal{Y}_1 \times \ldots \times \mathcal{Y}_k$. Suppose that each variable is either observed or completely unobserved. If the pattern of missingness is such that there is an ordering of the variables, say $(1) < (2) < \ldots < (k)$, and such that if $Y_{(i)}$ is missing (i.e. $Y^*_{(i)} = \mathcal{Y}_{(i)}$) then all $Y_{(j)}$ are missing as well as long as $(i) < (j)$, the pattern is called a monotone missing pattern (see e.g. Little, Rubin 2002: Chap. 7.4). A typical element of the set $\mathcal{Y}^*$ is

$$\{(y_{(1)}, \ldots, y_{(i-1)})\} \times \mathcal{Y}_{(i)} \times \ldots \times \mathcal{Y}_{(k)}$$

and the union of these sets over all tuples $(y_{(1)}, \ldots, y_{(i-1)})$ forms a partition of $\mathcal{Y}_{(1)} \times \ldots \times \mathcal{Y}_{(k)}$. Furthermore, if $\mathcal{P}_i$ is such a partition with smallest index $(i)$ such that all the values of $Y_{(j)}$ with index $(j)$ larger than $(i)$ are missing, the partition $\mathcal{P}_j$ with $(j) > (i)$ is a refinement of $\mathcal{P}_i$. Thus, monotone missing data patterns are hierarchically ordered.

In section 2.2, I will show that the CAR condition for hierarchically ordered models can be treated very much like a single variable which is either exactly observed or completely missing. But how can one justify the choice of such a simple model? In the case of monotone missingness the order of the variables may arise from an ordering of observations according to the time they were made. This is often the case in panel studies. Moreover, in that case, monotone missingness just excludes the possibility of returning to the panel population after some temporary withdrawal. This restriction is often part of the design of a panel study. In other cases, it might be seen as a reasonable approximation to the design.

But in many other models such an assumption seems to be rather far fetched. Consider regression models with several covariates where some or even all of the covariates are incompletely observed. That situation can be formalised quite similarly to the case of monotone missingness. Nevertheless, a regression context will nearly never allow to justify a particular order of incompleteness information. And even if one is
thought to be adequate, a few data may prove the assumption to be inadequate. Or, arguing the other way around, even if a given data set is consistent with a certain order of incompleteness, this may not be expected in general and not even for similar data sets. Still, it would be inadequate to decide on the choice of the model for the incompleteness structure solely based on one (or several) data sets. It is often much easier to envisage broader forms of incompleteness than those actually present in a data set since it may then be possible to largely reduce the modelling burden. The choice of the incompleteness structure will therefore always contain preliminary modelling decisions that are informed by data aspects but are not determined by it.

Since hierarchically ordered incomplete data structures and their prototype of missing values in single variables share many nice properties, the study of the value of the “simple device” might in fact be prejudiced when only these special cases are considered. A less demanding data structure is given by a union of several partitions that need neither be refinements nor coarsenings of each other. An example in point would be data grouped into several types of categories which may overlap. Example 2 provides another example. $\mathcal{Y}^*_2$ is not hierarchically ordered since $\{1, 2\} \cap \{2, 3\} \neq \emptyset$ but neither $\{1, 2\} \subseteq \{2, 3\}$ nor $\{2, 3\} \subseteq \{1, 2\}$. But it is the union of several partitions, since one may write $\mathcal{Y}^*_2 = \{\{1, 2, 3\}\} \cup \{\{1\}, \{2, 3\}\} \cup \{\{3\}, \{1, 2\}\} \cup \{\{1\}, \{2\}, \{3\}\}$. Another prominent example is given by current status data where for a given point in time it is only known whether an event did or did not happen. Then the incomplete data of event times are of the form $\{0, \ldots, y\}$ if the event happened before time $y$. Or they take the form $\{y + 1, \ldots, \tau\}$ when the event did not yet happen at inspection time. Clearly, this forms a partition of $\mathcal{Y}$ for each inspection time $y$ and thus $\mathcal{Y}^*$ is a union of partitions. But these partitions are not refinements of each other.

The third example, $\mathcal{Y}^*_3$, is neither hierarchically ordered nor even the union of partitions of $\mathcal{Y}$ since for the element $\{2\}$ the only disjoint set in $\mathcal{Y}^*_3$ is $\{1\}$ but there union is not all of $\mathcal{Y}$. The consideration of such examples will provide the background against which several formulations of CAR can be compared. It will transpire that modelling decisions
pertaining only to the choice of \((\mathcal{Y}, \mathcal{Y}^*)\) have important consequences for the evaluation of the “simple device”.

2.1.2. Graphical Representations of Coarsened Data

Depicting the order of the elements of \(\mathcal{Y}^*\) will as well exhibit which elements of \(\mathcal{Y}\) belong to which sets in \(\mathcal{Y}^*\) if the latter contains all one-element sets. If this is not the case, one would need to indicate this additional information, e.g. by providing the incidence matrix with rows presenting the elements of \(\mathcal{Y}\) and columns the elements of \(\mathcal{Y}^*\). In the matrix, the \((i, j)\)-th element is 1 if the \(i\)-th element of \(\mathcal{Y}\) is an element of the \(j\)-th element of \(\mathcal{Y}^*\), 0 otherwise. In the first example one gets

\[
\begin{array}{cccccc}
\{\text{1}\} & \{\text{2}\} & \{\text{3}\} & \{\text{2}, \text{3}\} & \{\text{1}, \text{2}, \text{3}\} \\
\text{1} & 1 & 0 & 0 & 0 & 1 \\
\text{2} & 0 & 1 & 0 & 1 & 1 \\
\text{3} & 0 & 0 & 1 & 1 & 1 \\
\end{array}
\]

![Figure 2.2](image)

Figure 2.2.: Levi graph for the sets \(\mathcal{Y}_1^*\), \(\mathcal{Y}_2^*\) and \(\mathcal{Y}_3^*\). The top circles correspond to the set \(\{1, 2, 3\}\), the bottom circles to the sets \(\{1\}\), \(\{2\}\), \(\{3\}\). The squares correspond to the elements of \(\mathcal{Y}\), 1,2,3.

This could be visualised by a hypergraph whose vertices are the elements of \(\mathcal{Y}\) and the edges indicate the composition of the elements of \(\mathcal{Y}^*\). An equivalent representation is the Levi graph, a bipartite graph with the first set of vertices being the elements of \(\mathcal{Y}\) and the second set being the elements of \(\mathcal{Y}^*\). There is an edge between vertices of the two sets iff \(y \in y^*\). The Levi graphs of the three examples are given in Figure 2.2.
The Levi graph and the hypergraph of an incidence matrix are obviously in one-to-one correspondence.

Jaeger (2005b) argued that it is better to use the dual of the hypergraph where the elements of $\mathcal{Y}^*$ are the vertices and the edges indicate which elements of $\mathcal{Y}$ belong to them. The resulting hypergraphs for the three examples are depicted in Figure 2.3.

![Figure 2.3: Hypergraphs for the sets $\mathcal{Y}_1^*$, $\mathcal{Y}_2^*$ and $\mathcal{Y}_3^*$. The layout is the same as in the previous Figure, the top nodes correspond to the set $\{1, 2, 3\}$, the bottom nodes to the sets $\{1\}$, $\{2\}$, $\{3\}$. The edges indicate the elements of $\mathcal{Y}$.](image)

Taking an extreme example one may consider

$$\mathcal{Y}_4^* := \{\{1, 2\}, \{2, 3\}\}$$

so that the only information available would be whether someone was employed or unemployed, or unemployed or out of labour force. There is no ordering between the two sets. There are no singletons either and so the information on which element of $\mathcal{Y}$ belongs in which of the two coarsened sets in $\mathcal{Y}^*$ must be added. The corresponding hypergraph is given in Figure 2.4.

In this hypergraph, the edge representing 2 (the outer ellipse) contains two proper subsets, the edges representing 1 and 3 (the inner circles around the two circles representing the sets $\{1, 2\}$ and $\{2, 3\}$). Jaeger (2005b) argued that whenever the hypergraph contains properly nested edges then the CAR condition can not hold (except for degenerate cases). The non-existence of CAR models would then follow just from the
2.1. The Structure of Incomplete Data

![Hypergraph](image)

Figure 2.4.: Hypergraph for the set $Y^*_4$. The innermost circles represent the sets $\{1, 2\}$ and $\{2, 3\}$. The outer circles represent the elements 1 and 3, and the out-most oval represents the element 2.

structure of the incomplete data and without considering any more details like the distribution of $Y^*$.

There is an interesting connection of this example with the famous Monty Hall problem: A contestant at the Monty Hall show has to choose between three doors 1, 2, 3, behind one of which there is a price. Behind the others there is a goat. If the contestant has chosen door 2, say, the host opens one of the other doors which reveals a goat. The information that the contestant can expect after his choice of the door is that the price is behind either the doors $\{1, 2\}$ or $\{2, 3\}$. In particular, the Monty Hall problem has the structure given in Figure 2.4. Next, the host reveals which of the two sets from $Y^*_4$ obtain. Given this information, the candidate can either change his choice of a door or stick with the original one. But this opening of a door by the host is informative for the location of the price: Informally, if the rules of the game were uninformative about the location of the price, the opening of, say, door 3 by the host should only reveal that the price is either behind door 2 or behind door 1 ($Y^* = \{1, 2\}$). Put differently, the decision to reveal $\{1, 2\}$ ought not depend on the location of the price. Now if the price is behind door 1, the rules of the game force the host to open door 3. But then an uninformative rule would force him to open door 3 also when the price is behind door 2. By the same argument, the host must open door 1 (revealing $Y^* = \{2, 3\}$) when the price is behind door 3. An uninformative rule would lead him to open door 1 also when the price is behind door 2. In consequence, if an uninformative rule existed it would lead to inconsistent behaviour (opening both door 1 and door 3 when the price is behind door 2). Thus there is no uninformative rule and the CAR
2. Probabilistic Models of Incomplete Data

condition certainly does not hold. This example provides a connection between the incomplete data literature and that on probability dynamics in the Bayesian literature.

2.1.3. Distributions Consistent with Coarsened Data

The last piece of information required for the representation of incomplete data is a distribution on $\mathcal{Y}^*$. This might either be the distribution of a statistical variable $Y^*$ representing the observed frequencies of incomplete data. Or, as in the Monty Hall example, it might be a probability distribution representing the “rules of the game”. The task is to infer something useful about the distribution of $Y$ on $\mathcal{Y}$ from a distribution on $\mathcal{Y}^*$.

In any case, instead of taking a particular subset $\mathcal{Y}^*$ of the power set of $\mathcal{Y}$ as the support of the distribution, one might as well take the full power set and use zero probability assignments for certain elements of the power set to define the structure of incompleteness. However, I prefer to separate the determination of a distribution of $Y^*$ from the determination of its support encoded by a subset of the full power set, namely $\mathcal{Y}^*$. Often the distribution of $Y^*$ is determined by the empirical distribution. But the structure of $\mathcal{Y}^*$ is only partly determined by the data. It might be chosen much larger than the data suggest, so that a particularly simple structure emerges. Or it might, as in the Monty Hall example, encode the rules of the game. There are also technical problems:

---

2 This rather informal discussion presupposes that the price may be behind any of the doors with positive probability in order to derive a contradiction. If the price is always behind door 1 (and the contestant has chosen door 2 preliminarily), then there is a unique strategy of the host that is ‘uninformative’ in the above sense: He is forced to reveal $\{1, 2\}$. Similarly, if the price is always behind door 2, the host is free to either reveal $\{1, 2\}$ or $\{2, 3\}$. This will always be ‘uninformative’ whatever his strategy to choose between $\{1, 2\}$ and $\{2, 3\}$. These ‘uninformative’ possibilities are only ruled out by a further elaboration of the rules of the game.


4 The sets $y^* \in \mathcal{P}(\mathcal{Y})$ are called focal elements by Shafer (1976: 40). His notion of core as the union of all focal elements is a subset of my $\mathcal{Y}^*$. 

---
With infinite sets $\mathcal{Y}$, it is in general impossible to assign probabilities to all subsets of $\mathcal{Y}$. In such applications, subsets of the power set must in any case be determined before a probability distribution can be assigned. Consequently, in the following a particular subset $Y^*$ of the power set of $\mathcal{Y}$ is used as the support of the distribution of $Y^*$. The distribution may, however, assign zero probabilities to some further elements of $\mathcal{Y}^*$. On the other hand, it is much easier to work with probabilities defined on all of $\mathcal{P}(\mathcal{Y})$ so that I will without further mentioning use the extension of a probability on $\mathcal{Y}^*$ by stipulating $\Pr(Y^* = A) = 0$ for all $A \subseteq \mathcal{Y}$ such that $A \not\in \mathcal{Y}^*$.

The distribution with support $\mathcal{Y}^*$ together with the consistency condition will imply restrictions on the distribution of the corresponding variable $Y$. Since $Y(.) \in Y^*(.)$ for all arguments of the variables, $\{Y^* = y^*\} \subseteq \{Y \in y^*\}$ for all $y^* \in \mathcal{Y}^*$ because if $Y(\omega) = y^*$, then necessarily $Y(\omega) \in y^*$ for (almost) all $\omega \in \Omega$. In particular,

$$\Pr(Y^* = y^*) \leq \Pr(Y \in y^*) \quad (2.3)$$

for all $y^* \in \mathcal{Y}^*$. A useful consequence is

$$\Pr(Y^* = y^*) = \Pr(Y^* = y^*, Y \in y^*) \quad (2.4)$$

Since also $\{Y^* = y'^*\} \subseteq \{Y \in y^*\}$ for all subsets $y'^* \subseteq y^*$, it follows that $\bigcup_{y'^* \subseteq y^*} \{Y^* = y'^*\} \subseteq \{Y \in y^*\}$ and the inequality can be strengthened to

$$\sum_{y'^* \subseteq y^*} \Pr(Y^* = y'^*) = \Pr(Y^* \subseteq y^*) \leq \Pr(Y \in y^*) \quad (2.5)$$

The same reasoning applies not only to $y^*$ but to arbitrary subsets $A \subseteq \mathcal{Y}$. Thus one finally arrives at

$$\sum_{y'^* \subseteq A} \Pr(Y^* = y'^*) = \Pr(Y^* \subseteq A) \leq \Pr(Y \in A) \quad (2.6)$$

for all $A \subseteq \mathcal{Y}$. 
On the other hand, \( \{ Y = y \} \subseteq \cup_{y^* \exists y} \{ Y^* = y^* \} \) since if \( Y(\omega) = y \) then necessarily all possible values of \( Y^*(\omega) \) must contain \( y \) (the empty set is excluded from \( \mathcal{Y}^* \)) and therefore

\[
\Pr(Y = y) \leq \sum_{y^* \exists y} \Pr(Y^* = y^*) = \Pr(Y^* \cap \{ y \} \neq \emptyset) \tag{2.7}
\]

for all \( y \in \mathcal{Y} \). An immediate consequence is that

\[
1 = \sum_{y^* \exists y} \Pr(Y^* = y^* \mid Y = y) \tag{2.8}
\]

for all \( y \) with \( \Pr(Y = y) > 0 \) since the right hand side is just the probability of the set \( \cup_{y^* \exists y} \{ Y^* = y^* \} \) conditioned on its subset \( \{ Y = y \} \). The equality states that the support of the conditional probabilities given \( \{ Y = y \} \) is the subset of elements of \( \mathcal{Y}^* \) that contain \( y \).

The inequality can be generalised to give

\[
\Pr(Y \in y^*) \leq \sum_{y^* \subseteq y^* \neq \emptyset} \Pr(Y^* = y^*) = \Pr(Y^* \cap y^* \neq \emptyset) \tag{2.9}
\]

This is not, in general, a strengthening of (2.7). In the case of \( \mathcal{Y}^*_4 \), \( \Pr(Y = 1) \leq \Pr(Y^* \in \{1, 2\}) \) but trivially \( \Pr(Y \in \{1, 2\}) \leq \Pr(Y^* = \{1, 2\}) + \Pr(Y^* = \{2, 3\}) = 1 \). In hierarchical data structures, however, the upper bound of \( \Pr(Y = y) \) stays the same when the event \( \{ Y = y \} \) is replaced by the possibly larger event \( \{ Y \in y^* \} \) when \( y^* \) is the unique element of the finest partition of \( \mathcal{Y} \) that contains \( y \). Thus in hierarchical incomplete data structures it is always sufficient to consider the pairs of inequalities (2.5) and (2.9).

Since there is no best upper bound on subsets of \( \mathcal{Y} \) in general, one has to check that

\[
\Pr(Y \in A) \leq \sum_{y^* \subseteq A \neq \emptyset} \Pr(Y^* = y^*) = \Pr(Y^* \cap A \neq \emptyset) \tag{2.10}
\]

for all \( \emptyset \neq A \subseteq \mathcal{Y} \). Combining this with the respective lower bounds, the full set of inequalities becomes:

\[
\sum_{y^* \subseteq A} \Pr(Y^* = y^*) = \Pr(Y^* \subseteq A) \leq \Pr(Y \in A)
\]
2.1. The Structure of Incomplete Data

\begin{equation}
\leq \Pr(Y^* \cap A \neq \emptyset)
\end{equation}
\begin{equation}
= \sum_{y^* \cap A \neq \emptyset} \Pr(Y^* = y^*)
\end{equation}

for all \( \emptyset \neq A \subseteq \mathcal{Y} \).

While the extension to all non-empty subsets of \( \mathcal{Y} \) introduces many redundant restrictions, it allows for a concise and symmetrical formulation. Setting

\[
F(A) := \Pr(Y^* \subseteq A) = \sum_{y^* \subseteq A} \Pr(Y^* = y^*)
\]

\[
T(A) := \Pr(Y^* \cap A \neq \emptyset) = \sum_{y^* \cap A \neq \emptyset} \Pr(Y^* = y^*)
\]

the inequalities can be abbreviated to

\begin{equation}
F(A) \leq \Pr(Y \in A) \leq T(A)
\end{equation}

But \( \{Y^* \subseteq A\} = \{Y^* \cap A^c = \emptyset\} = \{Y^* \cap A^c \neq \emptyset\}^c \). Thus \( F(A) = 1 - T(A^c) \). In particular, the upper bound for the set \( A^c \), \( \Pr(Y \in A^c) \leq T(A^c) \), is equivalent to \( \Pr(Y \in A) = 1 - \Pr(Y \in A^c) \geq 1 - \Pr(Y^* \cap A^c \neq \emptyset) = F(A) \), the lower bound for the set \( A \). Since the inequalities are to hold for all non-empty \( A \subseteq \mathcal{Y} \), the consistency conditions can be shortened to the one sided condition

\begin{equation}
F(A) \leq \Pr(Y \in A) =: \Pi(A) \quad \forall \emptyset \neq A \subseteq \mathcal{Y}
\end{equation}

One may think of \( F(A) \) as a distribution function of \( Y^* \): It is monotone increasing with \( F(\emptyset) = 0 \) (since \( \emptyset \) is excluded from \( \mathcal{Y}^* \) by definition, \( \Pr(Y^* = \emptyset) = 0 \)) and \( F(\mathcal{Y}) = 1 \). Moreover, knowing \( F(\cdot) \), one gets back the probabilities of all \( y^* \in \mathcal{Y}^* \) by

\begin{equation}
\Pr(Y^* = y^*) = \sum_{y^* \subseteq y^*'} (-1)^{|y^* \setminus y^*'|} F(y^*)
\end{equation}

\footnote{This is the Möbius function from the combinatorics of partially ordered sets. See Aigner (1997: Chap. 4) who stresses the similarity with differentiation and integration and who presents many further connexions.}
This is similar to the formula determining the probability of rectangles from distribution functions in higher dimensions.

In fact, $F(.)$ enjoys a stronger monotonicity property. It is 2-	extit{monotone} in that for $A, B \subseteq \mathcal{Y}$

$$F(A \cup B) \geq F(A) + F(B) - F(A \cap B)$$

since

$$F(A \cup B) = \Pr(Y^* \subseteq A \cup B) \geq \Pr(Y^* \subseteq A \lor Y^* \subseteq B)$$
$$= F(A) + F(B) - F(A \cap B)$$

It is even 	extit{infinitely monotone} since for any $k$ and any subsets $A_1, A_2, \ldots, A_k$ of $\mathcal{Y}$

$$F\left(\bigcup_{i=1}^{k} A_i\right) \geq \sum_{\emptyset \neq I \subseteq \{1,2,\ldots,k\}} (-1)^{|I|+1} F\left(\bigcap_{i \in I} A_i\right)$$

It can be shown that all infinitely monotone set functions on $\mathcal{Y}$ with $F(\emptyset) = 0$ and $F(\mathcal{Y}) = 1$ arise from a probability distribution of a random variable $Y^*$ taking values in $\mathcal{Y}^*$.\(^6\)

This characterisation of $F(.)$ as a normed infinitely monotone set function connects the necessary consistency condition with many other concepts suggested to deal with incomplete information. In particular, such an $F(.)$ is a 	extit{belief function} in the sense of Shafer (1976). Furthermore, Strassen (1964) and Huber (1976) suggested to use either 2-monotone or infinitely monotone distribution functions to express limits of measurement precisions in statistics.\(^7\)

\(^6\) This result is known as Choquet’s theorem. See Nguyen (2006: 38) for a simple proof in the case of finite $\mathcal{Y}$. The general case is thoroughly discussed in the classical monograph of Phelps (2001) which was first published in 1965. Molchanov (2005) provides an updated review.

\(^7\) See Weichselberger (2001) and Walley (1991) for broad reviews. There is also an interesting connexion with game theory where the set of probability measures $\Pi(A) \geq F(A)$ is called the core of a game. Characterisations of the core are discussed by Chateauneuf/Jaffray (1989) from a combinatorial point of view. Wallner (2007)
2.1.4. The Geometry of Consistent Distributions

The consistency conditions in the form of (2.13) pertain only to either the distribution of the random variables \((Y, Y^*)\) or their statistical counterparts. As such, they do not encode whether in fact \(Y(.) \in Y^*(.)\) or not. They merely record distributional consequences of the consistency requirement \(Y(.) \in Y^*(.)\). But even these weak necessary conditions provide rather strong information about the set of possible distributions of \(Y\).

Looking again at the first example and writing \(p_{\{1\}}, \ldots, p_{\{1,2,3\}}\) for the density of the variable \(Y^*\), \(p_1, p_2, p_3\) for the density of the variable \(Y\), the (non-vacuous) consistency requirements become

\[
\begin{align*}
  p_{\{1\}} &\leq p_1 \leq p_{\{1\}} + p_{\{1,2,3\}} \\
  p_{\{2\}} &\leq p_2 \leq p_{\{2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\
  p_{\{3\}} &\leq p_3 \leq p_{\{3\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\
  p_{\{2\}} + p_{\{3\}} + p_{\{2,3\}} &\leq p_2 + p_3 \leq p_{\{2\}} + p_{\{3\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\
\end{align*}
\]

The last restriction is generally non-redundant since its upper bound can be smaller than the upper bound available from adding together the upper bounds for \(p_2\) and \(p_3\). Similarly, the lower bound will be larger than the lower bound implied by the second and third line.

For the next two examples, the restrictions are

\[
\begin{align*}
  p_{\{1\}} &\leq p_1 \leq p_{\{1\}} + p_{\{1,2\}} + p_{\{1,2,3\}} \\
  p_{\{2\}} &\leq p_2 \leq p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\
  p_{\{3\}} &\leq p_3 \leq p_{\{3\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \\
\end{align*}
\]

__provides a further connexion to interval probabilities where monotonicity need not hold.\__

The set function \(T(.)\) is called \textit{capacity functional} in the literature on random sets and \textit{upper probability} in the literature on imprecise probabilities. It has complementary properties to the ones of the set function \(F(.)\). Molchanov (2005: Chap. 1) provides a comprehensive review. Applied to singletons, \(T(\{y\})\) is the \textit{inclusion probability} of sampling theory.
2. Probabilistic Models of Incomplete Data

\[ p_{\{1\}} + p_{\{2\}} + p_{\{1,2\}} \leq p_1 + p_2 \]
\[ \leq p_{\{1\}} + p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \]
\[ p_{\{2\}} + p_{\{3\}} + p_{\{2,3\}} \leq p_2 + p_3 \]
\[ \leq p_{\{2\}} + p_{\{3\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \]

and

\[ p_{\{1\}} \leq p_1 \leq p_{\{1\}} + p_{\{1,2\}} + p_{\{1,2,3\}} \]
\[ p_{\{2\}} \leq p_2 \leq p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \]
\[ p_{\{1\}} + p_{\{2\}} + p_{\{1,2\}} \leq p_1 + p_2 \]
\[ \leq p_{\{1\}} + p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \]
\[ p_{\{2\}} + p_{\{2,3\}} \leq p_2 + p_3 \]
\[ \leq p_{\{2\}} + p_{\{1,2\}} + p_{\{2,3\}} + p_{\{1,2,3\}} \]

It is instructive to use a few numerical values for the distribution of \( Y^* \) to get an impression about the dependence of the range of distributions of \( Y \) implied by restrictions on the support of \( Y^* \). For the three examples, I will use the following values:

Example 1

\[ \mathcal{Y}_1^* = \{\{1\}, \{2\}, \{3\}, \{2, 3\}, \{1, 2, 3\}\} \]
\[ p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{3\}} = 0.1, p_{\{2,3\}} = 0.1, p_{\{1,2,3\}} = 0.2 \]

Example 2

\[ \mathcal{Y}_2^* = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\} \]
\[ p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{3\}} = 0.1, p_{\{1,2\}} = 0.1, p_{\{2,3\}} = 0.1, p_{\{1,2,3\}} = 0.1 \]

Example 3

\[ \mathcal{Y}_3^* = \{\{1\}, \{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\} \]
\[ p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{1,2\}} = 0.1, p_{\{2,3\}} = 0.1, p_{\{1,2,3\}} = 0.2 \]
A closer look at the first example provides further insight into the structure implied by the consistency condition. Each of the inequalities defines a closed half-space in $\mathbb{R}^3$. The set of points satisfying all the inequalities is the intersection of these half-spaces. Thus the set satisfying the consistency condition must be a polytope (Ziegler 1995: Chap. 1).\footnote{Ziegler’s book (1995) provides a vivid introduction to the theory of convex polytopes.} Since all the lower and upper bounds for $(p_1, p_2, p_3)$ are in the interval $[0,1]$, the set of points satisfying the first three constraints is a closed cuboid with sides parallel to the axes in the unit cube. The last constraint may chop off the outer edges parallel to the first axes. The polytope of solutions is depicted in Figure 2.5.

The intersection of this polytope with the simplex $p_1 + p_2 + p_3 = 1, p_i \geq 0$ gives the set of probability distributions compatible with the constraints. The intersection could only be empty if the polytope would lie completely on either side of the simplex. This can be checked by computing the vertices of the polytope from the constraints given in (2.15). Using the probabilities suggested above, the vertices are

\[p\{1\} = 0.5, p\{2\} = 0.1, p\{3\} = 0.1, p\{2,3\} = 0.1, p\{1,2,3\} = 0.2.\]

\[p\{1\} = 0.5, p\{2\} = 0.1, p\{3\} = 0.1, p\{2,3\} = 0.1, p\{1,2,3\} = 0.2.\]
Here, the first two points on the left lie below the probability simplex, the last two on the right above it. Thus the set of probability distributions subject to the constraints given by the consistency requirements (2.15) is non-empty. As an intersection of half-spaces with the probability simplex, it is a closed polytope. In particular, the solution set is convex. The set of probability distributions of \( Y \) compatible with the distribution of \( Y^* \) given above is depicted in Figure 2.6. The resulting set of probability distributions \( p_i \) is shown in Figure 2.7 together with the solution sets of the other two examples.

Geometrically, the inequalities of the consistency requirements generally lead to a closed convex polytope as the set of compatible values of the distribution of \( Y \). Requiring additionally that the values are in fact a probability distribution translates into intersecting the polytope with the simplex of probability distributions giving ones again a closed and convex polytope. The extreme points of this polytope are in Example 1
In the other two examples, the extreme points of the polytope turn out to be

\[
\begin{align*}
(0.5, 0.2, 0.3) & \quad (0.6, 0.1, 0.3) & \quad (0.6, 0.1, 0.3) & \quad (0.8, 0.2, 0.0) \\
(0.5, 0.4, 0.1) & \quad (0.7, 0.2, 0.1) & \quad (0.5, 0.2, 0.3) & \quad (0.5, 0.5, 0.0) \\
(0.7, 0.1, 0.2) & \quad (0.8, 0.1, 0.1)
\end{align*}
\]

As an immediate consequence of this geometric interpretation of the consistency conditions, a measure of the informativeness of the distribution of \(Y^*\) for the distribution of \(Y\) suggests itself: The volume of the solution set in relation to the volume of the simplex. In the three examples above, the volume of the possible distributions relative to the volume of the simplex is 8\%, 7\%, and 14\%, respectively.\(^9\) This may be compared with the simple version of missing versus exact answers in the case of the

\(^9\) The computation of the volumes depends heavily on ‘nice’ triangulations of the solution sets, and these may be difficult to find. Furthermore, constructing the solution sets becomes rather complicated both with an increasing number of constraints and an increasing number of elements of \(\mathcal{Y}\). The general formulation (2.13) suggests that the complexity may increase exponentially with the size of \(\mathcal{Y}\). Even in very restricted cases, where the number of binding inequalities may be much less than (2.13), programs such as polymake (www.math.tu-berlin.de/polymake) must be used.
National Longitudinal Survey of Youth where the observed information reduces the possible range of percentages and thus the one dimensional volume to only 18%.

With dichotomous variables, it is only possible to distinguish between either exact observations or no answer at all. In that case, the simple bounds of the introductory example characterise the set consistent distributions. But with larger numbers of states \(|\mathcal{Y}|\), the complexity of incomplete data models increases. At the same time, using the partial information (not just assuming that partial, inexact information is completely missing) may dramatically increase the information that can be extracted from the distribution on \(\mathcal{Y}^*\). To illustrate the possible gain from using also the partial information, Figure 2.8 compares the ranges of consistent distributions implied by the distributions on \(\mathcal{Y}_1^*, \mathcal{Y}_2^*, \mathcal{Y}_3^*\) with those that arise when all partial information is treated as completely missing. In the latter case, the set of consistent distributions is always a simplex whose vertices are the probabilities of the singletons in \(\mathcal{Y}^*\) where the mass of the missing data is added to each element in turn. The effect of including the partial information is to reduce the range of compatible distributions by excluding the grey areas from consideration. The gain of information (the relative reduction of the area of consistent distributions) is 12%, 28%, and 14%, respectively.

2.1.5. Constructing Consistent Distributions

The consistency conditions in form of the inequalities (2.12) or (2.13) provide a simple and intuitive view of the form of the implied set of distributions of \(Y\) as an intersection of half-spaces with the unit-simplex. Before one can use the result, however, it must be checked whether the solution set always is nonempty. That is, given a structure \(\mathcal{Y}^*\) and an arbitrary probability distribution on it, is there always at least one

to compute the solution sets. polymake, in turn, relies on cdd (ftp://ftp.ifor.math.ethz.ch/pub/fukuda/cdd/cddman/cddman.html) which may be used to compute the vertices of a polytope from its description in terms of an intersection of half-spaces as in (2.13). See the cdd home page for details on the algorithms employed.
distribution on $\mathcal{Y}$ consistent with the distribution on $\mathcal{Y}^*$? While the geometric interpretation in terms of half-spaces is suggestive, arguments in terms of explicit expressions for the vertices would make the answer obvious. But the connection between the half-space representation of polytopes and their representation as convex hull of their vertices is not immediate. Fortunately, in the case of finite $\mathcal{Y}$, a direct argument can be given: The simplest version would be to fix arbitrary numbers $p(y \mid y^*)$ such that they form, for fixed $y^*$, a probability distribution on $\mathcal{Y}$ and such that $p(y \mid y^*) = 0$ for $y \notin y^*$. One would then have to check that the implied probabilities $\Pr(Y \in A) := \sum_{y \in A} \sum_{y^* \in \mathcal{Y}^*} p(y \mid y^*) \Pr(Y^* = y^*)$ satisfy the consistency constraints (2.11) or (2.13).

While this is rather easy, a somewhat more involved construction gives probability distributions that equal the lower or upper bounds for some set $A$ and otherwise satisfy the constraints. This construction will give further insight into extreme members of the set of probabilities consistent with some distribution of $Y^*$. For a fixed set $\emptyset \neq A \subseteq \mathcal{Y}$ put

$$
\pi_A(y) := \begin{cases} 
\sum_{y^* \subseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|} & \text{for } y \in A \\
\sum_{y^* \not\subseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} & \text{for } y \notin A
\end{cases} \quad (2.16)
$$

In Example 1 with $A = \{1\}$, one obtains

$$
\pi_{\{1\}}(1) = p_{\{1\}} = 0.5
$$
\[ \pi_{\{1\}}(2) = p_{\{2\}} + \frac{p_{\{2,3\}}}{2} + \frac{p_{\{1,2,3\}}}{2} = 0.25 \]
\[ \pi_{\{1\}}(3) = p_{\{3\}} + \frac{p_{\{2,3\}}}{2} + \frac{p_{\{1,2,3\}}}{2} = 0.25 \]

and with \( A = \{2, 3\} \), the definition yields
\[ \pi_{\{2,3\}}(1) = p_{\{1\}} + p_{\{1,2,3\}} = 0.7 \]
\[ \pi_{\{2,3\}}(2) = p_{\{2\}} + \frac{p_{\{2,3\}}}{2} = 0.15 \]
\[ \pi_{\{2,3\}}(3) = p_{\{3\}} + \frac{p_{\{2,3\}}}{2} = 0.15 \]

Thus \( \pi_{\{1\}} \) attains the lower bound in the first equation in (2.15) while \( \pi_{\{2,3\}} \) attains the upper bound. Also, \( \pi_{\{2,3\}} \) attains the lower bound in the last line in (2.15) while \( \pi_{\{1\}} \) attains the upper bound. The other inequalities are fulfilled by both \( \pi_{\{1\}} \) and \( \pi_{\{2,3\}} \). The second and third sets of inequalities are attained by \( \pi_{\{2\}} \) and \( \pi_{\{3\}} \), and \( \pi_{\{1,3\}} \) and \( \pi_{\{1,2\}} \), respectively.

In the general case, it must be demonstrated that

a) (2.16) defines a probability density on \( \mathcal{Y} \) for all \( \emptyset \neq A \subseteq \mathcal{Y} \).

b) \( F(A) = \sum_{y \in A} \pi_A(y) =: \Pi_A(A) \), i.e. the bound for \( A \) in (2.13) is attained by choosing \( \pi_A \) as a density on \( \mathcal{Y} \). Here I write \( \Pi_A(.) \) for the measure corresponding to the density \( \pi_A(.) \).

c) The density \( \pi_A \) satisfies all other consistency conditions in that for all subsets \( \emptyset \neq B \subseteq \mathcal{Y} \)
\[ F(B) \leq \sum_{y \in B} \pi_A(y) = \Pi_A(B) \]

First note that
\[ \Pi_A(\mathcal{Y}) = \Pi_A(A) + \Pi_A(A^c) \]
or in terms of the density
\[ \sum_{y \in \mathcal{Y}} \pi_A(y) = \sum_{y \in A} \pi_A(y) + \sum_{y \notin A} \pi_A(y) \]
2.1. The Structure of Incomplete Data

For the first term on the right hand side

\[
\sum_{y \in A} \pi_A(y) = \sum_{y \in A} \sum_{y^* \subseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|} \\
= \sum_{y^* \subseteq A} \frac{\Pr(Y^* = y^*)}{|y^*|} \sum_{y \in y^*} 1 \\
= \sum_{y^* \subseteq A} \Pr(Y^* = y^*) = F(A)
\]

where in the second equation the order of summation is interchanged. It follows that the lower bound in the consistency condition (2.13) is attained for the set \( A \) by this construction. Thus requirement b) holds true. Further,

\[
\sum_{y \notin A} \pi_A(y) = \sum_{y \notin A} \sum_{y^* \notin A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} \\
= \sum_{y} \sum_{y^* \notin A} \mathbb{1}[A^c](y) \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} \\
= \sum_{y^* \notin A} \sum_{y} \mathbb{1}[A^c \cap y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|} \\
= \sum_{y^* \notin A} \Pr(Y^* = y^*)
\]

Therefore,

\[
\Pi_A(\mathcal{Y}) = \sum_{y \in \mathcal{Y}} \pi_A(y) = \sum_{y \in A} \pi_A(y) + \sum_{y \notin A} \pi_A(y) \\
= \sum_{y^* \subseteq A} \Pr(Y^* = y^*) + \sum_{y^* \notin A} \Pr(Y^* = y^*) = 1
\]

and, since certainly \( \pi_A(y) \geq 0 \), \( \pi_A(.) \) is a probability density on \( \mathcal{Y} \) (requirement a). It remains to show that \( \pi_A \) satisfies all the other constraints implied by the consistency conditions (2.13). Let \( \emptyset \neq B \subset \mathcal{Y} \) be
another set that might appear in the consistency conditions. Writing \( \Pi_A(B) := \sum_{y \in B} \pi_A(y) \) for the measure corresponding to \( \pi_A \),

\[
\Pi_A(B) = \Pi_A(A \cap B) + \Pi_A(A^c \cap B)
\]

Now

\[
\Pi_A(A \cap B) = \sum_{y^* \subseteq A} \sum_{y \in A \cap B \ y ^* \subseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|}
\]

\[
= \sum_{y^* \subseteq A} \sum_{y \in A \cap B} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|}
\]

\[
\geq \sum_{y^* \subseteq A \cap B} \sum_{y \in A \cap B} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|y^*|}
\]

\[
= \sum_{y^* \subseteq A \cap B} \Pr(Y^* = y^*)
\]

and

\[
\Pi_A(A^c \cap B) = \sum_{y^* \nsubseteq A} \sum_{y \in A^c \cap B \ y ^* \nsubseteq A} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|}
\]

\[
= \sum_{y^* \nsubseteq A} \sum_{y \in A^c \cap B} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|}
\]

\[
\geq \sum_{y^* \nsubseteq A} \sum_{y \in A^c \cap B} \mathbb{1}[y^*](y) \frac{\Pr(Y^* = y^*)}{|A^c \cap y^*|}
\]

\[
= \sum_{y^* \nsubseteq A \ y^* \subseteq B} \Pr(Y^* = y^*)
\]

so that

\[
\Pi_A(B) = \Pi_A(A \cap B) + \Pi_A(A^c \cap B)
\]

\[
\geq \sum_{y^* \subseteq A \cap B} \Pr(Y^* = y^*) + \sum_{y^* \nsubseteq A \ y^* \subseteq B} \Pr(Y^* = y^*)
\]
2.1. The Structure of Incomplete Data

\[ \sum_{y^* \subseteq B} \Pr(Y^* = y^*) = F(B) \]

Hence, also requirement c) is satisfied.

It follows that to every structure of incomplete data \( Y^* \) and to every distribution of the corresponding \( Y^* \) there exists a non-empty set of probability distributions on \( Y \). The set is restricted only by the distributional consequences of the consistency requirement \( Y(\cdot) \in Y^*(\cdot) \). Geometrically, the set of consistent distributions of \( Y \) is a non-empty, closed, convex polytope.

This answers one of the questions on the “simple device” formulated in the introduction, namely whether by introducing a probabilistic framework the set of underlying possible values consistent with the incomplete data is restricted beyond the consistency requirement. This does not happen, the “simple device” does not prejudice certain solutions against others simply by adding the constraints of a probability model.

2.1.6. Selectors and Allocations

But the previous construction gives only a marginal distribution consistent with the distributional consequences of the requirement \( Y \in Y^* \). What must be shown in order to be sure that there is no systematic bias introduced by adopting a probability model is that the requirement \( Y(\omega) \in Y^*(\omega) \) for almost all \( \omega \in \Omega \) is at most as strong as the distributional consistency requirement (2.13). Note that when dealing with statistical variables as functions of population members, the question is one pertaining to facts. It may be simply wrong that \( Y(u) \in Y^*(u) \) for all \( u \in U \). But given \( Y(u) \in Y^*(u) \) as a matter of fact, all (empirical) marginal distributions of \( Y \) satisfying (2.13) are possible distributions of interest with no further restrictions.

However, in a probability model one must be able to construct random variables \( (Y, Y^*) \) defined on a common probability space \( (\Omega, \mathcal{B}, \lambda) \) such that \( \Pr(Y \in Y^*) = 1 \) and such that the marginal distribution is prescribed by \( \Pr(Y^* = y^*) \) for all \( y^* \in Y^* \). If such a construction succeeds, the
constructed random variable $Y$ is called a selector of the set-valued variable $Y^*$. Thus $Y$ is a selector of $Y^*$ if and only if $Y^*$ is a coarsening of $Y$.

The construction in the previous subsection exhibits one possible distribution on $\mathcal{Y}$ consistent with the distributional consequences of the requirement $Y(\cdot) \in Y^*(\cdot)$. Moreover, $\Pi_A(A)$ attains the lower bound $F(A)$. But comparing the values of $\pi_A$ in the example with the vertices of the polytope of consistent distributions shows that these are not the extreme points of the polytope. Thus, a concise description of the polytope of consistent distributions is also still missing.

Both problems, that of the extent of the set of probability models satisfying $\Pr(Y \in Y^*) = 1$ and that of characterising the set of distributionally feasible marginal distributions, can be attacked using a very simple device. Let

$$\alpha: \mathcal{Y} \times \mathcal{P}(\mathcal{Y}) \rightarrow [0, 1]$$

such that

$$\Pr(Y^* = y^*) = \sum_{y \in y^*} \alpha(y, y^*) \quad \forall y^* \in \mathcal{Y}^*$$

(2.18)

Such a function is called an allocation in the literature on random sets. It is an obvious candidate for a joint density of $(Y, Y^*)$, at least when $\alpha(y, y^*) = 0$ for $y \notin y^*$.\(^{10}\) In fact, if one defines the marginal density of $Y$ as

$$\pi_\alpha(y) := \sum_{y^* \ni y} \alpha(y, y^*) = \sum_{y^* \in \mathcal{Y}^*} \mathbb{I}[y^*](y) \alpha(y, y^*)$$

the marginal density of $Y^*$ stays just $\Pr(Y^* = y^*)$. Furthermore,

$$\sum_{y^* \in \mathcal{Y}^*} \sum_{y \in \mathcal{Y}} \mathbb{I}[y^*](y) \alpha(y, y^*) = \sum_{y^* \in \mathcal{Y}^*} \sum_{y \in y^*} \alpha(y, y^*)$$

\(^{10}\) It is sometimes convenient not to require $\alpha(y, y^*) = 0$ for $y \notin y^*$. In that case, $\alpha$ need not be a density. It can be made into one by multiplying $\alpha$ by $\mathbb{I}[y^*](y)$ since this modification does not change the defining property (2.18). This is the reason to call $\alpha$ an allocation, not a joined density. Whether or not $\alpha(y, y^*) = 0$ for $y \notin y^*$ is assumed should be clear from the context.
Thus $\mathbb{1}[y^*](y)\alpha(.,.)$ is a joined density of $(Y, Y^*)$ with the prescribed marginal distribution of $Y^*$. Moreover,

$$
\Pr(Y \notin Y^*) = \sum_{y^* \in Y^*} \sum_{y \notin y^*} \mathbb{1}[y^*](y)\alpha(y, y^*) = 0
$$

so that $\Pr(Y \in Y^*) = 1$. In particular, the marginal measure $\Pi_\alpha$ derived from the density $\pi_\alpha$ satisfies the consistency equations (2.13) so that $F(A) \leq \Pi_\alpha(A)$ for all $A \subseteq \mathcal{Y}$.

By the above construction, there always exists a common probability space $(\Omega, \mathcal{B}, \lambda)$ for two variables $(Y, Y^*)$ such that $Y$ is a selector for $Y^*$ and $Y^*$ has any given marginal distribution. In particular, one may choose $\Omega = \mathcal{Y} \times \mathcal{P}(\mathcal{Y})$, $\mathcal{B} = \mathcal{P}(\Omega)$ and the density of $\lambda$ as $\mathbb{1}[.,.](\cdot)\alpha(.,.)$ for any allocation $\alpha(.,.)$. The pair of random variables $(Y, Y^*)$ can then be taken as the identity $(Y, Y^*)(\omega) = (Y, Y^*)((y, y^*)) := (y, y^*)$. In probability theory, such constructions of pairs of random variables on a common probability space are called couplings.\(^{11}\)

It remains to exhibit allocations and, if possible, construct some versions with properties that elucidate the structure of incomplete data problems. A simple construction starts out with any probability density, say $\pi(.)$ on $\mathcal{Y}$ such that $\pi(y) > 0$ for all $y$. Then one may put

$$
\alpha(y, y^*) := \pi(y) \frac{\Pr(Y^* = y^*)}{\Pi(y^*)}
$$

(2.19)

where $\Pi$ is the measure corresponding to the density $\pi$. Note that with this definition $\alpha(y, y^*) > 0$ for all $y$ as long as $\Pr(Y^* = y^*) > 0$. To provide a joint density, one may simply put

$$
\alpha'(y, y^*) := \mathbb{1}[y^*](y)\alpha(y, y^*)
$$

\(^{11}\) See Thorisson (2000) or Lindval (1992) who both provide lucid examples and a wealth of further constructions.
This construction will be of particular importance when different versions of the CAR condition are discussed. To illustrate it with the data from the first example, set \( \pi(y) := 1/3 \) for all \( y \in \mathcal{Y} = \{1, 2, 3\} \). Then \( \alpha \) becomes

<table>
<thead>
<tr>
<th></th>
<th>{1}</th>
<th>{2}</th>
<th>{3}</th>
<th>{2, 3}</th>
<th>{1, 2, 3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0.1</td>
<td>0.1</td>
<td>0.05</td>
<td>0.2/3</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.1</td>
<td>0.1</td>
<td>0.05</td>
<td>0.2/3</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
<td>0.1</td>
<td>0.1</td>
<td>0.05</td>
<td>0.2/3</td>
</tr>
</tbody>
</table>

which clearly is no density. But the corresponding \( \alpha' \) is given by

<table>
<thead>
<tr>
<th></th>
<th>{1}</th>
<th>{2}</th>
<th>{3}</th>
<th>{2, 3}</th>
<th>{1, 2, 3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2/3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0.05</td>
<td>0.2/3</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.05</td>
<td>0.2/3</td>
</tr>
</tbody>
</table>

which is a density with the marginal \( \pi_{\alpha'}(1) = 0.5 + 0.2/3 \) and \( \pi_{\alpha'}(2) = \pi_{\alpha'}(3) = 0.1 + 0.2/3 \) which clearly satisfies (2.15).

Thus, there always exists at least one pair of random variables \( (Y, Y^*) \) that the “simple device” requires for its operation. For finite \( \mathcal{Y} \), the existence of such a coupling satisfying the additional consistency requirement is almost trivial. It is, of course, always possible to choose an element from a finite, non-empty set as is required here.\(^\text{12}\)

But it is not yet clear how large the set of possible joined distributions is. In the case of statistical variables, it coincides with the set of consistent distributions characterised in the previous sections since \( Y(u) \in Y^*(u) \) as a matter of fact. Either it is the case that all respondents answer

\(^\text{12}\) The proof of the existence of almost sure selectors for infinite (even countable) \( \mathcal{Y} \) is much more involved. A direct construction of a selector will not work. Nevertheless, for general Polish spaces, the existence of selectors is guaranteed by the Kuratowski-Ryll-Nardzewski selection theorem. Moreover, in this setting, one has to be prepared to deal with the dependence of the set of allocations on the particularly chosen underlying probability space. Molchanov (2005: 32) provides an example. In consequence, the “simple device” has to deal with classes of selectors that are not unique. It turns out, however, that the (weak) closure of the set of selectors is in fact uniquely defined.
consistently, or it is not. But consistency is not automatically guaranteed for random variables. If the “simple device” is to work, the use of random variables \((Y, Y^*)\) should not restrict the set of compatible marginal distributions of \(Y\) for a given marginal distribution of \(Y^*\) beyond the distributional requirements (2.13). Put differently, one has to show that the marginal distributions of all possible selectors of a random set \(Y^*\) with a prescribed distribution exhausts the set of consistent distributions.

One possible approach is to investigate the set of all allocations, preferably constructing extreme ones that in some (vague) sense chart the boundary of the set of all allocations. Instead of distributing the mass \(\Pr(Y^* = y^*)\) according to some probability \(\pi\) on the elements of \(\mathcal{Y}\) one may try to distribute as much of the mass to a selected point, say \(y_{|\mathcal{Y}|}\), distribute as much as possible of the remaining mass on a second point, say \(y_{|\mathcal{Y}|-1}\), and so on. More precisely, suppose that \(\mathcal{Y}\) is ordered as \((y_1, y_2, \ldots, y_{|\mathcal{Y}|})\). Then put \(\alpha(y, y^*) := \Pr(Y^* = y^*)\) if \(y\) is the maximal element in \(y^*\) according to this ordering and set \(\alpha(y, y^*) = 0\) for all other \(y\). Then the marginal density of \(Y\) becomes

\[
\pi_a(y_i) = \sum_{y^* \ni y_i} \alpha(y_i, y^*) = \sum_{y^* \ni y_i, y^* \subseteq \{y_1, y_2, \ldots, y_i\}} \Pr(Y^* = y^*)
\]

\[
= F(\{y_1, y_2, \ldots, y_i\}) - F(\{y_1, y_2, \ldots, y_{i-1}\})
\]

where as before \(F(.)\) is the distribution function of \(Y^*\) with \(F(A) = \Pr(Y^* \subseteq A)\).

With the ordering \((1, 2, 3)\), this construction in Example 1 leads to the joint density

<table>
<thead>
<tr>
<th></th>
<th>{1}</th>
<th>{2}</th>
<th>{3}</th>
<th>{2, 3}</th>
<th>{1, 2, 3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.1</td>
<td>0.2</td>
</tr>
</tbody>
</table>

so that the marginal density becomes \(\pi_a(1) = 0.5, \pi_a(2) = 0.1, \pi_a(3) = 0.4\). Comparing this with the extreme points of the feasible distributions given in section 2.1.4, this solution is extreme in the set of consistent
distributions. In fact, computing \( \pi_\alpha \) for each permutation of (1, 2, 3) leads to all the extreme points of the polytope of consistent distributions: The ordering (2, 1, 3) leads to the same \( \pi_\alpha \) as that arising from (1, 2, 3), the orderings (1, 3, 2) and (3, 1, 2) both lead to \( \pi_\alpha = (0.5, 0.4, 0.1) \), and the orderings (2, 3, 1) and (3, 2, 1) lead to \( \pi_\alpha = (0.7, 0.1, 0.2) \) and (0.7, 0.2, 0.1).

Now the construction (2.20) depends on the ordering of the elements of \( \mathcal{Y} \). For each permutation \( \sigma \) of \( (y_1, y_2, \ldots, y_{|\mathcal{Y}|}) \) one gets an allocation and a corresponding marginal density of \( Y \) which one might index with the permutation \( \sigma \), say \( \pi_\sigma \). It can be shown that all \( \pi_\sigma \) are extreme points of the set of densities \( \pi \) on \( \mathcal{Y} \) that satisfy the consistency requirements \( F(A) \leq \Pi_\sigma(A) \) for all \( A \subseteq \mathcal{Y} \). Moreover, there are only these extreme points.\(^{13}\)

This result has many important consequences. First note that one gets an upper bound on the number of extreme points as \(|\mathcal{Y}|!\). But of much greater importance is the following consequence: The set \( \{ \Pi \mid F(A) \leq \Pi(A) \forall A \subseteq \mathcal{Y} \} \) of measures that satisfy the (distributional) consistency conditions is equal to the set of distributions that come from some allocation \( \alpha(.,.) \). This is because by construction, \( \{ \Pi_\alpha \mid \alpha \text{ is an allocation} \} \subseteq \{ \Pi \mid F(A) \leq \Pi(A) \forall A \subseteq \mathcal{Y} \} \). On the other hand, since \( \{ \Pi_\sigma \mid \sigma \} \) are the extreme points of \( \{ \Pi \mid F(A) \leq \Pi(A) \forall A \subseteq \mathcal{Y} \} \), the latter set is the convex hull \( \text{conv}(\{ \Pi_\sigma \}) \) of the extreme allocations. But convex combinations of allocations are again allocations so that also \( \{ \Pi \mid F(A) \leq \Pi(A) \forall A \subseteq \mathcal{Y} \} \subseteq \{ \Pi_\alpha \mid \alpha \text{ is an allocation} \} \). Therefore to each of the measures \( \Pi \) satisfying (2.13) there exists a pair of random variables \((Y, Y^*)\) where \( Y \) is a selector for \( Y^* \) and such that the marginal measure of \( Y \) is \( \Pi \).

This answers another possible concern with the “simple device”: Using

---


However, this construction of the vertices (extreme points) of the feasible set of marginal distributions on \( \mathcal{Y} \) is mainly of theoretical importance. It is numerically much faster and requires less memory to use the classical double description method from the theory of linear optimisation.
statistical variables, the set of possible distributions of \( Y \) consistent with
the observations \( Y^* \) is just the set of distributions satisfying \( F(A) \leq \Pr(Y \in A) \) since \( Y(u) \in Y^*(u) \) for all \( u \) as a matter of fact. But it is
now clear that the same applies to a probability model. Thus, the set of
probability models \( (\Omega, \mathcal{B}, \lambda, (Y, Y^*)) \) compatible with some prescribed
distribution of \( Y^* \) is only constrained by the same requirement that must
hold for statistical variables as well and by nothing beyond.
2. Probabilistic Models of Incomplete Data

2.2. Coarsening at Random

Problems of incomplete data and particularly the special case of either completely missing or exactly observed data have been discussed since the inception of statistics. But such problems were dealt with using either special extensions to algorithms developed within the framework of classical statistics (and typically restricted to completely missing observations) or by exploiting the structure of special types of incomplete data. The classical approach in the former tradition is well summarised in a series of articles by Afifi and Elashoff (1966, 1967, 1969). The treatment of partially complete observations can be traced back to the very first contributions to statistics in its present form. E.g., Ronald Fisher in his famous essay on maximum likelihood of 1922 deals at length with the problem of grouped data. As far as I know, however, a systematic treatment unifying the different approaches was never attempted before the 1990’s.

The introduction of the “simple device” had to wait until 1976 when Rubin’s account appeared in Biometrika. Xiao-Li Meng’s editorial for the Statistica Sinica special issue on missing data (vol. 16, no. 4, 2006) recounts details of the reservations of referees and editors before the paper was finally accepted. The resistance the paper met early on was certainly not only based on technicalities, as Meng seems to imply. As I indicated in the introduction, scepticism towards the “simple device” was widespread and well articulated.

Rubin not only gave a systematic account of missing data based on the “simple device”, he also introduced the notion of ‘missing at random’ (MAR), a condition on the joint distribution of \((Y, Y^*)\) that allows statistical methods to proceed as if no observations were missing. He also discussed the consequences of the MAR condition for inference within the different statistical paradigms. The first contribution made the “simple device” very attractive to applied statisticians, for obvious reasons. The general approach also sparked the interest of most statisticians and probabilists since it promised a principled treatment where previously ad hoc methods and subjective judgements prevailed. Thus, Rubin’s 1976
paper marks the starting point of the modern treatment of missing data problems.

Even though the “simple device” was accepted rather rapidly and even though some of the techniques developed within the framework became common practice in many applied fields, it took another 15 years until the approach was extended to partially complete observations. A first attempt at a general formulation was made by Heitjan and Rubin in 1991. They termed conditions similar to the MAR conditions coarsening at random (CAR) conditions. The notion was further developed by Heitjan (1993, 1994), Jacobsen and Keiding (1995), Gill et al. (1997), Nielsen (2000), Grünwald and Halpern (2003), Jaeger (2005a, b), Lu and Copas (2004), de Cooman and Zaffalon (2004), and by Cator (2004).\(^{14}\)

In the last section I have shown that a versatile representation of partially incomplete data can be obtained when incomplete data are taken to be subsets of the range of statistical or random variables. If a variable \(Y\) takes values in \(\mathcal{X}\), then partial knowledge of the value of the variable can be encoded by the subset of \(\mathcal{X}\) to which it necessarily belongs. Thus, incomplete data can be represented as derived variables \(Y^*\) with values in the power set \(\mathcal{P}(\mathcal{X})\), excluding the empty set. Such a representation presupposes that for all elements either of a population \(\mathcal{U}\) or a probability field \(\Omega\), \(Y(.) \in Y^*(.)\). In the case of data for a population, this consistency condition is satisfied as a matter of fact. In the case of a probability model, it was demonstrated in the previous section that the consistency requirement can always be satisfied for any given distribution of the variable \(Y^*\). Thus one can always construct an abstract probability space on which both \(Y\) and \(Y^*\) are defined. In particular, one may always assume that there is a joint distribution of \(Y\) and \(Y^*\).

The missing at random condition of Rubin as well as the general coarsening at random condition is formulated as an additional constraint

\(^{14}\) While this rather late development may seem surprising, it may have been due to the fact that partially complete data have been discussed under different headings in different applications, every discipline developing their own jargon. In fact, even today the analysis of censored, grouped, truncated or weighted data is often presented in monographs and articles accessible only to experts in a particular subject area. But this explanation is partial at best, since many unifying concepts were taken up rapidly in other areas of statistics.
on this joint distribution. It can be expressed most conveniently using conditional probabilities. One of its versions requires that for all \( y^* \) and all \( y \in y^* \):

\[
Pr(Y = y \mid Y^* = y^*) = Pr(Y = y \mid Y \in y^*)
\]

(2.21)

In words, the answers of respondents, \( y^* \), just tell that the random variable \( Y \) takes a value in \( y^* \) and nothing more. Or in Bayesian terms, the fact that \( \{ Y^* = y^* \} \) was observed allows to update probabilities only to the same extent as an observation of \( \{ Y \in y^* \} \) and does not provide any further clues.

The marginal distribution of \( Y \) can be obtained from combining the conditional distribution of \( Y \) given \( Y^* \) with the marginal distribution of \( Y^* \), i.e.

\[
Pr(Y = y) = \sum_{y^* \in Y^*} Pr(Y = y \mid Y^* = y^*) \cdot Pr(Y^* = y^*)
\]

but from (2.21), this equals

\[
Pr(Y = y) = \sum_{y^* \in Y^*} Pr(Y^* = y^*) \cdot Pr(Y = y \mid Y \in y^*)
\]

\[= \sum_{\substack{y^* \in Y^* \\ y^* \ni y \\ Pr(Y^* = y^*) > 0}} Pr(Y^* = y^*) \cdot \frac{Pr(Y = y)}{Pr(Y \in y^*)} \]  

(2.22)

since terms involving either sets \( y^* \) with \( y \notin y^* \) or with \( Pr(Y^* = y^*) = 0 \) are zero by definition. Furthermore, if \( Pr(Y^* = y^*) > 0 \), then because of (2.3), \( Pr(Y \in y^*) > 0 \) so that the last term is well defined.

Note that if \( Pr(Y^* = y^*) = 0 \) but \( Pr(Y \in y^*) > 0 \), then (2.21) may be seen as an arbitrary definition of the left hand side. If both sets \( \{ Y^* = y^* \} \) and \( \{ Y \in y^* \} \) have probability 0, then nothing is required.\(^{15}\) That is,

\[^{15}\] In particular, the explicit restriction of the CAR requirement (2.21) to sets \( \{ Y^* = y^* \} \) of positive probability as in Jaeger (2005a: 1969) is superfluous.

Also, in (2.22) one can not, in general, divide out \( Pr(Y = y) \) since it may be zero.
the CAR condition requires that the probability mass \( \Pr(Y^* = y^*) \) of a \( y^* \) is distributed among the \( \{ Y = y \} \) according to their conditional distribution given \( Y \in y^* \).

Returning to the introductory example of the National Longitudinal Survey of Youth, suppose that an answer to the employment question is either correct or is completely missing. In my notation, missing answers are indicated by the set \( y^* = \{ 0, 1 \} \). Suppose that the CAR condition (2.21) holds. Then \( \Pr(Y = y \mid Y^* = \{ 0, 1 \}) = \Pr(Y = y \mid Y \in \{ 0, 1 \}) \). But since it was supposed that all respondents are either employed or unemployed, the last quantity reduces to \( \Pr(Y = y) \). Consequently, to a Bayesian the answer \( Y^* = \{ 0, 1 \} \) would not change his prior probabilities \( \Pr(Y = y) \). And within all other approaches to statistics, the information \( Y^* = \{ 0, 1 \} \) is of no use in “inferential” procedures either. Thus, at least for the case of either completely missing or exact observations the condition (2.21) reproduces Rubin's MAR condition in that completely missing observations can safely be ignored.

In this simple example, the conditional probabilities are

\[
\begin{align*}
\Pr(Y = 1 \mid Y \in \{ 1 \}) &= 1 \\
\Pr(Y = 1 \mid Y \in \{ 0 \}) &= 0 \\
\Pr(Y = 1 \mid Y \in \{ 0, 1 \}) &= \Pr(Y = 1)
\end{align*}
\]

Thus

\[
\Pr(Y = 1) = \Pr(Y^* = \{ 1 \}) + \Pr(Y = 1) \Pr(Y^* = \{ 0, 1 \})
\]

leading to

\[
\Pr(Y = 1) = \frac{\Pr(Y^* = \{ 1 \})}{1 - \Pr(Y^* = \{ 0, 1 \})} = \frac{\Pr(Y^* = \{ 1 \})}{\Pr(Y^* \neq \{ 0, 1 \})} \tag{2.23}
\]

Similarly \( \Pr(Y = 0) = \Pr(Y^* = \{ 0 \})/ \Pr(Y^* \neq \{ 0, 1 \}) \). Within the model the probability of being employed can be deduced form the probability of reporting to be employed.
2.2.1. Does CAR Apply to Survey Non-response?

The textbook approach connecting the probability model with the sampled responses adds the assumption that the \((Y(u, .), Y^*(u, .))\) are identically and independently distributed across the population. It further adds that in some sense the mean value of observations should be close to the expectation derived from a reasonable probability model, i.e.

\[
m(Y, S) := \frac{1}{|S|} \sum_{u \in S} Y(u) \approx \mathbb{E}(Y(u_1, .)) = \Pr(Y(u_1, .) = 1) \quad (2.24)
\]

The precise nature of the connection between random variables and observations is of no immediate concern here. A different or more elaborate approach would lead to essentially the same procedure. The main point is that within the statistical approach, the search for reasonable values for \(m(Y, S)\) can be replaced by a search for a reasonable value of the expectation \(\mathbb{E}(Y(u_1, .))\) or, equivalently, for the probability \(\Pr(Y(u_1, .) = 1)\). One may proceed similarly to identify (approximately) the observed frequencies of the variables \(Y^*\) with their probability counterpart.

Since the variables \((Y(u, .), Y^*(u, .))\) are assumed to be independent and identically distributed, it is possible to drastically reduce the notational burden. The reference to the individuals \(u\) or the population \(U\) can be dropped completely. And, as is the custom in most applied statistical work, reference to the elements of the probability space \(\omega\) will also be dropped. The “simple device” leaves us with a very simple mathematical structure, namely the random tupel \((Y, Y^*)\) and its joint distribution. The original problem, that of connecting the employment rate \(m(Y, S)\) to the incomplete answers of the respondents, is now replaced by the problem of characterising the extent to which the distribution of the variable \(Y\) can be identified from knowledge of the distribution of \(Y^*\) alone.

The “simple device” of introducing a probability model plus the missing at random condition can now be translated back into an “answer” to the original question using a translation principle similar to (2.24). This
leads to the naive use of only the complete answers from the survey:

\[
\Pr(Y = 1) = \frac{\Pr(Y^* = \{1\})}{\Pr(Y^* \neq \{0, 1\})} \\
\approx \frac{\frac{1}{|S|} \sum_{u \in S} \mathbb{1}[Y^* = \{1\}](u)}{\frac{1}{|S|} \sum_{u \in S} \mathbb{1}[Y^* \neq \{0, 1\}](u)} \\
= \frac{1}{|S^*|} \sum_{u \in S^*} \mathbb{1}[Y^* = \{1\}](u)
\]

with \( S^* := \{u \mid u \in S \land Y^*(u) \neq \{0, 1\} \} \). Here, \( S^* \) is a subset of the sample with valid responses to the employment question. Using the data from the National Longitudinal Survey of Youth, one in fact arrives at a percentage of 78\% employed.

But what is the precise connection between the results of a survey and its translation into a probability model? And what justifies or at least motivates (2.24)? And what status has the CAR ‘assumption’?

The various textbook approaches are rather unclear and even potentially misleading. Even the very first step, the introduction of a probability model that somehow refers to the information provided by the interviewees, is generally glossed over. And if there are detailed accounts at all, the expositions given are neither enlightening nor cogent.

Consider Haavelmo’s influential paper “The Probability Approach to Econometrics” where he wrote

> The question is not whether probabilities exist or not, but whether—if we proceed as if they existed—we are able to make statements about real phenomena that are “correct for practical purposes”. (Haavelmo 1944: 43)\(^{16}\)

The literature on stochastic models is similarly obscure. E.g. Nelson in his introduction to “Stochastic Modeling” writes:

\(^{16}\) See Morgan (1990: Chap. 8) for an emphatic account of the papers impact on econometrics. A more critical account is given by Tryfos (2004: Chap. 7).
When the data are obtained by observing a real system, then we can treat the data as if they were simulation output. There is no need to differentiate “real” data from the output of a simulation. … Probability … is useful for deriving statements about the data that a completely specified model might generate if it were simulated. (Nelson 2002: 24)

And Bartholomew in his equally influential “Stochastic Models for Social Processes” writes:

Our contention is not that the employee actually uses a chance device to make the decision but that the group behaves as if its individual members did use such a method. The function of probability theory is thus simply to describe observed variability; it carries no implications about the freedom, or otherwise, of human choice. It is a fact of experience that ‘choice may mimic chance’. (Bartholomew 1967: 6)

The ‘as if’ rhetoric dominated also much of the history of probabilistic reasoning in the social sciences. Historically, the domination of the ‘as if’ rhetoric may not be too surprising in view of the roots of social statistics in the 19th century, where it was assumed by many that nearly all social phenomena might be amenable to probabilistic analysis.

The philosophical literature on models, while quite extensive, is not very useful either. Even those works directly dealing with statistics like

---

18 Hacking’s (1988) account of the origin of experimental design and randomisation from curiosity in telepathy provides an interesting case study of the sort of uses made of probability models in the late 19th century.
19 Frigg and Hartmann (2005) provide a short review. A somewhat more relevant view is expressed by Knuuttila and Voutilainen (2005). The classical study of the ‘as if’ rhetoric is Vaihinger’s “Philosophie des Als-Ob” (1911). He clearly distinguishes between ‘hypotheses’ and ‘fiction,’ writing “Der Verifizierung der Hypothese entspricht die Justifizierung der Fiktion. Muß jene durch Erfahrung bestätigt werden, so muß diese gerechtfertigt werden durch die Dienste, welche sie der Erfahrungswissenschaft schließlich leisten.” (1911/1923: 91). But he indiscriminately treats all forms of ‘fiction’ alike, whether it be mathematical or juridical or philosophical concepts. He therefore is unable to explain the role of
Lenhard (2006) take the probability model for granted. In the statistics literature, when models are discussed at all, then the role of probability is not questioned (e.g. Cox 1990 or Lehmann 1990).

On the other hand, survey statisticians were always aware of the problematic connection between the real answers of interviewees and probabilistic models of them. Tore Dalenius’ remark cited in the introduction voices that concern.

Many probabilists were also rather dismayed by the obscure ‘as if’ rhetoric. Even Kolmogorov in a review chapter that was first published in 1956 stated that

\[
\text{…there exists no event which is absolutely random; an event is random or is predetermined depending on the connection in which it is considered, but under specific conditions an event may be random in a completely non-subjective sense, i.e., independently of the state of knowledge of any observer. (Kolmogorov 1999: 249)}
\]

While Kolmogorov denounced the observer dependence of randomness but stressed that at least the range of “events” considered has to be taken into account, Jaynes insists on the dependends of the notion on “human information”:

Belief in the existence of ‘stochastic processes’ in the real world; i.e. that the property of being ‘stochastic’ rather than ‘deterministic’ is a real physical property of a process, that exists independently of human information, is another example of the mind projection fallacy: attributing one’s own ignorance to Nature instead. (Jaynes 2003: 506)

Matheron, whose “Random Sets and Integral Geometry” provided much inspiration for the treatment of incomplete data, also rejects the ‘as if’ rhetoric:

When we deal with a unique phenomenon and a probabilistic model, that is a space \((\Omega, \alpha, P)\) which is put in the ‘as if’ rhetoric within a model.
correspondences with this unique reality, the same kind of illusion incites us to say that everything happens, after all, as if the realised event had been “drawn at random” according to law $P$ in the sample space $\Omega$. But this is a misleadingly clear statement, and the underlying considerations supporting it are particularly inadequate. What is the mechanism of this “random choice” that we invoke …? This “random draw” myth, for it is one, (in the pejorative sense), is both useless and gratuitous. (Matheron 1989: 23)

He rather emphasises the role of models and their uses and states:

In fact there is not, nor can there be, any such thing as probability in itself. There are only probabilistic models. In other words, randomness is in no way a uniquely defined, or even definable property of the phenomenon itself. It is only a characteristic of the model or models we choose to describe it, interpret it, and solve this or that problem we have raised about it. (Matheron 1989: 4)

Similarly, Dawid (2004: 44) argued:

I regard “probability” as a purely theoretical term, inhabiting the intellectual universe and without any direct physical counterpart… [W]e should regard probabilities as entering our scientific theories as instrumental terms, the link between theoretical probabilities and the physical universe being indirect. This approach to interpreting probabilistic models avoids many potential philosophical pitfalls. In particular, by treating probabilities as purely theoretical terms with only indirect implications for the behavior of observables, it is able to eschew deep but ultimately irrelevant and distracting philosophical inquiry into the “true nature of Probability”.

This emphasis on models allows to see the role of probability in a different light, transcending the objectivist-subjectivist divide. Matheron distinguishes several steps in the choice of a model:
[T]here is first an epistemological choice: it has been decided to use probabilistic techniques to represent the phenomenon.... This is a decision, not a hypothesis. It is a constitutive decision. (It ‘constitutes’ the forest as an object of study, it defines the general framework within which we shall operate and determines the choice of the tools we use). ... At this level, we shall speak of a constitutive model.... (Matheron 1989: 52)

Therefore, statements made within a probabilistic framework do not directly refer either to social facts or to states of mind. They only work within a probability model. And the latter is not forced upon us by some obscure ‘as if’ similarity but created by a decision to use such a model. In fact, the classical theory of sampling does not rely on a probability model for the variables (Y, Y*). Rather, these variables are treated as fixed though unknown quantities ascribable to the individuals in the population. And probability only appears in the picture in as far as it is introduced by the researcher himself, by deliberately using random sampling designs.

There are some further elements of probability models used in the social sciences that can be regarded as constitutive in the sense of Matheron. In particular, in the context of surveys this is the assumption of independence and of identical distribution. They determine the choice of statistical and probabilistic tools but are not determined from any aspect of social facts.\(^\text{20}\) As Kolmogorov (1950: 9) noted: “…one of the most important problems in the philosophy of the natural sciences is … to make precise the premises which would make it possible to regard any given real events as independent.” It is argued here that it is more fruitful not to search for such premises but to take them as part of the decision to use a particular type of model. In fact, classical statistical tests of the independence assumption are self-refuting when carried out on a given data set: If a test does not reject independence, then there must be some

\(^{20}\) Recently, Humphreys (2008) argued similarly that the basic probability setup is a purely mathematical artefact, a mathematical template. He also counts probability models and classes of distributions as such templates. It seems to me, however, that he overlooked the fundamental role of independence assumptions.
dependence in the joint distribution of the data given the observed value of the test statistic.\footnote{See Hennig (2007) for some further comments and Leeb/Pötscher (2006) on some general impossibility results for model selection and the distribution of estimators conditional on specification tests.}

From this perspective one needs to reconsider the approximation suggested in (2.24). In so far as the left hand side is treated as a function whose behaviour is determined within the model, it is plain that it can only be correct when several caveats are added. In particular, if $\mathcal{Y}$ is finite, then (2.24) is true, but only with high probability. That is, (2.24) can only be expressed when probability is presupposed in the expression of the approximation.

If $\mathcal{Y}$ is unbounded, then (2.24) is not even true in probability since for any sample size $|S|$, there exist distributions such that the difference between the two sides can be made as large as one pleases. This is true even if the existence of all moments or similarly strong regularity conditions are stipulated. Simply put $Y = 0$ with probability $\epsilon$, but $Y = (1/\epsilon)^{k+1}$. Then $\mathbb{E}(Y) = (1/\epsilon)^k$. But with $\epsilon \ll 1/|S|$, most samples will contain only zeros, thus being in error by $(1/\epsilon)^k$. Or it does contain at least a value of $(1/\epsilon)^{k+1}$, then the estimated mean on the left is much larger than the “true” value. Examples like these show that there are (within a probability model) no consistent estimators, tests or confidence intervals of expectations when all probability models are taken into account, even within an i.i.d. setup.\footnote{See Bahadur/Savage (1956), Gleser/Hwang (1987), Lehmann/Loh (1990), Pfanzagl (1998), and Romano (2004) for further details.}

This may seem to contradict traditional wisdom from the statistical textbooks. But the textbooks rely on very narrow ‘assumptions’ that provide optimality results at fixed parameter points. Uniformity of results for all sets of parameters is rarely considered.

Motivated by similar examples, Davies (1995) argued to acknowledge the “approximate nature of probability models” by taking all probability models as adequate models of a data set if data simulated from the...
models are “very much like the sample actually obtained”\textsuperscript{23} If this is to work in the above example, one has to insist on a known bound on the support of the probability model. Otherwise, any value of an expectation (or any other moment) would be an ‘adequate’ model for any given data set.\textsuperscript{24}

What is expressed by (2.24) then is not a truth deducible within the probability model, i.e. when the left hand side refers to a realisation of the model variables. But it is also obviously wrong when the left hand side is taken to be the empirical mean over some sample, the right hand side is interpreted as a mean over the population and nothing is ‘assumed’ about the values in the population.\textsuperscript{25} In contrast to both positions, the one that makes (2.24) a consequence of the model and the one that naively sees it as an empirical justification of probability models, it seems better to see (2.24) as an expression of a further constitutive element of a particular type of a probabilistic model. It expresses the intent to treat empirical means to gauge the performance of subsets of probability models. It singles out a certain ‘estimation principle’ that can be used to criticise a given probabilistic model. But it in no way implies that it is the only way to do so. Nor does its acceptance imply that there is some ‘true’ model.

The CAR condition (2.21), on the other hand, can not be used in such a way simply because there is no empirical way to express it. It is also not constitutive in Matheron’s sense. It singles out one particular set of solutions from all probability distributions based on a probability concept, at least when (2.24) is accepted. But it can not, as an ‘assumption’, justify that particular choice. Nor does it constitute the form of models or the set of tools that are to be used within a certain type of models. And it does not provide means to argue about the model. It rather provides a reference point for further speculations about the situation within a

\textsuperscript{23} Note that this reverses the roles of probability models and observations as expressed by Nelson(2002) cited above.

\textsuperscript{24} See also some further comments in section 4.

\textsuperscript{25} It is at most a different figure of speech if ‘population’ here is interpreted within a probability model as a naive frequentist might suggest, whether directly referring to an ‘infinite population’ or via a super-population model. Since limiting frequencies do not express anything about any finite subsequence, counterexamples are easily produced.
probability model. Such speculations cannot be completely arbitrary, they are at least constrained by the mathematical rules of probability. But they are clearly different from assumptions about reality as well.

2.2.2. CAR in Hierarchically Structured Models

Before looking into the use of such principled speculations I still have to show whether the CAR formulation (2.21) is general enough to deal with all types of incomplete data beyond the simple case of either completely missing or exact observations.

Suppose that there is some partial information on \( Y \) beyond that of \( Y \) being either completely missing (\( Y^* = \mathcal{Y} \)) or exactly given (\( Y^* = \{y\} \)) for some \( y \in \mathcal{Y} \). Example 1 is a case in point since it differentiates between ‘employed’, ‘unemployed’, and ‘out of labour force’ (coded as \( \mathcal{Y} = \{1, 2, 3\} \)) and allows for partial answers: if a respondent does not differentiate between ‘unemployed’ and ‘out of labour force’, this answer is coded as \( y^* = \{2, 3\} \). In this case, the information \( y^* = \{2, 3\} \) cannot simply be ignored. But the CAR condition now says that the only information supplied by the respondent is that her employment status is in fact either ‘unemployed’ or ‘out of labour force’ and must be treated as such. There is then no need to contemplate why she has chosen to answer in the way she did, at least not as long as one is interested in the probabilities of ‘employment’, ‘unemployment’, and ‘out of labour force’.

Using the formulation (2.21) in conjunction with (2.22) and the previously defined short hand notation, plugging in the values for \( \Pr(Y^* = .) \) from Example 1, one arrives at:

\[
\begin{align*}
\hat{p}_1 &= p_{\{1\}} + p_{\{1,2,3\}}p_1 = 0.5 + 0.2p_1 \\
\hat{p}_2 &= p_{\{2\}} + p_{\{2,3\}} \frac{p_2}{p_2 + p_3} + p_{\{1,2,3\}}p_2 \\
&= 0.1 + 0.1 \frac{p_2}{p_2 + p_3} + 0.2p_2 \\
\hat{p}_3 &= p_{\{3\}} + p_{\{2,3\}} \frac{p_3}{p_2 + p_3} + p_{\{1,2,3\}}p_3 
\end{align*}
\]
$= 0.1 + 0.1 \frac{p_3}{p_2 + p_3} + 0.2p_3$

The first equation is easily solved, giving $p_1 = 5/8 = 0.625$. From this it follows that $p_2 + p_3 = 3/8$ so that

$$p_2 = 0.1 + \frac{0.1}{3/8}p_2 + 0.2p_2$$
$$p_3 = 0.1 + \frac{0.1}{3/8}p_3 + 0.2p_3$$

giving $p_2 = p_3 = 3/16 = 0.1875$.

This example, while exceedingly simple, is typical for hierarchically structured incomplete data. In this case, one can always proceed from top down through the hierarchy of partitions to provide simple linear equations for the probabilities of $Y$ in subsets of $\mathcal{Y}^*$. To see this, note that for a subset $A \in \mathcal{Y}^*$

$$\Pr(Y \in A) = \sum_{y \in A} \Pr(Y = y)$$
$$= \sum_{y \in A} \sum_{y^* \in \mathcal{Y}^*} \Pr(Y = y \mid Y^* = y^*) \Pr(Y^* = y^*)$$
$$= \sum_{y \in A} \sum_{y^* \in \mathcal{Y}^*} \mathbb{1}[y^*](y) \Pr(Y = y \mid Y^* = y^*) \Pr(Y^* = y^*)$$
$$= \sum_{y^* \in \mathcal{Y}^*} \sum_{y \in A} \mathbb{1}[y^*](y) \Pr(Y = y \mid Y \in y^*) \Pr(Y^* = y^*)$$
$$= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y \in A \cap y^* \mid Y \in y^*) \Pr(Y^* = y^*)$$
$$= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \frac{\Pr(Y \in A \cap y^*)}{\Pr(Y \in y^*)} \Pr(Y \in y^*) > 0$$

(2.25)

generalising (2.22). Since $\{Y^* = y^*\} \subseteq \{Y \in y^*\}$, the fractions in the last expression are well defined. Note that (2.25) can be written
alternatively as
\[
\Pr(Y \in A) = \sum_{y^* \in \mathcal{Y}^* \atop \Pr(Y^* = y^*) > 0} \Pr(Y^* = y^* | Y \in y^*) \Pr(Y \in A \cap y^*) \tag{2.26}
\]

In the hierarchically structured case, i.e. when $$\mathcal{Y}^*$$ is the union of refinements of a partition, if $$A$$ is an element of $$\mathcal{Y}^*$$ and if $$y^* \cap A \neq \emptyset$$ for some $$y^*$$, then either $$y^* \subseteq A$$ or $$A \subseteq y^*$$. Now if $$y^* \subseteq A$$, then the fraction on the right hand side is 1. That is, at most terms of a coarser partition than that to which $$A$$ belongs contribute non-trivially to the sought for probabilities $$\Pr(Y \in A)$$.

Starting at the coarsest non-trivial partition (i.e. ignoring the uninformative level comprising all of $$\mathcal{Y}$$, the set indicating completely missing information) in $$\mathcal{Y}^*$$, there are subsets $$A_1, \ldots, A_k \in \mathcal{Y}^*$$ forming a partition of $$\mathcal{Y}$$. For these sets, (2.25) takes the form
\[
\Pr(Y \in A_i) = \sum_{y^* \in \mathcal{Y}^* \atop \Pr(Y^* = y^*) > 0} \Pr(Y^* = y^*) \frac{\Pr(Y \in A_i \cap y^*)}{\Pr(Y \in y^*)} = \Pr(Y \in A_i) \Pr(Y^* = \mathcal{Y}) + \sum_{y^* \subseteq A_i} \Pr(Y^* = y^*)
\]
resulting in the explicit solution
\[
\Pr(Y \in A_i) = \frac{\sum_{y^* \subseteq A_i} \Pr(Y^* = y^*)}{\Pr(Y^* \neq \mathcal{Y})}
\]
which is always defined unless $$\Pr(Y^* = \mathcal{Y}) = 1$$. But in the latter case, obviously no information about $$\Pr(Y \in \cdot)$$ can be derived from the distribution of $$Y^*$$.

Proceeding down the hierarchy one level, one finds a refinement of the partition $$\{A_1, \ldots, A_k\}$$, say $$\{A_{1i}, \ldots, A_{ki}, \ldots, A_{ki}, \ldots, A_{kk}\}$$. But for this refinement, the denominators of the fractions $$\Pr(Y \in A_{ij} \cap y^*)/\Pr(Y \in y^*)$$ are either known from the previous computation (when
where it is assumed that \( \Pr(Y \in A_i) > 0 \) (otherwise, the term is simply dropped from the sum). But the denominator of the second term, \( \Pr(Y \in A_i) \), is known from the previous step. Once again, the resulting simple linear equations can be solved directly and one may proceed further down to the finest partition.

Consequently, the CAR condition together with the consistency condition reduces to a sequence of equations that can be solved recursively. In particular a CAR model always exists and is unique down to the finest partition of \( \mathcal{Y} \), i.e. the distribution of \( Y, \Pr(Y \in \cdot) \), can be determined for all sets in \( \mathcal{Y}^* \). If the latter contains all singletons with positive probabilities, then the distribution of \( Y \) is uniquely identified from the coarsened data together with the CAR condition.

This will be true even when the information from the distribution of \( Y^* \) is extremely scarce. Suppose that \( \Pr(Y^* = \mathcal{Y}) = 1 - \epsilon \) and \( \Pr(Y^* = \{y_i\}) = \epsilon_i \) with \( \sum_i \epsilon_i = \epsilon \) and \( \epsilon, \epsilon_i > 0 \) for some small \( \epsilon \). Then the set of consistent distributions of \( Y \) is a simplex whose volume is arbitrarily close to the full probability simplex. In other words, the consistency conditions do not restrict the set of consistent probabilities beyond their being positive. But even in this case, the CAR condition leads to a unique result. The effectiveness of the CAR condition to single out one particular distribution from the many possible ones simply does not depend on the informativeness of the distribution of \( Y^* \).

A noteworthy further consequence of the CAR condition in the hierarchical case is that it leads to a rational solution when the known \( \Pr(Y^* = y^*) \)
are rational. Thus, if the probabilities of the coarsened sets are taken to be relative frequencies, the CAR solution can be interpreted as a relative frequency as well. In particular, the construction does not necessarily lead beyond an interpretation based on the classical sampling theory.

From the CAR condition (2.21) and the marginal distribution of $Y^*$ one can also write down the joint distribution of $(Y, Y^*)$. The joined density follows from

$$
\Pr(Y = y, Y^* = y^*) = \Pr(Y = y \mid Y \in y^*) \Pr(Y^* = y^*)
$$

$$
= \prod [y^*(y)] \Pr(Y^* = y^*) \frac{\Pr(Y = y)}{\sum_{y \in y^*} \Pr(Y = y)}
$$

(2.27)

for $\Pr(Y^* = y^*) > 0$. Note that the consistency condition (2.3) implies that the denominator in the fraction above is $> 0$. Furthermore, $\Pr(Y = y, Y^* = y^*) = 0$ for $\Pr(Y^* = y^*) = 0$. Note also that this is an allocation of the form given in (2.19).

In Example 1, the joined density is readily seen to be

<table>
<thead>
<tr>
<th></th>
<th>{1}</th>
<th>{2}</th>
<th>{3}</th>
<th>{2, 3}</th>
<th>{1, 2, 3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/8</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/10</td>
<td>0</td>
<td>1/20</td>
<td>3/80</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1/10</td>
<td>1/20</td>
<td>3/80</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>1/10</td>
<td>1/10</td>
<td>1/10</td>
<td>2/10</td>
</tr>
</tbody>
</table>

2.2.3. Non-Hierarchical CAR

Suppose next that $Y^*$ is no longer hierarchical. A simple example is given by my Example 2 where

$$
Y^*_2 = \{\{1\}, \{2\}, \{3\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}
$$

$$
p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{3\}} = 0.1, p_{\{1, 2\}} = 0.1, p_{\{2, 3\}} = 0.1, p_{\{1, 2, 3\}} = 0.1
$$


Here the equations (2.22) become

\[
\begin{align*}
  p_1 &= 0.5 + 0.1 \frac{p_1}{p_1 + p_2} + 0.1p_1 \\
  p_2 &= 0.1 + 0.1 \frac{p_2}{p_1 + p_2} + 0.1 \frac{p_2}{p_2 + p_3} + 0.1p_2 \\
  p_3 &= 0.1 + 0.1 \frac{p_3}{p_2 + p_3} + 0.1p_3
\end{align*}
\]

where the restriction \( p_1 + p_2 + p_3 = 1 \) was used in the last terms. These equations are no longer linear. Multiplying through by \( p_1 + p_2 \), \((p_1 + p_2)(p_2 + p_3)\), and \( p_2 + p_3 \), one is led to a system of polynomial equations which in this case simplifies because the first and last equations depend only on \( p_1, p_2 \) and \( p_2, p_3 \), respectively. In fact, the system can be triangularised resulting in\(^{26}\)

\[
\begin{align*}
  243p_3^3 - 261p_3^2 + 73p_3 - 6 &= 0 \\
  7p_2 - 27p_3^2 + 33p_3 - 6 &= 0 \\
  7p_1 - 162p_3^2 + 135p_3 - 22 &= 0
\end{align*}
\]

The first equation has the three real solutions \( p_3 = 2/9 \) or \( p_3 = (23 \pm \sqrt{205})/54 \). Plugging the rational solution 2/9 into the second and third equation forces \( p_1 = p_2 = 0 \) so that this is no probability distribution. Using \((23 + \sqrt{205})/54\) gives a negative solution in the second equation. Only the last solution, which obviously is irrational, provides a solution that also respects the restrictions of a probability model. Thus, the solution is still unique and is given by \( p_1 = (95 - \sqrt{205})/126 \approx 0.64033 \), \( p_2 = (5\sqrt{205} - 34)/189 \approx 0.19888 \), \( p_3 = (23 - \sqrt{205})/54 \approx 0.16078 \).

\(^{26}\) There is a further triangularisation which, however, violates the requirement \( p_1 + p_2 + p_3 = 1 \). The solution was computed using the Singular package (http://www.singular.uni-kl.de) using lexicographical ordering of the monomial terms. Any other software that allows to compute Gröbner bases and resultants can obviously be used as well. Further pertinent free software packages for such computations include Macaulay 2 (http://www.math.uiuc.edu/Macaulay2/) and CoCoA (http://cocoa.dima.unige.it/). Another program that uses homotopy methods is PHCpack (http://www.math.uic.edu/~jan/download.html).
The elements of the solution vector, however, are no longer rational numbers.\(^{27}\) Consequently, there is no sample size and there are no relative frequencies such that the CAR requirement (2.21) holds exactly, whatever the number of observations.

If the CAR condition is supposed to hold, however, the implied joined density of \((Y, Y^*)\) is given (using (2.27)) by


density = \begin{bmatrix}
{1} & {2} & {3} & {1, 2} & {2, 3} & {1, 2, 3} \\
1 & 1/2 & 0 & 0 & 0.0763 & 0.0640 & 0.6403 \\
2 & 0 & 1/10 & 0 & 0.0237 & 0.0553 & 0.0199 & 0.1989 \\
3 & 0 & 0 & 1/10 & 0 & 0.0447 & 0.0161 & 0.1608 \\
1/2 & 1/10 & 1/10 & 1/10 & 1/10 & 1/10 & 1/10
\end{bmatrix}

In the case of Example 3, where

\[ Y_3^* = \{\{1\}, \{2\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\} \]

\[ p_{\{1\}} = 0.5, p_{\{2\}} = 0.1, p_{\{1, 2\}} = 0.1, p_{\{2, 3\}} = 0.1, p_{\{1,2,3\}} = 0.2 \]

the condition (2.22) gives

\[
\begin{align*}
p_1 &= 0.5 + 0.1 \frac{p_1}{p_1 + p_2} + 0.2p_1 \\
p_2 &= 0.1 + 0.1 \frac{p_2}{p_1 + p_2} + 0.1 \frac{p_2}{p_2 + p_3} + 0.2p_2 \\
p_3 &= 0.1 \frac{p_3}{p_2 + p_3} + 0.2p_3
\end{align*}
\]

If one adds the requirement \(p_1 + p_2 + p_3 = 1\) to the equations, the triangularised set of equations becomes

\[
\begin{align*}
16p_3^2 - 9p_3 &= 0 \\
7p_2 + 9p_3 - 2 &= 0
\end{align*}
\]

\(^{27}\) Whether there is an algorithm to decide whether a given system of polynomials with rational coefficients has rational solutions is Hilbert’s 10th problem. Such an algorithm does not exist for integer solution. Whether there is one for rational solutions is still unknown.
\[ p_1 + p_2 + p_3 - 1 = 0 \]

where the two real solutions to the first equation are \( p_3 = 0 \) and \( p_3 = 9/16 \). The second solution leads to a negative value for \( p_2 \). Thus the solution is given by \( p_1 = 5/7 = 0.71429, p_2 = 2/7 = 0.28571, p_3 = 0 \) which is now on the boundary of the distributions compatible with the consistency constraints, though again rational.

The joined density of \((Y, Y^*)\) for Example 3 is given by

<table>
<thead>
<tr>
<th></th>
<th>{1}</th>
<th>{2}</th>
<th>{3}</th>
<th>{1, 2}</th>
<th>{2, 3}</th>
<th>{1, 2, 3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/2</td>
<td>0</td>
<td>0</td>
<td>1/14</td>
<td>0</td>
<td>2/14</td>
</tr>
<tr>
<td>( p_1 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/10</td>
<td>0</td>
<td>1/35</td>
<td>1/10</td>
<td>2/35</td>
</tr>
<tr>
<td>( p_2 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( p_3 )</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>1/10</td>
<td>0</td>
<td>1/10</td>
<td>2/10</td>
<td></td>
</tr>
</tbody>
</table>

In general, non-hierarchical coarsening patterns will lead to polynomial equations for the probabilities of interest. One may try to gain insight in the solutions of such systems by using the methods of Sturmfels (2002) and Cox et al. (2002).\(^{28}\) An interesting method that uses only the structure of \( Y^* \) to bound the number of solutions from above is related to the (mixed) volume of certain polytopes.\(^ {29} \)

As far as I can see, the bound is of no immediate help for the solution of CAR equations. But the polytopes provide another graphical method represent the structure of \( Y^* \). To illustrate, one can write out the system of polynomial equations for Example 2, this time without using \( p_1 + p_2 + p_3 = 1 \) and multiplying with all denominators. This leads to the following three polynomials:

\[
\begin{align*}
p_1^3 + 2p_1^2p_2 + p_1^2p_3 - 0.7p_1^2 + p_1p_2p_3 + p_1p_2^2 - 1.2p_1p_2 - 0.6p_1p_3 \\
- 0.5p_2^2 - 0.5p_2p_3 \\
p_2^4 + 2p_2^3p_3 + 2p_1p_2^3 - 0.2p_2^3 + p_2^2p_3^2 + 3p_1p_2^2p_3 - 0.6p_2^2p_3 + p_1^2p_2^2
\end{align*}
\]

\(^{28}\) A short introduction with a view towards applications in statistics is Pistone et al. (2001).

\(^{29}\) Rojas (2003) provides an accessible introduction.
- 0.6p_1p_2^2 + p_1p_2p_3^2 - 0.2p_2p_3^2 + p_1^2p_2p_3 - 0.6p_1p_2p_3 - 0.2p_1^2p_2
- 0.1p_1^2p_3^2 - 0.1p_1^2p_3

p_3^3 + 2p_2p_3^2 + p_1p_3^2 - 0.1p_3^2 + p_2^2p_3 + p_1p_2p_3 - 0.2p_2p_3 - 0.2p_1p_3
- 0.1p_2^2 - 0.1p_1p_2

The exponents of the monomial terms of the polynomials encode to how many of the sets in \( \mathcal{Y}^* \) a particular \( y \) belongs. They also encode how often a pair \((y, y')\) is a subset of the sets in \( \mathcal{Y}^* \) etc. Writing down the exponents as vectors of integers in the order \( p_1, p_2, p_3 \) gives the following three sets of points:

\[
A_1 = \{ (3, 0, 0), (2, 1, 0), (2, 0, 1), (2, 0, 0), (1, 1, 1), (1, 2, 0), (1, 1, 0), (1, 0, 1), (0, 2, 0), (0, 1, 1) \}
\[
A_2 = \{ (0, 4, 0), (0, 3, 1), (1, 3, 0), (0, 3, 0), (0, 2, 2), (1, 2, 1), (0, 2, 1), (2, 2, 0), (1, 2, 0), (1, 1, 2), (0, 1, 2), (2, 1, 1), (1, 1, 1), (2, 1, 0), (1, 0, 2), (2, 0, 1) \}
\[
A_3 = \{ (0, 0, 3), (0, 1, 2), (1, 0, 2), (0, 0, 2), (0, 2, 1), (1, 1, 1), (0, 1, 1), (1, 0, 1), (0, 2, 0), (1, 1, 0) \}
\]

Taken as points in \( \mathbb{R}^3 \) with integer coordinates, one may construct the convex hull of the points in \( A_1, A_2, A_3 \). These are the Newton polytopes of the respective polynomials. Now the Newton polytope of the product of polynomials is just the Minkowski sum of the Newton polytopes. Thus in so far as the Newton polytope of a polynomial encodes information about the zeros of the polynomial, all Minkowski sums that can be constructed from the three sets should give information on the zeros of the system of equations. Figure 2.9 demonstrates the constructions.\(^\text{30}\)

\(^\text{30}\) The mixed volume in this case is defined as \( \text{MV}(A_1, A_2, A_3) = \text{Vol}(\text{conv}(A_1) + \text{conv}(A_2) + \text{conv}(A_3)) - \text{Vol}(\text{conv}(A_1) + \text{conv}(A_2)) - \text{Vol}(\text{conv}(A_1) + \text{conv}(A_3)) - \text{Vol}(\text{conv}(A_2) + \text{conv}(A_3)) + \text{Vol}(\text{conv}(A_1)) + \text{Vol}(\text{conv}(A_2)) + \text{Vol}(\text{conv}(A_3)) \). A multiple of this number bounds the number of the non-zero isolated complex solutions to the system of polynomial equations. But what is needed here, of course, is a bound on the number of real solutions in the probability simplex. It is still unclear whether the geometric methods can provide insight into this problem.
The actual solution to the CAR equations, however, depends not only on the structure of the set $Y^*$ but also on the coefficients $\Pr(Y^* = y^*)$. It is, moreover, not obvious how to include the inequalities of the consistency requirements into the framework of systems of polynomial equations. One may add slack variables to represent the constraints. But as far as I can judge from simple examples, such a move tends to complicate matters in many cases without any real gain of insight.

\[\text{Figure 2.9: The Newton polytope } \text{conv}(A_1) \text{ and the Minkowski sum } \text{conv}(A_1) + \text{conv}(A_2) + \text{conv}(A_3) \text{ from Example 2.}\]

2.2.4. CAR is Everything

Numerical solutions are much easier (though less explicitly) computed using (2.22) directly. After all, that equation expresses a fixed point property which may be used as a recipe to compute a stationary point satisfying both the consistency requirements and the additional CAR condition. Starting from any assignment of positive values to $\Pr(Y \in \cdot.)$ one can use the right hand side of (2.22) to update the probability so that the consistency constraints are always satisfied. The procedure can be iterated and the solutions will converge to a fixed point of the equations.
In section 2.1.6 it was shown that any choice of a probability measure with \( \Pr(Y = y) > 0 \) for all \( y \) will give rise to an allocation respecting the consistency requirements via (2.22). In fact, the identity (2.22) is just the marginal distribution of the selector corresponding to the allocation (2.19). Starting values for iterated (2.22) can therefore be chosen more generally from all selectors of allocations.

Note that the equations (2.22) are the self-consistency equations introduced by Efron (1967) and discussed by Tsai and Crowley (1985). Since

\[
\Pr(Y = y) = \mathbb{E}_{Y^*} \left( \mathbb{E}_{Y|Y^*} \left( \mathbb{I}[Y = y] \mid Y^* \right) \right)
\]

by (2.21) one gets

\[
\Pr(Y = y) = \mathbb{E}_{Y^*} \left( \Pr(Y = y \mid Y^*) \right) = \mathbb{E}_{Y^*} \left( \Pr(Y = y \mid Y \in Y^*) \right) = \sum_{\substack{y^* \in Y^* \\text{if } \exists y \in Y \text{ s.t. } \Pr(Y = y) > 0}} \Pr(Y^* = y^*) \frac{\Pr(Y = y)}{\Pr(Y \in y^*)}
\]

Turning to the existence of a solution, note that the right hand side of (2.22) is a continuous function of \( \Pr(Y = y) \) since the denominators in the equation are strictly positive and will stay so by the consistency requirement. Moreover, the image under this map of any consistent probability on \( Y \) (an element of the compact convex set of consistent

---

31 Numerical applications of the algorithm provide only approximations to actual fixed points. The numerical solutions may not even be close to actual solutions of (2.22). As usual, results must be checked by other means.

probabilities) is necessarily again an element of the set of consistent probabilities. Therefore, there always is at least one solution to the equations (2.22).\(^{33}\)

Because there is always a solution satisfying CAR for any given distribution on \(\mathcal{Y}^*\), CAR can not be ruled out empirically. Gill et al. (1997) summarised the finding with the slogan ‘CAR is everything.’ It is always possible to take the coarsened data at face value in the sense of the CAR condition (2.21). Put still differently, no distribution on \(\mathcal{Y}^*\) will contradict the CAR condition and, conversely, the CAR condition does not restrict the distribution of \(Y^*\).\(^{34}\)

The only caveat so far was that CAR may result in irrational probabilities for some of the \(y\), even if all \(\Pr(Y^* = y^*)\) are rational. In this sense, the “simple device” of Rubin and Little takes one beyond the strict limits of a finite sample approach.

2.2.5. Uniqueness of CAR

Certainly, solutions to the equations will not be unique in general. In particular, a distribution that puts zero probability on some \(y \in \mathcal{Y}\) but having \(\Pr(Y \in y^*) > 0\) for all \(y^*\) with \(\Pr(Y^* = y^*) > 0\) may satisfy the equations as well as some other distribution with \(\Pr(Y = y) > 0\). This will certainly happen when the partition of \(\mathcal{Y}\) induced by the sets in \(\mathcal{Y}^*\) does contain a set, say \(A\), with cardinality \(|A| > 1\). If \(y, y' \in A\), then the sets in \(\mathcal{Y}^*\) do not allow to distinguish between \(y\) and \(y'\). In such a case, it can not be expected that CAR by itself identifies the probabilities of \(y\) and \(y'\). Modifying Example 1 by eliminating all singletons except \(\{1\}\) from \(\mathcal{Y}^*\) gives

Example 1a

\[
\mathcal{Y}_{1a}^* = \{\{1\}, \{2, 3\}, \{1, 2, 3\}\}
\]

\(^{33}\) This is the mean value theorem for \(|\mathcal{Y}| = 2\) and Brouwer’s fixed point theorem for general (finite) \(|\mathcal{Y}|\).

\(^{34}\) Molenberghs et al. (2008) formulate the equivalent result in a parametric setting, assuming parameter distinctness.
say. The equations (2.22) become

\begin{align*}
p_1 &= 0.5 + 0.2p_1 \\
p_2 &= 0.3 \frac{p_2}{p_2 + p_3} + 0.2p_2 \\
p_3 &= 0.3 \frac{p_3}{p_2 + p_3} + 0.2p_3
\end{align*}

whose solution set is given by \( p_1 = 5/8 \) and any partition of the remaining mass of 3/8 between \( p_2 \) and \( p_3 \). This is a closed, connected and convex subset of the probability simplex. The set of solutions within the set of consistent distributions is shown in Figure 2.12.

Restricting for the time being \( \mathcal{Y}^* \) to those elements \( y^* \) such that \( \Pr(\mathcal{Y}^* = y^*) > 0 \), a very special case arises when \( \mathcal{Y}^* \) is just a partition of \( \mathcal{Y} \). That happens when only simply grouped data are available. This is a hierarchically ordered model and thus \( \Pr(\mathcal{Y} \in y^*) \) is uniquely determined for all \( y^* \in \mathcal{Y}^* \). In particular, \( \Pr(Y \in y^*) = \Pr(Y^* = y^*) \). But for \( y^* \) with \( |y^*| > 1 \), this is the only restriction on the probabilities \( \Pr(Y = y) \) for \( y \in y^* \), i.e. the probabilities can vary between 0 and \( \Pr(Y \in y^*) \).

In this situation, the set of consistent distributions coincides with the set of CAR distributions. To see this, suppose that \( \Pr(Y \in \cdot) \) is a consistent distribution. Then \( \{Y^* = y^*\} \subseteq \{Y \in y^*\} \) by consistency and \( \{Y \in y^*\} \subseteq \{Y^* = y^*\} \) by construction. Thus, the CAR condition (2.21) follows. This is the only situation where the CAR condition is implied by the structure of \( (\mathcal{Y}, \mathcal{Y}^*) \) and is therefore always justified. There simply are no distributions consistent with the structure that could contradict CAR.\(^{35}\) Furthermore, in this case the probability distribution of \( Y \) is determined on the \( \sigma \)-algebra generated by the partition.

Gill et al. (1997: 262) claimed that the only form of non-uniqueness is the one exhibited by Example 1a, and that in particular the solution to the CAR conditions are always unique on sets \( y^* \) with \( \Pr(Y^* = y^*) > 0 \). This is obviously the most one can hope for. Suppose now that \( \Pr(Y \in y^*) \) is

\(^{35}\) See the discussion by Grünwald/Halpern (2003: Proposition 4.1) and De Cooman/Zaffalon (2004: Sect. 3).
in fact uniquely identified from the distribution of $Y^*$. Then also the probability of complements, $\Pr(Y \in y^c)$, is identified. In other words, the probabilities of subsets $\{y^* \in \mathcal{P}(\mathcal{Y}) \mid y^* \cup y^{*c} \in \mathcal{Y}^*\}$ are identified. Furthermore, disjoint unions and set differences $y^* \setminus y'^*$ for $y'^* \subseteq y^*$ are identified. This set of subsets of $\mathcal{Y}$ will be denoted by $\mathcal{D}(\mathcal{Y}^*)$.\[^{36}\]

In the hierarchically ordered case, this implies that the probabilities of the finest partition (and their unions) are known. Since the elements of the finest partition together with all their unions forms a $\sigma$-algebra, as is the case for simply grouped data, $\mathcal{D}(\mathcal{Y}^*) = \sigma(\mathcal{Y}^*)$.

In all other cases, however, the sets $y^*/y'^*$ for which $\Pr(Y \in y^*/y'^*)$ would be identified from the uniqueness of CAR for $\Pr(Y^* = y^*) > 0$ will not form a $\sigma$-algebra. Consider Example 5 below:

Example 5
Suppose $\mathcal{Y} = \{1, 2, 3, 4\}$ and

$$\mathcal{Y}_5^* = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\}$$

$$p_{\{1,2\}} = p_{\{1,3\}} = p_{\{2,4\}} = p_{\{3,4\}} = 1/4$$

The set of solutions of the CAR equations (2.22) is given by $p_1, 1/2 - p_1, 1/2 - p_1, p_1$ for $p_1 \in [0, 1/2]$, a line segment joining $(0, 1/2, 1/2, 0)$ to $(1/2, 0, 0, 1/2)$. This is a closed, convex set. Moreover, $\Pr(Y \in \{1, 2\}) = 1/2 = \Pr(Y \in \{3, 4\}) = \Pr(Y \in \{2, 4\}) = \Pr(Y \in \{1, 3\})$ as suggested by the Gill et al. conjecture.

The hypergraph corresponding to this example is given in the following Figure 2.10. Its symmetry suggests why the the probability of the above mentioned sets is identified from the distribution of $Y^*$ together with the CAR condition.

\[^{36}\] The elements of $\mathcal{D}(\mathcal{Y}^*)$ form what is known as a $\lambda$-system in measure theory, a set of subsets of $\mathcal{Y}$ containing the empty set and being closed under the formation of complements and disjoint unions. See Pollard (2002: 42). Note that Grünwald and Halpern (2003: 251) included also intersections into what they termed $\mathcal{R}$-atoms, the building blocks for their algebraic characterisation of CAR.
Looking at the structure of subsets whose probability is uniquely determined by the CAR condition, neither \( \Pr(Y = 1) \) nor \( \Pr(Y \in \{1, 2, 3\}) \) are identified unless one of the \( \Pr(Y^* = \cdot) \) equals 0. But \( \{1\} = \{1, 2\} \cap \{1, 3\} \) so that \( \{1\} \) is an element of the \( \sigma \)-algebra generated by \( \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{3, 4\}\} \). Similarly, \( \{1, 2, 3\} = \{1, 2\} \cup \{1, 3\} \) is an element of the \( \sigma \)-algebra but its probability is not identified. The identified probabilities cannot be extended to the \( \sigma \)-algebra because the unions and intersections here are taken among non-nested elements of \( Y^* \) and such that these unions and intersections are not themselves elements of \( Y^* \). Therefore, in this example \( D(Y^*) \subsetneq \sigma(Y^*) \).

I have already shown that if \( Y^* \) is hierarchically ordered, then necessarily \( D(Y^*) = \sigma(Y^*) \). Conversely, suppose that \( D(Y^*) = \sigma(Y^*) \) and let \( y^{*'} \) be an arbitrary element of \( \sigma(Y^*) \). Being an element of \( \sigma(Y^*) \), \( y^{*'} \) is the union of elements of \( Y^* \) or their complements, or the complement of such a union. Write \( y^{*'} = \bigcup y^{*''} \) with either \( y^{*''} \in Y^* \) or \( y^{*''} \sqsubseteq Y^* \). For this to be also an element of \( D(Y^*) \), any pair of sets \( y_1^{*''} \) and \( y_2^{*''} \) must satisfy either \( y_1^{*''} \subseteq y_2^{*''} \) or \( y_2^{*''} \subseteq y_1^{*''} \) or \( y_1^{*''} \cap y_2^{*''} = \emptyset \) so that \( \sigma(Y^*) \) is itself hierarchically ordered. If this holds true for \( \sigma(Y^*) \), it must be true for \( Y^* \), a subset of \( \sigma(Y^*) \). Thus, \( Y^* \) is hierarchically ordered. This clarifies the relation between the structure of the set \( D(Y^*) \) (which according to the Gill et al. conjecture would be the largest set of subsets whose probabilities are uniquely determined by CAR) and the structure of \( Y^* \).

But is the form of non-uniqueness exemplified by Example 1a and Example 5 really the only one that is to be expected? Do we only have to
care about non-uniqueness on sets of the above form? A variation of Example 3 shows that the claim of Gill et al. is false, at least when (2.21) is used as a definition of CAR.

In Example 3 it is the fact that \( \Pr(Y^* = \{2\}) > 0 \) from which \( \Pr(Y = 2) > 0 \) follows. But this implies \( \Pr(Y = 3) = 0 \). If the example is modified such that \( \Pr(Y^* = \{2\}) = 0 \), then it may seem to be difficult to identify separately the probabilities of \( \{Y = 2\} \) and \( \{Y = 3\} \). On the other hand, since \( \Pr(Y^* = \{1\}) > 0 \), \( \Pr(Y = 1) \) should be uniquely identified. And from \( \Pr(Y^* = \{1, 2\}) > 0 \), \( \Pr(Y \in \{1, 2\}) \) should be identified. Therefore \( \Pr(Y = 2) = \Pr(Y \in \{1, 2\} \setminus \{1\}) = \Pr(Y \in \{1, 2\}) - \Pr(Y = 1) \) is identified together with \( \Pr(Y = 3) \). Thus, \( \Pr(Y = .) \) is completely identified. In fact, CAR in that case implied a unique solution for the distribution of \( Y \).

Now suppose that Example 3 is replaced by

**Example 3a**

\[
V_{3a}^* = \{\{1\}, \{1, 2\}, \{2, 3\}, \{1, 2, 3\}\}
\]

\[
p_{\{1\}} = 0.5, p_{\{1, 2\}} = 0.1, p_{\{2, 3\}} = 0.1, p_{\{1, 2, 3\}} = 0.3
\]

As in Example 3, since \( \{2\} = \{1, 2\} \setminus \{1\} \), all the singletons should have unique probability and thus \( \mathcal{D}(V^*) = \mathcal{P}(V) \) when the Gill et al. conjecture was true. The probability distribution of \( Y \) would be uniquely determined. This was the case in Example 3.

The hypergraph of the modified example is given in Figure 2.11. It shows that the edges representing \( \{Y = 2\} \) and \( \{Y = 3\} \) are properly nested. This is already the case in Example 3 (see Figure 2.3) so that there is no reason to expect additional trouble.

The CAR condition (2.22) gives

\[
p_1 = 0.5 + 0.1 \frac{p_1}{p_1 + p_2} + 0.3p_1
\]

\[
p_2 = 0.1 \frac{p_2}{p_1 + p_2} + 0.1 \frac{p_2}{p_2 + p_3} + 0.3p_2
\]
2. Probabilistic Models of Incomplete Data

Figure 2.11.: Hypergraph for Example 3a. The innermost circles represent the sets \{1\}, \{1, 2\}, \{2, 3\}, and \{1, 2, 3\}. The outer ellipses represent the elements 1, 2, and 3.

\[ p_3 = 0.1 \frac{p_3}{p_2 + p_3} + 0.3p_3 \]

But now there are two different solutions to the equations:

\[ p_1 = \frac{5}{6}, p_2 = \frac{1}{6}, p_3 = 0 \text{ and } p_1 = \frac{6}{7}, p_2 = 0, p_3 = \frac{1}{7} \]

The implied joined densities are

<table>
<thead>
<tr>
<th></th>
<th>{1}</th>
<th>{1, 2}</th>
<th>{2, 3}</th>
<th>{1, 2, 3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/2</td>
<td>5/60</td>
<td>0</td>
<td>15/60</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/60</td>
<td>6/60</td>
<td>3/60</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>1/10</td>
<td>1/10</td>
<td>3/10</td>
</tr>
</tbody>
</table>

and

<table>
<thead>
<tr>
<th></th>
<th>{1}</th>
<th>{1, 2}</th>
<th>{2, 3}</th>
<th>{1, 2, 3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/2</td>
<td>7/70</td>
<td>0</td>
<td>18/70</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>7/70</td>
<td>3/70</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>1/10</td>
<td>1/10</td>
<td>3/10</td>
</tr>
</tbody>
</table>

so that it is easy to check that both solutions indeed satisfy (2.21).
But now, contrary to the case of Example 3 and the conjecture by Gill et al. there is no proper subset of \( \mathcal{Y} \) whose probability is uniquely determined by CAR. None of the singletons nor any of the two-element subsets of \( \mathcal{Y} \) have identical probabilities under both solutions. Thus, the only subsets whose probabilities are uniquely determined are trivially \( \emptyset \) and \( \mathcal{Y} \). On the other hand, \( D(\mathcal{Y}^*) = \mathcal{P}(\mathcal{Y}) \) so that the probability of \( Y \) ought to be uniquely identified if the conjecture was correct. Thus either the conjecture is false or one has to use a stronger condition than the CAR condition (2.21) to further reduce the solution set.

The same phenomenon as in the previous example occurs in the Monty Hall problem. If one puts \( \Pr(Y^* = \{1, 2\}) = \Pr(Y^* = \{2, 3\}) = 1/2 \), then the CAR equations result in

\[
\begin{align*}
p_1 &= 0.5 \frac{p_1}{p_1 + p_2} \\
p_2 &= 0.5 \frac{p_2}{p_1 + p_2} + 0.5 \frac{p_2}{p_2 + p_3} \\
p_3 &= 0.5 \frac{p_3}{p_2 + p_3}
\end{align*}
\]

Again, there are two different solutions to the equations:

\[
p_1 = 0, p_2 = 1, p_3 = 0 \quad \text{and} \quad p_1 = 1/2, p_2 = 0, p_3 = 1/2
\]

Note that only the second solution depends on the probability distribution of \( Y^* \): if one sets \( p_{\{1,2\}} := \Pr(Y^* = \{1, 2\}) \) arbitrarily, then the second solution becomes \( p_1 = p_{\{1,2\}}, p_2 = 0, p_3 = 1 - p_{\{1,2\}} \). The joined distributions are given by

<table>
<thead>
<tr>
<th></th>
<th>{1, 2}</th>
<th>{2, 3}</th>
<th></th>
<th>{1, 2}</th>
<th>{2, 3}</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1/2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1/2</td>
<td>1/2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td>1/2</td>
<td>1/2</td>
<td></td>
<td>1/2</td>
<td>1/2</td>
</tr>
</tbody>
</table>

Once again, condition (2.21) is easily verified.
The solutions to the CAR equations of the Examples 1a, 3a, and the Monty Hall solutions are depicted within the consistency region of the probability simplexes in Figure 2.12.

![Figure 2.12: The consistency polytopes and the solutions of the CAR equations for Examples 1a, 3a, and the Monty Hall problem. CAR solutions are indicated by \( \times \) in the latter two cases.](image)

But by the Gill et al. conjecture, since \( \Pr(Y^* = \{1, 2\}) > 0 \) identification of \( \Pr(Y \in \{1, 2\}) \) and therefore of \( \Pr(Y = 3) \) should follow. Similarly, \( \Pr(Y = 1) \) should be identified from \( \Pr(Y^* = \{2, 3\}) > 0 \). Thus \( \Pr(Y = .) \) should be completely identified. Once again, however, the only sets with identical probabilities for both solutions are trivially \( \emptyset \) and \( Y' \).

What are the difficulties encountered when there are CAR solutions that differ on \( D(Y^*) \)? Are these difficulties related to the definition of CAR or to the rather optimistic conjecture of Gill et al.?

It is instructive to look at the Monty Hall problem in some more detail. I have argued in section 2.1.2 rather intuitively that there can not be a CAR solution in that example or in any example with the same hypergraph structure. Moreover, the two solutions presented above using the CAR condition (2.21) look strange in that they force zero probabilities on some alternatives that on a priori grounds would be taken to be possible. Even though the two solutions are contradictory, any one of them rules out a possibility solely based on a theoretical condition which therefore must be termed dogmatic.

Furthermore, the non-uniqueness is not of the kind that is necessarily present in all incomplete data situations. It does not arise from asking
more than can reasonably be expected. If $$\Pr(Y \in y^*)$$ for both $$y^* = \{1, 2\}$$ and $$y^* = \{2, 3\}$$ were identified, then the distribution of $$Y$$ would be identified as well. And as long as $$\Pr(Y^* = \{1, 2\}) \notin \{0, 1\}$$, all distributions of $$Y^*$$ result in two distinct solutions. The difficulty is not related to the distribution of $$Y^*$$, but to the structure of ($$\mathcal{Y}, \mathcal{Y}^*$$). The result does more to undermine the CAR condition as formulated in (2.21) than to question the reasonable expectation of Gill et al.

If these considerations are reasonable, one would need a more restrictive CAR formulation so that not all of the solutions consistent with (2.21) would be permitted under the label of CAR. This would rule out counter-intuitive results as the ones presented above.

But it must then also be allowed that no CAR solution exists. This would contradict the slogan that CAR is everything. Even more importantly, it would contradict the general believe, reiterated in nearly all expositions of the “simple device”, that CAR is an ‘untestable assumption’. In particular, the Monty Hall problem would not have a CAR solution and this can be asserted even without looking at the actual data.\(^3^7\)

If despite all the difficulties CAR in the formulation (2.21) was adopted, the slogan ‘Car is everything’ would be vindicated and it would indeed be an ‘untestable assumption’. But one then must be prepared to deal with nasty complications. CAR could be compatible with contradictory probabilistic accounts involving in extreme cases all non-trivial sets of the form $$\{Y \in y^*\}$$ for $$y^* \in \mathcal{Y}^*$$ simultaneously. And even in less extreme cases, elements of $$\mathcal{D}(\mathcal{Y}^*)$$ may not be unequivocally identified by CAR alone. Consequently, the concept would be of only limited use for either frequentist inference or Bayesian kinematics.

2.2.6. Further CAR Formulations

The CAR condition (2.21) can be reformulated in a number of useful ways that will also shed more light on the difficulties encountered in the previous section. One possible reformulation reverses the roles of

\(^3^7\) The only prerequisite is that the set $$\mathcal{Y}^*$$ contains only elements with positive probability.
Y and \( Y^* \) in the conditional probabilities and is thus very helpful as a complement to (2.21). It can be written as

\[
\text{Pr}(Y^* = y^* \mid Y = y) \text{ is constant on } y \in y^*
\]

(2.28)

for all \( y^* \in \mathcal{Y}^* \) and all \( y \in y^* \) such that \( \text{Pr}(Y = y) > 0 \). Looking at the problem from this direction is somewhat closer to the emergence of incomplete data. The employment status of the respondents and possibly many other characterisations of the respondent and the interview situation are the background against which the respondents decide to participate in the survey and to answer its questions. The conditional probability of a particular response given (a subset) of the background variables may be used to model this. In particular, (2.28) says that the conditional probability of refusing to answer the question of employment status \( (Y^* = \{1, 2, 3\} \text{ in Example 1}) \) does not depend on that status. An immediate consequence is that

\[
\text{Pr}(Y^* = y^* \mid Y = y) = \text{Pr}(Y^* = y^* \mid Y \in y^*)
\]

(2.29)

for all \( y^* \in \mathcal{Y}^* \) and all \( y \in y^* \) such that \( \text{Pr}(Y = y) > 0 \), a further form of the condition that is often useful.

Obviously, if (2.29) holds, then \( \text{Pr}(Y^* = y^* \mid Y = y) \) is constant for all \( y \in y^* \) and \( \text{Pr}(Y = y) > 0 \) so that (2.28) follows. In the other direction, fix an \( y \in y^* \) with \( \text{Pr}(Y = y) > 0 \) and suppose that (2.28) holds. Then

\[
\text{Pr}(Y^* = y^* \mid Y \in y^*) = \frac{\sum_{y' \in y^*} \text{Pr}(Y^* = y^* \mid Y = y') \text{Pr}(Y = y')}{\sum_{y' \in y^*} \text{Pr}(Y = y')}
\]

\[
= \text{Pr}(Y^* = y^* \mid Y = y) \frac{\sum_{y' \in y^*} \text{Pr}(Y = y')}{\sum_{y' \in y^*} \text{Pr}(Y = y')}
\]

\[
= \text{Pr}(Y^* = y^* \mid Y = y)
\]

where terms with \( \text{Pr}(Y = y') = 0 \) are zero and where the second equality uses (2.28). Note that the denominator must be positive since \( \text{Pr}(Y = y) > 0 \) by assumption.

Now (2.29) can equivalently be written as

\[
\text{Pr}(Y^* = y^* \mid Y = y \land Y \in y^*) = \text{Pr}(Y^* = y^* \mid Y \in y^*)
\]
since \( \{Y = y\} \cap \{Y \in y^*\} = \{Y = y\} \) (and ignoring again \( y \) with \( \Pr(Y = y) = 0 \)). In this form the condition can be re-expressed as a conditional independence condition:

\[
\{Y = y\} \perp \perp \{Y^* = y^*\} \mid \{Y \in y^*\} \text{ or } \{Y^* = y^*\} \perp \perp \{Y = y\} \mid \{Y \in y^*\}
\]

(2.30)

where \( \perp \perp \) denotes stochastic independence and where the second version follows from the symmetry of conditional independence. Rewriting this again in the form of conditional probabilities one arrives at

\[
\Pr(Y = y \mid Y^* = y^* \land Y \in y^*) = \Pr(Y = y \mid Y \in y^*)
\]

But as a consequence of the consistency condition, if \( \{Y^* = y^*\} \) obtains then so must \( \{Y \in y^*\} \). Thus, \( \{Y^* = y^*\} \subseteq \{Y \in y^*\} \) (see (2.3)) and so the previous equality is equivalent to

\[
\Pr(Y = y \mid Y^* = y^*) = \Pr(Y = y \mid Y \in y^*)
\]

which is just the first form of the CAR condition, (2.21). Therefore, all the conditions (2.21), (2.28), (2.29), and (2.30) are equivalent.

Gill et al. presented a further reformulation of CAR from which, as they informally argued, both the existence of CAR solutions without further conditions as well as the uniqueness on \( D(Y^*) \) would follow (Gill et al. 1997: Sect. 2).

This reformulation starts from the consistency requirement \( \{Y^* = y^*\} \subseteq \{Y \in y^*\} \) to arrive at

\[
\Pr(Y^* = y^*) = \Pr(Y^* = y^*, Y \in y^*) = \Pr(Y \in y^*) \Pr(Y^* = y^* \mid Y \in y^*)
\]

(2.31)

But the CAR version (2.29) requires that

\[
\Pr(Y^* = y^* \mid Y \in y^*) = \Pr(Y^* = y^* \mid Y = y)
\]

for all \( y^* \in Y^* \) and all \( y \in y^* \) with \( \Pr(Y = y) > 0 \). It follows that one can write down the distribution of \( Y^* \) as a product of two terms

\[
\Pr(Y^* = y^*) = P(y^*)\pi(y^*)
\]

(2.32)
where $P(.)$ is a probability on $\mathcal{Y}$ and where the $\pi(y^*)$ satisfy

$$\sum_{y^*: y \in y^*} \pi(y^*) = 1 \quad \forall y \in \mathcal{Y} : \Pr(Y = y) > 0 \quad (2.33)$$

encoding version (2.29) of CAR.

Suppose that the decomposition (2.32) together with the restriction (2.33) holds. Then one may define the conditional distribution of $Y$ given $Y^*$ by

$$\Pr(Y = y \mid Y^* = y^*) := \frac{P(\{y\})}{P(y^*)}$$

for $y \in y^*$ and $\Pr(Y^* = y^*) > 0$. This is well defined since (2.32) together with $\Pr(Y^* = y^*) > 0$ implies $P(y^*) > 0$. If $y \notin y^*$, one sets $\Pr(Y = y \mid Y^* = y^*) = 0$.

Now the marginal distribution of $Y$ is

$$\Pr(Y = y) = \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \Pr(Y = y \mid Y^* = y^*)$$

$$= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \frac{P(\{y\})}{P(y^*)}$$

$$= \sum_{y^* \in \mathcal{Y}^*} P(y^*) \pi(y^*) \frac{P(\{y\})}{P(y^*)}$$

$$= \sum_{y^* \in \mathcal{Y}^*} \pi(y^*) P(\{y\}) = P(\{y\})$$

since the sum over the $\pi(y^*) = 1$ by (2.33). Note that there is no need to define $\Pr(Y = y \mid Y^* = y^*)$ if $\Pr(Y^* = y^*) = 0$. It follows that

$$\Pr(Y = y \mid Y^* = y^*) = \frac{P(\{y\})}{P(y^*)} = \frac{\Pr(Y = y)}{\Pr(Y \in Y^*)} = \Pr(Y = y \mid Y \in y^*)$$
as long as \( \Pr(Y^* = y^*) > 0 \). If \( \Pr(Y^* = y^*) = 0 \), then either \( P(y^*) = 0 \) or \( \pi(y^*) = 0 \). In the first case, both sides can be left unspecified. If however \( P(y^*) > 0 \), the right hand side of the equality can be taken as a definition of the left hand side without affecting the joint distribution of \((Y, Y^*)\). Lastly,

\[
\Pr(Y \in Y^*) = \sum_{y^* \in Y^*} \Pr(Y \in y^* \mid Y^* = y^*) \Pr(Y^* = y^*)
\]

\[
= \sum_{y^* \in Y^*} \frac{P(y^*)}{P(y^*)} \Pr(Y^* = y^*) = 1
\]

so that the consistency condition is met as well. In summary, if there is a decomposition of the distribution of \( Y^* \) as in (2.32) which obeys the restriction (2.33), then there is a joined distribution of \((Y, Y^*)\) that is consistent and satisfies the CAR condition (2.21).

Conversely, suppose that the joined distribution of \((Y, Y^*)\) satisfies the CAR condition (2.21) and that \( \Pr(Y \in Y^*) = 1 \). Then one can define

\[
P(y^*) \coloneqq \Pr(Y \in y^*) \quad \forall y^* \in Y^*
\]

and

\[
\pi(y^*) \coloneqq \begin{cases} 
\Pr(Y^* = y^* \mid Y \in y^*) & \Pr(Y^* = y^*) > 0 \\
0 & \text{else}
\end{cases}
\]

With these definitions,

\[
\Pr(Y^* = y^*) = P(y^*) \pi(y^*)
\]

follows since

\[
P(y^*) \pi(y^*) = \Pr(Y^* = y^* \land Y \in y^*) = \Pr(Y^* = y^*)
\]

by consistency if \( \Pr(Y^* = y^*) > 0 \) or by definition if \( \Pr(Y^* = y^*) = 0 \). Moreover, for a fixed \( y \) with \( \Pr(Y = y) > 0 \)

\[
\sum_{y^* \ni y} \pi(y^*) = \sum_{y^* \ni y \atop \Pr(Y^* = y^*) > 0} \Pr(Y^* = y^* \mid Y \in y^*)
\]
\[
\sum_{y^* \ni y} \Pr(Y^* = y^* \mid Y = y) \frac{\Pr(Y^* = y^*)}{\Pr(Y = y)} = \sum_{y^* \ni y} \Pr(Y^* = y^* \mid Y = y) = 1
\]

where the first equality uses the definition of \( \pi(.) \), the second uses (2.29), the third the definition of conditional probability, and the last consistency again.

Thus (2.32) together with (2.33) are in fact equivalent to the other four forms of the CAR condition. The equations (2.32) are now linear in the probabilities of interest, \( P(y^*) \) and the difficulties of solving the system are hidden in the restrictions (2.33) and the requirement that \( P(.) \) must be a probability distribution.

There is a direct interpretation of the reformulated CAR conditions (2.32) and (2.33). Since \( \pi(y^*) \) must equal \( \Pr(Y^* = y^* \mid Y = y) \) for all \( y \in y^* \) with \( \Pr(Y = y) > 0 \), the second term in the decomposition encodes the way incomplete data are constructed in the model, at least when \( \Pr(Y = y) > 0 \). Given \( \{ Y = y \} \), choose \( Y^* \) according to this conditional distribution. The decomposition thus states that the distribution of \( Y^* \) factors in a term that only depends on the distribution of interest (the distribution of \( Y \)), and in a term that describes the generation of the incomplete version \( Y^* \). This may be translated into the statistical language by writing down the expected log-likelihood (with respect to the distribution of \( Y^* \)) for the parameters of interest \( p_y := \Pr(Y = y) \) and the nuisance parameters \( \pi_{y^*} \):\(^{38}\)

\[
\ell(p_y, \pi_{y^*}; y \in \mathcal{Y}, y^* \in \mathcal{Y}^*)
= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \log (P(y^*) \pi_{y^*})
= \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \log \left( \sum_{y \in y^*} p_y + \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \log \pi_{y^*} \right)
\]

\(^{38}\) I write here \( \pi_{y^*} \) instead of \( \pi(y^*) \) to emphasise their role as a set of parameters. These \( \pi_{y^*} \) should not be confused with the same symbol used in section 2.1.
where the summands are taken to be 0 if \( \Pr(Y^* = y^*) = 0 \). The likelihood factors into a part that depends on the parameters of interest only and in a part that depends on the nuisance parameters. The log-likelihood can be maximised separately in both parts subject to the restriction (2.33) and the requirement that \( p_y \) be a probability density. In consequence, if one is only interested in the density \( p_y \) one may simply maximise the first term of the log-likelihood ignoring the way incomplete data distributions are supposed to have emerged.\(^{39}\)

This reformulation of the CAR conditions into statistical language provides an alternative though somewhat tedious way to show that to any given distribution of \( Y^* \) one can always construct a joint distribution of \( (Y, Y^*) \) such that both \( \Pr(Y \in Y^*) = 1 \) and CAR holds. It is nevertheless instructive to follow through the proof suggested by Gill et al. (1997) since as a consequence of their proof they claim uniqueness of the CAR solution for all sets \( y^* \) such that \( \Pr(Y^* = y^*) > 0 \). It is the uniqueness that was disputed at the end of the last section. And as I have indicated, non-uniqueness raises potentially serious problems for both the application of the CAR condition and the general acceptability of the “simple device”.

Note first that if there is a decomposition of the distribution of \( Y^* \) as in (2.32) that satisfies the constraint (2.33) and such that \( P(.) \) is a probability distribution, then this decomposition certainly maximises the expected likelihood (2.34). This is a simple consequence of Jensen's inequality\(^ {40}\) since for any other values \( p_y', \pi_{y^*}' \) that satisfy the restrictions one has

\[
\ell(p_y', \pi_{y^*}'; y \in \mathcal{Y}, y^* \in \mathcal{Y}^*) - \ell(p_y, \pi_{y^*}; y \in \mathcal{Y}, y^* \in \mathcal{Y}^*)
\]

\(^{39}\) This is rather obviously true only if the nuisance parameters \( \pi_{y^*} \) are not related to the parameters of interest \( p_y \). This is sometimes called ‘parameter distinctness’. The concept of weak exogeneity in econometrics (Engle et al. 1983) and that of a cut (Barndorff-Nielsen 1978: 50; Barndorff-Nielsen/Cox 1994: 38) relate parameter distinctness to inferential procedure. While frequentist inference procedures are generally free to define their parameters arbitrarily (which many see as a threat to such ‘inferences’), the mere existence of nuisance parameters will influence any frequentist inference procedure. I am not going to discuss these problems here, but Jaeger (2005a) provides some details in the case of coarsened data.

\(^{40}\) The special case is known as the information inequality in coding and information theory.
\[ = \sum_{\gamma^* \in \mathcal{Y}^*} \Pr(Y^* = \gamma^*) \log \frac{P(y^*)' \pi_{\gamma^*}'}{P(y^*) \pi_{\gamma^*}} \]
\[ \leq \log \sum_{\gamma^* \in \mathcal{Y}^*} \Pr(Y^* = \gamma^*) \frac{P(y^*)' \pi_{\gamma^*}'}{P(y^*) \pi_{\gamma^*}} \]
\[ = \log \sum_{\gamma^* \in \mathcal{Y}^*} \Pr(Y^* = \gamma^*) \frac{P(y^*)' \pi_{\gamma^*}'}{\Pr(Y^* = \gamma^*)} \]
\[ = \log \sum_{\gamma^* \in \mathcal{Y}^*} P(y^*)' \pi_{\gamma^*} \leq 0 \]

The last inequality follows from
\[ \sum_{\gamma^* \in \mathcal{Y}^*} P(y^*)' \pi_{\gamma^*}' = \sum_{\gamma^* \in \mathcal{Y}^*} \sum_{\gamma \in \mathcal{Y}} \mathbb{1}[\gamma^*](y) p_y' \pi_{\gamma^*}' \]
\[ = \sum_{y \in \mathcal{Y}} p_y' \sum_{\gamma^* \in \mathcal{Y}^*} \mathbb{1}[\gamma^*](y) \pi_{\gamma^*}' = \sum_{y \in \mathcal{Y}} p_y' = 1 \]

where the restriction (2.33) is used for \( \pi_{\gamma^*}' \) and \( p_y' \) is assumed to be a probability density. Note that the second equality requires (2.33) to hold only for all \( y \in \mathcal{Y} \) with \( p_y' > 0 \).

In the other direction, the log-likelihood certainly has a maximiser since it is continuous and the set of parameters \( (p_y, \pi_{\gamma^*}; y \in \mathcal{Y}, \gamma^* \in \mathcal{Y}^*) \) satisfying the restrictions is convex and compact.

Starting with the first summand of the log-likelihood that involves only \( p_y \), one may show using a Lagrange multiplier ensuring \( \sum_y p_y = 1 \) that a maximiser of this summand must satisfy
\[ \sum_{y^* \in \mathcal{Y}^*} \Pr(Y^* = y^*) \frac{\Pr(Y^* = y^*)}{P(y^*)} = 1 \text{ if } p_y > 0 \quad (2.35) \]

Thus one can define
\[ \pi_{\gamma^*} := \frac{\Pr(Y^* = \gamma^*)}{P(y^*)} \]
for all those $y^*$ that contain at least one $y$ with $p_y > 0$. For such sets $y^*$, certainly $\Pr(Y^* = y^*) = P(y^*)\pi_{y^*}$ and
\[
\sum_{y^* \ni y} \pi_{y^*} = 1 \quad \forall y : p_y > 0
\]
so that (2.33) is satisfied. Moreover, one can extend the definition of $\pi_{y^*}$ to all elements of $\mathcal{Y}^*$ by setting it to 0 on sets $y^*$ that only contain $y$ with $p_y = 0$. For these sets, both $P(y^*) = 0$ and $\Pr(Y^* = y^*) = 0$ so that the decomposition holds for all elements of $\mathcal{Y}^*$. This set of parameters is a maximiser of the expected log-likelihood simply because it is a decomposition of the probabilities $\Pr(Y^* = y^*)$.

In summary, the argument shows that there is always a maximum of the expected log-likelihood and this must entail a decomposition (2.32). In other words, for any given probability distribution $\Pr(Y^* \in .)$ there is always a CAR solution, a result that was already obtained along a different route in section 2.2.4.

But what does this method tell about the uniqueness of CAR? Gill et al. (1997: 264) claim that their method provides uniqueness of $P(.)$ at least for those $y^*$ with $\Pr(Y^* = y^*) > 0$. Moreover, they claim that the restriction (2.33) can be extended to hold for all elements $y \in \mathcal{Y}$, whether or not $p_y > 0$. Unfortunately, both claims are unfounded and it turns out that the problem is related to both claims simultaneously.

The examples of the previous section can be used to illustrate what goes wrong. They show that there may be several maxima of the log-likelihood function in the restricted variables $p_y$ and $\pi_{y^*}$. Moreover, not all of them admit an extension of the restrictions on $\pi_{y^*}$ to all of $\mathcal{Y}$.

Consider again the Monty Hall problem. As previously noted, there are two solutions given by

1) $\Pr(Y = 2) = 1$ and $\pi_{\{1,2\}} = \pi_{\{2,3\}} = 1/2$

2) $\Pr(Y = 1) = \Pr(Y = 3) = 1/2$ and $\pi_{\{1,2\}} = \pi_{\{2,3\}} = 1$

when $\Pr(Y^* = \{1, 2\}) = \Pr(Y^* = \{2, 3\}) = 1/2$. Both solutions do provide a decomposition, since in both cases $\Pr(Y^* = \{1, 2\}) = \Pr(Y \in$
\{1, 2\} \pi_{\{1,2\}} \) and similarly for \( \Pr(Y^* = \{2, 3\}) \). Thus both do satisfy the first part of the CAR condition (2.32). Also, they satisfy the restriction on \( \pi_{y^*} \) since in the first case, where only \( y = 2 \) must be considered, \( \pi_{\{1,2\}} + \pi_{\{2,3\}} = 1 \) and in the second case for both \( y = 1 \) and \( y = 3 \) only one summand with value 1 exists. Thus the solutions satisfy both (2.32) and (2.33). And obviously both solutions lead to the same value of the expected log-likelihood.

Nothing of this contradicts the arguments given in the proof above. But then Gill et al. (1997: 263–264) try to show that it is possible to construct the \( \pi_{y^*} \) so that their sum over all \( y^* \) that contain a given \( y \), whether or not \( p_y > 0 \), must equal 1. They try to construct such a \( \pi_{y^*} \) using an extension of (2.35) to the case \( p_y = 0 \). Note that this results in a much stronger restriction than the one used in (2.33). Nevertheless, if the extension was possible, then there were \( |\mathcal{Y}| \) valid constraints on the \( \pi_{y^*} \). Thus the \( 2^{|\mathcal{Y}|} − 1 \) equations \( \Pr(Y^* = y^*) = P(y^*)\pi_{y^*} \) are exactly matched by the \( |\mathcal{Y}| \) parameters \( p_y \) and the \( 2^{|\mathcal{Y}|} − 1 − |\mathcal{Y}| \) parameters \( \pi_{y^*} \), making uniqueness a reasonable conjecture.

However, such an extension is impossible in both solutions to the Monty Hall problem. In the first solution, if \( y = 1 \), then \( \pi_{\{1,2\}} \) is the only set to which it belongs, but its value is 1/2. Similarly for \( y = 3 \). Note that in this situation one might be tempted to define \( \pi_{\{1\}} = \pi_{\{3\}} = 1/2 \) so that the argument goes through nevertheless. But this strategy does not work for the second solution where for \( y = 2 \) we have \( \pi_{\{1,2\}} + \pi_{\{2,3\}} = 2 \). Since the \( \pi_{y^*} \) must be non-negative, one can not extend \( \mathcal{Y}^* \) so that the sum condition can be met.

The only situation where the strategy to prove uniqueness will work is when \( \Pr(Y^* = \{y\}) > 0 \) for all \( y \in \mathcal{Y} \). In that case, the original Gill et al. proof shows that there must be a unique CAR distribution. It is still an open question how to characterise those situations where the only form of non-uniqueness of CAR is of the simple form Gill et al. conjectured. There are not even easily stated conditions which imply the existence of multiple solutions of the type exemplified by the Monty Hall problem and Example 3a.
2.2.7. Strong CAR?

With the present CAR condition and its various re-expressions there is always at least one joint distribution of \((Y, Y^*)\) such that a given incompleteness \(\{Y^* = y^*\}\) implies nothing more about the distribution of interest than that \(\{Y \in y^*\}\) (this is condition (2.21)) or such that incompleteness does not depend on the actual value of the random variable \(Y\) within the coarsening set \(y^*\) (condition (2.28)), or that the coarsened variables are constructed without regard to particular values of the variable of interest (condition (2.30)), or that the likelihood factors into a part describing the distribution of interest and the ‘missing mechanism’ (condition (2.32)). All these CAR formulations turned out to be equivalent mathematical conditions for which a solution always exists. That is, for any given distribution of \(Y^*\) there is a joined distribution of \((Y, Y^*)\) that is consistent and compatible with the CAR condition. The CAR condition does not restrict the set of possible distributions of the coarsened variable.\(^{41}\)

This is what one ought to hope for since the “simple device” can prove useful only to the extend that its central concepts are not special mathematical artifacts. The definition of CAR depends crucially on the introduction of a probability model. Its usefulness as a regulative idea would be impaired if its definition would depend on particular data constellations.

On the other hand, the CAR condition may lead to contradictory results even for all non-trivial \(y^* \in \mathcal{Y}^*\), as Example 3a and the Monty Hall problem demonstrate. This situation is at least as threatening to the use of CAR since it would imply contradictory results for many or even all sets of interest. One possible solution, albeit a rather radical one, is to strengthen the CAR definition in such a way that contradictory

---

\(^{41}\) This is no longer true when infinite sets are considered. See Gill et al. (1997: 264) for an example. They conjecture that this problem can be alleviated by “compactifying both the sample space and all the observed random sets in a careful way”. A discussion of this type of difficulty will not be pursued here. Cator (2004: 1962) provides another type of example where the difficulty lies in the fact that the incompleteness cannot be represented in terms of subsets of the set \(\mathcal{Y}\). A discussion of this type of problem will be postponed to the next section.
results are ruled out. One may then hope that the situations that are not CAR according to the strengthened definition in fact turn out to be ‘pathological’ exceptions. An attractive possibility is opened up by the CAR formulations of the previous subsection.

While all these definitions are equivalent with CAR in the form (2.21), the change of the conditioning variable from \( Y^* \) in (2.21) to \( Y \) suggests that one may ask for (2.28)–(2.30) to hold for all \( y \) and not only for those with \( \text{Pr}(Y = y) > 0 \). This is a vacuous strengthening if \( \text{Pr}(Y^* = \{y\}) > 0 \) for all \( y \in Y \) because then \( \text{Pr}(Y = y) > 0 \) by the consistency requirement. In all other cases, however, it is strictly stronger than the CAR condition (2.21). This is an attractive approach since one may argue that (2.28)–(2.30) refer to a ‘mechanism’ producing the form of incompleteness from any given ‘input’ \( y \). If such a ‘mechanism’ can reasonably be posited, it should not depend on whether or not in fact \( \text{Pr}(Y = y) > 0 \) holds true. It should be possible to define the ‘mechanism’ of missingness without paying attention to particulars of a given situation. They ought to be irrelevant for the definition of a predetermined procedure.

The exclusion of situations where such a ‘mechanism’ does not exist and thus the argument would have no force might be tolerable if this would happen only in rare or exceptional circumstances. But one of the arguments of the critics of the “simple device” must be born in mind. It challenges that there is such a thing as a probabilistic ‘mechanism’ or an ‘answer probability’ of interviewees in a survey at all. Such a position has some empirical support for social science surveys, and it certainly pertains to observational studies where the degree of incomplete information depends on the deliberation of respondents.

On the other hand, the modelling strategy might have some value, at least if it does not presuppose facts about surveys or the interviewees. It should only exclude unreasonable mathematical artifacts from consideration that would not have been considered by informed survey statisticians either. It would then not touch directly on the question whether the ‘mechanism’ must be supposed to ‘work’ for every interviewee, or whether its preconditions are ever encountered by any member of the population.

In short, the suggestion is to require that (2.28) and the other versions
should hold not only for all \( y \) that actually occur but for all \( y \in \mathcal{Y} \). This version of CAR has been termed strong CAR by Jaeger (2005a) and Grünwald/Halpern (2003). If this direction is pursued, then (2.28) is strengthened to

\[
\Pr(Y^* = y^* \mid Y = y) \text{ is constant on } y \in y^*
\]  

(2.36)

for all \( y^* \in \mathcal{Y}^* \) and all \( y \in y^* \), even when \( \Pr(Y = y) = 0 \).

It turns out that both the Monty Hall example and Example 3a do not satisfy strong CAR. This follows in the case of the Monty Hall example along the same lines as indicated at the end of section 2.1.2: If the price is behind door 1 (and the contestant has chosen door 2), then Monty is forced to open door 3. Thus \( \Pr(Y^* = \{1, 2\} \mid Y = 1) = 1 \). From strong CAR, it follows that \( \Pr(Y^* = \{1, 2\} \mid Y = 2) = 1 \). But \( \Pr(Y^* = \{2, 3\} \mid Y = 3) = 1 \) as well, from which \( \Pr(Y^* = \{2, 3\} \mid Y = 2) = 1 \) would follow. In consequence, \( 1 = \Pr(Y^* \in \mathcal{Y}^* \mid Y = 2) = \Pr(Y^* = \{1, 2\} \mid Y = 2) + \Pr(Y^* = \{2, 3\} \mid Y = 2) = 2 \), a contradiction. The contradiction is only avoided when \( \Pr(Y = 2) \in \{0, 1\} \) and simultaneously conditional distributions with conditioning events whose probability is zero are not assumed to take a definite value.

Similarly, in Example 3a

\[
\Pr(Y^* = \{2, 3\} \mid Y = 2) = a = \Pr(Y^* = \{2, 3\} \mid Y = 3)
\]

\[
\Pr(Y^* = \{1, 2, 3\} \mid Y = 2) = b = \Pr(Y^* = \{1, 2, 3\} \mid Y = 3)
\]

so that \( a + b = 1 \). But this forces \( \Pr(Y^* = \{1, 2\} \mid Y = 2) = 0 = \Pr(Y^* = \{1, 2\} \mid Y = 1) \), contradicting \( \Pr(Y^* = \{1, 2\}) = 0.1 > 0 \).

In general, strong CAR will be impossible whenever there are properly nested edges in the hypergraph of \((\mathcal{Y}, \mathcal{Y}^*)\). In that case, the sum over all nodes of the conditional distributions given the inner edge must equal the sum of the conditional distributions given the outer edge and both must equal 1. But this can not be the case when the nodes unique to the outer edge have non-vanishing probability. For suppose that all elements of \( \mathcal{Y}^* \) have positive probability and that for some \( y, y' \in \mathcal{Y} \)

\[
A := \{y^* \in \mathcal{Y}^* \mid y \in y^*\} \subsetneq \{y^* \in \mathcal{Y}^* \mid y' \in y^*\} =: B
\]
so that the edge of the hypergraph corresponding to \( y \) is nested within the one corresponding to \( y' \). Then \( \Pr(Y^* = y^* \mid Y = y) = \Pr(Y^* = y^* \mid Y = y') \) and \( \sum_{y^* \in A} \Pr(Y^* = y^* \mid Y = y) = 1 \). It follows that \( \Pr(Y^* = y^* \mid Y = y') = 0 \) for all \( y^* \in B \setminus A \). Fix one such \( y^* \), say \( y^*'. \) By the strong version of (2.29) \( \Pr(Y^* = y^*' \mid Y \in y^*') = 0 \). But now \( \Pr(Y^* = y^*') = 0 \) follows from consistency, in contradiction to the assumption that all elements of \( \mathcal{Y}^* \) have positive probability. Thus the existence of properly nested edges of the hypergraph implies that strong CAR does not hold.

In the Monty Hall example, strong CAR fails and in that setting, the idea of a well defined ‘mechanism’ that produces the coarsened data can be based on the expectation that the rules of the game are in fact well defined for all situations that might arise. In this case the argument for strong CAR based upon a ‘mechanism’ defined independently of the distribution of \( Y \) has much force. Similarly, in Example 3a, the contradictory solutions \((5/6, 1/6, 0)\) and \((6/7, 0, 1/7)\) should be ruled out as well. But strong CAR excludes many more situations. Consider e.g. Example 3. Here there is a unique CAR solution \((5/7, 2/7, 0)\), but the hypergraph given in Figure 2.3 shows nested edges for \( Y = 2 \) and \( Y = 3 \) so that there is no strong CAR solution. On the other hand, the interpretation given at the beginning of section 2.1 can certainly not be ruled out by referring to a ‘mechanism’. It would be simply too much to ask for a ‘mechanism’ that produces coarsened data according to strong CAR in a survey context. Even if the illusion of respondents answering according to some probability model would be taken for granted, it would be unreasonable to expect that they should be able to react to circumstances and situations that nobody ever encountered. But this would be needed to define a ‘mechanism’.

There is another argument in favour of strong CAR. It is based on the fact that a CAR solution that is not strong CAR as well must be such that \( \Pr(Y \in D) = 0 \) for some element \( D \in \mathcal{D}(\mathcal{Y}^*) \). As Examples 1a and 5 demonstrate, strong CAR may hold even when some solutions imply 0 probabilities for a subset of \( \mathcal{Y} \). Moreover, \( \Pr(Y^* = y^*) > 0 \) for all \( y^* \in \mathcal{Y}^* \) implies \( \Pr(Y \in y^*) > 0 \) by consistency (2.6). The difference between CAR and strong CAR arises solely if CAR implies 0 probability for at least one element \( D \in \mathcal{D}(\mathcal{Y}^*) \setminus \mathcal{Y}^* \).
Grünwald and Halpern (2003) have argued that degenerate solutions (zero probabilities for some sets) are the main source of difficulties encountered with CAR solutions. They counter the slogan ‘Car is everything’ by ‘sometimes CAR is nothing’ (Grünwald/Halpern 2003: 256).\(^{42}\) There are obviously circumstances where degenerate results indicate doubtful results. If to a given \(\mathcal{Y}\) a further element, say \(*\), is added and if \(\mathcal{Y}^*\) is modified so that all \(y^*\) contain \(*\), then obviously \(*\) can not be distinguished from any other element of \(\mathcal{Y}\).\(^{43}\) But the CAR condition now leads to either \(\Pr(Y = *) = 1\) or \(\Pr(Y = *) = 0\). Moreover, there is no strong CAR solution since all other edges of the hypergraph are nested within that of \(*\), and some will be properly nested. Here CAR would seemingly state much more than just that conditioning on \(\{Y^* = y^*\}\) can be replaced by conditioning on \(\{Y \in y^*\}\). It says something about the existence (or non-existence) of certain events. From a Bayesian perspective, one is forced to assign probability zero to events which were a priori considered possible. The disturbing feature of the example is that from (weak) CAR there seems to follow something definite even though nothing must be known or assumed about the particular situation, not even probabilities for the coarsened variables. This seems to suggest that there is more in (weak) CAR than the assertion that coarsened variables can be taken at face value.

But adding an extra symbol \(*\) to a model may be ruled out either by an appeal to parsimony or because it would generally lead to ridiculous models. There is no need to apply to strong CAR. Consider the survey statistician who, after the survey, considers that labour market conditions of many respondents are much more complicated than the simple distinction between ‘employment’, ‘out of labour force’, and ‘unemployment’. He could add further categories to his \(\mathcal{Y}\) and surmise that respondents are kind enough to answer to the original question even if they do not consider themselves to be in any of the categories that were being asked. Redoing his analysis he finds that either nobody or everybody was in his additional categories just by invoking CAR. Since this can be done with

\(^{42}\) Grünwald and Halpern (2003: Ex. 4.6) provide examples where the degenerate solutions do not arise from incomplete identification and characterise such situations for the case \(|\mathcal{Y}^*| = 3\).

\(^{43}\) This is a form of Jaeger’s (2005a) second example.
all categories one might ever come up with, the result of such a strategy is self-defeating.

The argument concerning degenerate solutions is compelling in the case of an additional element * or in the Monty Hall example. But in cases like Example 3 it has no more force than the argument based on ‘mechanisms’. Note that in Example 3, CAR always forces \( \Pr(Y = 3) = 0 \) as long as \( \Pr(Y^* = \{3\}) = 0 \) (the structure of \( \mathcal{Y}^* \) does not change) and the other probabilities of elements in \( \mathcal{Y}^* \) are positive. And \( \{3\} = \mathcal{Y} \setminus \{1, 2\} \in \mathcal{D}(\mathcal{Y}^*) \setminus \mathcal{Y}^* \) so that CAR puts 0 probability on the ‘wrong’ subset. But that CAR implies a particular distribution only shows that CAR is a very strong condition: It can identify a single distribution from a large set of consistent distributions. That this particular solution puts 0 probability on some elements of \( \mathcal{Y} \) is as such no argument against that solution.

In summary, the CAR condition in the form (2.21) can be satisfied for any given distribution on \( \mathcal{Y}^* \). Therefore, weak CAR is untestable as it should be if the modelling strategy is not to be an artifact of the “simple device”. On the other hand, strong CAR may fail. Note that the strong version of CAR in (2.36) refers to conditional distributions and is required to hold for all \( y^* \in \mathcal{Y}^* \) and all \( y \in \mathcal{y}^* \). The condition must hold irrespective of whether \( \{Y = y\} \) has positive probability or not. But for those points \( y \in \mathcal{Y} \) that have 0 probability, the joint probability of \( \{Y = y, Y^* = y^*\} \) must be 0 as well. But then the conditional probabilities in the condition (2.36) are not needed to define the joined distribution. Using (2.36) or equivalent formulations as a CAR condition thus uses more than just the joined distribution of \( (Y, Y^*) \).

One possible justification of such a move is the requirement of a ‘mechanism’ of coarsening whose definition should not depend on any aspects of \( Y \). Another argument is based on the possibly strong ontological commitment that may follow from degenerate probability assignments. But in many relevant cases in the social sciences, neither argument is compelling. In these cases, strong CAR seems to be much too strong, excluding even situations as Example 3 which otherwise looks perfectly natural. Moreover, strong CAR is not neutral as a regulative modelling strategy since the structure of \( (\mathcal{Y}, \mathcal{Y}^*) \) on which it hinges is at least partly determined by the data.
Lastly, the main problem with weak CAR is not the existence of degenerate solutions for elements of $\mathcal{D}(Y^*)$, the situation addressed by both arguments. It is the existence of contradictory solutions as in the case of Example 3a and the Monty Hall problem where there are two values of $\Pr(Y \in D)$ for some $D \in \mathcal{D}(Y^*)$. This follows from the existence of disconnected solution sets implied by weak CAR. Excluding such situations, a condition strictly weaker than the strong CAR condition, may turn out to be a reasonable compromise that almost saves the slogan ‘CAR is everything’.

### 2.3. Coarsening Variables and Coarsening Completely at Random

Even if the probability model is taken as the natural framework to discuss incomplete observations in a general context, it is by no means implied that the formalisations of the concept of coarsening at random, and therefore an appreciation of departures from it, are to be framed in terms of conditional distributions as was done up to now. In the literature on missing data problems one finds many variants of the missing at random conditions, often formulated after introducing additional random variables. Before the CAR conditions can justifiably be used within the probability model, the status of different versions of the conditions must be discussed. In particular, it must be shown that no further structure is implied by different formulations of the conditions.

Of particular importance is the relation of the present formulation to the now classical formulation used by Little and Rubin (2002). The latter exploits additional random variables that together with the variable of interest determine the incomplete variable $Y^*$. Then the missing at random (or coarsening at random) conditions are expressed in terms of conditional densities involving the additional variables.\footnote{It seems that the origins of this way to formulate missing or coarsening at random conditions must go back at least to the early 20th century. It is used implicitly in Fisher’s 1925 article on Maximum Likelihood, inter alia.}
The next subsection discusses the relation of this classical formulation to the formulation used in the previous sections. It turns out that the classical formulation potentially adds further structure to the “simple device” so that the two formulations may provide contradictory results. Furthermore, the classical formulation allows to define the concept of “completely coarsening at random” or CCAR rather naturally. But it turns out that the concept is rather difficult to formulate in general. Some of the consequences are discussed in section 2.3.2. On the other hand, the additional structure can sometimes be exploited to express information present in incomplete data problems that can not be expressed in the simple framework used previously. Moreover, there are situations that are naturally described as incomplete data problems but that can not be expressed in the language of subsets of the original sample space. This is discussed in the last subsection.

2.3.1. Coarsening Variables

The early development of methods for censored duration data was based nearly exclusively on the random censorship model where a censoring variable, say \( R \), was stipulated such that the coarsened (censored) observations are represented by \( (Y', \delta) := (\min(Y, R), \mathbb{I}[Y < R]) \). There is a natural one-to-one translation into the previous setup by setting \( y^* = \{y'\} \) if \( \delta = 1 \) and \( y^* = \{y', \ldots\} \) otherwise. In the other direction, one would put \( y' = \min y^* \) and \( \delta = 0 \) if \( |y^*| > 1 \), and \( y' = \min y^* \), \( \delta = 1 \) otherwise. But it is not so obvious what the role of \( R \) is. In particular, its value is not determined by the value of \( y^* \): If \( (Y' = y', \delta = 1) \), the value of \( R \) can not be inferred. Its inclusion extends the setup previously discussed and it is plain that the additional structure needs justification.

In the case of missing data, the additional variable is an indicator, say \( R \), which takes the value 1 if the data are complete, 0 otherwise. The incomplete variable is then a function \( Y^* = g(Y, R) \) such that \( g(y, 1) = \{y\} \) and \( g(y, 0) = \mathcal{Y} \). Knowing the underlying value of \( Y \) and the realised value of \( R \), the value of the observation can be computed by a known function. In this case, one can indeed infer the value of \( r \) from the value of \( y^* \) so that nothing essential is added to the structure.
Rubin’s article of 1976, following the tradition, uses the existence of an indicator $R$ to define his version of MAR. In general, the idea is to express MAR/CAR as a property of a procedure that from the random variable $Y$ plus a randomisation variable $R$ produces the observed $Y^*$. Such a formulation is obviously very closely connected to the idea of a ‘mechanism’ and the formulation of CAR in (2.28).

The method can easily though rather vacuously be adapted to deal with all partially complete data defined as set-valued observations. To see this, define a random variable $R$ as

$$R: \Omega \longrightarrow \mathcal{R} := \mathcal{Y}^*$$

where for ease of translation the range of $R$ is chosen to equal $\mathcal{Y}^*$. Furthermore, the conditional distribution of $R$ given $Y$ can be defined as

$$\Pr(R = y^* \mid Y = y) := \Pr(Y^* = y^* \mid Y = y)$$

for all $y$ such that $\Pr(Y = y) > 0$. This defines a joined distribution of $(Y, R)$. Now the function $g(., .)$ can be defined as

$$g: \mathcal{Y} \times \mathcal{R} \longrightarrow \mathcal{Y}^*$$

$$g(y, r) := r$$

using just the second argument. Since $(Y, Y^*)$ is consistent, $\Pr(Y \in g(Y, R)) = \Pr(Y \in R) = \Pr(Y \in Y^*) = 1$ so that $(Y, R, g)$ is consistent as well.

This is a trivial reformulation of the setup in the previous sections, even though it introduces further elements that seem to be extraneous to the situations. However, it is probably the easiest way to express the idea that the incomplete data arise from the underlying data by some randomisation expressed by $R$. Moreover, it may be used as a first translation of the previous formulation into the more familiar formulation used in the classical literature. The construction shows that to each incomplete data model in terms of $(Y, Y^*)$ there is an equivalent one in terms of $(Y, R, g)$ though the translation need not be unique.

In the other direction, to any triple $(Y, R, g)$ there is a tuple $(Y, Y^*)$ defined by $(Y, Y^*) := (Y, g(Y, R))$. One may call the joined distribution
of \((Y, R)\) CAR if the joined distribution of \((Y, Y^*)\) implied by \((Y, R)\) and \(g(., .)\) is CAR. It is now natural to inquire whether CAR can be directly formulated in terms of \((Y, R, g)\). Using (2.28) as a natural starting point, \((Y, R, g)\) might tentatively be called CAR if for all \(r \in \mathcal{R}\) and all \(y\) of positive probability

\[
\Pr(R = r \mid Y = y') \text{ is constant on } \{y' \mid y' \in g(y, r), \Pr(Y = y') > 0\}
\]

With the trivial construction of a \((Y, R, g)\) from the tupel \((Y, Y^*)\) given above, (2.40) certainly implies CAR. However, the appeal of the classical formulation derives from a much simpler structure of \(R\). In the case of missing data, \(R\) is just an indicator. In the case of censored data, it is a number giving censoring times. Both versions of \(R\) are much easier to handle than the brute force translation that simply identifies \(R\) with the set-valued variable \(Y^*\). Moreover, the classical formulation justifies the use of the face value likelihood by invoking stochastic independence of their \(R\) from the underlying \(Y\), a condition that is much easier to treat than the respective CAR conditions or the clumsy (2.40).

However, this ease of characterising CAR can be deceptive in general incomplete data models. Returning again to Example 1, one might define a random variable \(R\) taking the values 0, 1, 2 that specifies the degree of incompleteness. Then one can define the variable \(Y^* := g(Y, R)\) via

\[
\begin{align*}
g(y, 2) &= \{y\} \\
g(y, 1) &= \begin{cases} 
\{2, 3\} & y \in \{2, 3\} \\
\{1, 2, 3\} & y = 1
\end{cases} \\
g(y, 0) &= \mathcal{Y}
\end{align*}
\]

With this definition of \(g(., .)\) and \(\mathcal{R}\) one recovers the joined CAR distribution of \((Y, Y^*)\) by setting the joined distribution of \((Y, R)\) to

<table>
<thead>
<tr>
<th>(Y)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1/2</td>
<td>8/80</td>
<td>8/80</td>
<td>56/80</td>
<td>58/80</td>
<td>60/80</td>
<td>62/80</td>
<td>64/80</td>
<td>66/80</td>
</tr>
</tbody>
</table>

100
Note that the joined distribution of \((Y, R)\) and therefore the marginal of \(R\) is not unique for one might as well choose

\[
\begin{array}{c|ccc|c}
Y & 1 & 2 & 3 \\
R & 0 & 0 & 3/80 & 3/80 & 6/80 \\
   & 1 & 1/8 & 4/80 & 4/80 & 18/80 \\
   & 2 & 1/2 & 8/80 & 8/80 & 56/80 \\
\hline
   & 10/16 & 3/16 & 3/16 \\
\end{array}
\]

without changing the joint distribution of \((Y, Y^*)\).

Obviously, the choice of \(g(., .)\) is not unique either, even when the setting of \(R := Y^*\) is discarded in favour of more concise representations. In Example 1, another possible choice is

\[
g(y, 2) = \{y\} \\
g(y, 1) = \begin{cases} 
{2, 3} & y \in \{2, 3\} \\
\{1\} & y = 1 
\end{cases} \\
g(y, 0) = \mathcal{Y}
\]

where now the joined distribution of \((Y, R)\) might be

\[
\begin{array}{c|ccc|c}
Y & 1 & 2 & 3 \\
R & 0 & 1/8 & 3/80 & 3/80 & 16/80 \\
   & 1 & 0 & 4/80 & 4/80 & 8/80 \\
   & 2 & 1/2 & 8/80 & 8/80 & 56/80 \\
\hline
   & 10/16 & 3/16 & 3/16 \\
\end{array}
\]

The obvious non-uniqueness of the \((Y, R, g)\) representation of incomplete data creates a problem for the definition of CAR in terms of a missingness indicator. If the first choice of \(g(., .)\) is combined with either of the two joined distributions of \((Y, R)\) proposed above, then both violate the condition (2.40), because

\[
\Pr(R = 0 \mid Y = 1) = \frac{5/80}{50/80} = \frac{1}{10} \neq \Pr(R = 0 \mid Y = 2) = \frac{3/80}{15/80} = \frac{1}{5}
\]
in the first case and
\[
\Pr(R = 0 \mid Y = 1) = 0 \neq \Pr(R = 0 \mid Y = 2) = \frac{3/80}{15/80} = \frac{1}{5}
\]
in the second case. Thus there are \((Y, R, g)\) that imply a CAR distribution
for \((Y, Y^*)\) but that are not CAR according to (2.40).

On the other hand, the second choice of \(g(., .)\) yields
\[
\Pr(R = 0 \mid Y = 1) = \frac{2/16}{10/16} = \frac{1}{5} = \Pr(R = 0 \mid Y = 2) = \frac{3/80}{3/16}
\]
and similarly for \(\Pr(R = 0 \mid Y = 3)\). Furthermore,
\[
\Pr(R = 1 \mid Y = 2) = \frac{4/80}{3/16} = \frac{4}{15} = \Pr(R = 1 \mid Y = 3)
\]
covering all cases required by (2.40). Not too surprisingly, therefore, condition
(2.40) strongly depends on the choice of \(R\) and \(g(., .)\). Even if
the joined distribution of \((Y, Y^*)\) is CAR according to the definitions in
the previous sections, this need not be the case according to 2.40 based
on the conditional distribution of \(R\) given \(Y\).

The case of missing data is an exception in that in this simple case the
formulations in terms of a missing indicator and in terms of the tuple
\((Y, Y^*)\) are in fact equivalent. When Heitjan and Rubin (1991) intro-
duced the term CAR, they defined it in terms of the joined distribution
of \((Y, Y^*)\) using (2.28) as definition, even though the coarsening was
defined by a function such that \(Y^* = g(Y, R)\). Their corollary 1 then
stated that for ‘consistent’ \(g(., .)\) and those \(R\) that are functions of \(Y^*\) the
formulation (2.40) is equivalent to CAR.

There have been several attempts to explicate ‘consistent’ choices of \(g(., .)\)
such that the relation to CAR can be stated in terms of \((Y, R, g)\). First,
the consistency requirement \(\Pr(Y \in Y^*) = 1\) translates into
\[
\Pr(Y \in g(Y, R)) = 1 \quad \text{(2.43)}
\]
If \(Y\) is finite, then this is equivalent to
\[
\Pr(g(y, R) \ni y \mid Y = y) = 1 \quad \text{(2.44)}
\]
for all $y \in \mathcal{Y}$ with $\Pr(Y = y) > 0$. If the roles of $Y$ and $R$ are reversed, one obtains
\[ \Pr(Y \in g(Y, r) \mid R = r) = 1 \] (2.45)
for all $r \in \mathcal{R}$ such that $\Pr(R = r) > 0$.

Jaeger (2005b: Definition 2.5) explicitly strengthened the condition to
\[ y \in g(y, r) \] (2.46)
for all $y$ such that $\Pr(Y = y) > 0$ and all $r \in \mathcal{R}$. This is a strictly stronger condition than (2.43). In particular, (2.46) is violated by the construction in (2.37)–(2.39). But it is sometimes easier to work with this deterministic condition than with (2.43). Therefore much of the literature concentrates on this case.\(^{45}\)

Jaeger (2005b: Definition 2.5) further assumes that
\[ y' \in g(y, r) \implies g(y', r) = g(y, r) \] (2.47)
for all $r$ and all $y, y'$ of positive probability. Note that (2.47) holds trivially for the construction (2.37)–(2.39) since $g(., .)$ depends only on $r$. Both conditions also hold for the case of missing and censored data when the latter are expressed in terms of $y^*$ instead of the more usual $(y', \delta)$.

However, the conditions severely restrict the models that can be represented by the triple $(Y, R, g)$. In fact, the set $\mathcal{Y}^*$ corresponding to such an $(Y, R, g)$ must be a union of partitions of $\mathcal{Y}$. To see this, fix an $r \in \mathcal{R}$ and set
\[ \mathcal{Y}_r^* := \{y^* \in \mathcal{Y}^* \mid \exists y \in \mathcal{Y} : g(y, r) = y^*\} \]
But $g(., r)$ trivially induces a partition on $\mathcal{Y}$ and the conditions allow to identify it with $\mathcal{Y}_r^*$. More precisely, by condition (2.46)
\[ \mathcal{Y} = \bigcup_{y \in \mathcal{Y}} g(y, r) = \bigcup_{y^* \in \mathcal{Y}_r^*} y^* \]
\(^{45}\) Heitjan and Rubin (1991: 2247) use a similar condition. See Gill et al. (1997: 282) for some remarks.
Suppose further that there is a \( y \in y_1^* \cap y_2^* \) for \( y_1^*, y_2^* \in \mathcal{Y}_r^* \). Then \( g(y, r) = y_1^* \) since there is a \( y' \) with \( g(y', r) = y_1^* \) by the definition of \( \mathcal{Y}_r^* \) and by (2.47), \( g(y, r) = g(y', r) = y_1^* \). Similarly, \( g(y, r) = y_2^* \). Consequently, \( y_1^* = y_2^* \). Therefore, \( \mathcal{Y}_r^* \) is a partition of \( \mathcal{Y} \) and hence \( \mathcal{Y}_r^* = \cup_r \mathcal{Y}_r^* \) a union of partitions. Thus, with the restrictions (2.46) and (2.47), the variable \( R \) simply indicates a particular partition chosen to produce the coarsening \( y^* \).

Even though the restrictions (2.46) and (2.47) exclude \( \mathcal{Y}^* \) that are not unions of partitions of \( \mathcal{Y} \) (e.g. Example 3), they provide a closer connection between the \((Y, R, g)\) formulation and the implied pair \((Y, Y^*)\). So one might try to define the CAR condition in terms of \((Y, R, g)\) by

\[
\Pr(R = r \mid Y = y) \text{ is constant on } \{y \mid g(y, r) = y^*, \Pr(Y = y) > 0\} \tag{2.48}
\]

or

\[
\Pr(R = r \mid Y = y) = \Pr(R = r \mid Y \in \{y \mid g(y, r) = y^*\}) \tag{2.49}
\]

or

\[
\{Y = y\} \perp \perp \{R = r\} \mid \{Y \in \{y \mid g(y, r) = y^*\}\} \tag{2.50}
\]

for all \( y^* \in \mathcal{Y}^* \) where the conditions are only required to hold for \( y \) of positive probability. These conditions parallel the CAR conditions formulated in (2.28), (2.29), and a combination of (2.21) and (2.29), respectively. I will call any of these equivalent conditions R-CAR.

The example of \( g(., .) \) in (2.41) does not satisfy condition (2.47), but (2.42) satisfies both conditions (2.46) and (2.47). It also satisfies the R-CAR condition (2.48) as shown above. Thus one might hope that the implied tuple \((Y, Y^*)\) is CAR if there exists an \((Y, R, g)\) model satisfying (2.46) and (2.47) that is R-CAR. This is in fact true, since

\[
\Pr(Y^* = y^* \mid Y = y) = \Pr(R \in \{r \in \mathcal{R} \mid g(y, r) = y^*\} \mid Y = y) = \sum_{r: g(y, r) = y^*} \Pr(R = r \mid Y = y) \tag{2.51}
\]
But by (2.47), \( g(y', r) = y^* \) for all \( y' \in y^* \) so that the summation is over the same set of values of \( r \). From R-CAR (2.48), the \( \text{Pr}(R = r \mid Y = y) \) are constant on \( y^* \). It follows that \( \text{Pr}(Y^* = y^* \mid Y = y) \) is constant on \( y \in y^* \) and thus CAR.

Jaeger (2005b: Theorem 2.9) claimed that the converse holds as well, namely that if the distribution of \( (Y, Y^*) \) is CAR, then there is an R-CAR model. But this is wrong since the conditions (2.46) and (2.47) rule out the representation of CAR models where \( \mathcal{Y}^* \) is not a union of partitions. As Example 3 demonstrates, even in this case there is a joined distribution of \( (Y, Y^*) \) that is CAR. The inconsistency arises because Jaeger uses the trivial \( (Y, R, g) \) construction given at the beginning of this section to prove the existence of R-CAR. However, this construction violates restriction (2.46) which Jaeger presupposes in his Definition 2.5.

There is another way in which \( (Y, R, g) \) representations together with the conditions (2.48)–(2.51) may fail to be equivalent to CAR. The restrictions (2.46) and (2.47) insure that \( g(., .) \) is sensitive to the \( y \) argument so that examples like the trivial construction at the beginning of the section are ruled out. In that case, \( g(., .) \) did not even depend on \( y \). But it may as well be that \( g(., .) \) is insensitive to values of \( r \). In that case, the conditions (2.48)–(2.51) would be either void (using any independent distribution of \( R \)) or false (using a distribution of \( R \) contradicting (2.48)–(2.51)) without changing the implied joined distribution of \( (Y, Y^*) \). A simple example was provided by Jaeger (2005b: 895). Suppose that \( Y = (Y_1, Y_2) \) is a two-dimensional binary variable and that the corresponding \( \mathcal{Y}^* \) is just \{\{(0, 0), (0, 1)\}, \{(1, 0)\}, \{(1, 1)\}\}. Here, \( Y_2 \) is missing if \( Y_1 = 0 \). Otherwise, all values are present. Now one may choose

\[
g((0, 0), r) := g((0, 1), r) := \{(0, 0), (0, 1)\} \\
g((1, 0), r) := \{(1, 0)\} \\
g((1, 1), r) := \{(1, 1)\}
\]

so that \( g(., .) \) does not depend on \( r \) at all. If \( \mathcal{R} = \{r\} \), then (2.48)–(2.51) are trivially true. But with \( \mathcal{R} = \{r_1, r_2\} \) one may easily produce examples that falsify the R-CAR conditions. E.g., one might choose

\[
\text{Pr}(R = r_1 \mid Y = (1, 0)) = \text{Pr}(R = r_1 \mid Y = (1, 1)) = 1
\]
Since \( g(.,.) \) does not depend on \( r \), the implied distribution of \((Y, Y^*)\) stays the same, whatever distribution is assumed for \( Y \). Now

\[
\Pr(R = r_1 \mid Y = (0, 1)) = 1/3 \neq \Pr(R = r_1 \mid Y = (0, 0)) = 2/3
\]

and also

\[
\Pr(R = r_2 \mid Y = (0, 1)) = 2/3 \neq \Pr(R = r_1 \mid Y = (0, 0)) = 1/3
\]

so that (2.48) is violated.

In consequence, one might try to restrict the setup even further by asking that \( g(y, .) \) is injective for all fixed \( y \). Then \( r \) is a function of \( y^* \) such that for \( y \in y^* \)

\[
y^* = g(y, r) \iff r = h(y^*)
\]

for some function \( h(.) \).

In this case, (2.52) reduces to

\[
\Pr(Y^* = y^* \mid Y = y) = \Pr(R = h(y^*) \mid Y = y)
\]

so that if the left hand side is constant on \( y^* \) then so is the right hand side and vice versa. In this special case, CAR and R-CAR are equivalent: If there is a representation of a \((Y, Y^*)\) pair in terms of a triple \((Y, R, g)\) satisfying (2.46) and (2.47), and such that \( R \) is a function of \( Y^* \), then the joined distribution of \((Y, Y^*)\) is CAR if and only if \((Y, R, g)\) is R-CAR.

However, this is a very special situation. Not only is \( Y^* \) the union of partitions indexed by \( r \), say \( Y^*_r \), but the intersection of the \( Y^*_r \) must be empty. Measurements at different levels of precision, i.e. unions of strict refinements of partitions provide an example. On the other hand,

---

46 If the condition holds, then Jaeger (2005b: 894) calls \( R \) ‘observable’. But this is an unfortunate choice of term since it conflates the “simple device” and its modelling methods with the observations obtained in a survey.
\( \mathcal{Y}^* = \{\{1, 2, 3\}\}, \{\{1\}\}, \{\{2, 3\}\}, \{\{1\}, \{2\}, \{3\}\} \) from Example 1 neither is a union of strict refinements nor does it permit a function \( h(.) \) since \( \{1\} \) appears in two of the partitions and there is no other way to arrange the partitions. Hence, a function \( h(.) \) that would select a partition based only on \( y^* \) can not exist. Similarly, neither Example 2 nor Example 3 are of this form, nor are censored data.\(^{47}\)

Thus, the CAR and R-CAR conditions are guaranteed to be equivalent only if \( R \) is a function of \( Y^* \). But this condition severely restricts the structure of \( \mathcal{Y}^* \). If the restriction does not hold or if another \( (Y, R, g) \) structure is deliberately chosen, then there are CAR situations that are R-CAR or not depending on the choice of \( R \) and \( g(., .) \). If this variable is just an addition to the original probability model, a randomisation device that only involves the distribution of \( (Y, Y^*) \) and that simply serves as a shortcut in computations involving the joined distribution, then there is no criterion to choose between competing possible versions and the choice can be based on the equivalence given above. However, more often than not, \( R \) is given some more or less intuitive interpretation. In that case, one must be aware that the statement of the CAR condition can not safely be formulated within the \( (Y, R, g) \) framework alone.

A particularly interesting case arises from censored data where the censoring is generally constructed from an extra variable \( R \) of censoring times such that

\[
g(y, r) = \begin{cases} 
\{y\} & r > y \\
\{r, \ldots, \tau\} & r \leq y
\end{cases}
\]

(2.54)

so that the variable is censored if \( R \leq y \). Here, \( R \) is not a function of \( Y^* \), but \( g(., .) \) at least satisfies (2.46) and (2.47). This last requirement also motivates the definition of censored sets by \( r \leq y \) even though the

\(^{47}\) Note that unions of strict refinements form a graded partially ordered set with respect to set inclusion in that all maximal linearly ordered subsets (flags) of \( \mathcal{Y}^* \) have the same number of elements \( |\mathcal{R}| \). Of course, unions of strict refinements are not the only possible forms of \( \mathcal{Y}^* \). The requirement (2.53) says only that no element of \( \mathcal{Y}^* \) is contained in more than one of the partitions \( \mathcal{Y}^*_r \). A situation somewhat dual to that of strict refinements arises when no element of a partition \( \mathcal{Y}^*_r \) is the union of elements from another partition. \( \mathcal{Y}^* = \{\{1\}, \{2, 3\}\} \cup \{\{2\}, \{1, 3\}\} \) provides an example. Obviously, there are intermediate cases as well.
more traditional choice would take \( r < y \). But with the latter convention, (2.54) would be violated when \( y = r \). With this choice, if the model is assumed to be R-CAR, then the implied distribution of \((Y, Y^*)\) must be CAR.

In the history of the analysis of censored data, the most prominent model certainly is that of a random censoring variable \( R \) that is independent of \( Y \). The presumption is that there is a random variable \( R \) such that \( R \perp \perp Y \) and such that \( Y^* = g(Y, R) \). Within this setup, CAR follows since \((Y, R, g)\) trivially satisfies (2.48), so it is R-CAR, and since \( g(., .) \) is of the right kind, it is also CAR. In consequence, the “simple device” sanctions the use of the face value likelihood that ignores how censored data come about.\(^{48}\)

There are many drawbacks of this formulation that actually hampered the development of a clear understanding of censored data methods. A rather obvious problem occurs when the event of interest in a clinical study is the time of death of a patient. In that case, ‘censoring after death’ (the case \( R > Y \)) generally makes no sense when \( R \) refers to drop-out decisions or depends (possibly by protocol) on particular values of clinical variables. A possible remedy is to redefine \( R \) by setting \( R := \tau + 1 \) if \(|Y^*| = 1\) (\( Y \) is exactly observed) and \( R := \min Y^* \) otherwise. This convention does not change the status of R-CAR since R-CAR is trivially true when \(|Y^*| = 1\). Moreover, the redefined \( R \) is a function of \( Y^* \) so that if this model is not R-CAR then it cannot be CAR by the previous equivalence result. However, the redefinition destroys the global condition \( R \perp \perp Y \), a condition that makes working with the \((Y, R, g)\) model so attractive. In particular, if \( Y^* = \{y\} \) then \( R = \tau + 1 \) so that independence cannot hold.\(^{49}\)

\(^{48}\) Remember that the present framework is completely nonparametric so that problems connected with parameter dependence do not arise.

\(^{49}\) See the remarks in van der Laan and Robins (2003: 23–24) as well as Rubin (2006) and the discussion of his article. Further discussion can be found in Egleston et al. (2007). A particular case is taken up by Frangakis et al. (2007) and Imai (2008). Those articles include references to arguments about counterfactual models and their connexion to the “simple device”.

A superficially similar example with quite different consequences occurs when time to divorce (say \( Y \)) is of interest and death of a partner is treated as ‘censoring.’ This is just the reverse of the situation contemplated previously, where the value of an
As shown in the previous section, to any distribution of \( Y^* \) there is a joined distribution of \((Y, Y^*)\) that is CAR. Since for censored data the set \( \mathcal{Y}^* \) is hierarchically ordered, the solution of the CAR equations is straightforward. Moreover, the solution is unique on the finest partition with positive \( y^* \) probability. Of course, the latter two properties hold only when all elements of \( \mathcal{Y}^* = \{\{0\}, \ldots, \{\tau\}, \{0, \ldots, \tau\}, \{1, \ldots, \tau\}, \ldots\} \) have positive probability. Note that Example 1 has the required structure except that other forms of \( g(\cdot, \cdot) \) were considered.

Reformulating the example slightly by using \( \mathcal{Y} = \{0, 1, 2\} \) instead of \( \{1, 2, 3\} \) and giving all elements of \( \mathcal{Y}^* \) the probability 1/5, the joined CAR distribution of \((Y, Y^*)\) is:

\[
\begin{array}{cccccc}
{0} & {1} & {2} & {1, 2} & {0, 1, 2} \\
0 & 1/5 & 0 & 0 & 2/40 & 1/4 \\
1 & 0 & 1/5 & 0 & 1/10 & 3/40 & 3/8 \\
2 & 0 & 0 & 1/5 & 1/10 & 3/40 & 3/8 \\
\end{array}
\]

In this case, it is easy to see that both distributions imply the joined CAR distribution of \((Y, Y^*)\) and that both are R-CAR. However, only in the left table are \( R \) and \( Y \) stochastically independent, while in the right table \( \Pr(Y = 0, R' = 1) = 4/40 \neq \Pr(Y = 0) \Pr(R' = 1) = 1/4 \cdot 3/10 \) and similarly for \( \Pr(Y = 0, R' = 2) \). Here, \( R \) is a function of \((Y, Y^*)\) on \( \{Y \in \{1, 2\}\} \times \mathcal{Y}^* \cup \{Y = 0\} \times \{0, 1, 2\} \) so that only

\[\text{assumed } R \text{ was not well defined. It is now } Y \text{ that is not a well defined quantity because it presupposes that a time to divorce } Y \text{ is defined even if one of the partners died previously. But this is more an example of incompetent modelling than a problem of the random censoring model.}\]
the probabilities \( \Pr(R = 1 \mid Y = 0) \) and \( \Pr(R = 2 \mid Y = 0) \) are undefined by the requirement that the joined distribution of \((Y, Y^*)\) should be reproduced by the \((Y, R, g)\) model. Hence, all feasible \((Y, R, g)\) models are R-CAR in this situation, but some do not arise from independent censoring.

In fact, the situation presented above and its solution hold generally: When \( g(., .) \) is of the form (2.54), then to any \( Y^* \) representing censored data (i.e. when \( Y^* \) contains only sets of the form \( \{y, \ldots, \tau\} \) and the singletons) and any distribution of \( Y^* \) there are random variables \( R \) such that \((Y, R, g)\) is R-CAR and implies the CAR distribution of \((Y, Y^*)\). To see this, one has only to note that CAR and R-CAR are restrictions only for the censored sets in \( Y^* \). But for these (and \( g(., .) \) of the given form), \( R \) is a function of \( Y^* \) so that the argument for strictly invertible \( g(., .) \) applies here as well. The general condition of invertibility (2.53) is too strong. It is enough to require that \( r \) is a function of \( y^* \) only for those \( y^* \) with \( |y^*| > 1 \).

Furthermore, the conditional independence of these R-CAR models implied by (2.51) can be extended to unconditional independence (just counting equations) and this extension defines an independent censoring model (i.e. \( R \perp \perp Y \)). Thus to any censored data model there is one that can be represented by an independent censoring variable \( R \).

This simple fact, while known for special cases (e.g. Berman 1963), was in some generality only observed in 1977 by Williams and Lagakos. In particular, they showed that a CAR distribution of \((Y, Y^*)\) (in the absolutely continuous case) can always be represented by an independent censoring model (Williams and Lagakos 1977: Theorem 3.1). Their constant sum condition has prompted the development of quite a few reformulations of conditions where likelihood inference based on the face value observations is valid (Kalbfleisch/MacKay 1979, Ebrahimi et al. 2003). The existence of several R-CAR distributions in the \((Y, R, g)\) model was also noted for interval censoring (when only a random interval is known during which the event of interest happened) and for current status data (when it is only known whether an event happened before a random inspection time or not). See Oller et al. (2004, 2007) and Lawless (2004) for the case of interval censoring and Betensky (2000)
for current status models. For competing risks, Langberg et al. (1978) noted the usefulness of constructing independent representations of the type \((Y, R, g)\) for the handling of many questions in the description of stochastic systems. But only much later, Crowder (1996, 1997, 2000) clarified the situation also for discrete and mixed time models.

The relation of the \((Y, R, g)\) formulation to CAR becomes even more complex when multidimensional censored times are considered. In this case, one may consider a \(g(., .)\) as in (2.54) applied coordinate-wise: If \((Y_1, Y_2)\) and \((R_1, R_2)\) are times and censoring times, then

\[
g(y_1, y_2, r_1, r_2) = \begin{cases} 
\{(y_1, y_2)\} & r_1 > y_1, r_2 > y_2 \\
\{r_1, \ldots, \tau_1\} \times \{y_2\} & r_1 \leq y_1, r_2 > y_2 \\
\{y_1\} \times \{r_2, \ldots, \tau_2\} & r_1 > y_1, r_2 \leq y_2 \\
\{r_1, \ldots, \tau_1\} \times \{r_2, \ldots, \tau_2\} & r_1 \leq y_1, r_2 \leq y_2 
\end{cases}
\]

But now it may happen that a CAR distribution cannot be represented by independent pairs \((Y_1, Y_2) \perp (R_1, R_2)\). For an example, suppose that \((Y_1, Y_2)\) are times to consecutive events so that the event times are \((Y_1, Y_1 + Y_2)\). Let \(R\) be an independent variable that censors \((Y_1, Y_1 + Y_2)\) in calendar time. Then either \(R > Y_1 + Y_2\) so that \(Y^* = (Y_1, Y_1 + Y_2)\), or \(Y_1 < R \leq Y_1 + Y_2\) so that \(Y^* = \{Y_1\} \times \{R, \ldots, \tau\}\), or, lastly, \(R \leq Y_1\) so that \(Y^* = \{R, \ldots, \tau\} \times \{R, \ldots, \tau\}\). When \(g(., .)\) is chosen as in (2.55), then one may say that \((R, R)\) censors \((Y_1, Y_1 + Y_2)\). This model is trivially \(R\)-CAR and since \(r\) is a function of \(y^*\) for the censored elements of \(Y^*\) it is also CAR.

If one is interested in the joined distribution of \((Y_1, Y_2)\) (each episode starts from time 0), then the correspondingly transformed \(Y^*\) consists of either complete observations \((y_1, y_2)\), or the half lines \(\{y_1\} \times \{y_2, \ldots, \tau_2\}\), or the half spaces \(\{y_1, \ldots, \tau_1\} \times \mathcal{Y}_2\). The form of incomplete elements of \(\mathcal{Y}^*\) is illustrated in Figure 2.13.

The model clearly is still CAR. After all, it is just a one-to-one transform of the original model. If one constructs an \((R_1, R_2)\) for the \(g(., .)\) in (2.55), then one gets \(R_1 \coloneqq R, R_2 \coloneqq \max(0, R - Y_1)\). But if \((Y_1, Y_2) \perp R\), then
this can not be the case for \((Y_1, Y_2)\) and \((R_1, R_2)\): \(R_2\) is a function of \(Y_1\) unless \(\Pr(\{R < Y_1\} \cup \{R > Y_1 + Y_2\}) = 1\).

It follows that there are CAR models that can not be represented by independent censoring variables \((R_1, R_2)\). Modelling bivariate censored data by an independent censoring model \((Y, R, g)\) restricts the class of CAR models. Gill et al. (1997: 270–272) note the statistical consequences: When only CAR is assumed, then one must use the non-parametric maximum likelihood estimator (see also Chapter 8). On the other hand, when only the smaller set of models that assume independent censoring according to (2.55) is contemplated, then there are competing, non-equivalent estimators that may be even better (or at least more stable) than the non-parametric maximum likelihood estimator in small samples.

Moreover, \(R_1\) is not a function of \(y^*\) when \(y^*\) is the half line \(\{y_1\} \times \{r_2, \ldots, \tau_2\}\). One may therefore change the conditional distribution of \(R_1\) given \(Y\) on \(\{R_1 > y_1\}\) so that it depends on the values of \(y_2 \in \{r_2, \ldots, \tau_2\}\) without changing the joined distribution of \((Y, Y^*)\). Thus there can be \((Y, R, g)\) models that are not R-CAR according to (2.48) but that imply a joined distribution of \((Y, Y^*)\) that is CAR. R-CAR is not a necessary condition for CAR. Working within a \((Y, R, g)\) model for multidimensional censored data makes it difficult to decide whether the CAR condition can be stipulated, even for rather restricted structures of
2.3. Coarsening Variables and Coarsening Completely at Random

\( Y^* \).

2.3.2. Completely Coarsening at Random

A further consideration in connection with the \((Y, R, g)\) model of considerable recent interest is a strengthening of the R-CAR condition. It is a condition that has been termed coarsening completely at random (CCAR). It is a rather obvious extension in the case of missing data where it simply states that incompleteness does not depend on any other information in the model. The standard example is that of a missing dependent variable in a regression model where CCAR says that the probability of a missing dependent variable does not depend in any way on the values of the covariates. The extension to general incomplete data models was suggested by Heitjan (1994).

The CCAR condition can be easily defined in terms of \((Y, R, g)\) models by strengthening condition (2.48) to

\[
\Pr(R = r \mid Y = y) \text{ is constant} \quad (2.56)
\]

or equivalently—strengthening condition (2.51)—as

\[
\{Y = y\} \perp \perp R \quad (2.57)
\]

for \(y\) of positive probability. When CCAR is defined using the \((Y, R, g)\) model I call any one of the conditions the R-CCAR condition.

The condition (2.57) asking for stochastic independence of \(Y\) and \(R\) is arguably the most famous notion of CAR cited in the applied literature. The fame of this strong version of R-CAR stems from the fact that far reaching conclusions can be drawn from it with ease.

In view of the difficulties of R-CAR conditions to provide an unanimous definition of CAR it would be very convenient to have a formulation that could be based on the distribution of \((Y, Y^*)\) without recourse to some \(R\). It turns out that this is an extremely difficult task. A little contemplation
reveals that for a strengthening of the condition (2.28) one would need to require
\[
\text{Pr}(Y^* = y^* \mid Y = y) \text{ is constant }
\] (2.58)
which is self-contradictory unless the variable \(Y^*\) would only take the value \(Y\) so that the incomplete variable carries no information at all. The reason is that for some \(y\) and an \(y^*\) not containing \(y\) the probability must be 0. It follows that the probability is 0 for all \(y\) and thus the marginal probability of any \(y^* \neq Y\) must be 0. The same problem shows up in all the alternative formulations of the CAR conditions.

There were a few attempts to remedy the situation. An obvious one was suggested by Jaeger (2005b) who in parallel with the CAR situation defined a model distribution to be CCAR if there was at least one R-CCAR version of \((Y, R, g)\) compatible with \((Y, Y^*)\).

I will argue here that CCAR is properly viewed as a condition on the structure of \(Y^*\) that extends CAR and not as a stronger version of the CAR condition itself. If CAR is equivalent to a factorisation of the fully non-parametric incomplete data likelihood as in (2.32) then there can be no further information that would ameliorate the estimation problem. CAR without any further conditions permits to use only the variable of interest part of the likelihood while disregarding the coarsening model, at least for likelihood inference. In fact, most examples for R-CCAR in contrast to CAR introduce further variables into a problem and equate R-CCAR then with the ability to drop the relation of these extra variables to the variable \(R\) and thus the relevant likelihood factors from consideration. But this is more like a re-expression of the standard CAR condition extended to more complicated models. Examples of R-CCAR that are not just barely disguised versions of CAR hint in another direction: Loosely speaking, \(R\) indicates the degree of incompleteness. Thus the R-CCAR condition says that it is possible to determine the degree of incompleteness without knowledge of the actual \(y\) drawn in a procedure. Therefore, CCAR should single out a certain class of procedures to generate incomplete data. These procedures should not depend in any way on the procedures used to create the values of \(Y\). If that was the case, then the outcome of combining the independent procedures to create
2.3. Coarsening Variables and Coarsening Completely at Random

the coarsening and to create the variable of interest would be a special case of CAR, or so one would hope.

Such independent coarsening procedures certainly exist: If $\mathcal{Y}^*$ is a union of partitions $\mathcal{P}_1, \ldots, \mathcal{P}_k$ one can do the following: select one of the partitions with probability $p^r$. Next generate an $y$ according to the marginal distribution of $Y$ and report as $y^*$ the element of the chosen partition into which $y$ falls. As Jaeger (2005b: Theorem 2.14) shows, this is also the only way R-CCAR may come about: It follows from the restrictions (2.46) and (2.47) on $g(.,.)$ already discussed in the previous subsection. As shown there, any R-CAR model $(Y, R, g)$ where $g(.,.)$ satisfies (2.46) and (2.47) implies that $\mathcal{Y}^*$ is a union of partitions indexed by $r \in \mathcal{R}$. If the model is R-CCAR as well, then $R$ and therefore the partition chosen is independent of $Y$ and so the above algorithm applies. In the other direction, one needs only to set $\mathcal{R} = \{1, \ldots, k\}$ and $\Pr(R = r \mid Y = y) = p^r$ for all $y$ of positive probability. This is obviously R-CCAR.

2.3.3. Coarsening Variables as Additional Data

Even though the $(Y, R, g)$ model and the R-CAR conditions can only be made equivalent with CAR under very restrictive conditions on $\mathcal{Y}^*$ and the form of $g(.,.)$ and $\mathcal{R}$, there is a sense in which it might be more general than the model solely based on $(Y, Y^*)$. Consider the case of censored data represented by a censoring time $R$. Occasionally it is possible to identify such a variable with further information in the situation. This might be the case when all censoring events are recorded even when they happen after the event of interest. In that case the value of $R$ would always be known and should be introduced into the model. This cannot be represented in terms of the original variables $(Y, Y^*)$ alone. One must at least expand the set of contemplated observations to some $\mathcal{Z}$, say. Then CAR and similar concepts must relate to the common distribution of $(Y, Z)$ where $Z$ is a random variable with values in $\mathcal{Z}$. To connect this general setup with models of incomplete data, there must be a function

$$\alpha: \mathcal{Z} \longrightarrow \mathcal{P}(\mathcal{Y}) \quad (2.59)$$
that presents the coarsened values of $Y$ such that consistency holds, i.e.

$$\Pr(Y \in \alpha(Z)) = 1$$

The generalisation fits well with the $(Y, R, g)$ formulation of the previous section. Now one would set

$$g : \mathcal{Y} \times \mathcal{R} \longrightarrow \mathcal{Z}$$

and put

$$\alpha(z) := \{y \in \mathcal{Y} | \exists r \in \mathcal{R} : g(y, r) = z\}$$

An obvious advantage of using a general space $\mathcal{Z}$ together with a function $\alpha(.)$ instead of $\mathcal{Y}^*$ is the ability to use mathematically convenient spaces in place of sets of subsets. In the censored data case I have already used informally the standard notation $(Y', \delta)$. The corresponding $\alpha(.)$ is the one-to-one transformation with $\alpha(y, 1) = \{y\}$ and $\alpha(y, 0) = \{y, \ldots, \tau\}$. Furthermore, the move allows to work with situations where the set of subsets is too large to allow a reasonable probabilistic treatment. An example is the arbitrary grouping of a continuous variable to subsets of the real line. If the subsets can be taken to be intervals, one can replace the sets by their upper and lower bounds so that one can conveniently work in $\mathbb{R}^2$.

Mathematical convenience aside, the impact of additional ‘information’ on the concept of CAR is of primary interest here. A direct generalisation of the CAR conditions (2.21), (2.28), and (2.30) gives either

$$\Pr(Y = y \mid Z = z) = \Pr(Y = y \mid Y \in \alpha(z))$$

\[ \tag{2.60} \]

---

50 The translation into manageable random variables $Z$ was used by Gill et al. (1997: Sections 6–9) to formulate CAR when $\mathcal{Y}$ is infinite, and in particular when $\mathcal{Y} = \mathbb{R}$. The most appropriate way to do so is still debated. See Jacobsen and Keiding (1995), Nielsen (2000), and Cator (2004) for some critical remarks.

51 The term ‘information’ is used in quotes here, since random variables and conditional distributions have properties that sometimes contradict the properties the term carries in informal use. See Dubra and Echenique (2004) for an example similar to Billingsley’s (1979: Example 33.1).
or

\[
\Pr(Z = z \mid Y = y) \text{ is constant on } y \in \alpha(z) \tag{2.61}
\]

or

\[
\{Y = y\} \perp \{Z = z\} \mid \{Y \in \alpha(z)\} \tag{2.62}
\]

for \(y\) of positive probability. The equivalence of these conditions follows as in the previous, simple case. I will call this generalised CAR condition augmented CAR or A-CAR.

From the last condition, since \(Y^*\) is a (measurable) function \(\alpha(.)\) of \(Z\), the simple CAR condition (2.30) \(\{Y = y\} \perp \{Y^* = y^*\} \mid \{Y \in y^*\}\) follows from the condition (2.62). The requirements (2.60)–(2.62) can therefore be strictly stronger than the CAR condition discussed thus far. Consequently, one cannot expect that the slogan ‘CAR is everything’ will hold in this situation. There may well be distributions of \(Z\) (and functions \(\alpha(.)\)) such that no joined distribution of \((Y, Z)\) satisfying (2.60)–(2.62) exists. It is clear that adjoining further ‘information’ to \(Y^*\) — whether in the form of an explicit random variable \(R\) in a generalised model \((Y, R, g)\) or simply as a joined distribution of \((Y, Z)\)—will generally destroy the conditional independence relation that can always be established for simple CAR. After all, knowing \(\{Z = z\}\) might entail much more detailed ‘information’ than knowing \(\{Y^* = \alpha(z)\}\).

The Monty Hall problem illustrates the expressiveness of A-CAR: Until now I took as starting point the moment when the contestant has announced his choice of door 2. In that case one might take \(\mathcal{Y} = \{1, 2, 3\}\) and \(Y^* = \{\{1, 2\}, \{2, 3\}\}\) as before. Now suppose that we step back to the situation just before the contestant chooses a door. Then there is a new element that needs to be represented in the model: The initial choice of the door number by the contestant. Denote this choice by \(C\). Furthermore, denote by \(R\) (the number of) the remaining door, the one not opened by Monty and not chosen by the contestant (this is known after the contestants choice and Monty’s reaction). Then, according to the rules of the game, the price must be either behind door \(C\) or behind door \(R\). One may define \(Z := (R, C)\) as the relevant information.
After all, both the choice of the contestant and the choice of the show master might be regarded as part of the probability model describing the Monty Hall problem. With both parts of ‘information’ available—the choice of the contestant and the choice of the host—one should be able to represent the outcome as \{((C = c, R = r))\}. The function \(\alpha(.)\) maps the ordered pair to the set \{\(R, C\)\} with the same interpretation as before: The doors that may hide the price after Monty opened the door with the goat. Suppose that \(C\), the choice of the contestant, is independent from \(Y\). Now

\[
\Pr(R = 1 \mid Y = 1, C = 2) = 1
\]

since Monty has no choice when the price is not behind the door chosen by the contestant. If he has a choice, suppose it is not deterministic so that

\[
\Pr(R = 1 \mid Y = 2, C = 2) =: p_1 \notin \{0, 1\}
\]

Then

\[
\Pr(Y = 1 \mid Z = (1, 2)) = \frac{\Pr(C = 2) \Pr(Y = 1)}{\sum_y \Pr(Z = (1, 2) \mid Y = y) \Pr(Y = y)} = \frac{\Pr(Y = 1)}{\Pr(Y = 1) + p_1 \Pr(Y = 2)} \neq \Pr(Y = 1 \mid Y \in \{1, 2\}) = \Pr(Y = 1 \mid Y \in \alpha((1, 2)))
\]

so that (2.60) is wrong unless \(\Pr(Y = 2) = 0\) or \(\Pr(Y = 1) = 0\). By a similar computation, \(\Pr(Y = 3 \mid Z = (3, 2)) \neq \Pr(Y = 3 \mid Y \in \{2, 3\})\) unless \(\Pr(Y = 3) = 0\) or \(\Pr(Y = 2) = 0\). Thus on \(\{C = 2\}\) A-CAR can only hold when \(\Pr(Y = 2) \in \{0, 1\}\). This is of course the same result as the one obtained earlier. After all, the derivation was basically conditional on \(\{C = 2\}\). But the previous discussion could only proceed unambiguously by conditioning on the choice of the contestant. Otherwise, using the possible locations of the price as the coarsened information could be misleading: The set \(\{1, 2\}\) might come about because the contestant has chosen door 2 and Monty had opened door 3, or because the contestant
2.3. Coarsening Variables and Coarsening Completely at Random

had chosen door 1 and Monty opened door 3, a difference that can not be expressed in the subset formulation. With A-CAR, we can now speculate explicitly about both C, the choice of the contestant, and R, Monty’s choice.

First, if one considers different choices of the contestant, the CAR (or A-CAR) condition requires \( \Pr(Y = 1) \in \{0, 1\} \) if \( C = 1 \), \( \Pr(Y = 2) \in \{0, 1\} \) if \( C = 2 \), and \( \Pr(Y = 3) \in \{0, 1\} \) if \( C = 3 \). If at least two of the three choices have positive probability, then the second CAR solution (\( \Pr(Y = 2) = 0 \) and \( \Pr(Y = 1) = \Pr(Y^* = \{1, 2\}) \), \( \Pr(Y = 3) = \Pr(Y^* = \{2, 3\}) \)) is ruled out. Note that I still use the weak version of CAR where the condition is trivially true for degenerate distributions of \( Y \).

Secondly, one may consider Monty’s behaviour. The above argument worked only if he decided non-deterministically. Now suppose he always chooses the larger number if he has a choice. So he opens door 3 (and never door 1) when the contestant has chosen door 2 and the price is behind that door. Then

\[
\Pr(R = 1 \mid Y = 2, C = 2) = 1
\]

and therefore

\[
\Pr(Y = 1 \mid Z = (1, 2)) = \frac{\Pr(C = 2) \Pr(Y = 1)}{\sum_y \Pr(Z = (1, 2) \mid Y = y) \Pr(Y = y)} = \frac{\Pr(Y = 1)}{\Pr(Y = 1) + \Pr(Y = 2)} = \Pr(Y = 1 \mid Y \in \alpha((1, 2)))
\]

in accordance with A-CAR. But now

\[
\Pr(R = 3 \mid Y = 2, C = 2) = 0
\]

so that if he does open door 1 and thus \( R = 3 \), the price must be behind door 3,

\[
\Pr(Y = 3 \mid Z = (3, 2)) = 1 \neq \frac{\Pr(Y = 3)}{\Pr(Y = 2) + \Pr(Y = 3)}
\]
unless $\Pr(Y = 2) = 0$ (or $\Pr(Y = 3) = 0$, in which case the condition is irrelevant), a weaker condition than the previous result. Such possibilities are generally overlooked when the Monty Hall problem is discussed using some kind of simulation.

While A-CAR can be much more expressive and easier to work with, it is not obvious what the role of $R$ should be. If it represents additional ‘information’, then presumably it should be incorporated into the model by adding it to $Y$. Here it should make no difference whether $R$ is completely specified ($R$ is ‘observable’ in Jaeger’s terminology) or only partial given as in the $(Y, R, g)$ formulation. In this perspective, A-CAR is not a generalisation of CAR. The introduction of $R$ or other auxiliary variables is simply an indication that the original model was chosen too small. In the properly enlarged model, A-CAR reduces to CAR.

On the other hand, such an argument disregards the requirement that $Y$ ought to represent particular facts within the model. When $Y$ and $\mathcal{Y}$ are used to express the social facts of interest, extensions of the definition of $Y$ can not be justified by particular considerations of aspects of the “simple device”. But now the additional random variable $R$ certainly introduces further structure beyond the one present in the simple probability model, at least when $R$ is not just a function of a possibly extended $Y^*$. And as was demonstrated in the last section, the corresponding R-CAR and A-CAR conditions need not be equivalent with CAR. In particular, when $R$ is used to represent genuine further ‘information’ that nevertheless should not be incorporated into a larger model, then A-CAR will often fail. The slogan ‘A-CAR is everything’ is certainly not correct. A-CAR can be rejected empirically. It can not be invoked in the same way as CAR can, as a simple extension of the “simple device”. It is not a modelling decision.

Note, however, that the decision to distinguish between $Y$ and $R$ is a modelling decision. It will almost always rest on an idea of a ‘mechanism’ that brings about incompleteness. $R$ is used to represent aspects of this ‘mechanism’ and, since the ‘mechanism’ is thought to operate after the fact represented by $Y$ came about, it can not be part of $Y$ itself. While I do not think that the idea of a ‘mechanism’ acting independently from $Y$ can be given a reasonable interpretation in social science applications,
the distinction it hints to might still be relevant when versions of A-CAR are contemplated.

2.4. Conclusions: CAR Modelling

Even though the connection with empirical investigations still needs clarification and the discussion up to now was confined to the model world, the results for the “simple device” are much stronger than could be expected from a discussion within the framework of classical sampling theory. First of all, I have shown that the introduction of a probability model does not restrict the possible underlying values beyond the consistency requirement so that the “simple device” does not prejudice certain solutions against others. Secondly, it is possible to show the mathematical equivalence of several formalisations of randomly coarsened data. This can’t be done in the sampling framework since in that case decompositions and independence relations can hold at most approximately. And it is possible to deduce that the distribution of randomly coarsened variables will imply a unique distribution of the variable of interest. The latter result is more or less obvious in the case of either exact or completely missing data where it may be formulated within the classical framework of sampling as well. But in all other cases, and thus in situations of much larger practical relevance, this could only be done within the probability model.

Since to any distribution of the coarsened variables there always exists a (weak) CAR model, it is impossible to reject the suggestion by referring to observations of the coarsened variables. CAR is always a viable choice in the model framework that can not be criticised on empirical grounds. CAR can therefore serve as a reference point for further speculations about the modelling within the probabilistic framework. Such speculations are not arbitrary. They are constrained by the very construction of probability models. But they clearly are not assumptions about reality. They can definitely only be formulated within the probability model. They thus presuppose the “simple device”.

121
It is however possible to judge the merits of such a model by appeal to other principles. The existence of a ‘mechanism’ may serve to exclude some cases where CAR would seemingly imply more than the validity of taking the coarsened variables at face value. When a ‘mechanism’ in the strong sense of a well defined transition kernel from \( \mathcal{Y} \) to \( \mathcal{Y}^* \) does not seem to be appropriate, one must be cautious when CAR implies several (degenerate) solutions. Similarly, extensions of CAR like R-CAR and A-CAR introduce further elements that might be used to judge model adequacy. But it turns out that these extensions can not be employed as a general reference point in the same way as simple CAR does.

While it is reassuring to learn that the “simple device” works without prejudicing certain solutions, it must be remembered that the critics refer to answer patterns of real people in real surveys. The theoretical analysis provided here can do no more than to show that the “simple device” is not inadequate by construction. An answer to the concerns of the critics, however, will also need a critically appraisal of empirical work.
3

A Case Study: Parent’s Length of Life

This Chapter studies a particular case of incomplete data. It does not aim to provide an appraisal of empirical work using the “simple device”. Rather, it provides a view of the working of the “simple device” in a complicated setting, a setting not covered by the now classical applications to missing, grouped, truncated, or censored data. Moreover, the relation of the “simple device” to classical survey statistics is examined.

The practical problem posed here is how to gain insight into mortality conditions during the early 20th century from current survey data. While official statistical agencies routinely provide life tables since the late 19th centuries, these are naturally restricted in scope and detail. Survey data may provide more background information and apply also to times and geographical areas not covered by the official statistics. More precisely, additional information about mortality in the early 20th century can be gained from current surveys when respondents were asked to provide information about their parents, in particular about their parent’s birth years, whether they were still alive at the interview date, and, if not, about their respective years of death. Both the German Life History Study (GLHS) and the Socio-economic Panel (SOEP) provide such information and one might try to use this type of information to enlarge the knowledge about mortality conditions in earlier periods.¹

¹ For previous analyses of the SOEP data about the life lengths of parents see Schepers
On the other hand, survey data on mortality of the parent generation naturally suffer from incompleteness for various reasons. Since the information about parents is supplied by their children, life length of childless persons is unavailable. It is not just missing: From survey data alone even the number of such persons can not be estimated. Moreover, the information is only available when the children themselves survived up to the survey date. And the chance of this to happen will also depend on the number of children of the parents. Thus, even though the focus is on the mortality of the parent generation, the mortality of the current generation must be taken into account. A further complication arises from the possibility that the same person is included several times in the survey because several of her or his children were questioned. This will in general be difficult or impossible to check using survey methodology. Relatedly, people with many children have a higher probability to be included in the survey. Lastly, it should not be forgotten that migration and expulsion as well as changes of territorial borders make it difficult to maintain the idea of a uniquely defined target population. In this Chapter, however, I will cling to the idea in order not to overburden the discussion.

3.1. The Structure of Descent

The first problem to be treated is that of possible dependence by descent and of unequal inclusion probabilities. In this Section and the next, I will abstract from all other aspects of the data and consider only what this type of incompleteness implies for the analysis strategies.

An extremely simplified model of the situation is depicted in Figure 3.1. Members of a generation are represented on horizontal lines by circles and Wagner (1989), and Klein (1993). In some cases, survey data about about life lengths of relatives or household members are the only source of information on mortality available. See Gakidou and King (2006) for further references to applications and a discussion of of standard methods.

124
(for women) and squares (for men).² The generations follow each other through time $t = -2, -1, 0$ down the Figure. Arrows point from parents to children. Thus all generations are exactly aligned, there is no overlap of generations. Moreover, members of a generation are distinguished by sex and the number of their offsprings only. Life length will be thought of as an additional attribute of the nodes of the graph, definable without regard to either the placement into a certain generation or the number of offsprings.

Looking forward in time, starting at $t = -2$ and following the arrows down the Figure allows to trace out all ancestral dependencies between members of different generations. I will call the graph representing the complete ancestral dependencies the *ancestral graph*.³ However, looking backward in time, starting at $t = 0$ and following the arrows in the opposite direction will recover only a part of the ancestral graph. In the Figure, the two empty circles cannot be reached from any member of the population at time $t = 0$. The resulting graph will be called *pedigree*.

² It should be obvious that the terms ‘generation’ as well as ‘parent’, ‘child’, ‘men’, etc. pertain to the model entities presented here. I have therefore abstained from making additional typographical distinctions.

³ This use of the word “ancestral graph” should not be confused with the ancestral graphs used in the theory of graphical models. See Richardson and Spirtes (2002, 2003) for the latter concept.
3. A Case Study: Parent’s Length of Life

Suppose the members of a generation can be listed in the form

\[ U_t := \{ u_{t,1}, \ldots, u_{t,N_t} \} \]

so that the size of a generation is \( |U_t| = N_t \). Also, let the generations be partitioned into their male and female members so that

\[ U_t = U_t^f \cup U_t^m \quad \text{and} \quad U_t^f \cap U_t^m = \emptyset \]

Further write \( U := \bigcup_t U_t \) for the set of members of all generations considered. Then the ancestral graph is technically speaking a directed, acyclic, bipartite graph whose nodes are the elements of \( U_f \cup U_m \). Translating the conditions of an ancestral graph into the language of graph theory, all nodes of positive indegree must have indegree exactly 2 (all people have two parents, up to the oldest generation considered). Furthermore, if there is an edge \((u_{t,i}, u_{t+1})\) from \( u_{t,i} \) to one \( u_{t+1} \) and a different edge \((u_{t,j}, u_{t+1})\) to \( u_{t+1} \), then either \( u_{t,i} \in U_t^f \) and \( u_{t,j} \in U_t^m \), or \( u_{t,i} \in U_t^m \) and \( u_{t,j} \in U_t^f \) (all people have a mother and a father).

Note that the subgraphs that arise from deleting either all nodes \( U_f \) or all nodes \( U_m \) and the respective edges gives a forest, a union of trees. Thus, restricting attention to either fathers or mothers simplifies analyses considerably.

In order to deal with the ancestral information from a statistical point of view, define a pair of statistical variables that record the mother and father of each member of a generation as

\[ m: U_{t+1} \longrightarrow U_t^f \]
\[ f: U_{t+1} \longrightarrow U_t^m \]

\[ \text{The reconstruction of pedigree or ancestral graphs from partial information is an important topic in biology. There is a large specialised literature dealing with superficially similar problems. See Tavaré (2004) for a vivid account of probability models for pedigree graphs and Steel and Hein (2006) for a particular application. However, the focus of that literature is on many generations and rarely touches statistical problems. There is also a formal connexion to the theory of random refinements of intervals (see e.g. Bertoin (2006)), which, however, also concentrates on long sequences.} \]
giving the mothers (elements of \(U_t^f\)) and fathers (elements of \(U_t^m\)) of all members of the generation \(U_{t+1}\). This is a well defined function since everyone has both a father and a mother. Now the set of parents can be written as

\[
U_1^p := \{m(u_0) \mid u_0 \in U_0\} \cup \{f(u_0) \mid u_0 \in U_0\}
\]

and one of the problems to be discussed can now be formulated as

\[
U_1^p \subsetneq U_1
\]

i.e., the parent generation \(U_1^p\) is in general a strict subset of the previous generation \(U_1\).

To make reference to children easier, it is convenient to introduce the set valued function

\[
c: U_t \longrightarrow \mathcal{P}(U_{t+1})
\]

\[
c(u_t) := \begin{cases} m^{-1}(\{u_t\}) & \text{if } u_t \in U_t^f \\ f^{-1}(\{u_t\}) & \text{if } u_t \in U_t^m \end{cases}
\]

Thus, \(c(u_t)\) is just the set of children of \(u_t\) and the set of parents can also be written as \(U_1^p = \{u_{-1} \in U_{-1} \mid c(u_{-1}) \neq \emptyset\}\).

To complete the description of this simplified model, a variable representing life length is attributed to each member of the population:

\[
T: U \longrightarrow T := \{0, 1, \ldots, \tau\}
\]

where in accordance with the discussion in Chapter 2 life length is supposed to take a finite number of values. Note that I here do not impose any restrictions on \(T(.)\), it may take any value. In particular, \(T(u) = 0\) and \(|c(u)| > 0\) is not ruled out. Also, \(T(u)\) is assumed to be defined for all \(u\), even for the currently living generation.

Finally, in the following I will drop the reference to the generation subscript whenever there is no danger of confusion. In particular, \(U_1^p\) will be denoted by \(U^p\) and \(U^c := U_0\) will denote the set of members of the current generation, the children of \(U^p\). Similarly, subscripts to the elements of these sets will be dropped when possible.
3.2. Sampling from the Children Generation

In this Section, I will ignore the problem that in general $U^P \subsetneq U$. Instead, I will pretend that interest actually centres on the parent generation. This redefines the original question by excluding childless persons from the target population. Then only two related difficulties remain: First, the more children someone of the parent generation has, the higher the probability of obtaining information about her or his life length. Secondly, he or she may be included as many times in the survey as she has children, but this will generally be unknown to the survey statistician.

To discuss the impact of these features on the sampling approach, let a sample be any subset of the population $U^c$. A collection of samples $S := \{s_1, s_2, \ldots\}$ together with a probability distribution defined on the elements of $S$ is called a design. Furthermore, let $S$ denote a random variable on some probability space, say $(\Omega^S, \mathcal{A}, \Pr)$, with values in $S$ whose distribution equals the design distribution. Thus

\[
S: \Omega^S \longrightarrow S = \{s_1, s_2, \ldots\}
\]

\[
\Pr(S = s) := \Pr(\{\omega \in \Omega^S \mid S(\omega) = s\})
\]

where measurability is taken for granted. In order to distinguish this probability from other distributions—specifically those used by the “simple device”—I will write $p(s)$ for $\Pr(S = s)$. Suppose for simplicity that the design is that of a simple random sample without replacement, i.e. all samples $s \in S$ are of equal size $|s| := n$, say, and all samples have the same probability. Then $\Pr(S = s) = p(s) = 1/(\binom{N_0}{n})$ and the inclusion probabilities of the first and second order are

\[
\pi(u) = \Pr(S \ni u) = \sum_{s \in S} \mathbb{I}_s(u)p(s) = \frac{n}{N_0}
\]

\[
\pi(u, u') := \Pr(S \ni u \cap S \ni u') = \frac{n(n-1)}{N_0(N_0-1)} \text{ for } u \neq u'
\]

These inclusion probabilities are the backbone on which most computations in the classical sampling theory rest.
When a sample is chosen according to this probability distribution (using some appropriate random device), one proceeds to ascertain the life lengths of the parents of all sampled persons, i.e. \( T(m(u)) \) and \( T(f(u)) \) for all \( u \in S \).^5

### 3.3. Induced Sampling from the Parent Population

A natural next step is to consider \( \{m(u) \mid u \in s\} \cup \{f(u) \mid u \in s\} \), the fathers and mothers of the people included in the sample \( s \), as a sample from \( \mathcal{U}^p \). This is superficially similar to sampling with replacement in that parents of several children may be included several times. The multiplicities of inclusion do not, however, result from a sequential drawing from the complete population as in most sampling designs with replacements.

I will denote the *set* of sampled parents by \( s^* \). Thus,

\[
s^* := \{m(u) \mid u \in s\} \cup \{f(u) \mid u \in s\}
\]

which I will call the *set-sample* from \( \mathcal{U}^p \). The multiplicities of elements \( u \) of \( s^* \) will be denoted by \( N(\{u\}) \). The set \( s^* \) together with the map \( u \mapsto N(\{u\}) \) for \( u \in s^* \) will be called a multiset, denoted by \( s^{**} \).^6

---

^5 Textbooks more often than not liken the values of sample statistics to random variables, possibly adding even stronger qualifications like independence or identical distributions. The reasoning seems to be that even if \( T(u) \) is a fixed quantity for each \( u \), whether it is included in a sample will only depend on the sampling design. In consequence, one might conceive of a random sample as a probability sample from \( T \), or so the reasoning goes. But on closer inspection, the reference should be to the function \( T^* : \Omega^S \to T^n \) which is the composition of \( T \) and \( S \). Since \( T \) is finite and \( S(.) \) is assumed to be measurable, the function \( T^* \) is in fact a well defined random variable. But since the values of this function will be unknown for all \( \omega \in \Omega^S \) except for the one used in the actual sample, nothing more can be said about this function. The use of the definite article when referring to such a function will be easily misleading. It seems much better to think of \( T^*(.) \) as a whole class of random variables.

^6 See Stanley (1997: Chap. 1.2) for some conventions about multisets. The multiset \( s^{**} \) is called an *ordered sample* by Särndal et al. (1992: 49) because they consider sequential draws from a population with replacement. Since this is not what actually
An example may make the situation clear. Suppose the following pedigree graph is given: Suppose further that a sample of size 3 is drawn without replacement from the 5 children. Then a possible sample comprises only the children of the first couple, \( s = \{u_{0,1}, u_{0,2}, u_{0,3}\} \) and this induces a sample \( s^{**} = \{u_{1,1}^3, u_{1,2}^3\} \) of the parent generation, where the multiplicities are noted as superscripts. This is obviously the sample of minimal effective sample size \( |s^*| = 2 \). If the sample is \( s = \{u_{0,1}, u_{0,4}, u_{0,5}\} \), then \( s^{**} = \{u_{1,1}^1, u_{1,2}^1, u_{1,3}^1, u_{1,4}^1, u_{1,5}^1, u_{1,6}^1\} \), with effective sampling size 6.

It is convenient to extend the definition of the counting function \( N(.) \) so that it counts the number of people included in the sample \( s^{**} \) from arbitrary subsets of \( A \subseteq U^p \), counting multiplicities. Any given sample of children \( s \) will then induce a counting measure, a function defined on all subsets of parents, by:

\[
N : \mathcal{P}(U^p) \rightarrow \mathbb{N}
\]

\[
N(A) := \sum_{u \in A} N(\{u\}) \quad \forall A \subseteq U^p
\]

that associates to a subset of parents \( A \) the total number of times they are included in the sample. Here, of course, \( N(\{u\}) = 0 \) if \( u \) is not included in the sample of parents \( s^* \).

\[\text{happens here, I prefer to call the resulting sample a multiset. The reduced set } s^* \text{ that only enumerates the unique elements of the sample is also called set-sample by Särndal et al. (1992: 49).}\]

\[\text{7 This justifies the rather clumsy notation } N(\{u\}) \text{ used above.}\]
Finally, the dependence of $N(.)$ on the underlying sample of children and therefore on $\omega \in \Omega^S$ must be made explicit. To do so, one can further extend the definition of $N(.)$ to the two-place function

$$N^* : \Omega^S \times \mathcal{P}(U^p) \longrightarrow \mathbb{N}$$

$$N^*(\omega, A) := \sum_{u \in A} |c(u) \cap S(\omega)| \quad \forall A \subseteq U^p \quad (3.1)$$

That is, $N^*(\omega, A)$ counts the number of times members of the subset $A$ of the parent generation are included in the sample $S(\omega)$. In this way, $N^*(., .)$ becomes a point process on $U^p$.

Note that since everyone in the sample of children has both a father and a mother, and since the sample of children is a simple random sample with fixed sample size $n$, the sample size $n^* := |s^{**}|$ (i.e. counting multiplicities) of the parent generation is

$$n^*(\omega) := \sum_{u \in U^p} N^*(\omega, \{u\}) = 2 \sum_{u \in S(\omega)} 1 = 2n$$

for all outcomes of the sampling procedure $\omega$.

The advantages of using the notion of point processes in the context of the induced sampling from the parent generation stem from the ease with which the dependencies between subsets of $U^p$ can be handled and the possibility to incorporate multisets of observations. In particular, one can define generalisations of the inclusion probabilities introduced in the previous Section. To do so, let

$$\nu(\{u\}) := \mathbb{E}(N^*(., \{u\})) = \int_{\Omega^S} N^*(\omega, \{u\}) \, d\Pr(\omega)$$

where the expectation is taken with respect to the probability distribution of the sampling design. Thus, $\nu(\{u\})$ is the expected number of times that $u \in U^p$ is included in the sample of parents. It would equal the

---

8 A point process is called simple if $N^*(., \{u\}) \leq 1$ with probability 1. Thus, we are dealing with non-simple point processes. This must be born in mind since many introductory texts on point processes concentrate on simple processes.
inclusion probabilities $\pi(.)$ if parents would be included at most once into the induced sample. More generally, for subsets $A \subseteq \mathcal{U}^p$ of parents, $\nu(A) = \mathbb{E}(N^*(., A))$ is called the intensity measure of the point process $N^*(., .)$.

To put the machinery of point processes into action, consider the problem of how to estimate the mean life time of all persons of the parent generation $\mathcal{U}^p$. Whether this is possible is not at all clear at the outset. After all, people with more children will be included with higher probability in the sample and the number of children will normally be connected to survival (or to life time). On the other hand, the sampling theory approach should be agnostic about any such relation between numbers of children and life lengths. In fact, for any parent $u$, the number of children of $u$, $|c(u)|$, and her or his life time $T(u)$, are not constrained in any way, not even probabilistically, in the present set-up.

Looking first at the sample total sum of life lengths of parents, one may consider the function

$$\hat{M}(\omega, T) := \sum_{u \in S^*(\omega)} T(u)N^*(\omega, \{u\}) = \sum_{u \in \mathcal{U}^p} T(u)N^*(\omega, \{u\})$$

This is a random function depending on the sampling design. The second expression extends the sum to all parents and makes the handling of $\hat{M}(., T)$ particularly simple and transparent. The expectation of $\hat{M}(., T)$ with respect to the design is

$$\mathbb{E}(\hat{M}(., T)) = \sum_{u \in \mathcal{U}^p} T(u)\nu(\{u\})$$

because of the additivity of expectations. To compute the intensity measure for singletons, note that

$$\nu(\{u\}) = \mathbb{E}(N^*(., \{u\})) = \sum_{k=0}^{|c(u)|} k \Pr(N^*(., \{u\}) = k)$$

---

9 This is a trivial version of Campbell’s theorem in point process theory (e.g. Reiss 1993: Chap. 5.3) or Robbin’s formula in the theory of random sets (e.g. Nguyen 2006: Chap. 2). Both are obviously related since samples of the parent generation are random (multi-) subsets of the underlying $\mathcal{U}^p$. 

132
To see how the probability of being exactly $k$ times included in the sample, $\Pr(N^*[, \{u\}] = k)$, can be computed, one may look back at the example presented at the beginning of the Section. Of the $10 = \binom{5}{3}$ different samples of children, exactly one induces the sample $s^{**} = \{u_{1,1}^3, u_{1,2}^3\}$ of parents and 3 samples (taking one of the children of the first couple and both the children of the second and third couple) of the form $s^{**} = \{u_{1,1}^1, u_{1,2}^1, u_{1,3}^1, u_{1,4}^1, u_{1,5}^1, u_{1,6}^1\}$. In the remaining 6 cases, the first couple has multiplicities 2 (2 of their 3 children are chosen) and one of the other couples appears once. Thus, $\nu(\{u_{-1,1}\}) = \nu(\{u_{-1,2}\}) = 1/10 \cdot 3 + 6/10 \cdot 2 + 3/10 \cdot 1 = 1.8$, while $\nu(\{u_{1,j}\}) = 6/10 \cdot 1 = 0.6$ for all other members of the parent generation. Note that indeed $\sum \nu(\{u\}) = 2 \cdot 1.8 + 4 \cdot 0.6 = 6$, so that the sum of the expected multiplicities gives twice the number of sampled children.

Now consider the slightly more general case where monogamy is no longer assumed, as in the following Figure 3.3. Here, $\nu(\{u_{1,1}\}) = 3/10 \cdot 2 + 6/10 \cdot 1 = 1.2$ but $\nu(\{u_{1,2}\}) = 1/10 \cdot 3 + 6/10 \cdot 2 + 3/10 \cdot 1 = 1.8$ and $\nu(\{u_{1,3}\}) = 3/10 \cdot 2 + 6/10 \cdot 1 = 1.2$ while $\nu(\{u_{1,j}\}) = 0.6$ for the rest of the parents. Once again, $\sum \nu(\{u\}) = 6$.

The general pattern follows from observing how the multiplicities of a given parent $u$ might arise: She will be included exactly $k$ times in the sample of parents ($k \leq |c(u)|$) if exactly $k$ of her children are included in the sample of children. There are $\binom{|c(u)|}{k}$ ways of selecting $k$ of her children. Setting these aside in sampling the children, one has only to consider the number of samples of size $n - k$ from all the $N_0 - |c(u)|$
children that are not hers. It follows that
\[
Pr(N^*(., \{u\}) = k) = \binom{|c(u)|}{k} \binom{N_0 - |c(u)|}{n - k} / \binom{N_0}{n}
\]
This is just the hypergeometric distribution. It follows that
\[
v(\{u\}) = \frac{n|c(u)|}{N_0} \quad \text{and} \quad \mathbb{E}(\hat{M}(., T)) = \frac{n}{N_0} \sum_{u \in \mathcal{U}^p} T(u)|c(u)|
\]
This suggests to use the estimator
\[
\hat{M}^*(\omega, T) = \frac{N_0}{n} \sum_{u \in \mathcal{U}^p} T(u)N^*(\omega, \{u\}) / |c(u)| \quad (3.2)
\]
which will be design unbiased for the population total \(\sum T(u)\). This is called the Hansen-Hurwitz estimator of a total (e.g. Tillé 2006: Chap. 2). It is not the better known Horvitz-Thompson estimator which would use weights equal to the inclusion probabilities \(\pi(.)\), which in this case would be \(Pr(N^*(., \{u\}) > 0)\). Since the Horvitz-Thompson estimator correspondingly uses only the set-sample \(s^*\), it can in general not be computed from the sample information available in general surveys: It would require either access to a complete list of labels of the parent generation (and an identification of the label from the respondents) or the possibility to identify siblings among the respondents.

In contrast, despite its apparent dependence on the frame \(\mathcal{U}^p\) and on \(N^*(., .)\), neither of which will be known in general, (3.2) can be computed from the sample information provided by the children alone. It therefore is a valid statistic for most sampling situations. The only prerequisite is that the number of siblings of the children (the respondents) is known. To see this, note that
\[
\hat{M}^*(\omega, T) = \frac{N_0}{n} \sum_{u \in \mathcal{U}^c} \mathbb{I}[S(\omega)](u) \left( \frac{T(m(u))}{c_m(u)} + \frac{T(f(u))}{c_f(u)} \right)
\]
where now $c_m(u)$ and $c_f(u)$ refer to the number of siblings of $u$ (including herself) on her mothers resp. her fathers side. The distinction between maternal and paternal siblings is of course necessary unless strict monogamy is assumed. To illustrate this point, reconsider the last example where the sample of children is $s = \{u_{0,1}, u_{0,2}, u_{0,3}\}$. Then the sample of parents is $s^{**} = \{u_{1,1}^2, u_{1,2}^3, u_{1,3}^1\}$. Thus $\hat{M}^*$ expressed in terms of the parents becomes

$$
\hat{M}^* = \frac{5}{3} \left( \frac{T(u_{-1,1}) \cdot 2}{2} + \frac{T(u_{-1,2}) \cdot 3}{3} + \frac{T(u_{-1,3}) \cdot 1}{2} \right)
$$

Now in terms of the children’s sample, the first term arises from both $u_{0,1}$ and $u_{0,2}$ reporting the age of their common mother, so it is reported twice. Furthermore, there are two siblings of this mother. The second term arises similarly. Finally, there is only one report on the age of $u_{-1,3}$ (from $u_{0,3}$), there is, however, a half-brother of $u_{0,3}$ (through his mother) that must be taken into account and so the denominator is also 2. Note that only the values of $c_m(.)$ and $c_f(.)$ for the respondents are needed in the calculation. On the other hand, neither the labels of the parent generation (the elements of $U^p$) nor the multiplicities of their inclusion in the sample (the values of $N^*(., \{u\})$) are needed for the calculation of $\hat{M}^*$.

To get an estimate of the mean life length of the parents, the size of the parent generation $N := |U^p|$ is needed. Now clearly

$$
\sum_{u \in U^c} \frac{1}{c_f(u)} = |U^m| \quad \text{and} \quad \sum_{u \in U^c} \frac{1}{c_m(u)} = |U_{-1,1}^f|
$$

so that a design unbiased estimator of $|U^p|$ is

$$
\sum_{u \in U^c} \mathbb{I}[S(\omega)](u) \frac{1}{\pi(u)} \left( \frac{1}{c_f(u)} + \frac{1}{c_m(u)} \right)
$$

Once again, for the property of design unbiasedness, it is not necessary to know the actual relatedness of the respondents.
The ratio estimator of the mean life time of the parent generation thus becomes

\[
\sum_{u \in S(\omega)} \left( \frac{T(m(u))}{c_m(u)} + \frac{T(f(u))}{c_f(u)} \right) \\
\sum_{u \in S(\omega)} \left( \frac{1}{c_f(u)} + \frac{1}{c_m(u)} \right)
\]

(3.3)

where the factor \( \pi(.) \) cancels since simple random sampling was assumed. This is no longer design unbiased, but will be approximately so, at least to first order and when both \( n \) and \( N_0 \) are large (Särndal et al. 1992: Chap. 5.6).

In summary, the classical sampling theory approach provides a design unbiased estimate of the total life time of the parent generation and an approximately design unbiased estimator of the mean life time even though nothing is assumed about the connection between survival times and number of children. But this result should not be overstated: First of all, design unbiasedness is not a panacea, and design unbiased estimators can behave rather erratic. Secondly, the classical sampling theory presupposes at least some knowledge of the underlying labels of the population. This is not in general available in the current problem so that many of the standard estimators (and/or evaluations of their performance) are unavailable. And lastly, up until now the estimators constructed from the sampling theory perspective presuppose an extremely simplified and restrictive model world to work at all.

3.4. Assumptions about Independence

One of the restrictions used at the outset in the constructions of the previous two Sections was that instead of the life lengths of the previous generation only the life length of the parent generation was of interest. But such inferences would only rarely be useful. To extend the results to all members of the previous generation, one may assume that mortality is in a reasonable sense ‘independent’ of whether or not someone of the
previous generation had children. Such an assumption would open the possibility to draw at least partial inferences about the mortality of the whole previous generation.\textsuperscript{10}

But how can the notion of ‘independence’ be made precise so that it can serve its purpose? What is meant when it is assumed that two variables are independent? There are basically two different approaches to an explication. A first one conceptually refers to the process that brings about the facts recorded by statistical variables. This might be what most statisticians and econometricians have in mind when they speak of a \textit{data generating process}. There is, however, some ambiguity in this expression because one needs to distinguish between processes which create data and processes which bring about the facts that, subsequently, can be recorded as data. I propose to use the expression ‘data generating process’ only in the former sense, for example, when referring to a survey, and speak of “social processes” when referring to processes that one can sensibly think of as creating facts.

A second approach avoids speculations about social processes and concentrates instead on the recorded facts. The approach follows closely the survey sampling tradition where one considers a two-dimensional variable

\[ (X, Y): \mathcal{U} \rightarrow \mathcal{X} \times \mathcal{Y} \quad (3.4) \]

In this context, the presumption is that \((X, Y)\) records the facts prevailing in the population \(\mathcal{U}\) according to some categorisation expressed by \(\mathcal{X} \times \mathcal{Y}\). This is a particular way to record the facts, a way which I have called a \textit{statistical variable}. The approach avoids to say anything about how facts are brought about. It only stipulates that the facts are reasonably recorded by using \(\mathcal{X} \times \mathcal{Y}\) and that the facts can be ascertained for any person in the population \(\mathcal{U}\).

The first approach has to add further structure to (3.4) in order to express ‘independence’. It must express the idea that values of \(X\) and \(Y\) result from some processes and then assume that these processes develop

\textsuperscript{10} The inferences will still be partial since it will not be possible to estimate survival probabilities for children and young adults below the minimum child bearing age.
3. A Case Study: Parent’s Length of Life

independently. In other words, a model for the social processes that create facts must be proposed. In addition to the assumption that the facts are reasonably described by $X \times Y$, one needs a particular way to model the emergence of these facts. If a probabilistic model—as in the “simple device”—is adopted, then the emergence of values can be described by a probability space and a reasonable explication of independence becomes stochastic independence between random variables defined on that probability space.

The mathematical definition of stochastic independence requires (after the introduction of a suitable probability space) that all joint probabilities are equal to the product of the respective marginal probabilities. This might be done by writing

$$(X, Y): \mathcal{U} \times \Omega \rightarrow \mathcal{X} \times \mathcal{Y} \quad (3.5)$$

where $\Omega$ is equipped with a set of subsets of $\Omega$, say $\mathcal{B}$, for which a probability $\Pr(.)$ is defined. Then a probability distribution for $(X, Y)$ can be introduced by setting

$$\Pr((X, Y)(u, .) \in A) := \Pr(\{\omega \in \Omega \mid (X, Y)(u, \omega) \in A\})$$

provided that $\{\omega \in \Omega \mid (X, Y)(u, \omega) \in A\}$ is an element of $\mathcal{B}$.

For each member $u$ of the population, stochastic independence can now be defined by

$$\forall A \subseteq \mathcal{X} \forall B \subseteq \mathcal{Y}: \Pr(X(u, .) \in A \mid Y(u, .) \in B) = \Pr(X(u, .) \in A) \quad (3.6)$$

where both $X(u, .)^{-1}(A)$ and $Y(u, .)^{-1}(B)$ are elements of $\mathcal{B}$ so that probabilities for these ‘events’ can be deduced. Note that stochastic independence can be defined for each member of the population $\mathcal{U}$. And it is an extra assumption that independence holds for all $u \in \mathcal{U}$. It must be stressed that independence of $(X, Y)(u, .)$ from $(X, Y)(u', .)$, while defined in parallel to the definition (3.6) in probabilistic approaches, is used for purposes quite different from the definition of the independence of $X$ from $Y$. The former is a constitutive decision about the model.
structure and must be distinguished from assumptions like independence of \( X(u, .) \) from \( Y(u, .) \) for all \( u \in \mathcal{U} \).\(^{11}\)

Mathematicians and statisticians often call elements of \( \mathcal{B} \) ‘events’ even though, of course, nothing has to happen for such ‘events’ to be defined. Rather, the notion of ‘events’ described by a system of subsets \( \mathcal{B} \) is a static notion quite far removed from dynamic concepts. While such parlance can only present stylised facts and not their bringing about, the mathematical definition of independence can be turned into a procedural explication by referring to random generators that work physically independent from each other. For example, one might create values for \( X \) with a die, and values for \( Y \) with another die, and assume that both processes work independently from each other. The procedural explication of stochastic independence refers to causally unrelated methods of producing real events that can be described by a probability model.

This approach presupposes a probabilistic model for the social processes that create facts. While the introduction of a probability space might seem to be an extraneous and artificial addition to the description of the emergence of social facts, the obvious advantage of the approach is that it can rely on the calculus of probability, including the many results known about independent variables.

On the other hand, the approach may be judged to be far too restrictive, presupposing that probability models accurately describe the constitution as well as the changes of statistical variables recording social facts. An alternative approach should therefore refer only to the eventually realised data, i.e. the data generating process in the restricted sense. Turning back to the formulation (3.4), a description of social facts, one may try to define independence not with respect to some probability space but with respect to the relative frequencies that describe social facts. To this

\(^{11}\) While the condition (3.6) seems innocuous and is often treated in textbooks as a basic definition that does not deserve much comment, the mathematical strength of the condition cannot be overemphasised. E.g., Holbrook (1981) shows that if \((X(\omega), Y(\omega))\) are independent uniform random variables defined on \( \Omega = [0, 1] \) with Lebesgue measure, then \((X, Y): \Omega \rightarrow [0, 1] \times [0, 1] \) is a space-filling curve. Further examples from many areas of mathematics are discussed by Kac (1959).
end, write \( P(X \in A) \) for the relative frequency, i.e.

\[
P: \mathcal{P}(\mathcal{X}) \longrightarrow [0, 1]
\]

\[
P(X \in A) := \frac{|\{u \in \mathcal{U} | X(u) \in A\}|}{|\mathcal{U}|} \quad \forall A \in \mathcal{P}(\mathcal{X}) \quad (3.7)
\]

This is, of course, a probability measure, albeit a special one. There is no additional probability space \( \Omega \) needed for its construction. Furthermore, it is defined not for each member of \( \mathcal{U} \) but only for the population \( \mathcal{U} \) as a whole. It simply counts subsets of the population \( \mathcal{U} \) and relies only on the statistical variables introduced in (3.4).

The intuition behind this approach, rather different from the probabilistic version, is that \( X \) is ‘independent’ from \( Y \) if the distribution of \( X \) does not depend on subsets of \( \mathcal{U} \) that can be selected by considering values of \( Y \) alone. The advantage of such a definition is that an explicit reference to the processes creating the values of \( X \) and \( Y \) would be superfluous.

One may then try to copy the definition of stochastic independence from the probabilistic version (3.6) by requiring

\[
\forall A \subseteq \mathcal{X} \forall B \subseteq \mathcal{Y}: \ P(X \in A | Y \in B) = P(X \in A) \quad (3.8)
\]

There is, however, a problem resulting from the fact that the equality formulated in (3.8) can not, in general, be satisfied. A small example will show this. Assume that \( \mathcal{X} = \mathcal{Y} = \{0, 1\}, \mathcal{U} \) has 10 members, and the marginal distributions are as follows:

\[
\begin{array}{c|cc}
Y = 0 & X = 0 & X = 1 \\
\hline
Y = 0 & n_{11} & n_{12} & 4 \\
Y = 1 & n_{21} & n_{22} & 6 \\
\hline
& 3 & 7 & 10
\end{array}
\quad (3.9)
\]

I now want to make the assumption that \( X \) and \( Y \) are independent. Following the definition given in (3.8), one would need to find, for example, values of \( n_{11} \) and \( n_{12} \) such that

\[
n_{11} + n_{12} = 4, \quad \frac{n_{11}}{4} = \frac{3}{10}, \quad \text{and} \quad \frac{n_{12}}{4} = \frac{7}{10}
\]
But no integral values of $n_{\text{11}}$ and $n_{\text{12}}$ can be found that satisfy these requirements. The simple consequence is that (3.8) can not, in general, be used to postulate assumptions about independence in finite populations.

In fact, there are four tables with the given marginal frequencies:

\[
\begin{align*}
\text{a) } & 0 & 4 & b) & 1 & 3 & c) & 2 & 2 & d) & 3 & 1 \\
\text{3} & 3 & 2 & 4 & 1 & 5 & 0 & 6
\end{align*}
\]  

(3.10)

The corresponding relative frequencies of these tables can be embedded in the simplex \( \{(p_1, p_2, p_3, p_4) \mid p_i \geq 0, \sum p_i = 1\} \). In Figure 3.4, the four points are shown together with the surface of all points that satisfy the independence formulation 3.8. It is plain that it will be the exception rather than the rule to find a point on the surface of independence for any given set of marginal frequencies.

![Figure 3.4: The simplex of probabilities in 2 × 2 tables, the surface of independence, and the 4 tables with margins given by Example 3.9.](image)

If one nevertheless wants to follow this second approach one needs a weaker formulation that allows for approximate equalities between conditional distributions. I will use the following definition:
3. A Case Study: Parent’s Length of Life

X is $\Delta$-independent from Y w.r.t. the partition $\mathcal{Y}_p$ of $\mathcal{Y}$ if

$$\max\{|P(X \in A \mid Y \in B) - P(X \in A)| \mid A \subseteq \mathcal{X}, B \in \mathcal{Y}_p\} \leq \Delta$$

(3.11)

The idea behind this definition is that conditional distributions of X should be approximately equal among the subsets of $\mathcal{U}$ induced by a given partition of $\mathcal{Y}$. The partition $\mathcal{Y}_p$ (instead of the singletons of $\mathcal{Y}$) figures in the definition since it will often only be necessary to distinguish less detail than could be achieved with the full information from Y. E.g. in the problem treated here, $Y(u)$ would correspond to the names (and numbers) of children of u. But what is needed, at least in this Section, is only the information whether $u$ had children or not ($|c(u)| > 0$ or $|c(u)| = 0$).

The earlier formulation in (3.8) can then be seen as a limiting case where $\Delta = 0$ and $\mathcal{Y}$ is partitioned into its one-element subsets. But, as the example has shown, these strong requirements are rarely fulfilled. On the other hand, given any partition of $\mathcal{Y}$, one can always find a minimal $\Delta$ such that $X$ is $\Delta$-independent from Y. To illustrate with the example, one finds for the table b) in (3.10):

$$|P(X = 0 \mid Y = 0) - P(X = 0)| = |1/4 - 3/10| = 0.050$$

$$|P(X = 0 \mid Y = 1) - P(X = 0)| = |2/6 - 3/10| = 0.033$$

$$|P(X = 1 \mid Y = 0) - P(X = 1)| = |3/4 - 7/10| = 0.050$$

$$|P(X = 1 \mid Y = 1) - P(X = 1)| = |4/6 - 7/10| = 0.033$$

The minimal $\Delta$ is therefore 0.05 w.r.t. the partition of $\mathcal{Y}$ into the sets $\{0\}$ and $\{1\}$.

Note the asymmetry of the definition: While X is 0.05-independent from Y, the minimum of $\Delta$ for Y given X follows from

$$|P(Y = 0 \mid X = 0) - P(Y = 0)| = |1/3 - 4/10| = 0.0667$$

$$|P(Y = 0 \mid X = 1) - P(Y = 0)| = |3/7 - 4/10| = 0.0286$$

$$|P(Y = 1 \mid X = 0) - P(Y = 1)| = |2/3 - 6/10| = 0.0667$$

142
\[ |P(Y = 1 \mid X = 1) - P(Y = 1)| = |4/7 - 6/10| = 0.0286 \]

Thus, \( Y \) is only 0.0667-independent from \( X \). The asymmetry may seem unfortunate. But the formulation accurately reflects what is needed in the application: There is information on the relative frequencies of life times of members of the previous generation given that they have children, i.e. \( P(T = . \mid \{c\} > 0) \) is (approximately) known, and this should provide a clue for the frequencies of life lengths of all members of the previous generation, \( P(T = .) \). When it can be stipulated that \( T \) is \( \Delta \)-independent of \( \{c\} > 0 \), then the frequencies are close by definition, \( P(T = . \mid \{c\} > 0) \approx P(T = .) \). If further \( T \) is finite, then the closeness of relative frequencies implies closeness of conditional means to marginal means and the same holds true for many other functions of the frequencies.

A few immediate implications of the definition are:

a) If \( X \) is \( \Delta \)-independent from \( Y \) w.r.t. a partition \( \mathcal{Y}_p \), then this is also true for all \( \Delta' \geq \Delta \).

b) For a fixed \( B \in \mathcal{Y}_p \),

\[
\max_{A \subseteq \mathcal{X}} \{|P(X \in A \mid Y \in B) - P(X \in A)|\} = \|P(X \in \cdot \mid Y \in B) - P(X \in \cdot)\|
\]

is often called the **total variation distance** between \( P(X \in \cdot \mid Y \in B) \) and \( P(X \in \cdot) \). This distance can be expressed in terms of the densities \( P(X = x) \) and \( P(X = x \mid Y \in B) \):

\[
\max_{A \subseteq \mathcal{X}} \{|P(X \in A \mid Y \in B) - P(X \in A)|\} = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(X = x \mid Y \in B) - P(X = x)|
\]

This shows that the computation of \( \Delta \) for a given \( B \) can be carried out running only through the singletons of \( \mathcal{X} \). There is no need to consider all subsets of \( \mathcal{X} \). Only \( |\mathcal{X}| \cdot |\mathcal{Y}_p| \) absolute differences must be computed.\(^{12}\)

\(^{12}\) Strasser (1985: 5–7) proved this and similar relations in a rather general setting.
c) If $X$ is $\Delta$-independent from $Y$ w.r.t. a partition $\mathcal{Y}_p$, and $B_1$ and $B_2$ are both elements of $\mathcal{Y}_p$, then

$$\max\{|P(X \in A \mid Y \in B_1 \cup B_2) - P(X \in A)| \mid A \subseteq \mathcal{X}\} \leq \Delta$$

This follows since if $B_1 \neq B_2$, then $B_1 \cap B_2 = \emptyset$ ($\mathcal{Y}_p$ is a partition) so that

$$P(X \in A \mid Y \in B_1 \cup B_2) = \frac{P(Y \in B_1)}{P(Y \in B_1) + P(Y \in B_2)} P(X \in A \mid Y \in B_1) + \frac{P(Y \in B_2)}{P(Y \in B_1) + P(Y \in B_2)} P(X \in A \mid Y \in B_2)$$

As an implication of (b) one finds: If $X$ is $\Delta$-independent from $Y$ w.r.t. a partition $\mathcal{Y}_p$, then $X$ is also $\Delta$-independent from $Y$ w.r.t. any coarser partition of $\mathcal{Y}$.\(^{13}\) In particular, if $X$ is $\Delta$-independent from $Y$ w.r.t. to the partition of $\mathcal{Y}$ into one-element subsets, it is $\Delta$-independent from $Y$ w.r.t. to any partition of $\mathcal{Y}$.

I therefore simply say that $X$ is $\Delta$-independent from $Y$ if this is true w.r.t. the partition of $\mathcal{Y}$ into one-element subsets.

As shown by the example of the $2 \times 2$ table, one can not expect a small $\Delta$ even when the data are as close as possible to stochastic independent. It is natural to ask for the smallest possible $\Delta$ achievable in a finite population of size $n$ and a given set of conditional frequencies $P(X = x \mid Y = y)$. I.e. given the distribution of the life lengths of the parent generation and the total size $n$ of the previous generation, what is the smallest achievable value of $\Delta$? To facilitate the discussion, suppose that $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ so that we deal with a $2 \times 2$ table. Suppose further that both $P(X = 0 \mid Y = 1)$ and $P(Y = 1)$ are fixed at some rational numbers and that $n$ is given. Now choose natural numbers closest to $nP(Y = 1)$ and $nP(X = 0, Y = 1)$ and choose $nP(X = 0, Y = 0)$ in such a way that $\Delta$ is minimised. It might be

\(^{13}\) A partition $\mathcal{Y}'_p$ is coarser than a partition $\mathcal{Y}_p$, denoted by $\mathcal{Y}'_p \subseteq_p \mathcal{Y}_p$, if for each $B \in \mathcal{Y}_p$ there is an element $B' \in \mathcal{Y}'_p$ such that $B \subseteq B'$. As a consequence, one can extend $\Delta$-independence w.r.t. $\mathcal{Y}_p$ to $\Delta$-independence w.r.t. the $\sigma$-algebra generated by the partition $\mathcal{Y}_p$.\[\]
conjectured that $\Delta$ converges rapidly with $n$ to 0. But this is not true, the convergence is far from uniform, and it is not as fast as is often suggested by statistical reasoning. Figure 3.5 shows the minimal $\Delta$ for $n$ from 50 to 300 together with an indication of the order of the estimated standard error. Also indicated is a case with $|\mathcal{X}| = 4$, where the behaviour is even less regular.

A further example may help to illustrate the conceptual difference between $\Delta$-independence and stochastic independence. Values of the variable $(X, Y)$ are created with a pair of dice such that the dice operate independently. Throwing the pair of dice 60 times, again independently, the following values might arise:

<table>
<thead>
<tr>
<th></th>
<th>$Y = 1$</th>
<th>$Y = 2$</th>
<th>$Y = 3$</th>
<th>$Y = 4$</th>
<th>$Y = 5$</th>
<th>$Y = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 1$</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$X = 2$</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>$X = 3$</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$X = 4$</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Now, from a procedural point of view, the processes creating the values of \( X \) and \( Y \) are clearly independent, and this provides one possible explication of the independence of the two variables: that their values are generated by independent processes. Note that this explication completely ignores the actually realised values. On the other hand, one might as well ask for the degree of independence exhibited by the realised data. Given a partition of \( Y \), I can ask to which degree \( X \) is \( \Delta \)-independent from \( Y \) w.r.t. this partition. Using the partition \( Y_p := \{\{1\}, \ldots, \{6\}\} \), the task is to find the smallest \( \Delta \) such that \( X \) is \( \Delta \)-independent from \( Y \) w.r.t. this partition. In a first step, one may consider the restricted maximum

\[
\max_{x \in X, y \in Y} \{|P(X = x \mid Y = y) - P(X = x)|\}
\]

that only takes into account one-element subsets of \( Y \). This is easy to calculate and yields 0.633 (arising from the entry \( X = 4, Y = 3 \)). Such a large value is hardly useful and the conditional density of \( X \) changes heavily with the value of \( y \) chosen.

To compute the final value of \( \Delta \), one has to compute the maximum with respect to all subsets of \( X \). By property b) above, it is enough to compute the sum of the absolute differences between the conditional and marginal densities for all fixed \( y \). This gives 0.115, 0.333, 0.650, 0.226, 0.316, 0.183. Thus, the smallest value of \( \Delta \) achievable is 0.65. The conditional distribution behaves only slightly worse than the conditional density.

In summary, \( \Delta \)-independence of one variable from another is conceptually different from stochastic independence. Statistical variables with small \( \Delta \)-independence obviously need not be created from stochastically independent random variables. In fact, \( \Delta \)-independence is defined without reference to an additionally introduced probability space \( \Omega \). It only refers to the recorded data.
Nor does stochastic independence imply $\Delta$-independence for a reasonably small $\Delta$. This is clear from the dice example. It is only within the model world, and only when further assumption are made, that $\Delta$-independence nearly follows from stochastic independence, using a form of the strong law of large numbers. If data are generated by an appropriate random device, if draws from the device are stochastically independent, if both $\mathcal{X}$ and $\mathcal{Y}_p$ are finite, if the number of draws is potentially infinite, and, of course, if $X \perp \perp Y$, then the event that

$$\max_{y_p \in \mathcal{Y}_p} \| P_n(X \in \cdot \mid Y \in y_p) - P_n(X \in \cdot) \| > \Delta$$

will have probability 0 for all $\Delta > 0$ and $n$ large enough. Here, $\| \cdot \|$ is the total variation distance and $P_n(.)$ is the empirical distribution of the first $n$ elements of a sequence $(X, Y)_1, (X, Y)_2, \ldots, (X, Y)_n, \ldots$ of realisations of $(X, Y)$. But this result has no operational meaning: The conclusion can only be formulated within the probability model.

On the other hand, the concept of $\Delta$-independence can be used easily to formulate assumptions about the distribution of life lengths of the previous generation based on knowledge of the distribution of the life lengths of the parent generation. If it can be argued that the distribution of life lengths $T$ is $\Delta$-independent from $\{|c(.)| > 0\}$ for a small enough $\Delta$, then knowledge of the life lengths of the parent generation implies knowledge of the distribution of life lengths of the previous generation up to an error of at most $\Delta$.

A claim of $\Delta$-independence is a claim pertaining to the facts recorded by the variables under consideration. Therefore, arguments in favour of the claim—and its criticism—must rest on references to other pertinent facts, e.g. information on life lengths of the generation from other sources. Arguments referring to genetic inheritance, living conditions, and historical circumstances are irrelevant in the discussion of such claims. They can not even be incorporated into the formulation of $\Delta$-independence.

---

14 It should be clear, however, that the arguments are constrained by the framework introduced in Section 3.1.
3. A Case Study: Parent’s Length of Life

3.5. The “Simple Device”

In contrast to the approach based on statistical variables, the “simple device”—when taken as the general advice to express everything in terms of a probability model—would require to account for the emergence of all the facts including the development of the ancestral graph. This is sometimes attempted in the biological sciences (e.g. Whittemore/Halpern 1994), but it would be certainly too much to require such a detailed model only to answer a simple question on the distribution of life lengths.

In fact, the “simple device” can be made to work within a much more limited scope. Considering only the previous generation (ignoring the sampling from the children’s generation), one may construct a family of random variables

\[(T, C) : U \times \Omega \longrightarrow T \times \{0, 1, \ldots, K\}\]

where \((T, C)(u, \cdot)\) represents life length and number of children of the element \(u \in U\), respectively. The probability space \((\Omega, \mathcal{B}, \Pr)\) and the random variable \((T, C)\) is such that there is at least one \(\omega\) such that \((T, C)(u, \omega)\) equals the life length and number of children for all \(u \in U\). Suppose further that the random variables constructed for different members \(u\) of the population are stochastically independent and identically distributed (a constitutive model assumption). Following the notation developed in Chapter 2, the coarsened variables (representing information on the parents only) can be written as

\[(T, C)^* : U \times \Omega \longrightarrow \{(0, 1), \{(1, 1), \ldots, \{(\tau, K)\}\} \cup \{T \times \{0\}\} \subset \mathcal{P}(T \times \{0, \ldots, K\})\]

i.e. \((T, C)(u, \cdot)\) is observed exactly if and only if \(C(u, \cdot) > 0\) \((u\) has at least one child). Otherwise, any value of \(T\) might have happened. The implied probabilities for \((T, C)\) can conveniently be arranged in a table

<table>
<thead>
<tr>
<th>(\Pr(T = 0, C &gt; 0))</th>
<th>(\ldots)</th>
<th>(\Pr(T = \tau, C &gt; 0))</th>
<th>(\Pr(C &gt; 0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(*)</td>
<td>(\ldots)</td>
<td>(*)</td>
<td>(\Pr(C = 0))</td>
</tr>
</tbody>
</table>

148
where the * indicate unknown probabilities and the values of $C(u, .)$ are coarsened to the set \{0, 1\} with 1 indicating $C(u, .) > 0$.

As introduced up to now, the “simple device”, by expressing the situation in terms of random variables, forces the assumption that the probability of having children is known. This follows since random variables are functions on the known domain $U \times \Omega$ so that both the number of elements of $U$ and the probability of $\{C(u, .) > 0\}$ is supposed to be known. Even the values of $C(u, \omega)$ must be supposed known for all $u \in U$ and the realised $\omega$. The assumption contradicts the original description of the problem where the size of the previous generation is unknown. Of course, the model may be amended by restricting attention to the subset of $U$ such that $C(u, .) > 0$. In probabilistic terms, one needs to consider the conditional distribution of $(T, C)(u, .)$ given the event $\{C(u, .) > 0\}$. The existence of (and reference to) elements of childless members of the previous generation then becomes part of the probability model itself, further separating the “simple device” from its sampling theory counterpart. However, I will discuss the conditional version of the model only in the next Section.

In order to facilitate the comparison with results from Chapter 2 and to ease the notational burden, I will for the rest of the Section drop the reference to individuals $u$ so that $(T, C)$ and $(T, C)^*$ are just pairs of random variables defined on a common probability space. Within the present model, the set of underlying distributions of $(T, C)$ compatible with a given distribution of $(T, C)^*$ is easily described: it is the simplex $\{\operatorname{Pr}(T = t, C = 0) \mid t \in \{0, \ldots, \tau\}\}$. It is also easy to see that there is a unique CAR solution: Remember from 2.21 that the joined distribution of $(Y, Y^*)$ satisfies CAR if

$$
\operatorname{Pr}(Y = y \mid Y^* = y^*) = \operatorname{Pr}(Y = y \mid Y \in y^*)
$$

In the present situation, $C(\omega) = 0$ (no children) is equivalent to $T^*(\omega) = T$, so that $\{C = 0\} = \{\omega \in \Omega \mid C(\omega) = 0\} = \{\omega \mid T^*(\omega) = T\}$. Now

$$
\operatorname{Pr}(T = t, C = 0 \mid (T, C)^* = T \times \{0\}) = \operatorname{Pr}(T = t \mid T^* = T, C = 0) = \operatorname{Pr}(T = t \mid C = 0)
$$
and

$$\Pr(T = t \mid C = 0) = \Pr(T = t) = \Pr(T = t \mid T \in T)$$

if and only if $T$ and $C$ are stochastically independent ($T \perp \perp C$). It follows that $((T, C), (T, C)^*)$ is CAR if and only if $T \perp \perp C$. Since the conditional distribution of $T$ given $\{C > 0\}$ is known, the CAR condition uniquely identifies the joined distribution of $(T, C)$ from the known (estimated, approximated, assumed, …) distribution of $(T, C)^*$.\(^{15}\)

Within the present rather simplistic probabilistic model, stochastic independence or, equivalently, the CAR condition takes the role of $\Delta$-independence in the approach based on relative frequencies. But what exactly is the role of the probability model? The model as presented does not pretend to refer to the process by which life lengths are determined. It does not even provide the means to express pedigree, or inheritance, or when people do have children. And certainly life lengths are not created by the roll of a die. An argument in favour of such a model must necessarily rely on an “as if” reasoning: that the distribution of life lengths may be treated for the current purpose “as if” generated by a random device. As argued in section 2.2.1, the adoption of a particular probabilistic model is best seen as decision taken by the researcher. Arguments in support of the decision for using a probabilistic model can not simply refer to reality or data. But what kind of arguments can provide justifications for the decision?

Jerzy Neyman (1960: 629) explicitly argued for the use of stochastic models acknowledging their “as if” character:

The essence of dynamic indeterminism in science consists in an effort to invent a hypothetical chance mechanism, called a “stochastic model”, operating on various clearly defined hypothetical entities, such that the resulting frequencies of the various possible outcomes correspond approximately to those actually observed.

\(^{15}\)But note that this joined distribution can not, in general, coincide with any frequency distribution realised using the same model (cf. Example 3.9).
He also emphasises that the term ‘indeterminism’ relates “not to the phenomena themselves but to our approach to these phenomena” (Neyman 1960: 626).

He then goes on to formulate criteria for the ‘scientific value’ of such models: ‘broad applicability’ and ‘identifiability’. He introduces these criteria only by example. But even a cursory reading shows that the model introduced here is neither broadly applicable nor identifiable in his sense. In particular, the “simple device”, expressing incomplete observations in a particular way, is the only aspect of ‘broad applicability’ of the model. However, this is a methodological recipe, a mere template. As such it has no ‘scientific value’. It must be amended within a given context. But then its merits must be judged separately in each application. This can be done in the present case. But it can only be done within a much more elaborate model where it is possible to argue about the details of the model based both on consequences of the model and the known facts. It is only in relatively rich models that the strengths of the “simple device”, the possibility to discuss in a principled way consequences of different assumptions of missingness, can be achieved. The “simple device” per se will not necessarily provide such a framework.

To be more precise, the set of distributions of \((T, C)\) compatible with \((T, C)^*\) is the simplex

\[
\{ \Pr((T, C) = (t, 0) \mid t \in \mathcal{T}) \mid \sum_{t=0}^{\tau} \Pr(T = t, C = 0) = \Pr((T, C)^* = \mathcal{T} \times \{0\}) \}
\]

In the absence of further model details, the CAR solution is but a particular point in this set. There is no way to argue persuasively for a subset of compatible distributions, let alone for one particular solution, except, of course, that the CAR condition would justify a particularly simple statistical treatment. But simplicity of statistical treatment can not be an argument in judging facts.

Furthermore, if the probability \(\Pr((T, C)^* = \mathcal{T} \times \{0\})\) is set equal to the observed frequency \(P(|c(.)| = 0)\), then the set of compatible distributions
of the “simple device” coincides with the set-valued sample based solution favoured by Manski (2003) and others. As argued in the previous Chapter, the equality is reassuring from a conceptual point of view: The probability model does not introduce additional aspects that prejudice the solution of the incomplete data problem. But many who think of the “simple device” as a mere statistical “method” for solving their problem will be dissatisfied with this equality since it generally generates rather large intervals of possible values.¹⁶

But the problem turns out to be worse. As witnessed by the dice example of the previous Section, the connection between an assumed probability model and (realised) values is rather weak, and in general, the connection is even weaker when the model is confronted with observations: Even if observations are generated from a given probability model with probability measure \( \Pr(.) \) and assuming that \( n \) observations are independently generated from this measure, then there is no function of the observations (i.e. a statistic) such that the maximum deviation of empirical frequencies (or kernel estimators etc.) from their theoretical counterparts converges to 0 in the variation distance with the number of observations.¹⁷ The result applies in particular to the “simple device”. No amount of observations will allow to narrow the gap between data and model. Neither the rather broad set of compatible distributions nor the CAR condition can be justified by reference to the observations alone. Therefore, arguments for adopting the CAR condition must be sought based on the probability model itself. This can be done, but it requires an elaborated model that goes beyond the mere introduction of some random variables. Note the contrast to the sampling perspective: there, the set of compatible frequencies can be directly compared to frequencies derived either from further data or from other sources. Accordingly, both, arguments for the assumption of \( \Delta \)-independence and its criticism

¹⁶ E.g., in the discussion of Manski and Horowitz (2000: 86), Raghunathan complained: “I am afraid that I agree with Cochran (1977) that such an approach is so conservative as to be of little value in most practical settings for inferential purposes.”

¹⁷ This might appear to be a somewhat trivial result since empirical distributions always have maximal variation distance from an absolutely continuous probability measure. The negative result is, however, not restricted to the divide between discrete and absolutely continuous distributions. See Section 2.2.1 for general comments. Devroye/Győrfi (1990) present further details on the general negative result.
can be assessed based on empirical information alone.

There is yet another problem with the “simple device” in the present case: As I argued in Section 3.3, sampling the children induces the possibility of including a parent several times without this being known to the surveyor. Even within a probabilistic model, one strategy to deal with the issue is to keep with the classical sampling approach. One would treat the sampling of the children (and the induced sampling of the parents) as a process independent of the underlying probability space of the “simple device”. This is what is regularly done in the model assisted approach to sampling.\(^{18}\) But since the argument in Section 3.3 includes a reference to the number of children, and since values of statistical variables are fixed numbers, one would be induced to treat that number as fixed as well in the probabilistic model of the “simple device”. This in turn would require a reformulation of the CAR condition, and of stochastic independence, where the number of children may now only appear on the right side of the conditioning bar.

One might, however, try to follow the opposite strategy and formulate the problem completely within a probabilistic framework. After all, not knowing the number of times a particular sample member is included in the sample might be seen as a further level of incompleteness. But this approach turns out to be difficult at best, without offering an appreciable advantage. In fact, in trying to do so one has to model the ancestral graph and its development. In consequence, one can not take the reference to individuals (the sets \(U_t\) and \(U_{t+1}\)) as fixed. Their coming about must become part of the model. One can then condition on the realised units of the children’s generation and look backwards in time to produce realisations of the complete previous generation (not only the parents). This would provide a model both for the multiplicities in a sample and the number (and life lengths) of childless people. But the childless people of the model are now just creations of the model without an identifiable counterpart in (partially) recorded data.\(^{19}\) It is at least difficult to see

\(^{18}\) See e.g. Särndal et al. (1992). For a more comprehensive discussion of a joint sampling and probabilistic approach see Rubin-Bleuer/Schiopu Kratina (2005) and Hansen et al. (1983).

\(^{19}\) McCullagh (2008) discusses problems arising from randomly labelled units in the context of binary random effects models. In his his article of 2005 he briefly refers to
3. A Case Study: Parent’s Length of Life

how even summary statistics about these phantoms could be related to known facts. Even if this problem is dismissed in favour of a purely model based interpretation, the problem of multiple inclusion into the sample still has no obvious solution. In particular, the analysis of the weighted solution (3.3) appears to be extremely difficult.

3.6. Left Truncated Data

In order to arrive at a more realistic model, one further aspect has to be incorporated into the description of the situation. Up to now, no restrictions between life lengths and number of children were considered. But children and young adults below the minimum child bearing age can not have children. Therefore, no deaths before the minimum child bearing age can be recorded in the parent sample and nothing can be said about the distribution of very short life lengths.\(^{20}\)

To discuss the issue, consider a slight extension of the previous model where the statistical variables \((T, C)\) are defined as

\[
(T, C): \mathcal{U} \rightarrow \mathcal{T} \times (\mathcal{T} \cup \{-1\})
\]  

(3.12)

As before, \(T(u)\) denotes \(u\)’s life length for each \(u \in \mathcal{U}\). In contrast to the previous Section, however, \(C(u)\) now records the age at which \(u\) became for the first time mother (or father) of a child. If \(u\) stays childless throughout his or her lifetime, \(C(u)\) is set to -1. To facilitate the discussion, I will subsequently only deal with women (and still denote the population by \(\mathcal{U}\), the population of mothers by \(\mathcal{U}^p\) etc.).\(^{21}\)

\(^{20}\) what he calls “block factors”, i.e. unlabelled factors in a regression that just indicate membership in an otherwise uncharacterised group. He shows that in both cases unlabelled units do create problems of reference and interpretation. The present case may be considered as even worse since the labels only exist within the model, while in McCullagh’s cases they do have a real counterpart.

\(^{21}\) I still assume that information from all children is available, ignoring their mortality up to the interview date.

\(^{21}\) The same reasoning, however, applies to men with obvious minor modifications.
The available data still refer to the subset of women that had at least one child, say

\[ \mathcal{U}^P := \{ u \in \mathcal{U} \mid C(u) \geq 0 \} \]

But with the introduction of the variable \( C(.) \) (the age at which a first child was born) it is now possible to discuss the impact of a lower truncation point of observed life times on the assessment of the distribution of life lengths.

Following the basic idea of the Kaplan-Meier procedure, one may use rates to assess the distribution of \( T \). Assuming complete observations it would be possible to define both a risk set

\[ \mathcal{U}_t := \{ u \in \mathcal{U} \mid T(u) \geq t \} \]

i.e. the set of all members of \( \mathcal{U} \) who are alive just before age \( t \), and an event set

\[ \{ u \in \mathcal{U}_t \mid T(u) = t \} \]

consisting of those members of \( \mathcal{U}_t \) who actually died at age \( t \).\(^{22}\) From these sets one can calculate rates

\[ r(t) := \frac{|\{ u \in \mathcal{U}_t \mid T(u) = t \}|}{|\mathcal{U}_t|} = P(T = t \mid T \geq t) \]

which can be used to find the survivor function

\[ P(T \geq t) =: G(t) = \prod_{j=0}^{t-1} (1 - r(j)) \]

But since the data only refer to \( \mathcal{U}^P \), neither the cardinality of the risk set nor that of the event set is known and consequently one cannot calculate the rates \( r(t) \). One can only try to estimate these rates, but this will require an assumption. The assumption will roughly be that mortality

\(^{22}\) I am once again slightly misusing notation by reusing the symbol \( \mathcal{U}_t \) to denote not the (previous) generation but a subset of \( \mathcal{U} \), the set at risk of dying.
does not depend on whether and when people became mothers and fathers. More precisely, the assumption is

\[ \tilde{r}^*(t) := P(T = t \mid T \geq t, 0 \leq C \leq t) = \frac{\left| \{u \in \mathcal{U}_t^P \mid T(u) = t\} \right|}{|\mathcal{U}_t^P|} \approx \frac{\left| \{u \in \mathcal{U}_t \mid T(u) = t\} \right|}{|\mathcal{U}_t|} = P(T = t \mid T \geq t) = r(t) \]

where the risk set on the left-hand side is now defined by

\[ \mathcal{U}_t^P := \{u \in \mathcal{U}^P \mid T(u) \geq t, 0 \leq C(u) \leq t\} \]

Looking at the expression of the rates in terms of conditional frequencies

\[ P(T = t \mid T \geq t, 0 \leq C \leq t) \approx P(T = t \mid T \geq t) \]

it becomes apparent that this is just a dynamic version of \( \Delta \)-independence introduced in Section 3.4. In other words, \( \{T = t\} \) is \( \Delta \)-independent from \( \{0 \leq C \leq t\} \) within the subset of women surviving at least to age \( t \).

Since both this risk set and the corresponding event set can be calculated from data restricted to \( \mathcal{U}^P \), it is possible to approximate the rates \( r(.) \) by \( \tilde{r}^*(.) \). Of course, this will be possible only for ages

\[ t \geq a^+ := \min\{C(u) > 0 \mid u \in \mathcal{U}^P\} \]

which implies that only the conditional survivor function \( G(. \mid T \geq a^+) \) can be estimated:

\[ G(t \mid T \geq a^+) \approx \prod_{j=a^+}^{t-1} (1 - \tilde{r}^*(j)) =: G^*(t) \quad (3.13) \]

Notice that in general

\[ \mathcal{U}_t^P = \{u \in \mathcal{U}^P \mid T(u) \geq t, 0 \leq C(u) \leq t\} \subset \{u \in \mathcal{U}^P \mid T(u) \geq t\} = \mathcal{U}^P \]

because a women in \( \mathcal{U}^P \) might have her first child later than \( t \). In order to create suitable risk sets \( \mathcal{U}_t^P \) one has to apply the same conditioning as
used for the event sets to meet the assumption that mortality does not depend on whether and when women become mothers.

The following example illustrates the reasoning. Assume that \( U \) refers to a set of 1000 women and consider, in turn, five age classes:

0 In the age class \( t = 0 \) all 1000 women are at risk of dying, and I assume that 100 women actually die.

1 There remain 900 women. 200 of them will bear a first child in the age class \( t = 1 \). Also, 100 of the 900 will die. Implied by the assumption that mortality does not depend on becoming a mother, approximately

\[
\frac{100}{900} \times 200 \approx 22
\]

of the mothers will die.

2 There remain 800 women in the age class \( t = 2 \). I assume that 200 of these women die and 300 women become mothers of a first child. The assumption of equal mortality implies that about \( 300 / 4 = 75 \) of these women also die. In addition, there are 178 women who became mothers in age class \( t = 1 \), and of these about \( 178 / 4 \approx 45 \) will die.

3 The remaining number of women is \( 800 - 200 = 600 \), and I assume that 200 of these women die in the age class \( t = 3 \). Again, 200 women bear children for the first time. Consequently, about \( 200 / 3 \approx 67 \) of these women will die. Of the 358 women who became mothers before \( t = 3 \), 119 will die.

4 Finally, the remaining 400 women are supposed to die in the oldest age class.

Since in this example the complete data are available, one may calculate
3. A Case Study: Parent’s Length of Life

the survivor function explicitly:

| $t$ | $|U_t|$ | $|\{u \in U_t \mid T(u) = t\}|$ | $r(t)$ | $G(t)$ |
|-----|--------|-------------------------------|--------|--------|
| 0   | 1000   | 100                           | 1/10   | 1.00   |
| 1   | 900    | 100                           | 1/9    | 0.90   |
| 2   | 800    | 200                           | 1/4    | 0.80   |
| 3   | 600    | 200                           | 1/3    | 0.60   |
| 4   | 400    | 400                           | 1      | 0.40   |

Obviously, the survivor function is simply proportional to the number of women in the risk set. In a next step, I assume that data are only available for $U^p$, that is, for women who gave birth to at least one child. In the example, there are altogether $200 + 300 + 200 = 700$ women. I now perform the same calculations for these women using the risk and event sets as defined above. This can be summarised in the following table:

| $t$ | $|U^p_t|$ | $|\{u \in U^p_t \mid T(u) = t\}|$ | $\tilde{r}^*(t)$ | $G^*(t)$ |
|-----|--------|-------------------------------|------------------|--------|
| 0   |        |                               | 0.110            |        |
| 1   | 200    | 22                            | 0.251            | 0.890  |
| 2   | 478    | 120                           | 0.333            | 0.667  |
| 3   | 558    | 186                           | 1.000            | 0.445  |
| 4   | 372    | 372                           |                  |        |

For $t = 0$, the risk set is empty and a death rate cannot be calculated. Consequently, the value of the survivor function for $t = 1$ is not estimable either. For $t > 0$ it is possible, however, to create risk and event sets and calculate corresponding rates $\tilde{r}^*(t)$. And these rates can finally be used to derive the values of the conditional survivor function

$$G(t \mid T \geq a^+) = \frac{G(t)}{G(a^+)} \approx G^*(t)$$

where $a^+ = 1$ and $G(a^+) = 0.9$ in the present example. Here, $G^*(.)$ refers to the conditional survivor function estimable from the risk sets $U^p_t$. Note in particular that the cardinality of risk sets may increase with increasing life times, a feature not encountered with right censored
data. In contradistinction to the latter case, the statistical precision of estimates may be rather low for the young ages. This fact may cause severe practical problems of identification and robustness. Some aspects of these difficulties are taken up in the next Section.

How does the “simple device” deal with the current problem? As a first step, again, a probability space is introduced and random variables in parallel with (3.12) are defined:

$$(T, C) : \mathcal{U} \times \Omega \longrightarrow \mathcal{T} \times (\mathcal{T} \cup \{-1\}) \quad (3.14)$$

If $T(u, .)$ was stochastically independent from $C(u, .)$ ($T(u, .) \perp \perp C(u, .)$), then the model restricted to the event $\{0 \leq C(u, .) \leq T(u, .)\}$ is called a random left truncation model.\(^{23}\)

Mandel (2007) provides a vivid discussion of the distinction between truncation and censoring. He states: “Whereas censoring is a model of missing observations ..., truncation is a model of selection bias.” (p. 322). Since my setup includes an explicit reference to members of the population and assumes that at least the cardinality of $\mathcal{U}$ is known, the truncation model has the additional consequence that the probability model must be extended to a model that includes at least the population size as a random variable.

For a single (possibly hypothetical) mother $u$ the joint distribution of $T(u, .)$ and $C(u, .)$ conditional on $\{a^+ \leq C(u, .) \leq T(u, .)\}$ ($u$ has at least one child) becomes

$$\Pr(T(u, .) = t, C(u, .) = s \mid a^+ \leq C(u, .) \leq T(u, .)) = \frac{\Pr(T(u, .) = t) \Pr(C(u, .) = s)}{\Pr(a^+ \leq C(u, .) \leq T(u, .))} \mathbb{1}[a^+ \leq s \leq t] \frac{\Pr(a^+ \leq C(u, .) \leq t)}{\Pr(a^+ \leq C(u, .) \leq T(u, .))}$$

\(^{23}\) More details on the left truncation model are presented in Chapter III.3 of Andersen et al. (1993). They also discuss conditions that are less restrictive than stochastic independence. Keiding/Gill (1990) re-parametrisate the left truncation model as a three state Markov model, enabling them to provide another relaxation of independence.
\[ \times \frac{\Pr(C(u, .) = s)}{\Pr(a^+ \leq C(u, .) \leq t)} \mathbb{1}[a^+ \leq s \leq t] \]

\[ = \frac{\Pr(T(u, .) = t) \Pr(C(u, .) \leq t \mid C(u, .) \geq a^+)}{\Pr(C(u, .) \leq T(u, .) \mid C(u, .) \geq a^+)} \]

\[ \times \Pr(C(u, .) = s \mid a^+ \leq C(u, .) \leq t) \mathbb{1}[a^+ \leq s \leq t] \]

where \( T(u, .) \perp \perp C(u, .) \) is used in the first equation. The first factor in the last line is the distribution of \( T(u, .) \) in the coarsened model, the second one the conditional distribution of \( C(u, .) \) given \( \{T(u, .) = t\} \cap \{a^+ \leq C(u, .) \leq T(u, .)\} \). This suggests a simple estimation strategy: The first factor can be estimated by the empirical distribution of the observed death times, i.e. from the distribution of \( T(u, .) \) in the coarsened data model. Since also the distribution function \( \Pr(C(u, .) \leq t \mid C(u, .) \geq a^+) \), the distribution of the age at first birth of all women who had at least one child, can be estimated, the empirical distribution of the observed death times \( T(u, .) \) weighted by the inverse of the latter distribution function should be a reasonable estimator of the marginal distribution of the death time \( T(u, .) \).

While the problem of reference to population members \( u \) with \( C(u, \omega) = -1 \) as discussed in the previous Section persists, there is an additional problem with the “simple device” as presented up to now: The “assumption” of stochastic independence, which by itself is “untestable”, is simply and obviously wrong, at least when the model is to be realistic. For then the random variables \((T(u, .), C(u, .))\) must satisfy both \( \Pr(C(u, .) \leq T(u, .)) = 1 \) (either children are born before the death of their mother or there are no children to that woman) and \( \Pr(C(u, .) = -1 \mid T(u, .) < a^+) = 1 \) (there are no children to women dying before the reproductive age).

It may help here to be more explicit about the distinction of “assumptions” within the “simple device” and assumptions about facts that are to be represented by the model. Suppressing again the reference to individuals, it is certainly possible to produce a joint distribution of a random vector \((T, C)\) such that \( T \perp \perp C \) and such that the requirements \( \Pr(C \leq T) = 1 \) and \( \Pr(C = -1 \mid T < a^+) \) hold. One possibility is to simply restrict the
range of \((T, C)\) such that the minimum of the range of \(T\) is larger than the maximum of the range of \(C\). In that case, \(T\) and \(C\) may be independent while still fulfilling the consistency requirements. That is, this form of “independence assumption” is not self-contradictory. It is, moreover, “untestable” for any given data set since by the very construction of the observation scheme neither \(\{T < a^+\}\) nor \(\{C = -1\}\) are ever observed. Nevertheless, this model of independence is plainly wrong. It is obvious that there are deaths before age \(a^+\) in human populations. And there are deaths during the reproductive ages. Thus, range restrictions on \((T, C)\) are counter to fact. But if the ranges \(C, T\) of \(T\) and \(C\) overlap and \(|C \cap T| \geq 2\), then independence \(T \perp \perp C\) forces \(\Pr(C > T) > 0\). But giving birth to a child after death is equally counter to fact. While in many cases model assumptions counter to fact do not rule out the usefulness of the models concerned, in the present case the very assumptions, even if taken only as an “approximation”, would undermine the very goal of estimating relevant aspects of the distribution of life lengths of a previous generation.

Thus the simple minded reference to stochastic independence will not do. One has to replace the condition of stochastic independence by a dynamic, local condition similar to the dynamic form of \(\Delta\)-independence used above. In the latter case, the construction of the risk sets \(\mathcal{U}_p^t\) (and the derived rates \(\tilde{r}^*(t)\)) take into account the temporal evolution of membership in \(\mathcal{U}^p\). Women become members of \(\mathcal{U}^p\) after the birth of her first child. This can be mimicked in a probability model by asking for a dynamic and local form of independence: \(^{24}\)

\[
\Pr(T(u, .) = t \mid T(u, .) \geq t) = \Pr(T(u, .) = t \mid T(u, .) \geq t \geq a^+, a^+ \leq C(u, .) \leq t)
\]

i.e. \(\{T(u, .) = t\} \perp \perp \{a^+ \leq C(u, .) \leq t\}\) given \(\{T(u, .) \geq t \geq a^+\}\). This condition might be used to justify the construction of consistent estimators of the conditional survivor function within the framework of the “simple device”. However, there is no obvious, non-trivial joint

\(^{24}\) Further details concerning concepts of dynamic or local independence are discussed in Chapter 5.
distribution of \((T(u, .), C(u, .))\) that obeys this form of local independence. The “simple device”, while illuminating in the context of simple (and important) examples, becomes rather cumbersome in the present context.

3.7. Selection by Survival

The problem of “left truncation” is not the last problem that must be tackled before a reasonable use of information from the children generation for inference on the distribution of the life length of the previous generation can be proposed. As already indicated in the introductory remarks to this Chapter, the available information from survey data will depend on the survival of the children up to the interview date. While the problem can be discussed in a similar way as was done up to now, in this Section I will propose a simulation model that simultaneously takes into account the problem of selection by survival and most of the previously discussed problems. The aim is to construct a background against which the “simple device” can be judged. The construction of the simulation model will be based on a probabilistic formulation. It is therefore favourable to the “simple device” and allows for the introduction of further probabilistic elements whose merits for arguments and speculations within the “simple device” can then be judged.

At the same time, I will try to build the many features characteristic for the present problem into the generating procedure. Moreover, data from official sources will be used as templates for the distributions employed for the data generation. This will allow for a comparison of the simulation results with the empirical results from survey data reported later in this Chapter.

3.7.1. The Simulation Model

The basic idea is to simulate data for a set of women according to a known survivor function and then to compare this known function with estimates based on information from the women’s children who survived
until some fixed interview date. In the first version of the model I refer to a set of \( N = 10000 \) women all born in the year \( t_0 := 1900 \); this set will be denoted by \( U \). I assume that these women survive according to the 1891–1900 period life table for Germany (see Table 3.3); the corresponding age-specific death rates will be denoted by \( \delta_f \). Further, the life tables for their children will be considered. The following table describes the official period life tables available for Germany:

<table>
<thead>
<tr>
<th>Period</th>
<th>Publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>1871 – 1880</td>
<td>Statistik des Deutschen Reichs, Vol. 246 (pp. 14*-17*).</td>
</tr>
<tr>
<td>1881 – 1890</td>
<td>Statistik des Deutschen Reichs, Vol. 246 (pp. 14*-17*).</td>
</tr>
<tr>
<td>1891 – 1900</td>
<td>Statistik des Deutschen Reichs, Vol. 246 (pp. 14*-17*).</td>
</tr>
<tr>
<td>1901 – 1910</td>
<td>Statistik des Deutschen Reichs, Vol. 246 (pp. 14*-17*).</td>
</tr>
<tr>
<td>1949 – 1951</td>
<td>Statistik der Bundesrepublik Deutschland, Vol. 75 and 173. Statistisches Jahrbuch für die Bundesrepublik Deutschland 1954 (pp. 62-63).</td>
</tr>
<tr>
<td>1960 – 1962</td>
<td>Statistisches Jahrbuch für die Bundesrepublik Deutschland 1965 (pp. 67-68).</td>
</tr>
</tbody>
</table>

All tables are period life tables. Until 1932 – 34, they refer to the territory of the former \( \text{Deutsches Reich} \); all other tables refer to the territory of
3. A Case Study: Parent’s Length of Life
the former FRG. Methods of table construction have slightly changed
throughout the years.
The following table 3.3 gives the surviving number of women at age t
from these life tables.
Table 3.3.: Female survivor functions. German period life tables.
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

164

100000 100000 100000 100000 100000 100000 100000 100000 100000 100000 100000
78260 79311 80138 82952 84695 90608 93161 95091 97222 98016 99298
73280 74404 76137 79761 82070 89255 92394 94749 97027 97888 99241
70892 72073 74482 78594 81126 88743 92026 94545 96922 97810 99201
69295 70514 73406 77867 80523 88422 91761 94390 96845 97745 99174
68126 69377 72623 77334 80077 88169 91535 94270 96782 97690 99153
67249 68537 72038 76924 79730 87975 91338 94177 96728 97641 99136
66572 67881 71577 76587 79445 87817 91160 94100 96682 97597 99119
66035 67358 71206 76301 79206 87683 91003 94041 96643 97558 99103
65599 66942 70903 76058 79001 87563 90870 93986 96609 97523 99088
65237 66601 70646 75845 78816 87452 90753 93937 96579 97492 99073
64926 66309 70420 75651 78642 87347 90650 93893 96552 97465 99058
64649 66049 70210 75467 78476 87243 90557 93850 96525 97439 99044
64390 65801 70003 75285 78311 87134 90467 93805 96498 97413 99029
64136 65555 69789 75094 78131 87013 90373 93756 96468 97384 99013
63878 65306 69562 74887 77930 86877 90270 93701 96434 97349 98995
63609 65045 69319 74661 77710 86719 90152 93637 96395 97305 98974
63322 64764 69060 74411 77470 86534 90016 93564 96351 97251 98947
63013 64468 68787 74143 77216 86319 89858 93484 96301 97189 98916
62681 64160 68500 73861 76945 86075 89680 93394 96246 97124 98881
62324 63838 68201 73564 76659 85808 89490 93295 96188 97059 98843
61941 63500 67888 73254 76362 85523 89287 93188 96128 96996 98806
61534 63142 67559 72929 76052 85226 89072 93073 96068 96934 98768
61102 62762 67212 72586 75730 84920 88849 92955 96008 96874 98731
60648 62360 66848 72225 75397 84602 88622 92834 95948 96815 98694
60174 61937 66467 71849 75043 84275 88390 92711 95884 96755 98657
59680 61497 66072 71463 74668 83943 88151 92586 95814 96694 98619
59170 61042 65666 71070 74283 83610 87904 92457 95739 96632 98579
58647 60570 65249 70669 73896 83274 87653 92324 95660 96567 98538
58111 60082 64822 70261 73513 82937 87397 92185 95575 96499 98493
57566 59584 64385 69848 73115 82597 87139 92039 95485 96429 98446
57010 59076 63937 69432 72703 82254 86876 91887 95390 96355 98395
56445 58554 63479 69008 72291 81909 86607 91729 95290 96276 98340
55869 58018 63010 68575 71876 81559 86329 91565 95184 96190 98280
55282 57473 62533 68132 71457 81205 86044 91396 95071 96098 98216
54685 56921 62047 67679 71020 80847 85754 91221 94949 95997 98146
54078 56360 61549 67215 70554 80482 85455 91039 94818 95886 98071
53462 55789 61041 66744 70080 80105 85145 90850 94676 95764 97988
52837 55215 60524 66266 69610 79720 84819 90651 94524 95632 97896
52207 54638 59998 65779 69139 79324 84481 90443 94360 95488 97796
51576 54054 59467 65283 68659 78917 84135 90225 94184 95331 97685


3.7. Selection by Survival

Table 3.3.: Female survivor functions. German period life tables.

|-----|------|------|------|------|------|------|------|------|------|------|------|

41 50946 53467 58931 64779 68172 78498 83779 93995 95161 97564
42 50320 52880 58391 64269 67689 78068 83410 93792 94975 97431
43 49701 52297 57848 63754 67194 77627 82630 93081 94308 96954
44 49090 51720 57302 63238 66692 78068 83410 93792 95161 97564
45 48481 51466 56751 62717 66187 77627 83027 93573 94773 97286
46 47870 50569 56195 62181 65661 78217 83779 94308 96954
47 47248 49983 55628 61628 65105 76704 82211 94308 97286
48 46605 49385 55040 61053 64510 75136 80767 97286
49 45939 48765 54423 60449 63883 74557 80213 97286
50 45245 48110 53768 59812 63231 73943 80166 97286
51 44521 47418 52354 58418 61827 72592 78990 97286
52 43767 46692 52354 58418 61827 72592 78990 97286
53 42981 45934 51594 57648 61048 71854 78990 97286
54 42162 45136 50791 56837 60219 71071 78990 97286
55 41308 44293 49938 55984 59350 70711 78990 97286
56 40414 43396 49938 55984 59350 70711 78990 97286
57 39472 42448 47054 53067 56498 67383 78990 97286
58 38476 41462 45748 51959 55400 68383 78990 97286
59 37418 40415 45971 52199 55400 68383 78990 97286
60 36293 39287 44293 50791 54000 65076 78990 97286
61 35101 38087 43582 49938 55984 60219 78990 97286
62 33843 36823 42272 48176 53067 56498 78990 97286
63 32521 35497 40880 45971 52199 57800 78990 97286
64 31140 34102 39398 43582 50791 54000 78990 97286
65 29703 32628 37828 43582 50791 57800 78990 97286
66 28117 30486 34460 40415 52199 57800 78990 97286
67 26686 29506 32628 35971 50791 57800 78990 97286
68 25118 27897 30266 34460 45971 57800 78990 97286
69 23521 26252 28917 30266 45971 57800 78990 97286
70 21901 24546 26956 30266 45971 57800 78990 97286
71 20265 22786 24982 30266 45971 57800 78990 97286
72 18617 21000 23755 30266 45971 57800 78990 97286
73 16960 19204 22938 30266 45971 57800 78990 97286
74 15307 17416 21044 30266 45971 57800 78990 97286
75 13677 15645 18900 30266 45971 57800 78990 97286
76 12090 13892 16919 30266 45971 57800 78990 97286
77 10569 12219 19050 30266 45971 57800 78990 97286
78 9131 10661 13165 30266 45971 57800 78990 97286
79 7795 9192 11417 30266 45971 57800 78990 97286
80 6570 7815 9773 30266 45971 57800 78990 97286
81 5464 6550 8252 30266 45971 57800 78990 97286
82 4479 5205 6869 30266 45971 57800 78990 97286
83 3614 4394 5626 30266 45971 57800 78990 97286
84 2867 3511 4524 30266 45971 57800 78990 97286
85 2232 2756 3686 30266 45971 57800 78990 97286
86 1705 2124 2764 30266 45971 57800 78990 97286
87 1276 1605 2104 30266 45971 57800 78990 97286
88 935 1189 1571 30266 45971 57800 78990 97286

165
Table 3.3.: Female survivor functions. German period life tables.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>671</td>
<td>862</td>
<td>1149</td>
<td>1571</td>
<td>1574</td>
<td>3128</td>
<td>3857</td>
<td>6335</td>
<td>8147</td>
<td>10245</td>
<td>20393</td>
<td></td>
</tr>
<tr>
<td>90</td>
<td>471</td>
<td>612</td>
<td>821</td>
<td>1131</td>
<td>1126</td>
<td>2356</td>
<td>2868</td>
<td>4815</td>
<td>6480</td>
<td>8016</td>
<td>16903</td>
<td></td>
</tr>
<tr>
<td>91</td>
<td>323</td>
<td>424</td>
<td>573</td>
<td>797</td>
<td>786</td>
<td>1736</td>
<td>2083</td>
<td>3567</td>
<td>4872</td>
<td>6139</td>
<td>13738</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>217</td>
<td>288</td>
<td>390</td>
<td>549</td>
<td>534</td>
<td>1256</td>
<td>1476</td>
<td>2571</td>
<td>3580</td>
<td>4597</td>
<td>10935</td>
<td></td>
</tr>
<tr>
<td>93</td>
<td>142</td>
<td>191</td>
<td>260</td>
<td>370</td>
<td>354</td>
<td>891</td>
<td>1019</td>
<td>1814</td>
<td>2571</td>
<td>3362</td>
<td>8511</td>
<td></td>
</tr>
<tr>
<td>94</td>
<td>90</td>
<td>123</td>
<td>169</td>
<td>244</td>
<td>228</td>
<td>620</td>
<td>683</td>
<td>1253</td>
<td>1805</td>
<td>2409</td>
<td>6468</td>
<td></td>
</tr>
<tr>
<td>95</td>
<td>56</td>
<td>78</td>
<td>107</td>
<td>157</td>
<td>142</td>
<td>423</td>
<td>445</td>
<td>846</td>
<td>1240</td>
<td>1671</td>
<td>4792</td>
<td></td>
</tr>
<tr>
<td>96</td>
<td>34</td>
<td>48</td>
<td>66</td>
<td>99</td>
<td>87</td>
<td>283</td>
<td>281</td>
<td>559</td>
<td>834</td>
<td>1134</td>
<td>3457</td>
<td></td>
</tr>
<tr>
<td>97</td>
<td>20</td>
<td>29</td>
<td>40</td>
<td>61</td>
<td>51</td>
<td>185</td>
<td>172</td>
<td>361</td>
<td>550</td>
<td>750</td>
<td>2425</td>
<td></td>
</tr>
<tr>
<td>98</td>
<td>11</td>
<td>17</td>
<td>24</td>
<td>38</td>
<td>29</td>
<td>119</td>
<td>101</td>
<td>227</td>
<td>356</td>
<td>483</td>
<td>1651</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>6</td>
<td>10</td>
<td>14</td>
<td>22</td>
<td>16</td>
<td>74</td>
<td>58</td>
<td>140</td>
<td>227</td>
<td>303</td>
<td>1090</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>3</td>
<td>6</td>
<td>8</td>
<td>13</td>
<td>9</td>
<td>45</td>
<td>31</td>
<td>84</td>
<td>142</td>
<td>185</td>
<td>697</td>
<td></td>
</tr>
</tbody>
</table>

Additionally, the birth of children must be considered. While women as well as men may become parents in a legal sense in many different ways, I will only consider women who gave birth to children. Age- and parity-specific birth rates are given by

$$\beta_{t,k} := \frac{\text{Number of women giving birth to a further child at age } t}{\text{Number of women aged } t \text{ and having } k \text{ children}}$$

In order to arrive at a simulation model that roughly corresponds to the historical situation these rates are calculated from a subsample of the census that took place in Germany in the year 1970.\(^{25}\)

The census of 1970 was conducted on May 27 of that year in the territory of the former FRG.\(^{26}\) As part of this census a subsample of 10% of the population was asked to provide additional information, in particular, all women with a German citizenship who participated in the 10% subsample were asked for dates of marriage and birth dates of all their marital children, regardless of their current marital status.\(^{27}\) Several

---

\(^{25}\) The birth rates normally reported by the statistical office refer only to marital births, starting parity counts afresh with each marriage. The non- or under-reporting of births per woman will at least continue to the next planned census in 2011.

\(^{26}\) For a detailed description, including a presentation of the questionnaire see Schubnell and Herberger (1970).

\(^{27}\) Some results from these additional questions were published, albeit in highly aggregated form, by the *Statistisches Bundesamt* in Fachserie A.
years ago, official statistics in Germany agreed to make available an anonymised 10% subsample of the 1970 census. This 1% subsample of all women who lived in May 1970 in the territory of the former FRG and had a German citizenship is used in the sequel. The number of cases is 314993. If multiplied by 100, this is roughly the number of women with a German citizenship living in the former FRG in May 1970.

For the calculation of age- and parity-specific birth rates I have used all of the women born between 1870 and 1925, altogether 131135 cases. While these rates may not be very accurate, are spanning very different historical contexts, and are themselves subject to selection- and migration problems (including possible differential mortality due to childbearing etc.), the main issue here is to provide a historically reasonable background to judge statistical methods, a task that may be accomplished by these numbers. An accurate account of the demography of the period is not intended here.

Table 3.4.: Parity-specific birth rates per 10000 women

<table>
<thead>
<tr>
<th>t</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>10</td>
<td>286</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>39</td>
<td>185</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>102</td>
<td>432</td>
<td>99</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>196</td>
<td>628</td>
<td>318</td>
<td>769</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>332</td>
<td>885</td>
<td>644</td>
<td>482</td>
<td>2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>455</td>
<td>999</td>
<td>776</td>
<td>651</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>22</td>
<td>598</td>
<td>1123</td>
<td>902</td>
<td>905</td>
<td>897</td>
<td>1250</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>23</td>
<td>702</td>
<td>1245</td>
<td>1040</td>
<td>1128</td>
<td>737</td>
<td>870</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>800</td>
<td>1299</td>
<td>1073</td>
<td>988</td>
<td>1045</td>
<td>923</td>
<td>2222</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>860</td>
<td>1317</td>
<td>1109</td>
<td>1090</td>
<td>897</td>
<td>1260</td>
<td>1304</td>
<td>4000</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>26</td>
<td>911</td>
<td>1278</td>
<td>1088</td>
<td>1121</td>
<td>1167</td>
<td>1133</td>
<td>816</td>
<td>1429</td>
<td>3333</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>27</td>
<td>905</td>
<td>1304</td>
<td>1009</td>
<td>1138</td>
<td>1238</td>
<td>1148</td>
<td>1237</td>
<td>556</td>
<td>0</td>
<td>10000</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>883</td>
<td>1255</td>
<td>1028</td>
<td>1150</td>
<td>1307</td>
<td>1395</td>
<td>1230</td>
<td>500</td>
<td>2000</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>29</td>
<td>844</td>
<td>1224</td>
<td>975</td>
<td>1103</td>
<td>1179</td>
<td>1589</td>
<td>954</td>
<td>1884</td>
<td>588</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>30</td>
<td>753</td>
<td>1176</td>
<td>922</td>
<td>1074</td>
<td>1263</td>
<td>1398</td>
<td>1835</td>
<td>1172</td>
<td>1818</td>
<td>1250</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>679</td>
<td>1078</td>
<td>886</td>
<td>1004</td>
<td>1191</td>
<td>1384</td>
<td>1442</td>
<td>1682</td>
<td>2031</td>
<td>1000</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

---

28 More information on these data sets are available from the Zentrum für Umfragen, Methoden und Analysen (ZUMA, Mannheim), Abteilung für Mikrodaten; see: http://www.gesis.org/Dauerbeobachtung/GML.

29 I thank Bernhard Schimpl-Neimanns (ZUMA) who prepared the basic counts used here. I still had to delete one woman with a negative age at birth.
3. A Case Study: Parent’s Length of Life

Table 3.4.: Parity-specific birth rates per 10000 women

<table>
<thead>
<tr>
<th>$t$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>597</td>
<td>1016</td>
<td>792</td>
<td>890</td>
<td>1177</td>
<td>1320</td>
<td>1675</td>
<td>2143</td>
<td>1333</td>
<td>2059</td>
<td>2727</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>510</td>
<td>937</td>
<td>740</td>
<td>914</td>
<td>1071</td>
<td>1304</td>
<td>1371</td>
<td>1628</td>
<td>2262</td>
<td>1231</td>
<td>0</td>
<td>5000</td>
</tr>
<tr>
<td>34</td>
<td>433</td>
<td>823</td>
<td>675</td>
<td>834</td>
<td>1074</td>
<td>1326</td>
<td>1480</td>
<td>1470</td>
<td>2087</td>
<td>1800</td>
<td>1515</td>
<td>1667</td>
</tr>
<tr>
<td>35</td>
<td>355</td>
<td>711</td>
<td>610</td>
<td>726</td>
<td>997</td>
<td>1198</td>
<td>1242</td>
<td>1587</td>
<td>1870</td>
<td>1985</td>
<td>1852</td>
<td>667</td>
</tr>
<tr>
<td>36</td>
<td>305</td>
<td>588</td>
<td>537</td>
<td>692</td>
<td>863</td>
<td>1133</td>
<td>1500</td>
<td>1476</td>
<td>1280</td>
<td>1905</td>
<td>2222</td>
<td>2000</td>
</tr>
<tr>
<td>37</td>
<td>259</td>
<td>499</td>
<td>466</td>
<td>626</td>
<td>772</td>
<td>993</td>
<td>1136</td>
<td>1529</td>
<td>1408</td>
<td>2247</td>
<td>1875</td>
<td>2619</td>
</tr>
<tr>
<td>38</td>
<td>197</td>
<td>373</td>
<td>422</td>
<td>523</td>
<td>713</td>
<td>914</td>
<td>998</td>
<td>1520</td>
<td>1574</td>
<td>1290</td>
<td>1981</td>
<td>2115</td>
</tr>
<tr>
<td>39</td>
<td>164</td>
<td>315</td>
<td>316</td>
<td>444</td>
<td>613</td>
<td>792</td>
<td>911</td>
<td>1352</td>
<td>1565</td>
<td>1486</td>
<td>1587</td>
<td>1803</td>
</tr>
<tr>
<td>40</td>
<td>122</td>
<td>243</td>
<td>245</td>
<td>355</td>
<td>510</td>
<td>642</td>
<td>868</td>
<td>1383</td>
<td>1126</td>
<td>1447</td>
<td>1600</td>
<td>1486</td>
</tr>
<tr>
<td>41</td>
<td>91</td>
<td>154</td>
<td>187</td>
<td>275</td>
<td>439</td>
<td>544</td>
<td>782</td>
<td>1017</td>
<td>1057</td>
<td>851</td>
<td>1494</td>
<td>930</td>
</tr>
<tr>
<td>42</td>
<td>59</td>
<td>102</td>
<td>133</td>
<td>179</td>
<td>275</td>
<td>451</td>
<td>539</td>
<td>765</td>
<td>930</td>
<td>1212</td>
<td>1371</td>
<td>1667</td>
</tr>
<tr>
<td>43</td>
<td>39</td>
<td>64</td>
<td>83</td>
<td>119</td>
<td>200</td>
<td>309</td>
<td>435</td>
<td>581</td>
<td>747</td>
<td>732</td>
<td>787</td>
<td>808</td>
</tr>
<tr>
<td>44</td>
<td>25</td>
<td>43</td>
<td>51</td>
<td>81</td>
<td>130</td>
<td>177</td>
<td>231</td>
<td>459</td>
<td>389</td>
<td>543</td>
<td>761</td>
<td>1143</td>
</tr>
<tr>
<td>45</td>
<td>14</td>
<td>20</td>
<td>28</td>
<td>48</td>
<td>60</td>
<td>141</td>
<td>159</td>
<td>214</td>
<td>255</td>
<td>548</td>
<td>316</td>
<td>707</td>
</tr>
<tr>
<td>46</td>
<td>8</td>
<td>12</td>
<td>11</td>
<td>23</td>
<td>27</td>
<td>42</td>
<td>62</td>
<td>122</td>
<td>211</td>
<td>28</td>
<td>432</td>
<td>400</td>
</tr>
<tr>
<td>47</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>12</td>
<td>20</td>
<td>23</td>
<td>31</td>
<td>47</td>
<td>31</td>
<td>56</td>
<td>0</td>
<td>306</td>
</tr>
<tr>
<td>48</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>6</td>
<td>14</td>
<td>12</td>
<td>13</td>
<td>9</td>
<td>15</td>
<td>84</td>
<td>110</td>
<td>0</td>
</tr>
<tr>
<td>49</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>9</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

From a starting set of identifying numbers $U$ of women born in 1900, their children are represented by a second set $U^c := \bigcup_{u \in U} (u)$ containing identification numbers of all children born to the women in $U$. The number of members of $U^c$ is not known in advance but depends on both the death rates $\delta_t$ and the birth rates $\beta_{t,k}$. Given these rates, I created two lists. One list contains for each woman in $U$ her identification number, her death year, and her number of children. Another list contains for each child in $U^c$ an identification number, the birth year, and the identification number of the mother. In addition, in order to simulate a retrospective survey, I assume that the children survive according to the 1960–1962 period life table for women (see Table 3.3); the corresponding age-specific death rates will be denoted by $\delta_{t,c}$. Using these rates, a simulated death year is added for each child in the second list.

Box 3.7-1 summarises the algorithm used to generate the simulated data. In the description, the $\varepsilon$ refer to draws from random numbers uniformly distributed in the unit interval. Using this algorithm I get a first list with
Box 3.7-1 Skeleton of the simulation model.

For each \( u \in \mathcal{U} \) do:
\[
    n(u) := 0; \quad \# \text{ counter for } u\text{'s children}
\]
For \( t = 0, \ldots, 100 \) {
    Generate a new, independent random number \( \epsilon_{t,b}(u) \);
    If \( (\epsilon_{t,b}(u) \leq \beta_{t,n(u)}) \)
        increase \( n(u) \) by 1; create a new entry in \( \mathcal{U}^c \),
        and record the mother’s identification number and age;
    Get, independently, another random number \( \epsilon_{t,d}(u) \);
    If \( (\epsilon_{t,d}(u) \leq \delta_{t}) \)
        goto L1;
}
L1:
Record that \( u \) died at age \( t \) and has given birth to \( c(u) \)
children, also record for all children the mother’s age at death;

For each \( u \in \mathcal{U}^c \) do: \( \# \) Simulate children’s survival
For \( t = 0, \ldots, 100 \) {
    Get another, independent random number \( \epsilon_{t,c}(u) \);
    If \( (\epsilon_{t,c}(u) \leq \delta_{t}) \)
        goto L2;
}
L2:
Record that \( u \) died at age \( t \);

\( N = 10000 \) entries that records the identification numbers of the women,
\( \mathcal{U} \), their age at death, and their number of children. In one run of the
algorithm, presented below, 4776 of these 10000 women have at least one
child.\textsuperscript{30} The second list contains entries for 11407 children. Figure 3.6
shows a frequency distribution of the years in which the women in \( \mathcal{U} \)
died on a historical time axis. Also shown are frequency distributions
of the birth and death years of the children. Note that the algorithm is
based on the assumption that women’s survival is locally independent of
their giving birth to children. The algorithm provides an operational

\textsuperscript{30} Note that according to the 1891–1900 period life table, only 68% of the women
survived age 20, and only 60% survived age 40.
3. A Case Study: Parent’s Length of Life

![Graph showing frequency distributions of birth and death years in the simulated dataset.](image)

**Figure 3.6:** Frequency distributions of birth and death years in the simulated data set.

Explication of this notion in discrete time: given survival to age $t$, the draws of random numbers $\varepsilon_{t,b}$ and $\varepsilon_{t,d}$ used to simulate births and deaths at age $t$ are stochastically independent. 

3.7.2. Considering Left Truncation

Before using the model to discuss the question of how to recover the survivor function of the members of $\mathcal{U}$ based on information resulting from a retrospective survey of their children, I illustrate the importance

---

31 Problems arising from a violation of this assumption can not be checked within this model. However, it can be adapted rather easily to allow for some form of (local) dependence. Also, dependence of life lengths of mothers and children (both local and global) might be incorporated. It is of some theoretical interest to investigate whether a simulation algorithm may be conceived that does not involve a fixed starting population $\mathcal{U}$ of potential mothers. After all, the surveyor will never have access to that population nor even direct information about its size. Such an algorithm ought to be able to reproduce all reasonable distributions of the life lengths of the previous generation without any extraneous assumptions on the dependence of life length across generations.
of correctly taking into account left truncated observations and discuss some complications arising from real data limitations.

The goal is to recover (some part of) the distribution of the statistical variable

$$T: \mathcal{U} \rightarrow \mathcal{T} := \{0, 1, \ldots, \tau\}$$

that records the life length of the members of $\mathcal{U}$ from observations on a selected subset of $\mathcal{U}$. The subset is not determined by some random device under the control of the survey administrator or some process that may be judged similar to such a device. Instead, the observations refer to the subset of members of $\mathcal{U}$ having given birth to at least one child. This subset will be denoted by $\mathcal{U}^p$. In parallel to the definition of $T$ one may define a statistical variable

$$T^*: \mathcal{U}^p \rightarrow \mathcal{T}$$

that records the life length of the members of $\mathcal{U}^p$. Then the statistical variable $T^*$ is the restriction of the function $T: \mathcal{U} \rightarrow \mathcal{T}$ to the subset $\mathcal{U}^p$.

In the simulation model of the previous Section, the distribution of $T$ is defined by the death rates $\delta^f_t$. The survivor function of $T$ is therefore given by

$$G(t) = \prod_{j=0}^{t-1} (1 - \delta^f_j)$$

One possibility to recover (some part of) this survivor function is to somehow estimate the death rates $\delta^f_t$. However, these death rates may be systematically different from the naively computed death rates derived from the statistical variable $T^*$

$$r^*(t) := \frac{|\{u \in \mathcal{U}^p \mid T^*(u) = t\}|}{|\{u \in \mathcal{U}^p \mid T^*(u) \geq t\}|} = \text{P}(T^* = t \mid T^* \geq t)$$

The rate function $r^*(.)$ refers to deaths among those women who ever became mothers. But in order to find estimates of $\delta^f_t$, one needs to take
3. A Case Study: Parent’s Length of Life

into account that women only become members of $U^p$ when they have given birth to a first child and not on the fact that they ever gave birth.

Recalling the framework of Section 3.6, consider the two-dimensional variable

$$(T, C): U \longrightarrow T \times (T \cup \{-1\})$$

As before, $T(u)$ denotes $u$’s life length and $C(u)$ refers to the age at which $u$ became for the first time mother of a child. If $u$ stays childless throughout her lifetime, $C(u)$ is set to -1. In a temporal perspective, therefore, the value of $C(u)$ is only determined after $u$’s death.

Remember further the definitions of the temporarily evolving populations

$U_t := \{u \in U \mid T(u) \geq t\}$ and $U^p_t := \{u \in U^p \mid T(u) \geq t, 0 \leq C(u) \leq t\}$

One may then define the rate function

$$\tilde{r}^*(t) := \frac{|\{u \in U^p_t \mid T(u) = t\}|}{|U^p_t|} = P(T = t \mid T \geq t, 0 \leq C \leq t)$$

which refers to the death rates in the population of mothers conditional on a previous birth.

Generally, $\tilde{r}^*(t) \geq r^*(t)$, and thus using $r^*(.)$ underestimates the rate and the survivor function is overestimated. This can be seen by comparing the denominators in the definitions of $r^*(.)$ and $\tilde{r}^*(.)$ (the numerators are equal):

$$|\{U^p \mid T(u) \geq t, 0 \leq C(u) \leq t\}| \leq |\{U^p \mid T(u) \geq t\}|$$

Now $\tilde{r}^*(t)$ is the death rate of women who actually are members of $U^p$ at the age of $t$ and, as I have construed the model, it is a reasonable estimate of $\delta^t$, at least in the extended probability model: The random variables corresponding to $(T, C)$ are

$$(T, C): U \times \Omega \longrightarrow T \times (T \cup \{-1\})$$
The simulation proceeds by stepping forward through time. At each age \( t \), the process ends if \( \varepsilon_{t,d}(u, \omega) \leq \delta_{t}^{f} \) and \( T(u, \omega) \) is set to \( t \), otherwise it proceeds to the next age \( t + 1 \). Thus the simulation provides realisations of life lengths \( T(u, .) \) with \( \Pr(T(u, .) = t \mid T(u, .) \geq t) = \delta_{t}^{f} \) and \( \Pr(T(u, .) \geq t) = G(t) \) for all \( u \). With \( T \) finite and \( U \) large, ‘most’ empirical distribution functions derived from the simulations can be expected to be close to \( G(.) \).

The simulation does not directly produce a random variable \( C(u, .) \). In order to use the results also for the discussion of more complicated situations, it proceeds to simulate data for all births of \( u \). However, the random variable \( C(u, .) \) as defined above is a simple function of the random variables \( n(u, .) \) and the birth dates recorded in the list of children. And since the simulation for any \( u \in U \) stops with a death event, \( C(u, .) \leq T(u, .) \) is ensured. Furthermore, also by construction

\[
\Pr(T(u, .) = t \mid T(u, .) \geq t) = \Pr(T(u, .) = t \mid T(u, .) \geq t, 0 \leq C(u, .) \leq t)
\]

But the former quantity is \( \delta_{t}^{f} \) and the latter quantity is the one estimated by \( \tilde{r}^{*}(.) \) so that \( \tilde{r}^{*}(.) \) should be close to \( \delta_{t}^{f} \) in reasonably large samples.

Figure 3.7.: Survivor functions, conditional on \( t \geq 21 \), from the 1891–1900 period life table (solid line) and from women with at least one child in the simulated data set (dotted line).
In principle, it would be possible to obtain estimates $\tilde{r}^*$ from the earliest observed age at first birth onward. Since this rate is zero up to the age of the first observed death in $\mathcal{U}_p$, one might as well start at this age, say $a^+$. On the other hand, due to the small number of cases in a sample of observations, $\tilde{r}^*(a^+)$ might not be a good estimate of $\delta^*(a^+)$ and one should condition on some later age. In fact, it might even happen that $\tilde{r}^*(a^+)=1$ so that one cannot find a reasonable estimate of a survivor function beginning at $a^+$.

Having chosen a reasonable $a^+$, the conditional survivor function derived from the estimates $\tilde{r}^*(.)$ becomes

$$\tilde{G}_{a^+}^*(t) := \prod_{j=a^+}^{t-1} (1 - \tilde{r}^*(j))$$

It might be taken as an estimate of the conditional survivor function $G(. \mid T \geq a^+)$. In the simulated data set, the earliest death occurs at age 19. However, this occurs only once, and at $t = 20$ there is no death at all. Therefore, I define $a^+ := 21$. The estimate $\tilde{G}_{a^+}^*(.)$ is shown in Figure 3.7 as a dotted line together with $G(. \mid T \geq a^+)$ calculated from the 1891-1900 period life table for women. Obviously, both curves agree quite well. On the other hand, if I had not taken into account the fact that women become members of $\mathcal{U}_p$ only after having given birth to a first child, but estimated the survivor function of $T^*$, the result would be systematically biased as a consequence of the inequality $r^*(t) \leq \tilde{r}^*(t)$. This is illustrated by Figure 3.8 where the solid line shows $\tilde{G}_{a^+}^*$ and the dotted line shows $P(T^* \geq . \mid T^* \geq a^+)$. The fact that women become members of $\mathcal{U}_p$ only after the birth of a child is formally equivalent to treating the observations as left truncated at the age at first birth. Of course, nothing is wrong with estimating the survivor function of $T^*$ instead of using $\tilde{G}_{a^+}^*(.)$. The argument has only shown that one should use the latter if the interest is in recovering part of the distribution of $T$. One might also notice that, while $T^*$ refers to a well-defined statistical variable, there is no statistical variable of which $\tilde{G}_{a^+}^*(.)$ is a conditional survivor function. I.e., the definition of $\tilde{G}_{a^+}^*(.)$ (based on $\tilde{r}^*(.)$) refers to a risk set that changes not only in
Figure 3.8.: Conditional survivor functions $\tilde{G}^*_a(.)$ (solid line) and $P(T^* \geq . \mid T^* \geq a^*)$ calculated from the simulated data set with $a^* = 21$.

accordance with the mortality of (a fixed subset of) women. It also depends dynamically on the fertility of the women. In fact, $\tilde{r}^*(.)$ is a mixture of rate functions defined for subsets of $U^p$. To see this, partition $U^p$ into subsets

$$U^p_{[a]} := \{ u \in U^p \mid C(u) = a \} = C^{-1}(\{a\})$$

consisting of those members of $U^p$ who had a first birth at the age $a$. The mortality experience of this subset can be summarised by the rate function

$$\tilde{r}^*_a(t) := \frac{\left| \{ U^p_{[a]} \mid T(u) = t \} \right|}{\left| \{ U_{[a]}^p \mid T(u) \geq t \} \right|} = P(T = t \mid T \geq t, C = a) \mathbb{1}[t \geq a]$$

The rate function $\tilde{r}^*(.)$ arises as a mixture of the rate functions $\tilde{r}^*_a(.)$:

$$\tilde{r}^*(t) = \sum_{a \leq t} \tilde{r}^*_a(t) w_a(t)$$

where the weights, defined as

$$w_a(t) := P(C = a \mid C \leq t, T \geq t) \mathbb{1}[t \geq a]$$
reflect the composition of the risk set at $t$. That there is in general no fixed subset of women in $U^p$ whose mortality experience is ‘representative’ of the mortality experience of $U$ reflects the difficulties created by the dynamic selection of $U^p$.

3.7.3. Using Information from Children

Turning to the question of how to estimate conditional survivor functions for the members of $U$ when only information from their children (members of $U^c$) is available, one needs to take into account the relationship between $U^c$ and $U^p$. Recall from Section 3.1 the definitions of a few functions that describe generational dependencies:

$$m: U^c \rightarrow U^p \cap U^f$$

where for each child $u \in U^c$, $m(u)$ refers to the mother of $u$ in $U^p$. Conversely, for each woman $u \in U^p$, $m^{-1}\{u\}$ is the set of her children in $U^c$.

Now let $s$ denote a simple random sample from $U^c$. This induces a random sample from $U^p$, namely

$$s^* := \{ u \in U^p \mid \text{there is an } u' \in s \text{ with } m(u') = u \}$$

But $s^*$ is not a simple random sample from $U^p$ because women with more children are more likely to be included in $s^*$. This should be taken into account when estimating $\tilde{r}^*(t)$ from information provided by the children in the sample $s$.

The further problem, considered in Section 3.7.2, of the temporal nature of the membership of women in $U^p$, can be traced from survey responses of children if the variable $C$ referring to the age at the first birth of the mothers (alternatively, the birth dates of all siblings) is recorded. Accordingly, if $u$ is included in $s$, one should not condition on her mother’s age when giving birth to $u$, but on the age of her mother’s first child-bearing. To illustrate the difference, I use the data from the simulation model and compare two fictitious samples: $s_1$ contains all first-born children from $U^c$, and $s_2$ contains all last-born children from $U^c$. 

---

176
Figure 3.9: Comparison of conditional survivor functions calculated from two different samples from the simulated data set.

$U^c$. Of course, both samples provide the same information about the life length of women in $U^p$. But there are now different ways to select truncation times. If I condition on the age of the mothers when giving birth to the children in the samples, I get the results shown in Figure 3.9. Obviously, conditioning on the mother’s age when giving birth to her last child would result in an extremely biased estimate.

In order to avoid this mistake, ideally, the following statistical variables should be available from a sample of children:

$$(T, C, N): s^* \subseteq U \longrightarrow \mathcal{T} \times \mathcal{T} \times \{1, 2, 3, \ldots\}$$

where $T(m(u))$ records the life length of $u$’s mother, $C(u)$ denotes mother’s age at first child-bearing, and $N(u)$ counts the mother’s number of children. Since $N$ will be used to provide weights for the observations in the sample $s^*$, this should be the number of children surviving up to the time when the sample is drawn. Assuming that this information is available from a simple random sample $s$ and in conformance with the discussion in Section 3.3 (now restricted to mothers), the rates $\tilde{r}^*$ can be estimated by:
Figure 3.10: Comparison of conditional survivor functions estimated with, and without, weights from the simulated data set.

\[
\tilde{r}^*(t) \approx \tilde{r}_w^*(t) := \frac{\sum_{u \in s} \frac{1}{N(u)} \mathbbm{1}[T = t, C \leq t](u)}{\sum_{u \in s} \frac{1}{N(u)} \mathbbm{1}[T \geq t, C \leq t](u)}
\]

To illustrate, I use again data from the simulation model. Figure 3.10 compares conditional survivor functions calculated from estimated rates \( \tilde{r}_w^*(t) \) and from analogously defined rates where the weights are dropped.\(^{32}\) The figure clearly indicates that one should use the weights \(1/N\) if this information is available.

However, general surveys often do not provide this information. It is therefore important to explore another way to arrive at reasonable estimates. In order to explain this possibility consider the risk set

\[
U_t^P = \{ u \in U^P \mid T(u) \geq t, C(u) \leq t \}
\]

\(^{32}\) In the calculation I have used all observations from \(U^c\), but basically the same differences would result from a simple random sample from \(U^c\).
at $t$. The death rates to be estimated can then be written as

$$
\tilde{r}^*(t) = \frac{\left| \{ u \in \mathcal{U}_t^p \mid T(u) = t \} \right|}{\left| \mathcal{U}_t^p \right|} = \frac{P(T = t \mid T \geq t, C \leq t)}{\left| \mathcal{U}_t^p \right|} \approx \delta_t^f
$$

By assumption, these rates do not depend on the number of children born of members of $\mathcal{U}_t^p$ until $t$, and also do not depend on the children’s birth dates. To make this explicit, partition the risk sets into subsets according to the number of children born until $t$. Let $K_t^*(u)$ denote the number of children born of $u$ until $t$. Each risk set $\mathcal{U}_t^p$ may then be written as a union of subsets

$$
\mathcal{U}_{t,k}^p := \{ u \in \mathcal{U}_t^p \mid K_t^*(u) = k \}
$$

taken over all possible values of $k$. Furthermore, I can define death rates for these subsets,

$$
\tilde{r}_k^*(t) := \frac{\left| \{ u \in \mathcal{U}_{t,k}^p \mid T^*(u) = t \} \right|}{\left| \mathcal{U}_{t,k}^p \right|}
$$

However, by assumption these rates are all (approximately) identical to the death rate $\tilde{r}^*(t)$. Consequently, I do not need weights when I only use information from children born until $t$. Instead, I can directly refer to the sets of children born of women in $\mathcal{U}_{t,k}^p$ which can be defined by

$$
\mathcal{U}_{t,k}^c := m^{-1}(\mathcal{U}_{t,k}^p)
$$

The death rates $\tilde{r}_k^*(t)$ may then be written as

$$
\tilde{r}_k^*(t) \approx \frac{\left| \{ u \in \mathcal{U}_{t,k}^c \mid T^*_c(u) = t \} \right|}{\left| \mathcal{U}_{t,k}^c \right|}
$$

and, since these rates are approximately identical across the subsets, I might finally write

$$
\tilde{r}^*(t) \approx \tilde{r}_c^*(t) := \frac{\left| \{ u \in \mathcal{U}^c \mid T^*_c(u) = t, S^*_c(u) \leq t \} \right|}{\left| \{ u \in \mathcal{U}^c \mid T^*_c(u) \geq t, S^*_c(u) \leq t \} \right|} \tag{3.15}
$$
Figure 3.11: Conditional survivor function estimated from the rates \( \tilde{r}_c^*(t) \), compared with a conditional survivor function from the 1891-1900 period life table.

where now \( S_c^*(u) \) is the age of \( u \)'s mother at the birth of \( u \). Notice that this approach does not require any weights and also requires no information about the mother's age at her first child-bearing.

To illustrate the argument I use again data from the simulation model. Taking into account all children in \( U^c \) but, for the calculation of the rates \( \tilde{r}_c^*(t) \) only use information from children born not later than \( t \). Of course, this simply means to use all information from \( U^c \) and, for each \( u \in U^c \), treat the observation about \( u \)'s mother as left truncated at \( S_c^*(u) \).\(^{33}\) Figure 3.11 shows the conditional survivor function calculated from the rates \( \tilde{r}_c^*(t) \). This function obviously agrees quite well with the 1891-1900 period life table that was used to generate the data. Of course, the result would be basically the same if I had used a simple random sample from \( U^c \).

\(^{33}\) One can use, therefore, any standard Kaplan-Meier procedure that allows for left truncated data. I have used TDA's dp1e procedure.
3.7.4. Retrospective Surveys

In the previous Section I assumed that I have data from a simple random sample from the complete set of children, \( \mathcal{U}^c \). However, the data actually result from a retrospective survey performed in some specific year, say \( \tau \), and I therefore have to take into account that not all members of \( \mathcal{U}^c \) survive until \( \tau \). Fortunately, the approach to estimate \( \delta(t) \) via the rates \( \tilde{\tau}^*_c(t) \) that was discussed in the previous Section can also be applied to a retrospective sample if I make the additional assumption that children's life lengths are independent of their mother's life length.\(^{34}\)

To explain the argument, let \( T^c \) denote the life length of children in the reference set \( \mathcal{U}^c \). On a historical time axis, if mothers are born in the year \( \tau_0 \), each child \( u \in \mathcal{U}^c \) survives until \( \tau_0 + S^*_c(u) + T^c(u) \) (as already introduced, \( S^*_c(u) \) is the age of the mother when \( u \) was born). The set of children who survive at least until the year \( \tau \) is therefore given by

\[
\mathcal{U}^c[\tau] := \{ u \in \mathcal{U}^c \mid \tau_0 + S^*_c(u) + T^c(u) \geq \tau \}
\]

In the simulation model introduced in Section 3.7.1 I assumed \( \tau_0 = 1900 \). Based on this assumption, Figure 3.6 shows the survival of children in historical time.

Now assume a retrospective survey performed in the year \( \tau \). The sample is then drawn from the reference set \( \mathcal{U}^c[\tau] \). Following the approach discussed in the previous Section, I can calculate rates

\[
\tilde{\tau}^*_{c,\tau}(t) := \frac{|\{ u \in \mathcal{U}^c[\tau] \mid T^*_c(u) = t, S^*_c(u) \leq t \}|}{|\{ u \in \mathcal{U}^c[\tau] \mid T^*_c(u) \geq t, S^*_c(u) \leq t \}|}
\]

which are defined analogously to the rates \( \tilde{\tau}^*_c(t) \) introduced in (3.15). In order to see that the rates \( \tilde{\tau}^*_{c,\tau}(t) \) are reasonable estimates of the rates \( \tilde{\tau}^*_c(t) \), their definition might be written in the following way:

\[
\tilde{\tau}^*_{c,\tau}(t) = \frac{|\{ u \in \mathcal{U}^c \mid T^*_c(u) = t, S^*_c(u) \leq t, S^*_c(u) + T^c(u) \geq \tau - \tau_0 \}|}{|\{ u \in \mathcal{U}^c \mid T^*_c(u) \geq t, S^*_c(u) \leq t, S^*_c(u) + T^c(u) \geq \tau - \tau_0 \}|}
\]

\(^{34}\) This assumption, already built into the simulation model in Section 3.7.1, is probably not completely true. However, for the moment I will base my argument on this assumption.
The further argument proceeds in terms of conditional frequencies. Using an abbreviated notation, one may write:

\[
\tilde{r}_{c,\tau}^*(t) = \frac{\Pr(T_c^* = t, S_c^* \leq t, S_c^* + T_c \geq \tau - \tau_0)}{\Pr(T_c^* \geq t, S_c^* \leq t, S_c^* + T_c \geq \tau - \tau_0)} = \frac{\Pr(S_c^* + T_c \geq \tau - \tau_0 \mid T_c^* = t, S_c^* \leq t)}{\Pr(S_c^* + T_c \geq \tau - \tau_0 \mid T_c^* \geq t, S_c^* \leq t)} \times \frac{\Pr(T_c^* = t, S_c^* \leq t)}{\Pr(T_c^* \geq t, S_c^* \leq t)} = \tilde{r}_c^*(t) \frac{\Pr(S_c^* + T_c \geq \tau - \tau_0 \mid T_c^* = t, S_c^* \leq t)}{\Pr(S_c^* + T_c \geq \tau - \tau_0 \mid T_c^* \geq t, S_c^* \leq t)}
\]

Now, given the assumption mentioned at the beginning, that, conditional on \( S_c^* \leq t \), the survival of children does not depend on the survival of their mothers, the last term on the right-hand side becomes approximately

\[
\frac{\Pr(S_c^* + T_c \geq \tau - \tau_0 \mid S_c^* \leq t)}{\Pr(S_c^* + T_c \geq \tau - \tau_0 \mid S_c^* \leq t)}
\]

and may be omitted.

There is, however, a further difficulty resulting from retrospective surveys. The later the year \( \tau \) in which the survey is performed, the smaller is the number of children who might participate in the survey, and consequently also the risk set to be used for the estimation of the death rates \( \tilde{r}_{c,\tau}^* \) becomes smaller. This is shown in Figure 3.12 which is based on the data from the simulation model. Shown are the functions

\[
t \longrightarrow \mathcal{U}_t^c[\tau] := \{u \in \mathcal{U}_t^c \mid T_c^*(u) \geq t, S_c^*(u) \leq t\}
\]

as they result from four fictitious retrospective surveys performed in the years \( \tau = 2000, 2010, 2015, \) and 2020. The possible problem concerns estimation with left truncated data. Contrary to the standard Kaplan-Meier procedure with right censored data only, the risk set is very small at the beginning and may not allow reliable estimates of the death rates. Due to the cumulative nature of the calculation of survivor functions from these rates, any imprecisions introduced at the beginning will then propagate to values of the survivor function at later ages. To illustrate,
Inferences from the GLHS and SOEP Data

Figure 3.12.: Sizes of the risk sets $U_c^\tau$, depending on $\tau$, calculated from four retrospective surveys of the simulated data set in the years $\tau = 2000$, 2010, 2015, and 2020.

I use the simulated data set and perform a retrospective survey in the year $\tau = 2010$. I assume that all children who survive this year, that is about 20% of the 11407 children in $U_c$, participate in the survey and provide information about their mothers. Nevertheless, I can only begin to estimate a conditional survivor function at $a^+ = 25$ as shown in Figure 3.13.

3.8. Inferences from the GLHS and SOEP Data

I now use the methods discussed in the previous Sections to draw some inferences from the GLHS and SOEP data.

The German Life History Study (GLHS) is a long-term project conducted by the Max Planck Institute for Human Development (Berlin). The main data source of this project is a series of retrospective surveys in which members of selected birth cohorts were asked to provide detailed information about their life courses. Part of these data are accessible for the general scientific public through the Zentralarchiv für empirische
Sozialforschung (Köln). Only the publicly accessible data are used in the following.

a) Data from the first survey (LV I) were sampled during the years 1981 – 83 and included 2171 members of the birth cohorts 1929 – 31, 1939 – 41, and 1949 – 51.

b) Data from a second survey (LV II) were sampled in two parts, both relating to persons born in the years 1919 – 21; a first part was conducted in 1985 – 86 and included 407 persons (LV IIA), a second part was conducted in 1987 – 88 and included 1005 persons (LV IIT).

c) Data from a third survey (LV III) were sampled in 1989 and included 2008 members of the birth cohorts 1954 – 56 and 1959 – 61.

All surveys were conducted in the territory of the former FRG. For the present study I include all female respondents from the surveys LV I, LV IIT, and LV III (only cohort 1959 – 61) having a German citizenship.
The number of cases and their distribution across the five cohorts is shown in the following table:  

<table>
<thead>
<tr>
<th>Birth cohort</th>
<th>Birth years</th>
<th>Male</th>
<th>Female</th>
<th>Interview date</th>
</tr>
</thead>
<tbody>
<tr>
<td>C20</td>
<td>1919 – 21</td>
<td>373</td>
<td>632</td>
<td>1987 – 88</td>
</tr>
<tr>
<td>C30</td>
<td>1929 – 31</td>
<td>349</td>
<td>359</td>
<td>1981 – 83</td>
</tr>
<tr>
<td>C40</td>
<td>1939 – 41</td>
<td>375</td>
<td>355</td>
<td>1981 – 83</td>
</tr>
<tr>
<td>C50</td>
<td>1949 – 51</td>
<td>365</td>
<td>368</td>
<td>1981 – 83</td>
</tr>
<tr>
<td>C60</td>
<td>1959 – 61</td>
<td>512</td>
<td>489</td>
<td>1989</td>
</tr>
</tbody>
</table>

3.8.1. Description of the Data

Of the 2171 respondents interviewed in LV I, 2120 respondents were able to provide a valid birth year of their mother. Of these mothers, 732 died before the interview date, 1386 were still alive, and for two mothers I have no information. Complete information is therefore available for 2118 mothers. In 8 cases this information is inconsistent or implausible, for example, the birth year of the respondent is greater than the death year of the mother. In addition to inconsistent cases I also exclude cases with a life length greater than 105 years. For women I also require that the age at which the women gave birth to her child (the respondent) is not greater than 51 years. If these cases are excluded, there finally remain 2110 cases in which the birth year of the mother, whether she died before the interview date, and, if she died, also her death year are known. Similarly, I get valid information for 2044 fathers.

The second study, LV II was conducted in two parts: LV IIA with interviews during 1985 – 86, and LV IIB with interviews during 1987 – 88. In the same way as explained for LV I I get valid information about the lifetimes of $387 + 956 = 1343$ mothers and $382 + 943 = 1325$ fathers. The third study, LV III, provides valid information about 1954 mothers and 1911 fathers.

---

35 Of the 632 women of birth cohort C20 three did not give valid birth years for their children and will be excluded in further calculations.
Table 3.5: Information about lifetimes of mothers and fathers available in the GLHS and SOEP data sets.

<table>
<thead>
<tr>
<th></th>
<th>LV I</th>
<th>LV IIA</th>
<th>LV II</th>
<th>LV III</th>
<th>SOEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Respondents</td>
<td>2171</td>
<td>407</td>
<td>1005</td>
<td>2008</td>
<td>8021</td>
</tr>
</tbody>
</table>

Mothers
- valid birth year | 2120 | 390 | 962 | 1954 | 7819 |
- still alive       | 1386 | 24  | 43   | 1766 | 4872 |
- known death year  | 732  | 366  | 919  | 188  | 2911 |
- no information    | 2    | 0    | 0    | 0    | 36   |
- complete information | 2118 | 390 | 962 | 1954 | 7783 |
- dismissed         | 8    | 3    | 6    | 0    | 37   |
- remaining cases   | 2110 | 387  | 956  | 1954 | 7746 |
- still alive       | 1385 | 24   | 43   | 1766 | 4854 |
- died              | 725  | 363  | 913  | 188  | 2892 |

Fathers
- valid birth year | 2062 | 386 | 955 | 1916 | 7699 |
- still alive       | 909  | 1   | 8   | 1460 | 3586 |
- known death year  | 1150 | 384 | 945 | 451  | 4053 |
- no information    | 3    | 1   | 2   | 5    | 60   |
- complete information | 2059 | 385 | 953 | 1911 | 7639 |
- dismissed         | 15   | 3   | 10  | 0    | 25   |
- remaining cases   | 2044 | 382 | 943 | 1911 | 7614 |
- still alive       | 909  | 1   | 8   | 1460 | 3577 |
- died              | 1135 | 381 | 935 | 451  | 4037 |

Comparable information is available from the third wave of the SOEP conducted in 1986. All members of subsample A of the SOEP were asked to provide information about birth years of their parents, whether parents died before the interview date and, if they died, about their death years. In order to get data comparable with the GLHS, I selected only persons with a German citizenship. As shown in Table 3.5, there are 8021 persons providing valid information about 7746 mothers and 7614 fathers. Taking the GLHS and SOEP data together, I finally have valid information about 13153 mothers and 12894 fathers.

I prepared two data files for further analysis, one for mothers and the
other one for fathers. Both files contain values of four variables:

\[
\begin{align*}
B^f & := \text{birth year of the mother} \\
P^f & := \text{birth year of the child (respondent)} \\
E^f & := 1 \text{ if mother died before the interview date, 0 otherwise} \\
D^f & := \text{mother’s death year, or the year of the interview, depending on the value of } E^f
\end{align*}
\]

Variables in the data file for fathers are defined accordingly and will be denoted by \(B^m, P^m, E^m, \text{ and } D^m\).

3.8.2. Survivor Functions of Parents

I now apply the method discussed in the previous Section to the data introduced in Section 3.8.1. Since mortality conditions have substantially changed during the last 100 years, I consider birth cohorts as defined in Table 3.6. To develop the argument I consider variables \(\hat{T}_c^f\) and \(\hat{T}_c^m\) representing the life length of women and men who belong to a birth cohort indexed by \(c\). Derivable from the variables introduced at the end of Section 3.8.1, available data are given by variables

\[
\begin{align*}
C_c^f & := P_c^f - B_c^f \quad \text{and} \quad C_c^m := P_c^m - B_c^m
\end{align*}
\]

which record the ages at which persons belonging to birth cohort \(c\) became mothers or fathers, and variables

\[
\begin{align*}
T_c^f & := D_c^f - B_c^f \quad \text{and} \quad T_c^m := D_c^m - B_c^m
\end{align*}
\]

which record the knowledge about the life length. If \(E_c(u) = 1\), \(T_c^f(u) = \hat{T}_c^f(u)\) is the known life length of \(u\); otherwise, the information is censored and I only know that \(\hat{T}_c^f(u) \geq T_c^f(u)\). For variables pertaining to men the interpretation is analogous.

---

\(^{36}\) Compared with the figures in Table 3.5 the total number of cases is slightly smaller because persons born before 1870 or after 1939 have been omitted.
3. A Case Study: Parent’s Length of Life

Table 3.6: Definition of birth cohorts used in the estimation of survivor functions.

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Birth years</th>
<th>Mothers</th>
<th></th>
<th></th>
<th></th>
<th>Fathers</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>died</td>
<td>alive</td>
<td>total</td>
<td></td>
<td>died</td>
<td>alive</td>
<td>total</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>1870 – 1879</td>
<td>271</td>
<td>0</td>
<td>271</td>
<td>528</td>
<td>0</td>
<td>528</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C2</td>
<td>1880 – 1889</td>
<td>1064</td>
<td>10</td>
<td>1074</td>
<td>1393</td>
<td>12</td>
<td>1405</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C3</td>
<td>1890 – 1899</td>
<td>1698</td>
<td>170</td>
<td>1868</td>
<td>1591</td>
<td>101</td>
<td>1692</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C4</td>
<td>1900 – 1909</td>
<td>1123</td>
<td>954</td>
<td>2077</td>
<td>1685</td>
<td>600</td>
<td>2285</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C5</td>
<td>1910 – 1919</td>
<td>438</td>
<td>1456</td>
<td>1894</td>
<td>907</td>
<td>1011</td>
<td>1918</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C6</td>
<td>1920 – 1929</td>
<td>272</td>
<td>2467</td>
<td>2739</td>
<td>464</td>
<td>2035</td>
<td>2499</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C7</td>
<td>1930 – 1939</td>
<td>123</td>
<td>2219</td>
<td>2342</td>
<td>196</td>
<td>1773</td>
<td>1969</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccc}
\hline
\text{t} & \text{(a)} & \text{(b)} & \text{(c)} & \text{(d)} \\
\hline
15 & 1 & 0 & 0 & 1.000 \\
16 & 2 & 0 & 0 & 1.000 \\
17 & 3 & 0 & 0 & 1.000 \\
18 & 18 & 0 & 0 & 1.000 \\
19 & 45 & 0 & 0 & 1.000 \\
20 & 100 & 1 & 0 & 1.000 \\
21 & 183 & 1 & 0 & 0.990 \\
22 & 266 & 2 & 0 & 0.985 \\
23 & 347 & 0 & 0 & 0.977 \\
24 & 435 & 2 & 0 & 0.977 \\
25 & 556 & 1 & 0 & 0.973 \\
26 & 690 & 3 & 0 & 0.971 \\
27 & 781 & 4 & 0 & 0.967 \\
28 & 928 & 1 & 0 & 0.962 \\
29 & 1075 & 3 & 0 & 0.961 \\
30 & 1217 & 3 & 0 & 0.958 \\
31 & 1327 & 2 & 0 & 0.956 \\
32 & 1427 & 5 & 0 & 0.954 \\
33 & 1512 & 9 & 0 & 0.951 \\
34 & 1597 & 5 & 0 & 0.945 \\
35 & 1677 & 6 & 0 & 0.942 \\
36 & 1740 & 5 & 0 & 0.939 \\
37 & 1800 & 5 & 0 & 0.936 \\
38 & 1864 & 9 & 0 & 0.934 \\
39 & 1903 & 13 & 0 & 0.929 \\
40 & 1929 & 12 & 0 & 0.923 \\
41 & 1945 & 2 & 0 & 0.917 \\
42 & 1952 & 5 & 0 & 0.916 \\
\hline
\end{array}
\]

\[
\begin{array}{cccc}
\text{t} & \text{(a)} & \text{(b)} & \text{(c)} & \text{(d)} \\
\hline
52 & 1901 & 4 & 0 & 0.878 \\
53 & 1897 & 15 & 0 & 0.876 \\
54 & 1882 & 13 & 0 & 0.869 \\
55 & 1869 & 21 & 0 & 0.863 \\
56 & 1848 & 7 & 0 & 0.854 \\
57 & 1841 & 17 & 0 & 0.850 \\
58 & 1824 & 14 & 0 & 0.843 \\
59 & 1810 & 18 & 0 & 0.836 \\
60 & 1792 & 20 & 0 & 0.828 \\
61 & 1772 & 19 & 0 & 0.818 \\
62 & 1753 & 15 & 0 & 0.810 \\
63 & 1738 & 23 & 0 & 0.803 \\
64 & 1715 & 18 & 0 & 0.792 \\
65 & 1697 & 33 & 0 & 0.784 \\
66 & 1664 & 23 & 0 & 0.769 \\
67 & 1641 & 36 & 0 & 0.758 \\
68 & 1605 & 32 & 0 & 0.741 \\
69 & 1573 & 24 & 0 & 0.727 \\
70 & 1549 & 41 & 0 & 0.715 \\
71 & 1508 & 39 & 0 & 0.697 \\
72 & 1469 & 54 & 32 & 0.679 \\
73 & 1383 & 53 & 53 & 0.654 \\
74 & 1277 & 56 & 57 & 0.629 \\
75 & 1164 & 53 & 43 & 0.601 \\
76 & 1068 & 55 & 30 & 0.574 \\
77 & 983 & 42 & 125 & 0.544 \\
78 & 816 & 57 & 96 & 0.521 \\
79 & 663 & 42 & 114 & 0.484 \\
\hline
\end{array}
\]
Table 3.7.: Mortality data for mothers belonging to birth cohort C4 in the merged GLHS and SOEP data set. (a) Size of risk set at age \( t \). (b) Number of deaths at age \( t \). (c) Number of censored cases at age \( t \). (d) Values of the conditional survivor function at age \( t \).

<table>
<thead>
<tr>
<th>( t )</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( d )</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>1957</td>
<td>7</td>
<td>0</td>
<td>0.914</td>
</tr>
<tr>
<td>44</td>
<td>1957</td>
<td>5</td>
<td>0</td>
<td>0.910</td>
</tr>
<tr>
<td>45</td>
<td>1956</td>
<td>13</td>
<td>0</td>
<td>0.908</td>
</tr>
<tr>
<td>46</td>
<td>1947</td>
<td>7</td>
<td>0</td>
<td>0.902</td>
</tr>
<tr>
<td>47</td>
<td>1943</td>
<td>10</td>
<td>0</td>
<td>0.899</td>
</tr>
<tr>
<td>48</td>
<td>1935</td>
<td>8</td>
<td>0</td>
<td>0.894</td>
</tr>
<tr>
<td>49</td>
<td>1928</td>
<td>8</td>
<td>0</td>
<td>0.891</td>
</tr>
<tr>
<td>50</td>
<td>1920</td>
<td>7</td>
<td>0</td>
<td>0.887</td>
</tr>
<tr>
<td>51</td>
<td>1913</td>
<td>12</td>
<td>0</td>
<td>0.884</td>
</tr>
<tr>
<td>( t )</td>
<td>( a )</td>
<td>( b )</td>
<td>( c )</td>
<td>( d )</td>
</tr>
<tr>
<td>80</td>
<td>507</td>
<td>33</td>
<td>85</td>
<td>0.454</td>
</tr>
<tr>
<td>81</td>
<td>389</td>
<td>20</td>
<td>78</td>
<td>0.424</td>
</tr>
<tr>
<td>82</td>
<td>291</td>
<td>12</td>
<td>70</td>
<td>0.402</td>
</tr>
<tr>
<td>83</td>
<td>209</td>
<td>21</td>
<td>48</td>
<td>0.386</td>
</tr>
<tr>
<td>84</td>
<td>140</td>
<td>12</td>
<td>34</td>
<td>0.347</td>
</tr>
<tr>
<td>85</td>
<td>94</td>
<td>5</td>
<td>41</td>
<td>0.317</td>
</tr>
<tr>
<td>86</td>
<td>48</td>
<td>0</td>
<td>39</td>
<td>0.300</td>
</tr>
<tr>
<td>87</td>
<td>9</td>
<td>0</td>
<td>5</td>
<td>0.300</td>
</tr>
<tr>
<td>88</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0.300</td>
</tr>
</tbody>
</table>

To illustrate the calculations I refer to women belonging to birth cohort C4. The data are shown in Table 3.7. The column labelled (a) shows the risk sets. As discussed in the previous Section, the risk set at age \( t \) contains all women who did not die before \( t \) and became a mother not later than \( t \).\(^{37}\) In this example, the youngest age for which I know of a child is 15; risk sets can therefore be calculated only for ages \( t \geq t^* = 15 \). The next column, labelled (b), shows the number of death events. Then follows column (d) providing the number of censored cases which are required to update the risk sets. As shown by the definition

\[
\mathcal{R}^*(t) := \{ u \mid T^f_c(u) \geq t, C^f_c(u) \leq t \}
\]

women belong to a risk set only until the maximal value of \( T^f_c \), that is, until a death event occurs or until the interview date (of their children).

The information in Table 3.7 suffices to calculate death rates. For example, \( r^*(20) = 1/100 \) and \( r^*(80) = 33/507 \). These rates can then be used to

\(^{37}\) Of course, from the data I do not know when women actually gave birth to a first child. Whether this has implications for the quality of the estimates will be discussed in a later Section.
estimate the survivor function

\[ G^*(t) = g_{t^*} \prod_{j=t^*}^{t-1} (1 - r^*(j)) \]

Of course, I do not know \( g_{t^*} \), that is, the proportion of women who survived age 14. So I can only estimate a conditional survivor function

\[ G(T_{cf}^f \mid T_{cf}^f \geq t^*) \approx \prod_{j=t^*}^{t-1} (1 - r^*(j)) \]

which is shown in the last column of Table 3.7 labelled (d).

Since this approach to estimate a conditional survivor function depends on a previous estimation of rates, one should also consider the question whether these rates can be reliably estimated. Formally, one can begin at age \( t^* \) which is 15 in the example. However, due to the small number of cases in the risk sets at ages under 20, one might question the reliability of these estimates. In fact, the estimation procedure implies estimated death rates having a value of zero during ages from 15 to 19. But given our knowledge about mortality and life tables from other sources, these estimates will clearly be wrong. Moreover, the reliability of estimates of death rates not only depends on the size of the risk sets but also on the number of death events that can be observed. With regard to the data in Table 3.7, it might be better to begin an interpretation of estimated death rates only at some later age, for example, at age 26 or even later.

Conditional survivor functions can be represented graphically in two possible ways: The function can be plotted beginning at some age \( t \) with arbitrary value \( g_t \); or one can try to find some estimate of \( g_t \) and then plot the conditional survivor function as part of a complete survivor function. In any case one needs to decide where to start the plotting. For the example I begin at age 26 and estimate \( g_{26} \) from the female survivor function of the German period life table for the period 1901–10 (see Table 3.3). Beginning at age 26, I therefore multiply all values of column (d) in Table 3.7 with the factor \( g_{26} = 0.71463/0.971 = 0.736 \). The result is shown in Figure 3.14. The dotted line represents the female survivor
function from the 1901–10 period life table; the solid line shows the adjusted conditional survivor function from Table 3.7. By definition, values are identical at age 26. The different development of both curves reflects the reduction of death rates that occurred during the period from about 1930 until the end of the century. So I might use the latest 1986–88 period life table for a further comparison. As can be estimated from Table 3.7, the death rate at age 80 is about 0.065. A corresponding estimate from the 1986–88 period life table is 0.066.\(^3\) One should note, however, that values of rates calculated from sample data for single years often show high fluctuations and it might be better, therefore, to use smoothed values based on larger age classes.

In the same way as has been discussed for women belonging to birth cohort C4 (1900–1909) one can estimate conditional survivor functions for all birth cohorts in Table 3.6. Results are shown in Figure 3.15. To allow for a comparison, all survivor functions are drawn conditional on \(t^* = 30\). The placement onto a historical time axis was done by using the centres of the birth cohort intervals. For example, the value of the

\(^3\) Calculated from the data in Table 3.3.
Figure 3.15: Conditional survivor functions, beginning at age 30, for men (solid lines) and women (dotted lines) belonging to specified birth cohorts.
3.8. Inferences from the GLHS and SOEP Data

conditional survivor function for birth cohort C1 at age 30 is shown in the year \(1875 + 30 = 1905\). The changing shapes of the survivor functions not only reflect a general tendency of decreasing death rates, both for men and women. Also clearly seen are period effects, especially the substantial increases of male death rates during the years of World War II. This seems not to be the case with regard to female death rates. An interpretation should consider, however, that the occurrence of death events might not be independent for mothers and their children, in particular during war time. The death events of mothers might therefore be substantially underrepresented in the data set.

3.8.3. Visualisation of Death Rates

In order to investigate period effects it is often preferable to directly plot the rates from which (conditional) survivor functions are derived. The only drawback is that rates calculated from small samples are often highly fluctuating. As an example I refer to death rates of men belonging to birth cohort C5 (1910 –1919). The solid line in Figure 3.16 shows the death rates as directly calculated from the data, that is, for each year of age, the number of deaths divided by the number of persons in the risk set. There obviously are big fluctuations. One should therefore apply some kind of smoothing procedure to provide a better view of the general shape of the rate function.

Many such smoothing procedures have been proposed in the literature. In the present context, smoothing will only serve to visualise rate functions. It might therefore suffice to simply use moving averages. Given a series of values \(r_t\), for \(t = t_1, \ldots, t_n\), each value is then substituted by a mean of neighbouring values. If the number of neighbours is denoted by \(2k\), the smoothed values are calculated as

\[
r_t^{(k)} = \frac{1}{2k + 1} \sum_{j=t-k}^{t+k} r_j
\]

At both ends of the series only the actually available values are taken into
account. The complete formula may then be written as follows:

\[ r_t^{(k)} := \frac{1}{\min\{t_n, t + k\} - \max\{t_1, t - k\} + 1} \sum_{j=\max\{t_1, t-k\}}^{\min\{t_n, t+k\}} r_j \]

where \( t_1 \) and \( t_n \) refer, respectively, to the first and last element of the series.

Choosing \( k = 2 \), this procedure was used to calculate values for the dotted line in Figure 3.16. It is seen how the smoothing removes the fluctuations but preserves the global shape of the rate function.

I now compare the death rates of men belonging to birth cohorts C1, \ldots, C6. The rate functions are shown in Figure 3.17 and placed onto a historical time axis. To support visibility, the rate functions are smoothed with the procedure just described (again, \( k = 2 \)). Compared with the survivor functions shown in Figure 3.15, the rate functions provide a much better view of the impact of World War II.
Figure 3.17: Smoothed death rates of men belonging to the indicated birth cohorts. (Moving averages with $k = 2$.)
3.9. Conclusion

The discussion in Section 3.7 and 3.8 was based on a rather simplistic simulation. The estimators derived from relative frequency analogues of dynamic locally dependence conditions discussed in Section 3.6 could be applied successfully within this context. The dynamic independence conditions is necessarily formulated within the context of probability models that provide the connection between the ‘full data’ and ‘left truncated data’. In fact, the dynamic independence condition was explicitly employed in the probabilistic description of the simulation algorithm used to gauge the performance of suggested estimators. It is therefore not surprising to find good agreement between the period survival rates and estimates based on the simulated birth and death events.

The fact that survival rates and their empirical analogues can be defined without recourse to the number of children born to a woman (see Section 3.7.3) illustrates the difference from a pure sampling based argument as developed in Section 3.3. While the latter does not depend on any form of probabilistic independence assumption and the correctness of its computations can be checked against empirical findings, Section 3.4 has shown that this will be impossible for the dynamic local independence condition incorporated in the construction of weights and estimators. An argument for the correctness of the dynamic independence conditions must necessarily refer to a probability model of the dynamic population process. It presupposes the use of a probability model. Therefore, the basic ingredients of any dynamic probability model are a necessary part of a constitutive decision in Matheron’s sense, the decision to model population dynamics in terms of dynamic probabilistic models. The appropriateness of the dynamic local independence assumption can only be judged within such a framework.

But the model used in the simulation is a rather simplistic one and misses one important ingredient inherent in any study of real populations: The simulation created sets $\mathcal{U}$ of potential mothers "born" at a fixed point in time, without reference to their parent generation. The arguments in Section 3.7 then showed how the local independence condition allows the development of reasonable estimators for such constructed cohorts.
In the applications, the reasoning was applied to cohorts artificially constructed from retrospective information. But real populations grow or shrink depending on both birth and death rates of previous generations. Even when ignoring possible dependencies between life length of parents and their children, the procedure can only be justified within a probability model that posits a simple random process creating parents. This has been proposed by Brillinger (1986) who tried to justify a Poisson process as a model of creating parents. A crucial consequence of such a proposal was pointed out by Hoem in the discussion of Brillinger’s paper.

Unfortunately, the assumption of Poisson births disregards essential aspects of the internal dynamics of real-life populations. Results based on the Poisson assumption must have little relevance for the analysis of statistics which involve data for several generations ... In normal populations, children are borne by population members ... Since real populations are subject to the internal dependencies which are made manifest by population waves, the assumption of Poisson births is untenable in multigenerational analyses.

The problem surfaces also in our problem of estimating parent’s length of life: While the number and and ages at birth of a mother are of no direct importance to the question posed, they necessarily appear in any analysis of the probability model of parent’s life length. It is this direct dependence of current populations on the timing and number of births of previous generations that makes the Poisson models untenable. At the same time, number of births and their timing are closely related to life length. Thus, they ought to be part of the probabilistic models. But their mere presence changes the form of reasonable estimators.³⁹

But reasonable probabilistic models are very difficult to construct since it turns out that Markov structures on a calendar time basis are inappropriate and intergenerational dependencies invalidate transition models

³⁹ In their discussion of the construction of likelihood functions, Bayarri et al. (1987) and Bayarri and DeGroot (1992) use selection models of the present type to demonstrate that various suggestions as to what ought to figure in a likelihood function are ill defined.
based on individuals alone.⁴⁰

If and in so far such an elaborated probabilistic model of population dynamics can be combined with dynamic local independence conditions, such models do satisfy the requirements for scientifically valuable stochastic models in the sense of Neyman. The repeated application of local dynamic independence conditions does lead to manageable estimators and manageable estimation strategies. But on its own the independence condition does not provide arguments for its support. Criticism of this condition must be based on a larger probability model of population dynamics as proposed by Jagers (1989) or similar ones.

---

⁴⁰ Jagers (1989) proposed a clever construction that allows to keep a Markov assumption for lines of descent. Life length, number of children, and other properties of interest can then be modelled as attributes of elements of these lines of descent. He himself describes the purpose of such a construction as "But one of the purposes of mathematical population dynamics is to serve as a bridge between individual properties and properties of the population as a whole. … This presupposes a theory of population evolution built upon a description of individual life. But individuals vary …, i.e. we need a stochastic description." Jagers (1989). My interest here is of course a different one.
4.1. Summary

Missing or incomplete data are a pertinent to any social survey. Interviewees do not answer some questions or refuse to participate at all. Nowadays, response rates in surveys rarely reach 40%. Can the answers of 40% of the selected individuals be treated as if one had answers from all of them? Probabilistic selection models can be used to speculate on the connexion between the attained information and the possible answers of the non-responders. These models can be extended to cover incomplete data such as grouped or censored data as well as processes of self-selection and the evaluation of social programmes. A brief review of some of these applications is given. Next, it is shown that the assumptions used in probabilistic selection models necessarily refer to concepts transcendentning empirically accessible social facts. However, the amount and impact of these assumptions can be reduced if possibly available additional incomplete information is fully used. Therefore, probabilistic selection models are formulated in a form that allows for grouped, censored and general coarsened data. Finally, several proposals for the sensitivity analysis of probabilistic selection models are presented and it is argued that approaches restricted to too small neighbourhoods of a given model can be quite misleading. Throughout, responses on
4. Probabilistische Selektionsmodelle

income questions obtained in the German ALLBUS 1996 are used as an illustration.

4.2. Zusammenfassung

4.3. Einleitung


4. Probabilistische Selektionsmodelle

In der Literatur zur Stichprobentheorie sind probabilistische Selektionsmodelle nur sehr zurückhaltend diskutiert worden. So schrieb Tore Dalenius, damals Präsident der International Association of Survey Statisticians:

I take a dim view of the usefulness of these endeavors on two grounds. (1) First, it appears utterly unrealistic to postulate ‘response probabilities’ which are independent of the varying circumstances under which an effort is made to elicit a response. …(2) …it seems unavoidable to introduce assumptions of unknown validity about probabilities. In summary, I am inclined to reject approaches to the non-response problem which involve ‘response probabilities’ (Dalenius in Madow und Olkin 1983, Band 3: 412).

Und Mohler et al. verweisen auf die großen Schwankungen in den Ausschöpfungsraten von Stichproben, „die sich nicht mehr auf statistischen Zufall zurückführen lassen“ (Mohler et al. 2003: 11). Daher hat man sich in dieser Tradition darauf beschränkt, Bereiche möglicher Schlussfolgerungen auszuweisen, die sich allein auf die Angaben der Befragten und Konsistenzannahmen stützen.

und Ausschlussrestriktionen früh kritisiert wurde (z.B. in Wainer 1986, 1989), haben sie sich in einigen Bereichen der Sozialwissenschaften als dominante analytische Methode durchgesetzt.


Wenn möglichst alle stochastischen Annahmen eines Selektionsmodells systematisch variiert werden, so zeigt sich, dass der Bereich möglicher Schlussfolgerungen probabilistischer Modelle sehr groß werden kann. Zudem deckt er sich häufig mit den Bereichen, die im Rahmen der klassischen Stichprobentheorie entwickelt wurden. Globale Sensitivitätsanalysen probabilistischer Selektionsmodelle führen daher zu Abschät-
zungen, die mit denen der Stichprobentheorie vergleichbar sind. Die Formulierung verschiedener Annahmen in probabilistischen Selektionsmodellen ermöglicht eine Diskussion über den Zusammenhang zwischen erhaltenen und nicht erhaltenen Angaben, die stichprobentheoretische Überlegungen ergänzen können.

4.4. Selektionsmodelle in den Sozialwissenschaften


The analysis here should also call attention to the other potential problem with any survey: nonresponse bias. … Consumers of public opinion surveys, as well as practitioners, must be reminded of this potential problem in order to avoid a future disaster like the Literary Digest poll of 1936.

Antwort jeder Person auf die Einkommensfrage, die über einen Probit-Link linear von allen Variablen (bis auf den Familienstand) abhängt. Außerdem nimmt sie an, das logarithmierte Einkommen aller Personen habe eine lineare, homoskedastische Regression auf alle Variablen (bis auf Alter, Demenzindikator und Interviewform) und folge einer bedingten Normalverteilung. Unter diesen Annahmen kann aus der bedingten Verteilung der beobachteten Einkommen auf die bedingte Verteilung der Einkommen aller Personen geschlossen werden. Engelhardt vergleicht die entsprechenden Ergebnisse mit einer linearen Regression, die nur die vollständigen Angaben berücksichtigt. Sie schließt,


206
4.4. Selektionsmodelle in den Sozialwissenschaften

wurden. Es folgt, dass die Person mindestens das Alter bei diesem Ereignis erreicht hat. Ist $Y$ das Lebensalter einer Person, dann ist $Y^*$ entweder der Wert von $Y$, falls ein Todesdatum registriert wurde, oder aber das Intervall $(T_{\text{max}}, \infty)$, wobei $T_{\text{max}}$ der Zeitpunkt des letzten Ereignisses ist, das registriert wurde. Wenn angenommen wird, diese Angaben sagten nichts anderes als das jemand älter als $T_{\text{max}}$ geworden ist, dann können Verfahren der Ereignisanalyse eingesetzt werden. Dagegen kann eingewandt werden, der Grund des Fehlens eines Todesdatums sei i.d.R. die Abwanderung der Personen. Dann wäre das Datum der Abwanderung eine untere Grenze für das erreichte Lebensalter und $T_{\text{max}}$ wäre immer kleiner als dieses Zensurereignis. Im Ergebnis würden Verfahren der Ereignisanalyse die Risikomengen unterschätzen und damit Sterberaten überschätzen. Selbst wenn es keine Angaben über Abwanderungen gibt, kann mithilfe eines probabilistischen Modells für die Zwischenereignisse sowie die Abwanderungszeiten über die Verteilung der Lebensdauern nachgedacht werden (Gill 1997; Jonker 2003).

In den Sozialwissenschaften sind Selektionsmodelle eher selten im Zusammenhang mit unvollständigen Angaben in Umfragen oder Registerdaten behandelt worden. Stattdessen dominieren Anwendungen, die sich auf Größen beziehen, die ihre Bedeutung nur im Rahmen eines vorab definierten Modells gewinnen. So untersuchen Diekmann und Wyder (2002) Reputationseffekte bei Internetauktionen, wobei sie auch die erzielten Preise der Auktionen heranziehen. Sie versuchen dabei auch diejenigen Auktionen einzubeziehen, für die gar kein Gebot abgegeben wurde und für die daher auch kein erzielter Preis existiert. Sie argumentieren:


Ein „Stichprobenauswahlfehler“ könnte aber nur vorliegen, wenn auch den Auktionen ohne Gebote ein „Preis“ zukäm. Diekmann und Wyder

In den Sozialwissenschaften werden Selektionsmodelle auch zur Evaluation sozialpolitischer Maßnahmen und zur Kausalanalyse herangezogen. Die Idee besteht darin, zunächst eine Variable \( Y \) festzulegen, die den Erfolg einer Maßnahme oder die Wirkung einer Ursache darstellen soll. Unterstellt wird dann die gleichzeitige Existenz der Variablen \( (Y_0, Y_1) \), wobei \( Y_0 \) den Wert von \( Y \) annimmt, der sich ergeben hätte, wenn jemand nicht an der Maßnahme teilgenommen hätte, \( Y_1 \) den bei Teilnahme. Der Erfolg der Maßnahme ließe sich dann etwa durch \( Y_1 - Y_0 \) ausdrücken.
Natürlich kann eine Person nur entweder an einer Maßnahme teilnehmen oder nicht teilnehmen. \( Y_0 \) und \( Y_1 \) können also nicht gleichzeitig beobachtet werden. Ist \( \mathcal{Y} \) der Wertebereich der Variablen \( Y \) und \( \mathcal{Y}^* \) der Wertebereich der beobachteten Variablen \( Y^* \), dann ist

\[
\mathcal{Y}^* = (\mathcal{Y} \times \{\mathcal{Y}\}) \cup (\{\mathcal{Y}\} \times \mathcal{Y})
\]

Entweder wird der Wert von \( Y_0 \) beobachtet, nicht aber der von \( Y_1 \), für den nur \( y_1 \in \mathcal{Y} \) bekannt ist. Die Beobachtung ist also \( (y_0, \mathcal{Y}) \). Oder der Wert von \( Y_1 \) wird beobachtet, nicht aber der von \( Y_0 \). Dann kann die Beobachtung durch \( (\mathcal{Y}, y_1) \) angegeben werden. In dieser Formulierung entspricht das Evaluationsproblem einem Problem unvollständiger Angaben. Wird ein passendes probabilistisches Selektionsmodell unterstellt, dann ist die gemeinsame Verteilung von \( (Y_0, Y_1) \) identifiziert. Andersherum ist aber klar, dass \( (Y_0, Y_1) \) außerhalb dieses Selektionsmodells gar nicht definiert ist. Die Abhängigkeit von willkürlich gesetzten Modellannahmen und die kontrafaktische Formulierung des Evaluationsproblems sind oft kritisiert worden (z.B. Dawid 2000). Die kontrafaktische Formulierung des Problems ist fragwürdig, weil sie einen wohldefinierten Wert für das Ergebnis von Ereignissen voraussetzt, die gar nicht stattgefunden haben. Und im Unterschied zu fehlenden Angaben in Umfragen oder Registerdaten kann die Abhängigkeit der Ergebnisse von den Annahmen des Selektionsmodells nie empirisch ergänzt oder kritisiert werden. Fehlen Angaben in Kirchenregistern, so können Angaben aus Lehns- und Pachtregistern, Handwerksrollen, Sippenbüchern und Gerichtsakten herangezogen werden. Interessiert die Verteilung von Einkommen, dann können Personen, die einmal die Auskunft verweigert haben, nochmals befragt werden. Zudem geben Steuerstatistiken, Sozialversicherungsmeldungen und Lohnstatistiken weitere Auskunft. Aber was in parallelen Welten geschehen würde, in denen alles bis auf die Teilnahme an bestimmten Maßnahmen gleich wäre, entzieht sich jedem Versuch empirischer Überprüfung. Die Betonung der formalen Äquivalenz zwischen beiden Situationen verwischt oft die inhaltlichen Unterschiede. Rubin führt zwei weitere Unterscheidungen an:

The formal equivalence of these problems ... is highly useful conceptually ..., but it is, I believe, not helpful to muddle the distinction when trying to generate sound, practical
statistical advice. The reasons for this conclusion are that (a) the estimands (the things we want to estimate) are fundamentally different for these situations, and (b) the processes that create the missing data are typically very different, both by investigators’ design and by nature’s devices. (Rubin in Wainer 1992: 183f).

Rubin’s Point (b) refers to the status of incomplete data in both cases. If someone refuses to provide information in an interview, it lies in the investigator’s discretion whether they find the interests and ideas of the researcher relevant. Conversely, data in the counterfactual model of evaluation as well as considerations of “willingness to pay” are only incomplete on the basis of the model assumptions of the researcher. Achieved prices in auctions as well as results of employment measures are at least in principle observable. “Willingness to pay” in the context of auctions or the success of a measure (defined as \( Y_1 - Y_0 \)) can only be observed. Even when this distinction is gradual, it must be clearly specified which statements will be made. Rubin’s Point (a) will be discussed in the following for incomplete statements in surveys.

### 4.5. Beispiel: Einkommensangaben im ALLBUS

The household net income is of central importance in many areas of social research, such as poverty research and household theory. Nonetheless, even the average household net income is empirically little known. The Statistisches Jahrbuch 2001 states that the basis of the income- and consumption survey (EVS) 1998 had a monthly average of 5115 DM in West Germany, 5346 DM in West Germany, 4059 DM in East Germany (Statistisches Bundesamt 2001: 566ff). Conversely, the data report 1999 based on the SOEP for the year 1996 an average household net income of 1978 DM (2061 DM West, 1644 DM East) (Statistisches Bundesamt 2000: 584). The two data sets (EVS and SOEP) differ significantly:

| s₁ | Bruttostichprobe | | s₁ | n₁ = 6491 |
| s₂ | Nettostichprobe | | s₂ | n₂ = 3518 |


4. Probabilistische Selektionsmodelle

| $s_3$ | Haushaltseinkommen angegeben | $|s_3| = n_3 = 2678$ |
| $s_4$ | genaue Angaben | $|s_4| = n_4 = 1772$ |
| $s'_3$ | Einkommen in [1,9999] | $|s'_3| = n'_3 = 2633$ |
| $s'_4$ | genaue Angaben in [1,9999] | $|s'_4| = n'_4 = 1748$ |
| $s_5$ | keine Angabe Haushaltseinkommen, aber persönliches Einkommen angegeben | $|s_5| = n_5 = 293$ |
| $s_6$ | keine Angabe Haushaltseinkommen, aber genaues persönliches Einkommen | $|s_6| = n_6 = 213$ |


tatsächlichen verfügbaren Haushaltsein kommen betrachtet werden. Die Rundung in den Einkommensangaben führt zu einer weiteren Unsicher - heit bei der Analyse der Daten, die im Folgenden aber nicht systematisch untersucht wird.

Folgt man dem üblichen Verfahren und ignoriert fehlende und gruppierte Angaben, so ergibt sich ein (ungewichtetes) mittleres Haushaltsnetto - einkommen von 3676 DM (3909 DM West, 3171 DM Ost). Beschränkt man sich bei der Berechnung des mittleren Haushaltseinkommens auf Einkommen unter 10000 DM und benutzt nur die genauen Angaben, also die Teilstichprobe $s'_4$, so ergibt sich ein Mittelwert von 3560.4

Wird das Stichprobendesign ignoriert und eine einfache Zufallsauswahl unterstellt, so kann ein 95%-Konfidenzintervall konstruiert werden: (3480, 3640). Dabei werden die unterschiedlichen Auswahlsätze für Ost- und Westdeutschland unterschlagen, obwohl sich die Mittelwerte in Ostdeutschland (3158 DM) und Westdeutschland (3749 DM) deutlich unterscheiden und ostdeutsche Personen deutlich überrepräsentiert sind. Das geht aber gar nicht anders. Denn weder können die Gewichte für die Teilstichprobe $s_2$ verwandt werden: dort beträgt der Anteil an Personen in Ostdeutschland 31,7%, während er in der Teilstichprobe $s'_4$ 32,0% beträgt.5 Noch können Gewichte benutzt werden, die sich aus dem Anteil in der Teilstichprobe $s'_4$ ergeben, denn diese würden sich

---


5 In der Teilstichprobe $s'_4$ ergibt sich sogar ein Anteil von 34,1%, ein „signikanter“ Unterschied zu 31,7%.
von Stichprobe zu Stichprobe ändern und erlaubten keine statistischen Aussagen. Man erhält also einen Mittelwert, der wegen der sehr unterschiedlichen Auswahlssätze in Ost- und Westdeutschland wohl nichts über das durchschnittliche Haushaltseinkommen in Deutschland sagt, und ein Konfidenzintervall, dessen Überdeckungseigenschaften schlicht unbekannt sind.⁶


---


Konfidenzintervall des letzten Absatzes. Um aber die Auswahlwahrscheinlichkeiten des ursprünglichen Stichprobendesigns verwenden zu können, muss zumindest die Teilstichprobe $s_2$ betrachtet werden, also zusätzlich die 885 Befragten, die gar keine Angaben machten.\footnote{Ich rechne die $n_3 - n_3' = 45$ Angaben außerhalb von [1, 9999] dazu.} Dann ergibt sich ein ungewichtetes Intervall der möglichen Mittelwerte von [2687, 5348]. Wird gar die Bruttostichprobe $s_1$ betrachtet, ergibt sich das Intervall [1457, 7479]. Für diese Intervalle könnten nun „korrekt“ gewichtete Versionen und Konfidenzintervalle ausgerechnet werden. Nur sind die Intervalle selbst schon viel zu groß, um von praktischem Interesse zu sein.

4.6. Stichproben und probabilistische Auswahlmodelle


4. Probabilistische Selektionsmodelle

Die Metaphorik des „als ob durch einen Zufallsprozess zustandegekommen“ erlaubt aber relativ kompakte Beschreibungen von empirischen Verteilungen durch die Parameter $\theta$ sowie einen Anschluss an die Stichprobentheorie, denn die realisierten Werte der Zufallsvariablen werden für die Stichprobenziehung als fix angenommen. Der Superpopulationsansatz hält also auf der Ebene der Stichprobenziehung an der Idee der Angaben der Befragten als fixem Datum fest. Allerdings wird das ursprüngliche Problem, Aussagen über die Verteilung eines Merkmals in der Gesamtheit $\mathcal{U}$ zu gewinnen, durch ein anderes ersetzt: Aussagen über $\theta$ zu gewinnen. Diese Parameter sind nur durch die Beziehung auf die unterstellte Modellklasse $\{F_\theta \mid \theta \in \Theta\}$ definiert. Ihnen entspricht kein Wert, der sich allein aus der Beobachtung der Werte der Variablen in der Gesamtheit $\mathcal{U}$ gewinnen ließe. Somit wird der realistische Ansatz der klassischen Stichprobentheorie unterlaufen.

Der Superpopulationsansatz geht aber noch nicht weit genug. Wenn man sich für das Antwortverhalten von Befragten interessiert, so müsste im Superpopulationsansatz angenommen werden, dieses Verhalten sei bereits vor jeder Befragung festgelegt, und zwar ganz unabhängig davon, ob jemand tatsächlich befragt wurde oder nicht. Ob also jemand auf die Frage nach dem Haushaltsinkommen gar nicht antwortet oder nur in gruppiertener Form, wäre vor jeder Befragung schon entschieden. Denn der Superpopulationsansatz unterstellt fixe Werte (Realisationen des Zufallsprozesses) in der Gesamtheit $\mathcal{U}$ zum Zeitpunkt der Stichprobenziehung. Im nächsten Schritt werden wie in der klassischen Theorie Stichproben aus diesen fixen Werten gezogen. Stichprobenfunktionen wie Mittelwerte hängen auf der Stufe der Stichprobenziehung allein davon ab, wer aus der Gesamtheit $\mathcal{U}$ in die Stichprobe gelangt. Dies ermöglicht den Anschluss an Ergebnisse der klassischen Theorie, hat aber zur Folge, dass das Antwortverhalten aller Mitglieder der Gesamtheit vor jeder Stichprobenziehung festgelegt sein muss.

Soll nicht nur die Tatsache unvollständiger Daten konstatiert, sondern auch ihr Zustandekommen reflektierbar gemacht werden, dann wird es notwendig, auch Variablen wie Interviewform, Merkmale der Interviewer und vieles mehr zu betrachten. Die Annahme, all diese Variablen seien vor jeder Stichprobenziehung für alle Personen der Gesamtheit festgelegt,
ist nicht nur fatalistisch und völlig unrealistisch, sondern würde auch die Spezifikation eines Stichprobendesigns wegen der notwendigen Details praktisch unmöglich machen.

Soll an probabilistischen Auswahlmodellen festgehalten werden, dann muss schließlich ganz auf Elemente der Stichprobentheorie verzichtet werden. Sowohl die interessierenden Größen wie das Haushaltseinkommen, die Stichprobenziehung und das Antwortverhalten der Befragten werden in einem einzigen probabilistischen Modell beschrieben. Ist \((\Omega, B, \lambda)\) ein hinreichend großer Wahrscheinlichkeitsraum, mit dem alle diese Variablen beschrieben werden können, dann lässt sich etwa das Haushaltseinkommen als Funktion von \(u \in U\) und \(\omega \in \Omega\) auffassen:

\[
Y: U \times \Omega \longrightarrow \{1, 2, 3, \ldots\} =: \mathcal{Y}
\]

\(Y(u, \omega)\) ist also das Haushaltseinkommen, das einer Person \(u \in U\) bei Realisierung von \(\omega \in \Omega\) zukommt. Eine Person \(u\) hat in Abhängigkeit von \(\omega\) verschiedene Einkommen. Aber es gibt ein \(\omega_0 \in \Omega\), für das \((Y(u, \omega_0), u \in U)\) den Haushaltseinkommen \((y(u), u \in U)\) in der BRD entspricht.

Der Zusammenhang mit Aussagen über Durchschnitte der \(Y(., \omega)\) über alle \(u \in U\) wird hergestellt, indem allen \(u\) gleiche Wahrscheinlichkeitsverteilungen zugeschrieben werden und die Unabhängigkeit von \(Y(u, .)\) und \(Y(u', .)\) für verschiedene \(u, u'\) angenommen wird.

Tatsächlich in der Stichprobe beobachtet wird aber nur eine mengenwerte Variable mit dem Merkmalsraum

\[
\mathcal{Y}^* := \{\{y\} | y \in \mathcal{Y}\} \cup \{[1, 400), [400, 800), \ldots, [15000, \infty)\} \cup \{\mathcal{Y}\}
\]

der genaue oder gruppierte Angaben bzw. gar keine Angabe darstellt. Es sei nun

\[
S: \Omega \longrightarrow \mathcal{P}(U) \setminus \{\emptyset\}
\]

eine Stichprobe, wobei \(\mathcal{P}(U)\) die Potenzmenge von \(U\) bezeichnet und \(S\) bzgl. \((\Omega, B)\) messbar sein soll. Für die Menge der befragten Personen \(u \in S(\omega)\) kann eine neue Variable mit dem Wertebereich \(\mathcal{Y}^*\) konstruiert
werden, die das „angegebene Haushaltseinkommen“ repräsentiert. Um das Problem zu umgehen, allen Personen unabhängig von der Befragung ein Antwortverhalten zuzuschreiben, wird zu $\mathcal{Y}^*$ noch ein Symbol „*“ hinzugefügt. Dann kann die Definition auf alle Personen $u \in \mathcal{U}$ ausgedehnt werden und es ergibt sich

$$Y^* : \mathcal{U} \times \Omega \longrightarrow \{\{y\} \mid y \in \mathcal{Y}\} \cup \{(1, 400), \ldots, (15000, \infty)\} \cup \mathcal{Y} \cup \{*\}$$

wobei $Y^*(u, \omega) = *$ für $u \notin S(\omega)$ gesetzt wird. Die Abbildung $Y^*$ repräsentiert das „in der Stichprobe s angegebene Haushaltseinkommen“. Mit dieser Konstruktion ist man nicht gezwungen, über die Antworten nicht befragter Personen zu spekulieren. Sowohl für eine gegebene Stichprobe $s$, also eingeschränkt auf die Menge $\{\omega \mid S(\omega) = s\}$, als auch auf ganz $\Omega$ sind $Y(u, .)$ und $Y^*(u, .)$ Zufallsvariablen im Sinn der Wahrscheinlichkeitsstatistik. Der Zusammenhang zwischen $Y(u, .)$ und $Y^*(u, .)$ lässt sich durch probabilistische Modelle darstellen. Sie beziehen sich zunächst auf eine Person $u$. Es muss zusätzlich angenommen werden, die Zufallsvariablen $(Y(u, .), u \in \mathcal{U})$ bzw. $(Y^*(u, .), u \in \mathcal{U})$ seien stochastisch unabhängig und identisch verteilt. In diesem Rahmen können nun Konsequenzen unvollständiger Angaben für statistische Aussagen abgeschätzt werden.

Die etwas aufwendige Notation ist notwendig, um Verwechslungen zwischen Durchschnitten über die Gesamtheit $\mathcal{U}$ und Verteilungen, Erwartungswerten etc. bezüglich des Wahrscheinlichkeitsraums $(\Omega, \mathcal{B}, \lambda)$ zu vermeiden. In der Literatur erscheint die Verwechslung häufig nach einem nicht kenntlich gemachten Übergang von stichprobentheoretischen zu probabilistischen Argumenten. So schreibt z.B. P. Holland: „A probability will mean nothing more nor less than a proportion of units in $\mathcal{U}$. The expected value of a variable is merely its average value over all of $\mathcal{U}$“ (Holland 1986: 945). Später verwendet er aber die stochastische Unabhängigkeit zwischen Variablen (1986: 948f), ohne zu bemerken, dass stochastisch unabhängige Variablen auf endlichen Räumen $\mathcal{U}$ nur selten existieren. Eine ähnliche Verwechslung findet sich noch bei Vytlačil (2002: 332).

Durchschnitte über $\mathcal{U}$ und Durchschnitte über den Wahrscheinlichkeitsraum $(\Omega, \mathcal{B}, \lambda)$ führen nicht nur zu unterschiedlichen numerischen
Ergebnissen, sie sind nach wie vor konzeptionell verbunden. Zwar garantieren asymptotische Aussagen wie starke Gesetze oder Ergodensätze, dass die beiden Durchschnitte im Grenzwert ($|U| \to \infty$ oder $\mathbb{E}(|S|)/|U| \to c \notin \{0, 1\}$, $|U| \to \infty$, etc.) gleich sind. Dies sind aber probabilistische Aussagen, die sich auf die Modellebene beziehen, also ein probabilistisches Modell auf $(\Omega, \mathcal{B}, \lambda)$ voraussetzen. Es sind mathematische Konstruktionen, die keine Aussage über empirische Verhältnisse wie Einkommensverteilungen wie Einkommensverteilungen begründen können. Und Grenzwertüberlegungen führen eine zusätzliche Abstraktionsebene ein, die über probabilistische Formulierungen von Antworten auf Fragen nach dem Einkommen hinausgehen. Le Cam und Yang schreiben hierzu:

It must be pointed out that the asymptotics of the ‘standard i.i.d. case’ are of little relevance to practical use of statistics, in spite of their widespread study and use. The reason for this is very simple. One hardly ever encounters fixed families $\{p_\theta | \theta \in \Theta\}$ with a number of observations that will tend to infinity. There are not that many particles in the visible universe! The use of such considerations is an abuse of confidence that has been foisted upon unsuspecting students and practitioners owing to the fact that we, as a group, possess limited analytical abilities and, perforce, have to limit ourselves to simple problems. … The use of asymptotics ‘as $n \to \infty$’ for the standard i.i.d. case seems to be based on an entirely unwarranted act of faith. (Le Cam und Yang 1990: 99f).

Selbst wenn asymptotische Argumente als relevant angesehen werden, so wird man konstatieren müssen, dass unterschiedliche Modelle für $F$ zu

---

Ergebnissen führen können, die offenbar nichts über den Durchschnitt von Werten aller \( u \in U \) sagen.

Die Abwendung von stichprobentheoretischen Konzepten zugunsten probabilistischer Modelle erfordert zudem eine Klärung der Annahmen, die in probabilistischen Modellen verwandt werden. Insbesondere die Annahme unabhängiger und identisch verteilter Zufallsvariablen ist keine Annahme, die verändert oder aufgegeben werden könnte, ohne den Rahmen des Modells zu sprengen. Sie verweist auf keine gesellschaftlichen Sachverhalte, ebenso wenig wie die Leinwand eines Gemäldes auf Eigenschaften der dargestellten Dinge verweist. Entsprechend gibt es auch keine empirischen Anhaltspunkte, aufgrund derer sich die Annahme zurückweisen ließe. Die Annahme ist weder wahr noch falsch, sondern ein Ausgangspunkt für alle probabilistischen Modelle. Wird in einem nächsten Modellierungsschritt eine Modellklasse, z.B. \( \{ F_\theta \mid \theta \in \Theta \} \) vorgeschlagen, so wird immer schon die Unabhängigkeit und identische Verteilung der so beschriebenen Zufallsvariablen unterstellt. Auch eine Modellklasse kann daher weder wahr noch falsch sein. Aber die Wahl einer Modellklasse kann sich bei einem Vergleich von Realisierungen der Zufallsvariablen mit empirischen Verteilungen als unangemessen erweisen. Eine solche Kritik von Modellvorschlägen, so notwendig und hilfreich sie ist, führt allerdings selbst unter idealisierten Bedingungen nicht zu der eindeutigen Wahl eines probabilistischen Modells. Der spekulative Spielraum, den probabilistische Modelle immer bieten, kann gerade bei der Behandlung von Daten mit unvollständigen Angaben produktiv genutzt werden. Denn dabei muss immer überlegt werden, was der Fall gewesen sein könnte. Probabilistische Modelle bieten einen Rahmen, eine Vielzahl alternativer Möglichkeiten einfach zu benennen und gegeneinander abzuwägen.
4.7. Ignorierbare Ausfälle: Parametrische und nichtparametrische Modelle

Am einfachsten wäre es, wenn eine Angabe $Y^*(u, \omega)$ nur die offensichtliche Information $Y(u, \omega) \in Y^*(u, \omega)$ enthalten würde. Dann bräuchte man sich bei Schätzungen keine Gedanken über den Zusammenhang von $Y(u, .)$ und $Y^*(u, .)$ zu machen. Der Ansatz sei an zwei Beispielen demonstriert: Zunächst sei die Verteilung $F_\theta$ von $Y(u, .)$ durch einen endlich-dimensionalen Vektor $\theta \in \mathbb{R}^k$ parametrisiert, etwa $Y(u, .)=\mathcal{N}(\mu, \sigma^2)$, also normalverteilten mit Erwartungswert $\mu$ und Varianz $\sigma^2$, $\theta=(\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. Der Beitrag einer Beobachtung $y(u)$ zur Likelihood $L(\theta; y(u), u \in s)$ ist dann $\phi(y(u); \mu, \sigma^2)$, wobei $\phi(.; \mu, \sigma^2)$ die Dichte der Normalverteilung mit Erwartungswert $\mu$ und Varianz $\sigma^2$ ist. Sind nun $\{Y(u, .), u \in s\}$ stochastisch unabhängig für gegebenes $s$, dann ist die Likelihood das Produkt der Dichten

$$L((\mu, \sigma^2); y(u), u \in s) = \prod_{u \in s} \phi(y(u); \mu, \sigma^2)$$

wobei unterstellt wird, der Stichprobenplan habe nichts mit den Variablen $Y(u, .)$ zu tun. Werden alle Informationen über das Zustandekommen einer Realisation $y^*(u)$ von $Y^*(u, .)$ vernachlässigt, wird also nur $y(u) \in y^*(u)$ berücksichtigt, und sind die $Y^*(u, .)$ weiterhin stochastisch unabhängig, dann wird die Likelihood zu

$$L(\theta; y^*(u), u \in s) = \prod_{u \in s} \int_{v \in y^*(u)} dF_\theta(v)$$

Im Folgenden wird unterstellt, dass die Befragten nicht „lügen“, also immer $Y(u, \omega) \in Y^*(u, \omega)$ gilt. Letzteres war, ohne einen probabilistischen Rahmen, bereits bei der Betrachtung intervallwertiger Statistiken unterstellt worden.

4. Probabilistische Selektionsmodelle

Im Fall der Angaben zum Haushaltseinkommen im ALLBUS ergibt sich

\[
L(\theta; y^*(u), u \in s_2) = \prod_{u \in s_4} \phi(y(u); \theta) \times \prod_{u \in s_3 \setminus s_4} \int_{v \in y^*(u)} \phi(v; \theta) \, dv \prod_{u \in s_2 \setminus s_3} \int_{v \in \mathbb{R}} \phi(v; \theta) \, dv
\]

Der erste Faktor repräsentiert den Beitrag der genauen Beobachtungen, der zweite den der gruppierten Angaben, und der letzte gibt den Beitrag der Verweigerungen (inklusive keine Angabe/weiß nicht) wieder. Der letzte Term ist konstant 1 und damit unabhängig von den Parametern, so dass man sich auf die ersten beiden Termen konzentrieren kann. Maximiert man diese Likelihoodfunktion, ergibt sich für die Nettostichprobe \(s_2\) und ohne Berücksichtigung der unterschiedlichen Auswahlsätze für Ost- und Westdeutschland \(\hat{\mu} = 3807\) und \(\hat{\sigma} = 2047\) sowie ein modellbasiertes 95%–Konfidenzintervall für \(\mu\) von \((3729, 3885)\). Der naive Ansatz, der die \(n_3 - n_4 = 906\) gruppierten Angaben ganz unberücksichtigt lässt und nur die Daten der Stichprobe \(s_4\) benutzt, ergibt einen Mittelwert von 3676 mit einem (ungewichteten) 95%–Konfidenzintervall \((3582, 3770)\). Der Wert 3676 liegt 131 DM unter dem Wert, der sich unter Berücksichtigung der gruppierten Angaben innerhalb des Normalverteilungsmodells ergibt, sogar ausserhalb des Konfidenzintervalls \((3729, 3885)\).

Anstelle einer Normalverteilung kann auch unterstellt werden, \(\log(Y(u, .))\) sei normalverteilt mit den Parametern \((\mu, \sigma^2)\). Die Maximierung der entsprechenden Likelihood führt zu \(\hat{\mu} = 8,113\) und \(\hat{\sigma} = 0,5371\). Da \(E_{\theta}(Y) = \exp(\mu + \sigma^2/2)\) für log-normalverteilte Zufallsvariablen \(Y\) ist, ergibt sich als Schätzung des Erwartungswerts 3849 mit dem (modellbasierten) Konfidenzintervall \((3765, 3932)\), berechnet mit der Deltamethode (Lehmann 1999: 85ff). Der Erwartungswert unter diesem Modell ist um 173 DM größer als der naive Durchschnitt. Wählt man schließlich die log-logistische Verteilung mit \(F_\theta(y) = \theta y / (1 + \theta y)\), dann ergibt sich \(\hat{\theta} = 2,9879 \times 10^{-4}\). Der Erwartungswert unter diesem Modell ist allerdings \(\infty\). Man würde auch einen Erwartungswert von \(\infty\) erhalten, wenn die Dichte etwa der log-normalen Verteilung rechts von einem beliebig großen Wert \(y_0\) durch die entsprechende Dichte der log-logistischen Verteilung ersetzt würde. Aber eine solche Veränderung in der Wahl
4.7. Ignorierbare Ausfälle: Parametrische und nichtparametrische Modelle

der Modellklasse lässt sich empirisch nicht beurteilen, weil \( y_0 \) immer größer als alle Beobachtungen gewählt werden kann. Der Mittelwert von Merkmalen einer endlichen Menge \( U \) ist dagegen sicher immer endlich. Hier zeigt sich der bereits angedeutete Perspektivenwechsel: An Stelle eines Durchschnitts über alle \( u \in U \) interessiert der Durchschnitt über die \( \omega \in \Omega \), durch die der Parameter \( \theta \) erst seine Bedeutung erhält. Es gibt aber zwischen dem Mittelwert einer endlichen Menge \( U \) und dem Erwartungswert einer Zufallsvariablen \( Y(u_0, .) \) keine notwendigen Beziehungen.

Es könnte scheinen, das Problem entstehe durch eine zu enge Wahl der Modellklasse und könne durch nichtparametrische Verfahren gelöst werden. Wird „nur“ angenommen, die \( (Y(u, .), u \in s) \) seien unabhängig und identisch verteilt, dann ist die nichtparametrische Likelihood für die Verteilung \( F \) bei Daten \( \{ Y^*(u, .) \mid u \in s \} \)

\[
L(F; y^*(u), u \in s) = \prod_{u \in s} \int_{v \in y^*(u)} dF(v)
\]

Im Fall von \( s_2 \) ergibt sich

\[
L(F; y^*(u), u \in s_2) = \prod_{u \in s_4} F(y(u)) - F(y(u)_-) \times 
\prod_{u \in s_3 \setminus s_4} \int_{v \in y^*(u)} dF(v) \prod_{u \in s_2 \setminus s_3} \int_{v \in \mathbb{R}} dF(v)
\]

wobei \( F(y) - F(y_-) \) die Sprunghöhe der Funktion \( F \) an der Stelle \( y \) ist. Werden nur diskrete Verteilungen \( F \) betrachtet, dann existiert häufig ein Maximum der Likelihoodfunktion, etwa \( \hat{F} \). \( \hat{F} \) wird nichtparametrischer Maximum-Likelihood-Schätzer (NPMLE) der Verteilungsfunktion \( F \) genannt. Als Schätzer des Erwartungswerts kann \( \hat{\mu} := \int v d\hat{F}(v) \) verwandt werden.

Im Fall der Haushaltseinkommen ergibt sich \( \hat{\mu} = 3777 \). Ein 95%-Konfidenzintervall, basierend auf 10000 Bootstrap-Replikationen, ist \( (3553, 3859) \).\(^{12}\) Auf den ersten Blick könnte es scheinen, als ergäbe sich immer

\(^{12}\) Die Verteilung der geschätzten Mittelwerte für die Replikationen ist multimodal. Daher ergibt sich ein sehr asymmetrisches Konfidenzintervall, wenn wir hier die Perzentile der Verteilung zur Konstruktion benutzt werden.
ein endlicher Schätzwert $\hat{\mu}$, der nun nicht mehr von der Wahl einer Modellklasse abhinge. Das ist bei unvollständigen Angaben aber nicht der Fall. Der NPMLE ist nicht eindeutig definiert, wenn es gruppierte Beobachtungen gibt, aber keine genauen Beobachtungen in dieses Intervall fallen. Dann kann $\hat{F}$ auf dem Intervall beliebig definiert werden, ohne den Wert der Likelihoodfunktion zu ändern. Fällt insbesondere keine genaue Beobachtung in die größte Einkommensklasse „Einkommen größer als 15000 DM“ während das Intervall wenigstens einmal genannt wird, dann kann die zugehörige Masse in $\hat{F}$ beliebig gegen $\infty$ verschoben werden. Das geschieht in den Bootstrap-Replikationen so oft, dass die obere Schranke des Konfidenzintervalls ehrlicherweise durch $\infty$ ersetzt werden müsste. In der Tat ist in der Berechnung des „Konfidenzintervalls“ der replizierte Datensatz einfach die Angabe „Einkommen größer als 15000 DM“ durch den Wert 22500 ersetzt worden, falls keine genauen Beobachtungen in das Intervall fielen. Jeder andere Wert $> 15000$ wäre aber genauso möglich. Auch unter nichtparametrischen Modellen ergibt sich ein beliebig großer Unterschied zwischen dem ursprünglich interessierenden Durchschnitt von Werten einer Gesamtheit $\mathcal{U}$ und dem Erwartungswert $\int v dF(v) = \int Y(u_0, \omega) d\lambda(\omega)$, einem Durchschnitt über $\omega \in \Omega$.

4.8. Ignorierbare Ausfälle: MAR, CAR und all das

Der Perspektivenwechsel führt zu einer Abkehr von dem Versuch, probabilistische Aussagen über den Zusammenhang eines Mittelwerts in einer Gesamtheit $\mathcal{U}$ mit Mittelwerten über Stichproben zu formulieren. Das mag gerechtfertigt sein, wenn stattdessen die Verwendung probabilistischer Modelle zum Verständnis von Effekten unvollständiger Angaben beiträgt, zumal über deren Zustandekommen nur spekuliert werden kann. Zunächst muss geklärt werden, unter welchen Bedingungen die Verwendung von $L(\theta; y^*(u), u \in s)$ im vorigen Abschnitt begründet werden kann. Dies kann nicht immer der Fall sein, selbst wenn für alle Personen $u \in \mathcal{U}$ identische Beziehungen zwischen $Y(u, .)$ und $Y^*(u, .)$ unterstellt werden.
4.8. Ignorierbare Ausfälle: MAR, CAR und all das

Abbildung 4.2.: Geschätztes mittleres Haushaltseinkommen und 95\% Konfidenzintervalle. a) Nichtparametrische Schranken auf \( s_3 \), b) Normalverteilung ohne gruppierte Angaben, c) Normalverteilung unter Einschluss gruppiert er Angaben, d) Lognormalverteilung unter Einschluss gruppiert er Angaben, e) Nichtparametrischer Mittelwert unter Einschluss gruppi erer Angaben

Sei z.B. \( \mathcal{Y} = \{1000, 1500, 2000\} \) und \( \mathcal{Y}^* = \{\{1000, 1500\}, \{1000\}, \{1500\}, \{2000\}\}. \) Ist nun \( y(u) = 1000 \), dann kann \( u \) entweder \{1000\} berichten, also den genauen Betrag nennen, oder aber mit \{1000, 1500\} antworten. Entsprechendes gilt für \( y(u) = 1500 \). Ist nun \( Y(u, .) \) auf \{1000, 1500, 2000\} gleichverteilt und berichten alle Personen \( y^*(u) = \{1000, 1500\} \) fällt \( y(u) = 1000 \), sonst aber immer den genauen Betrag, dann ergibt sich die folgende Verteilung auf \( \mathcal{Y}^* \):

\[
\begin{align*}
\Pr(Y^*(u, .) = \{1000, 1500\}) &= 1/3 \\
\Pr(Y^*(u, .) = \{1000\}) &= 0 \\
\Pr(Y^*(u, .) = \{1500\}) &= 1/3 \\
\Pr(Y^*(u, .) = \{2000\}) &= 1/3
\end{align*}
\]

Werden die Beobachtungen wie im letzten Abschnitt behandelt, so wird

\[
\Pr(Y^*(u, .) = \{1000, 1500\}) = \Pr(Y(u, .) = 1000) + \Pr(Y(u, .) = 1500)
\]

gesetzt. Ist die Verteilung von \( Y^*(u, .) \) bekannt, so ergibt sich als Verteilung von \( Y(u, .) \):

\[
\begin{align*}
\Pr(Y(u, .) = 1000) &= 0, \\
\Pr(Y(u, .) = 1500) &= 2/3,
\end{align*}
\]
4. Probabilistische Selektionsmodelle

\[ \Pr(Y(u, .) = 2000) = 1/3 \]

Für den Zusammenhang zwischen \( Y \) und \( Y^* \) wird

\[ \Pr(Y^*(u, .) = \{1000, 1500\} \mid Y(u, .) = 1000) = 1/2 \]
\[ = \Pr(Y^*(u, .) = \{1000, 1500\} \mid Y(u, .) = 1500) \]

angenommen. Die bedingten Wahrscheinlichkeiten von \( \{Y^*(u, .) = \{1000, 1500\}\} \) sind für beide möglichen Bedingungen \( \{Y(u, .) = 1000\} \) und \( \{Y(u, .) = 1500\} \) gleich. Kombiniert man die unterstellte Verteilung von \( Y(u, .) \) mit den beiden konditionalen Verteilungen, dann ergibt sich in der Tat die Verteilung von \( Y^*(u, .) \). Die Interpretation wäre: Ist \( Y(u, .) = 1500 \), dann antwortet \( u \) mit Wahrscheinlichkeit 1/2 entweder \( \{1500\} \) oder \( \{1000, 1500\} \). Die entscheidende Annahme ist offenbar die über die bedingten Verteilungen

\[ \Pr(Y^*(u, .) = \{1000, 1500\} \mid Y(u, .) = 1000) \quad \text{und} \quad \Pr(Y^*(u, .) = \{1000, 1500\} \mid Y(u, .) = 1500) \]

Die bedingten Verteilungen beschreiben den Zusammenhang zwischen \( Y(u, .) \) und \( Y^*(u, .) \).

Sei nun allgemein \( Y(u, .) \) eine Zufallsvariable mit Werten in der endlichen Menge \( \mathcal{Y} \) und \( Y^*(u, .) \) eine Zufallsvariable mit Werten in \( \mathcal{Y}^* \subseteq \mathcal{P}(\mathcal{Y}) \setminus \{\emptyset\} \) auf dem gemeinsamen Raum \( (\Omega, \mathcal{B}, \lambda) \). Dann soll \( Y^*(u, .) \) zufällige Vergrößerung (CAR, coarsened at random) von \( Y(u, .) \) heißen, wenn eine der folgenden äquivalenten Bedingungen für alle \( y^* \in \mathcal{Y}^* \) und für alle \( y \in y^* \) erfüllt ist:

\[ \Pr(Y^*(u, .) = y^* \mid Y(u, .) = y) \text{ ist konstant auf } y \in y^* \]  \hspace{1cm} (4.1)
\[ \Pr(Y^*(u, .) = y^* \mid Y(u, .) = y) = \Pr(Y^*(u, .) = y^* \mid Y(u, .) \in y^*) \]  \hspace{1cm} (4.2)
\[ \{Y^*(u, .) = y^*\} \perp \perp \{Y(u, .) = y\} \mid \{Y(u, .) \in y^*\} \]  \hspace{1cm} (4.3)
\[ \Pr(Y(u, .) = y \mid Y^*(u, .) = y^*) = \Pr(Y(u, .) = y \mid Y(u, .) \in y^*) \]  \hspace{1cm} (4.4)

Dabei bedeutet \( A \perp \perp B \mid C \) die bedingte stochastische Unabhängigkeit der Ereignisse \( A \) und \( B \) gegeben \( C \). \( \{\Pr(Y^*(u, .) = y^* \mid Y(u, .) = y) \mid y^* \in \mathcal{Y}^*, y \in y^*\} \) soll im Folgenden Selektionsmodell heißen.

226
Die Bedingungen (4.1) und (4.2) beschreiben die Situation ausgehend vom Wert $y(u)$ der zugrunde liegenden Variablen $Y(u, .)$. Bei gegebenem $y(u)$ müssen nur noch die Antwortmöglichkeiten $Y^*(u, .)$ in Erwägung gezogen werden. Ist etwa das tatsächliche Haushaltseinkommen eines Befragten 1234 DM, dann kann er im ALLBUS entweder 1234 oder das Intervall [1000, 1250) angeben, oder er antwortet gar nicht. Aus diesen Antwortalternativen wählt $u$ mit den bedingten Wahrscheinlichkeiten $\Pr(Y^*(u, .) = y^* \mid Y(u, .) = 1234)$. Ist dagegen $y(u) = 1123$, so erzwingt (4.1) eine Auswahl zwischen den drei Antwortmöglichkeiten 1123, [1000,1250) und [1, ..., $\infty$) mit den gleichen Wahrscheinlichkeiten wie im Fall $y(u) = 1234$. Für alle möglichen Einkommen im Bereich [1000, 1250) entscheidet sich $u$ nach den gleichen Wahrscheinlichkeiten zwischen einer genauen Angabe, der gruppierten Angabe oder gar keiner Angabe. Für Werte in einem anderen Gruppierungsintervall kann sich eine andere Aufteilung zwischen den Antwortmöglichkeiten „genau“ und „gruppiert“ ergeben. Die Bedingung (4.2) ist nur eine Umformulierung dieser Beschreibung. Denn (4.2) verlangt, dass die Entscheidung, kein Haushaltseinkommen anzugeben, unabhängig von dem tatsächlichen Einkommen getroffen wird, während die Entscheidung zwischen einer gruppierten oder genauen Angabe innerhalb eines Gruppierungsintervalls nicht von der tatsächlichen Höhe des Einkommens abhängt. Wenn bekannt ist, dass das tatsächliche Einkommen $Y(u, .) \in [1000, 1250)$ ist, dann verlangt (4.3) die Unabhängigkeit des Ereignisses „gruppierte Angabe“ von den tatsächlichen Einkommen innerhalb des Intervalls. Ist $y^* = Y = [1, 2, \ldots, \infty)$, dann ist die Bedingung $Y(u, .) \in [1, 2, \ldots, \infty)$ immer erfüllt und (4.3) verlangt die (unbedingte) stochastische Unabhängigkeit von $\{Y^*(u, .) = [1, 2, \ldots, \infty)\}$ und $Y(u, .)$. Ist dagegen $Y(u, .) \in y^* = \{y\}$, dann gilt $Y(u, .)) = y$ und die Bedingung (4.3) ist automatisch erfüllt.

4. Probabilistische Selektionsmodelle

worden.

Schreibt man \[ \Pr_\theta(Y(u, .) = y) \], um die Abhängigkeit der Verteilung von einem Parameter \( \theta \in \Theta \) anzugeben, und entsprechend \[ \Pr_\gamma(Y^*(u, .) = y^* \mid Y(u, .) = y) \] mit \( y \in \Gamma \) für das Selektionsmodell, dann kann die Verteilung der Beobachtungen \( Y^*(u, .) \) wie folgt aufgespalten werden:

\[
\Pr_{\theta,\gamma}(Y^*(u, .) = y^*) = \sum_{y \in Y^*} \Pr_{\theta,\gamma}(Y^*(u, .) = y^*, Y(u, .) = y) = \sum_{y \in Y^*} \Pr_\theta(Y(u, .) = y) \Pr_\gamma(Y^*(u, .) = y^* \mid Y(u, .) = y) = \Pr_\gamma(Y^*(u, .) = y^* \mid Y(u, .) \in y^*) \sum_{y \in y^*} \Pr_\theta(Y(u, .) = y) \Pr_\gamma(Y^*(u, .) = y^* \mid Y(u, .) \in y^*)
\]

Die dritte Gleichung folgt aus der CAR-Bedingung (4.1), die letzte Gleichung aus (4.2). Sind die Parameter \( \theta \) und \( \gamma \) variationsunabhängig, gibt es also zu jedem Element \((\theta, \gamma) \in \Theta \times \Gamma\) eine Wahrscheinlichkeitsverteilung \( \Pr_{\theta,\gamma} \), dann kann bei Likelihoodbetrachtungen für \( \theta \) der Selektionsteil \( \Pr_{\gamma}(Y^*(u, .) = y^* \mid Y(u, .) \in y^*) \) vernachlässigt werden. Es reicht,

\[
L(\theta; y^*(u), u \in s) = \prod_{u \in s} \Pr_\theta(Y(u, .) \in y^*)
\]

zu maximieren. Antworten die Befragten entweder mit einer genaunen Angabe oder gar nicht, dann ist \( \mathcal{Y}^* = \{ \{ y \} \mid y \in \mathcal{Y} \} \cup \{ \mathcal{Y} \} \). In diesem Fall impliziert CAR die Möglichkeit, sich nur auf die genauen Angaben beschränken zu können. Die CAR-Bedingung ist in diesem Zusammenhang auch MAR (missing at random) genannt worden.

Es kann gezeigt werden, dass es zu jeder vorgelegten Verteilung auf \( \mathcal{Y}^* \) immer eine Verteilung auf \( \mathcal{Y} \) und ein Selektionsmodell gibt, der die CAR-Bedingung erfüllt (Gill et al. 1997: 262; Heitjan 1994; Heitjan und Rubin 1991; Grünwald und Halpern 2003). Ist insbesondere \( \{ y \} \in \mathcal{Y}^* \) für alle \( y \in \mathcal{Y} \) und \( \Pr(Y^*(u, .) = \{ y \}) > 0 \), dann ist die Verteilung von \( Y(u, .) \) durch die CAR-Bedingung sogar eindeutig bestimmt. Ist die
4.8. Ignorierbare Ausfälle: MAR, CAR und all das

Verteilung von $Y^*(u,.)$ bekannt, dann kann immer ein CAR-Modell unterstellt werden. Mit anderen Worten: Keine noch so große Menge an Daten erlaubt es, zwischen einem ignorierbaren Selektionsmodell, der die CAR-Bedingung erfüllt, und nicht ignorierbaren Modellen (wie am Anfang des Abschnitts) zu unterscheiden. Selektionsmodelle sind nicht identifizierbar: Wird ein beliebiges Selektionsmodell vorgeschlagen, so kann immer ein CAR-Modell angegeben werden, der ebenso gut zu den Daten passt.

Wenn von einem CAR-Modell ausgegangen wird, dann braucht, basierend auf der Likelihoodtheorie, kein Selektionsmodell angegeben zu werden. Manchmal erscheint es aber sinnvoll, sich selbst in der CAR-Situation ein Bild des Selektionsprozesses zu machen. Wird ein (semi-)parametrisches Modell für den Selektionsprozess gewählt, dann hat dies empirische Konsequenzen, kann sich also als falsch erweisen. Denn unter der CAR-Bedingung sind auch die $\Pr(Y^*(u,.) = y^* | Y(u,.) \in y^*)$ eindeutig bestimmt, falls nur $\Pr(Y^*(u,.) = y^*) > 0$ ist. Die Annahme einer Klasse $\Pr_y(Y^*(u,.) = y^* | Y(u,.) \in y^*)$ von Selektionsmodellen kann unter der CAR-Bedingung zumindest potentiell aus empirischen Gründen zurückgewiesen werden, jedenfalls dann, wenn neben exakten und vollständig fehlenden Angaben auch partielle Angaben zur Verfügung stehen und modelliert werden.


---

13 Das Ergebnis gilt nicht nur für endliche Mengen $\mathcal{Y}$, sondern im wesentlichen auch in allgemeinen Räumen. Allerdings wird dann die Formulierung sehr aufwendig (Gill et al. 1997: 273ff).

4.9. Ignorierbare Ausfälle: Konditionale CAR-Modelle

In vielen Fällen gibt es neben den Angaben \((y^*(u), u \in s)\) zu den interessierenden Größen \((y(u), u \in U)\) weitere Informationen. Dabei kann es sich um Designvariablen handeln, deren Werte zumindest für die intendierte Stichprobe bekannt sind, um Angaben über die Kontaktaufnahme oder um Angaben des Befragten aus anderen Teilen des Interviews. Im ALLBUS gibt es z.B. Angaben für alle Befragten aus \(s_2\) zu Geschlecht, Alter, Haushaltsgröße, Staatsangehörigkeit und Befragungsgebiet (Ost/West). Nun kann die globale Annahme einer CAR-Bedingung unrealistisch erscheinen. Gleichwohl könnte die CAR-Bedingung getrennt für alle Teilmengen gelten, die durch die zusätzlichen Angaben \(X(u, .)\) definiert werden.

Wenn die zusätzlichen Angaben in einem Vektor \(X\) zusammengefasst werden, dann kann ein entsprechender Zufallsvektor konstruiert werden: \(X : U \times \Omega \rightarrow \mathcal{X}\). Die CAR-Bedingungen können konditional auf die Werte dieser Kovariablen formuliert werden:

\[
\Pr(Y^*(u, .) = y^* \mid Y(u, .) = y, X(u, .) = x) \text{ ist konstant auf } y \in y^* \quad (4.5)
\]

\[
\Pr(Y^*(u, .) = y^* \mid Y(u, .) = y, X(u, .) = x) \\
= \Pr(Y^*(u, .) = y^* \mid Y(u, .) \in y^*, X(u, .) = x) \quad (4.6)
\]
4.9. Ignorierbare Ausfälle: Konditionale CAR-Modelle

\[ \{ Y^*(u, .) = y^* \} \perp \{ Y(u, .) = y \} \mid \{ Y(u, .) \in y^* \}, X(u, .) \quad (4.7) \]

\[ \Pr(Y(u, .) = y \mid Y^*(u, .) = y^*, X(u, .) = x) = \Pr(Y(u, .) = y \mid Y(u, .) \in y^*, X(u, .) = x) \quad (4.8) \]

In vielen Texten wird suggeriert, die CAR-Annahme werde plausibler, wenn nur genügend „Informationen“ in Form von Kovariablen in das Selektionsmodell einbezogen werden, wenn also ein möglichst großer Vektor \( X(u, .) \) gewählt wird. So schreiben Little und Rubin:

…we believe that in situations where good covariate information is available and included in the analysis, the missing at random (MAR) assumption may often be a reasonable approximation to reality, thus obviating the need for a sensitivity analysis to model nonignorable nonresponse. (Little und Rubin in Scharfstein et al. 1999: 1130).

Die Argumentation beruht auf einer fehlerhaften Gleichsetzung von „Information“ mit bedingten Verteilungen. Während im umgangssprachlichen Gebrauch des Wortes eine „bessere Information“ immer zu einem besseren Verständnis einer Situation oder eines Ereignisses beiträgt, gilt dies nicht für die Einbeziehung zusätzlicher Kovariablen in den Bedingungen (4.5) – (4.8).\(^{14}\)

Eine Idee, die schon von Pearson Anfang des letzten Jahrhunderts formuliert wurde, geht davon aus, dass \( Y \) eine lineare Regression auf den

\(^{14}\) Sind z.B. \( U, V \) normalverteilte unabhängige Zufallsvariablen mit Erwartungswert 0 und Varianz 1, dann sind \( Y_1 := U + V \) und \( Y_2 := U - V \) unabhängig und normalverteilt, also \( Y_1 \perp \perp Y_2 \). Dagegen ist die bedingte Kovarianz, wenn die „Information“ \( U = u \) gegeben ist:

\[ \text{Cov}(Y_1, Y_2 \mid U = u) = \mathbb{E}(Y_1 Y_2 \mid U = u) - \mathbb{E}(Y_1 \mid U = u) \mathbb{E}(Y_2 \mid U = u) = \mathbb{E}((u + V)(u - V)) - u^2 = \text{Var}(V) - u^2 = 1 - u^2 \]. \( Y_1 \perp \perp Y_2 \mid U = u \) gilt also nur, falls \( U = 0 \) ist, ein Ereignis mit Wahrscheinlichkeit 0. Die Einführung der zusätzlichen „Information“ \( U = u \) führt von unabhängigen Variablen \( Y_1 \) und \( Y_2 \) zu korrelierten bedingten Variablen. Die Bedingung (4.7) kann durch die Einbeziehung weiterer Kovariablen also auch verletzt werden. Weitere Probleme bei der Interpretation von „Information“ bei bedingten Modellen diskutieren Dubra und Echenique (2004).
4. Probabilistische Selektionsmodelle

Vektor $X$ besitzt, $\mathbb{E}(Y \mid X) = X\beta$ (Lawley 1943). Die lineare Beziehung soll dabei sowohl für die Teilmenge der vollständigen Angaben als auch für die Gesamtheit $U$ mit dem jeweils gleichen $\beta$ gelten. An Stelle der Identität der Erwartungswerte in allen Teilstichproben wird also nur die Identität der Regressionsfunktion sowie eine konstante bedingte Varianz gefordert. Haben $Y \mid X$ und $Y \mid X$, $\mathbb{I}[R = 1]$ die gleiche Verteilung, dann ist das Modell sicherlich CAR. Wird nur die Gleichheit der Erwartungswerte und Varianzen gefordert, so ergibt sich ein etwas allgemeineres Selektionsmodell.

Im Pearson-Lawley Modell kann zunächst $\beta$ auf der Teilstichprobe mit vollständigen Angaben geschätzt werden; in einem zweiten Schritt aufgrund von $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y \mid X))$ ein Schätzer des Erwartungswerts von $Y$ durch

$$\hat{\mu} = \int \hat{\mathbb{E}}(Y \mid X = x) \, d\hat{F}_X(x) = \frac{1}{|s|} \sum_{u \in s} x(u)\hat{\beta}$$

konstruiert werden. Ist die Regression auch homoskedastisch, dann lassen sich Schätzer für die Varianzen angeben. Im Fall des ALLBUS ergibt sich $\hat{\mu} = 3763$ mit dem 95% Konfidenzintervall (3660, 3866), wenn als Kovariablen das Alter, Geschlecht, Haushaltsgröße und Staatsangehörigkeit benutzt werden. Der naive Mittelwert, der nur vollständige Angaben berücksichtigt, ist um 90 DM kleiner und befindet sich am unteren Rand des Konfidenzintervalls.

Im Rahmen des Pearson-Lawley Ansatzes können auch die gruppierten Angaben aus $s_3 \setminus s_4$ berücksichtigt werden, indem die Methoden des letzten Abschnitts auf die Residuen $y(u) - x(u)\beta$ angewandt werden. Außerdem kann an Stelle der linearen Regression ein beliebiges (parametrisches oder semiparametrisches) Regressionsmodell benutzt werden. Ein Nachteil der Methode ist aber ihre Abhängigkeit von Annahmen

\[15\] Die 6 Beobachtungen ohne Altersangabe wurden ausgeschlossen, die Schätzung von $\beta$ erfolgte auf $s_4$ ohne diese Beobachtungen. Die Haushaltsgröße bezieht sich auf die Anzahl der Personen im Haushalt, einschließlich der Kinder. Die Angabe wurde gruppiert, indem Haushalten mit mehr als 4 Personen der Wert 5 zugeordnet wurde. Bei der Staatsangehörigkeit wird nur unterschieden, ob jemand einen ausländischen Pass hat oder nicht.

Eine Möglichkeit, zumindest einige der theoretischen Probleme zu umgehen, ergibt sich aus einem Rückgriff auf eine Idee der Stichproben- theorie und verwendet gewichtete Schätzgleichungen. Wird nur zwischen vollständigen und fehlenden Angaben unterschieden, dann kann ein probabilistisches Modell für das Fehlen einer Angabe in Abhängigkeit von den Kovariablen $X$ formuliert werden. Wird $R(u, .) = 1$ gesetzt, falls $Y(u, .)$ beobachtet wurde, $R(u, .) = 2$ sonst, dann ist ein Modell für das Fehlen von Angaben über $Y(u, .)$ etwa durch $\pi(u, x) \coloneqq \Pr(R(u, .) = 1 \mid X(u, .) = x) = \Pr(R(u, .) = 1 \mid X(u, .) = x, Y(u, .) = y)$ bestimmt. Die Variable $\pi(u, X)$ wird häufig „Propensity Score“ genannt. Ignorierbarkeit des Selektionsmodells besteht gerade in der Unabhängigkeit des Propensity Scores $\pi(u, x)$ von $y$, also in der Annahme $R(u, .) \independent Y(u, .) \mid X(u, .)$. Ist die Auswahlwahrscheinlichkeit für alle Kovariablenwerte größer als
Daher ist

\[ \hat{\mu} = \frac{\sum_{u \in s} \mathbb{1}[R(u,.) = 1] Y(u,.)}{\sum_{u \in s} \mathbb{1}[R(u,.) = 1]} / \frac{\sum_{u \in s} \mathbb{1}[R(u,.) = 1]}{\sum_{u \in s} \mathbb{1}[R(u,.) = 1]} \]

ein erwartungstreuer Schätzer des Erwartungswerts von \( Y(u,.) \), und zwar ganz unabhängig von der Form des bedingten Erwartungswerts \( \mathbb{E}(Y \mid X) \).

Allerdings muss ein Schätzer für \( \pi(u, X) \) angegeben werden. Wird z.B. ein Logit-Modell \( \hat{\pi}(u, x(u)) = \exp(x(u)\hat{\beta})/1 + \exp(x(u)\hat{\beta}) \) verwendet, dann ergibt sich für das mittlere Haushaltseinkommen \( \hat{\mu} = 3765 \) mit einem Konfidenzintervall von (3599, 3931).

Das deutlich größere Konfidenzintervall des gewichteten Schätzers im Vergleich zum Pearson-Lawley-Schätzer ist eine Konsequenz sowohl der abgeschwächten probabilistischen Annahmen als auch der nur unvollständigen Nutzung der Verteilung der \( X(u,.) \) sowie der gemeinsamen Verteilung von \( Y(u,.) \) und \( X(u,.) \) auf \( \{u \in s \mid R(u,.) = 1\} \) für die Schätzung. Ein Teil der Information kann durch die Addition eines weiteren, von Funktionen der \( X(u,.) \) und \( \mathbb{1}[R(u,.) = 1] Y(u,.) \) abhängigen Terms zur Schätzgleichung zurückerhalten werden.\(^{16}\) Allerdings können im

\(^{16}\) Verschiedene Varianten der Einbeziehung dieser Information sind sowohl von Qin et al. (2002) als auch von Robins und seinen Mitarbeitern untersucht worden. Theoretisch können „optimale“ erweiterte Schätzfunktionen angegeben werden, die die (asymptotische) Varianz von \( \hat{\mu} \) minimieren. Die optimale Wahl einer Schätzfunktion hängt von der Spezifikation eines Modells der bedingten Verteilung von \( Y(u,.) \) gegebener \( X(u,.) \) bzw. gegebener \( X(u,.) \) sowie \( R(u,.) = 1 \) (Rotnizky und Robins 1997;
4.10. Nicht ignorierte Ausfälle


4.10. Nicht ignorierte Ausfälle

Bei der Diskussion von Seletionsmodellen wird überlegt werden müssen, was geschieht, wenn sie nicht die CAR-Bedingung erfüllen. In diesem Fall hängt \( \Pr(Y^*(u, .) = y^* \mid Y(u, .) = y) \) aufgrund von (4.1) von \( y \in y^* \) ab und diese bedingten Wahrscheinlichkeiten könnten modelliert werden. Andersherum kann von (4.4) ausgegangen und entsprechend \( \Pr(Y(u, .) = y \mid Y^*(u, .) = y^*) \) modelliert werden. Eine Reihe anderer Möglichkeiten der Konstruktion nicht ignoriarcerer Seletionsmodelle sind denkbar. Das wohl bekannteste Modell ist von Heckman (1976) vorgeschlagen und zumeist für den Fall vollständig fehlender Werte verwandt worden: \( \mathcal{Y}^* = \{ \{y\} \mid y \in \mathcal{Y} \} \cup \{\mathcal{Y}\} \). Sei \( R(u, .) \) eine Variable, die das Antwortverhalten von \( u \) bei der Frage nach dem Haushaltseinkommen darstellt, also \( R(u, .) := 1 \), wenn \( Y(u, .) \) genau angegeben wird, \( R(u, .) = 2 \) sonst. Wenn nun die Existenz einer Variablen \( R^*(u, .) \) mit \( R(u, .) = 1 \Leftrightarrow R^*(u, .) \geq 0 \) und ein Zusammenhang zwischen \( Y(u, .) \) und \( R^*(u, .) \) postuliert wird, dann ergibt sich für den Erwartungswert der Stichprobe mit vollständigen Antworten \( \mathbb{E}(Y(u, .) \mid R^*(u, .) \geq 0) \). Dies muss nicht mit dem unkonditionalen Erwartungswert \( \mathbb{E}(Y(u, .)) \) übereinstimmen, wenn \( R^*(u, .) \) und \( Y(u, .) \) stochastisch abhängig sind. Andererseits erfordert die CAR-Bedingung bei fehlenden Werten die Gleichheit von \( \mathbb{E}(Y(u, .)) \) und dem bedingten Erwartungswert von \( Y(u, .) \) gegeben \( R(u, .) = 1 \). Das Modell ist also sicher nicht CAR, falls ein stochastischer Zusammenhang zwischen \( R^*(u, .) \) und \( Y(u, .) \) unterstellt wird. Jede gemeinsame Verteilung von \( Y(u, .) \) und \( R^*(u, .) \) erzeugt ein

---

Selektionsmodell, das nur dann ignorierbar ist, wenn \( Y(u, .) \perp R^*(u, .) \) gilt. Denn der Ausdruck
\[
\Pr(Y^*(u, .) = Y \mid Y(u, .) = y) = \Pr(R(u, .) = 2 \mid Y(u, .) = y)
\]
\[
= \int_{-\infty}^{0} f(v \mid Y(u, .) = y) \, dv
\]
müsste aufgrund von (4.1) konstant in \( y \) sein, wenn es sich um ein CAR-Modell handelte. Es ist aber von vornherein klar, dass über ein solches nicht ignorierbares Selektionsmodell empirisch wenig zu sagen sein wird, da weder \( R^*(u, .) \) noch \( Y(u, .) \) tatsächlich beobachtet werden.

Sind \( Y(u, .) \) und \( R^*(u, .) \) gemeinsam normalverteilt mit
\[
\mu := \mathbb{E}(Y(u, .)), \quad \mu^* := \mathbb{E}(R^*(u, .))
\]
\[
\text{Var}(R^*(u, .)) := 1, \quad \sigma^2 := \text{Var}(Y(u, .))
\]
\[
\rho := \text{Corr}(Y(u, .), R^*(u, .))
\]
dann ist die bedingte Dichte von \( Y(u, .) \) gegeben \( R(u, .) = 1 \)
\[
\phi(y \mid R^*(u, .) \geq 0) = \frac{1}{\Pr(R(u, .) = 1)} \frac{1}{\sigma} \phi((y - \mu)/\sigma) \times
\]
\[
\Phi \left( \frac{1}{\sqrt{1 - \rho^2}} (\mu^* + \frac{\rho}{\sigma} (y - \mu)) \right) \quad (4.9)
\]
wobei \( \phi \) bzw. \( \Phi \) hier die Dichte bzw. Verteilungsfunktion der standardisierten Normalverteilung bezeichnen (Copas und Li 1997: 59f). Der letzte Term ist die bedingte Wahrscheinlichkeit für eine vollständige Angabe bei gegebenem Wert von \( Y(u, .) \)
\[
\Pr(R = 1 \mid Y(u, .) = y) = \Phi \left( \frac{1}{\sqrt{1 - \rho^2}} (\mu^* + \frac{\rho}{\sigma} (y - \mu)) \right)
\]
Diese bedingte Wahrscheinlichkeit entspricht der bedingten Wahrscheinlichkeit in (4.1). Die CAR-Bedingung gilt genau dann, wenn der Koeffizient von \( y \) 0 ist, wenn also \( \rho = 0 \) ist.\(^{17}\)

\(^{17}\) In der ökonometrischen Literatur wird Heckmans Modell oft als ein Beispiel für eine
Da $R(u, .)$ für alle $u \in s$ bekannt ist, kann wegen $\Phi(\mu^*) = 1 - \Phi(0; \mu^*) = \Pr(R^*(u, .) \geq 0) = \Pr(R(u, .) = 1)$ ein Schätzer von $\mu^*$ durch $\Phi^{-1}(\Pr(R = 1))$ konstruiert werden, wobei $\Pr(R = 1)$ den Anteil vollständiger Beobachtungen in der Stichprobe $s$ angibt. Dagegen müssen $\mu$, $\sigma$ und $\rho$ auf der Basis der vollständigen Beobachtungen geschätzt werden. Insbesondere der Parameter $\rho$, der die Abweichung von einem CAR-Modell darstellt, lässt sich nur aufgrund der Abweichung der Verteilung der vollständigen Beobachtungen von einer Normalverteilung identifizieren. Wenn im Fall des Haushaltseinkommens die genauen Angaben zur Berechnung der Parameter $\mu$, $\sigma$, $\rho$ verwandt werden, so erhält man die völlig unplausiblen Werte $\hat{\mu} = 1312$, $\hat{\sigma} = 3119$, $\hat{\rho} = 0, 987$. Der geschätzte Erwartungswert liegt selbst außerhalb des konsistenten Mittelwertintervalls auf der Bruttostichprobe $s_1$ (vgl. Abschnitt 4.5). Nach diesem Ergebnis hätte ein erheblicher Teil der Population ein stark negatives Haushaltseinkommen. Zudem wäre die „Antwortwortsbereitschaft“ $R^*(u, .)$ fast perfekt mit dem Einkommen korreliert. Benutzt man dagegen die logarithmierten Einkommen, ergibt sich als Erwartungswert 3849 DM und $\hat{\rho} = -0, 845$, also ein stark negativer Zusammenhang zwischen Einkommen und „Antwortbereitschaft“. Etwas stabilere Ergebnisse wird man nur erhoffen können, wenn Kovariablen mit großem Effekt auf $R(u, .)$ angegeben werden können, die keinen Einfluss auf $Y(u, .)$ haben. Aber selbst dann wird die Identifikation des Modells im wesentlichen durch Linearitätsannahmen im Modell für $R^*(u, .)$ ermöglicht. Denn da es zu gegebenen Daten immer ein ( konditionales) CAR-Modell gibt, kann der Koeffizient von $y$ in der bedingten Wahrscheinlichkeit $\Pr(R(u, .) = 1 \mid Y(u, .) = y, X(u, .) = x)$ immer als 0 angenommen werden, wenn nur der Einfluss der übrigen Kovariablen $X(u, .)$ allgemein modelliert wird. Die Identifikation des Koeffizienten von $y$ und damit der Korrelation $\rho$ erfordert eine parametrische Einschränkung der Wirkung der Kovariablen. Aber die Auswirkungen etwa von Linearitätsannahmen in $\Pr(R(u, .) = 1 \mid Y(u, .) = y, X(u, .) = x)$ sind ebenso wenig wie die Auswirkungen willkürlicher Verteilungsannahmen

„selection on unobservables“ angeführt (Nicoletti 2002: 3). Die Bezeichnung soll wohl auf die Abhängigkeit des Modells von der unbeobachteten Variablen $R^*$ hinweisen. Die Bezeichnung ist irreführend, denn wie die letzte Gleichung zeigt, hängt die Selektion nicht von unbeobachtbaren Größen ab, sondern von den Werten von $Y(u, .)$, etwa dem Haushaltseinkommen. Letzteres ist zwar nicht für alle Befragten $u \in s$ bekannt, aber es ist sicher nicht „unbeobachtbar“.
einfach zu überblicken.


### 4.11. Sensitivität

4.11. Sensitivität

in der unterstellten Modellwelt und ist zudem bisher nur für spezielle Statistiken formuliert worden.

Robins und Mitarbeiter haben vorgeschlagen, von den gewichteten Schätzfunktionen

$$\sum_{u \in s} \frac{\mathbb{1}[R(u,.) = 1]}{\pi(u, X)} (Y(u,.) - \mu)$$

auszugehen und den Propensity Score auch als Funktion von $Y(u,.)$ zu modellieren (Rotnitzky und Robins 1997; Robins 1997; Rotnitzky et al. 1998; Scharfstein et al. 1999; Scharfstein und Irizarry 2003). Da

$$E \left( 1 - \frac{\mathbb{1}[R(u,.) = 1]}{\pi(u, X, Y)} \right)$$

$$= 1 - E \left( \frac{1}{\pi(u, X, Y)} E \left( \mathbb{1}[R(u,.) = 1] \mid X(u,.), Y(u,.) \right) \right) = 0$$

ist, kann die Schätzfunktion erweitert werden:

$$\sum_{u \in s} \frac{\mathbb{1}[R(u,.) = 1]}{\pi(u, X, Y)} (Y(u,.) - \mu) + \left( 1 - \frac{\mathbb{1}[R(u,.) = 1]}{\pi(u, X, Y)} \right) \phi(X; \mu)$$

Wird etwa logit $\pi(u, x, y) = x(u)\beta + \alpha y(u)$ gesetzt und

$$\phi(X; \mu, \alpha) = \frac{E(Y(u,.) \exp(\alpha Y(u,.) \mid R(u,.) = 1, X(u,.) \}}{E(\exp(\alpha Y(u,.) \mid R(u,.) = 1, X(u,.))} - \mu$$

gewählt, dann ergibt sich ein *doppelt robuster* Schätzer, der auch dann konsistent ist, wenn die Auswahlgleichung fehlspezifiziert ist. Dabei kann der Koeffizient $\alpha$ von $Y(u,.)$ in der Auswahlgleichung $\pi(u, X, Y)$ nur aufgrund der Beobachtungen mit vollständigen Angaben zu $Y(u,.)$ geschätzt werden, wenn also $R(u,.) = 1$ ist. Da außerdem zu jedem Selektionsmodell auch ein (konditionales) CAR-Modell gebildet werden kann, wenn nur $\pi(u, X)$ flexibel genug gewählt wird, ist $\alpha$ auch in sehr großen Datensätzen kaum stabil schätzbar. Robins und Mitarbeiter haben daher vorgeschlagen, eine Reihe von Werten des Koeffizienten $\alpha$ fest zu wählen und die Auswirkungen auf die Schätzung von $\mu$ zu

Manskis Abschätzungen der Folgen von Selektivität (Manski 1993; Manski und Horowitz 2000; Manski 2003; Zaffalon 2002; Manski und Tamer 2002) erfolgen ähnlich wie die Abschätzungen in Abschnitt 4.5 und verzichten auf probabilistische Annahmen über den Selektionsprozess. Sie führen nur dann zu relativ engen Intervallen, wenn \( Y(u, .) \) beschränkt ist. Im Fall des Haushalteinkommens ergeben sich daher Intervalle für die möglichen Werte der Durchschnitte wie schon im Abschnitt 4.5. Diese Intervalle sind auch die theoretischen Grenzen der Methode von Robins et al., wenn ihr Koeffizient \( \alpha \) alle Werte zwischen \(-\infty \) und \( \infty \) durchläuft (Scharfstein et al. 1999: 1108). Raghunathan hat in der Diskussion der Arbeit von Manski und Horowitz (2000: 86) die Weite der Intervalle beklagt: "'I am afraid that I agree with Cochran (1977) that such an approach is so conservative as to be of little value in most practical settings for inferential purposes.'" Manski und Horowitz antworteten:

> The width of the bounds reflects the information available from the data per se about the population parameters of interest. The width also indicates the relative importance of the data and untestable assumptions in determining the values of point estimates. … Readers should be told that the point estimates are sensitive to untestable assumptions and that different assumptions could produce widely different results. (Mansi und Horowitz 2000: 87f).

### 4.12. Diskussion

Soziologen haben bisher die Möglichkeiten und vor allem die Grenzen von Selektionsmodellen selten zur Kenntnis genommen. Zwar gab es immer wieder Arbeiten in verbreiteten Zeitschriften, die das Thema aufgegriffen haben (Berk und Ray 1982; Stolzenberg und Relles 1997; Winship und Mare 1992). Aber in empirischen Arbeiten wird das Problem
oft vollständig ignoriert, mit einem kurzen Verweis auf MAR-Modelle abgetan oder mit einem Heckman Modell erledigt. Bei Ausschöpfungs
raten von 50% in Umfragen wird man es sich nicht auf Dauer leisten
können, Selektionsmodelle, ihre Grundlagen und ihre Konsequenzen zu
ignorieren.

Nun kann über die Entstehung fehlender Angaben immer nur spekuliert
werden. Man ist auf Informationen einschlägiger Untersuchungen über
Antwortverhalten angewiesen. Existieren aber gruppierte, zensierte oder
fehlklassifizierte Berichte der Befragten, dann müssen diese Angaben
ernst genommen und in die Berechnung von Statistiken einbezogen wer-
den. Es werden Verfahren benötigt, die auch diese partiellen Angaben
berücksichtigen. Selektionsmodelle erlauben es, über die Vorausset-
zungen und Konsequenzen solcher Verfahren nachzudenken. Zudem
sind in diesem Fall Spekulationen über Selektionsprozesse empirische
Grenzen gesetzt und Selektionsmodelle können auch ohne Rückgriff auf
zusätzliche Informationen beurteilt werden. Dazu können Methoden der
Sensitivitätsanalyse einen wesentlichen Beitrag leisten. In dieser Rolle
werden Selektionsmodelle auch für die empirische Sozialforschung an
Bedeutung gewinnen.

4.13. Postscriptum

Selection models have seen a tremendous development in recent years,
including the construction of new models, clever estimation strategies
and theoretical analyses. On a theoretical level, the interplay between
parametric specifications, model completeness, sufficient statistics, and
incomplete data has been clarified by Lu and Copas (2004). Furthermore,
the algorithmic description of CAR models has been completed by Gill
and Grünwald (2005).

The practical methods developed concentrate on imputation, propensity
scores and matching methods, and inverse probability weighted estimators.
Recent developments of imputation methodologies are reviewed by
Tang et al. (2005), Zhang (2003), Harel and Zhou (2007), Horton and
Lipsitz (2001), Kenward and Carpenter (2007), Chen et al. (2000), and
The role of proper imputation rules is discussed by Nielsen (2003a) and Shafer (2003). The possible role of imputation procedures for non-ignorable missing and incomplete data cases is surveyed by Demirtas (2005). Asymptotic and finite sample properties were further examined by Kim (2004). The possible use of imputations for model checking (Gelman et al. 2005 and Abayomi et al. 2008), disclosure control (Little et al. 2004), in binary data models (Demirtas and Hedeker 2007 and Münßich and Rässler 2005) and in structural models (Song and Belin 2004) were explored as well as more special aspects such as the bootstrap (Zhang 2004), sensitivity analysis (Carpenter et al. 2007), local procedures (Aerts 2002) and the effects of outliers (Elliott and Stettler 2007) and of rounding (Horton et al. 2003). There has also been much interest in applications in survey sampling (Beaumont 2005, Haziza and Rao 2005, Kim et al. 2006, Scheuren 2005, Skinner and Rao 2002, and Shao and Wang 2008). There have also been developments of conditional methods (van Buuren 2007), of empirical likelihood methods (Wang and Rao 2002), nearest neighbour methods (Wasito and Mirkin 2006) and of adaptive procedures (Reiter and Raghunathan 2007). There are by now many discussions of practical procedures in the context of well known data sets (Nicoletti 2006 discusses the ECHP, Frick and Grabka 2003 the SOEP). Moreover, there has been some interest of applying imputation techniques to survival data (Rubin and van der Laan 2005 and Hsu et al. 2006, 2007). Comparisons to other methods are provided by Kim and Park (2006) and Carpenter et al. (2006).

variables is taken up by Ichimura and Taber (2001). Kurt et al. discuss applications under non-standard conditions, Lu (2005) presents results about the use of propensity scores with time dependent covariates and McCaffrey et al. (2004) present first methods of boosted regression used in conjunction with propensity score techniques. Finally, Leon and Hedeker (2007) present results on using misspecified propensity scores. The latter problem is closely connected to the discussion of doubly robust estimators as developed by Robins and co-workers. A good recent review of constructions and properties of doubly robust estimators is given by Kang and Schafer (2007) and its discussion. The arguments of Neugebauer and van der Laan (2005) and van der Laan and Robins (2003) on the practical advantages of doubly robust procedures are also relevant here.

Methods of empirical likelihood for incomplete data models have been developed by Qin and Zhang (2007, 2008), Chen and Qin (2006), Leung and Qin (2006), and Wang and Rao (2002). General methods for sensitivity analyses are provided by Hines and Hines (2006, 2007) and Frank and Min (2007).

Progress for interval methods as surveyed by Manski (2003) is rather slow. Horowith and Manski (2006) discussed general identification problems. Dantsin et al. (2006) use interval methods for incomplete data to provide an algorithm for the variance if the interval lengths are small in comparison to the distances between the interval centers. This relates to the results discussed briefly in Section 4.5. Imbens and Manski (2004) suggest confidence bands for interval estimates, but their mix of fixed sample and probabilistic concepts is not too convincing. Blundell et al. (2007) give a very readable application of general interval methodology for the analysis of wage differentials. Scharfstein et al. (2004) provide another example.

There is by now a rather specialised literature treating missing and incomplete covariates in regression models and in censored regression models in particular. Recent contributions include Catchpole et al. (2008), Chen (2002), Chen and Little (1999, 2001), Chen et al. (2007), Didelez (2002), Dupuy and Mesbah (2004), Dupuy et al. (2006), Herring et al. (2004), Huang et al. (2005), Lee and Tang (2006), Liang (2008), Liang et al.
Causal Inference from Series of Events

5.1. Summary

Recent years have witnessed an increased interest, both in statistics and in the social sciences, in time dependent models as a vehicle for the causal interpretation of series of events. The Humean and empiricist tradition in the philosophy of science uses the constant temporal order of cause and effect as a decisive delimitation of causal processes from mere coincidences. To mimic the philosophical distinction, series of events are modelled as dynamic stochastic processes and the precedence of cause over effect is expressed through conditional expectations given the history of the process and the history of the causes. A main technical tool in this development is the concept of conditional independence.

In this article we examine some difficulties in the application of the approach within empirical social research. Specifically, the role of probabilistic concepts of causality and of conditional independence, the nature of events that reasonably qualify as causes or effects, and the time order used in empirical research are considered.
5.2. The Tradition of Causal Analysis in Sociology

The use of the concept of causality in sociology has been lingering between neglect and over-reliance.\(^1\) Even though the concept was never wholeheartedly accepted by sociologists, it became a cornerstone of arguments in favour of empirical research by the early 1970’s. The work of Lazarsfeld, Blalock, Coleman, Duncan, and many others led to the predominance of path models as the paradigmatic form of statistical analyses of causation, which was conceived as a relation between statistical variables. The flowering of structural equation modelling within sociology and psychology strengthened the technical applicability of the approach and gave impetus to the development of statistics in general.\(^2\) But the increased statistical sophistication was accompanied by a growing isolation from the rest of applied social research as well as from statistics (as a branch of applied mathematics). Critical discussions within sociology and across disciplinary boundaries were prematurely cut off. Even the debates in econometrics during the late 50’s and early 60’s\(^3\) on ‘autonomy’ of causes, the meaning of simultaneous equations, and the concept of exogeneity are rarely reflected in the textbooks on social statistics from the late 70’s onwards. Moreover, developments in statistics, even when originating from concerns for questions of causality from other disciplines, were largely ignored.\(^4\) And within sociology, the connection of causal analysis with certain statistical techniques was met by a general scepticism concerning the role of statistics and of ‘variables’ in general.

Sociologists outside the tradition of structural equation models often downplay the role of causality in the social sciences in favour of other forms of determination. In fact, the sociological literature abounds with examples of explanations that are not strictly causal in an empiricist or positivistic sense. Historical, functional, structural, teleological explanations—to name just a few of the distinctions used—are often

---

\(^1\) See Bernert (1983) for an account of its history in the American sociological literature.

\(^2\) See Clogg (1992) for a partial review, including other areas of social statistics.

\(^3\) See e.g. Epstein (1987).

\(^4\) Holland’s discussion of Clogg’s 1992 paper and Clogg’s rejoinder may serve as an illustration.
invoked. As far as causal reasoning is granted a place in sociology, many researchers agree with a view of causality that depends on subject-matter considerations. There seems to be wide consensus among sociologists that causality cannot simply and directly be inferred from empirical data, regardless of whether they are obtained from randomised experiments, collected through ingenious research designs or summarised by particularly advanced statistical models. Blumer’s well known early (1956) diatribe against statistical ‘variable analysis’ is but one example in the sociological tradition asking for more than a statistical analysis of the relation between dependent and independent variables.

More recently, however, statisticians, sociologists, and philosophers have begun to study the relation between statistics and causality more closely. The renewed interest was sparked by advances in the formalisation of concepts related to causation. Most of these developments are well documented in a special volume of Synthese (vol. 121, 1999) and — with emphasis on social science applications — in the proceedings “Causality in Crisis”, edited by McKim and Turner (1997). Most of the contributions in these recent publications concentrate on counterfactual or interventionist conceptions of causation. Complementary, we will here investigate the prospects of a classical empiricist criterion of causality in the Humean tradition: the time order of cause and effect.

5.3. Temporal and Probabilistic Criteria of Causation

Since many theories of causal interdependence rely on empiricist criteria as a prerequisite for the acknowledgement of causality, a strong argument against the assumption of a causal connection can be made if some of the empiricist criteria of causality do not hold. The empiricist conditions for the existence of a causal relation, based on Hume’s analysis, require a) spatial and temporal contiguity, b) constant conjunction between cause and effect, and c) temporal succession. We will argue that suitably modified versions of the requirements b) and c) can be used as starting points for empirical arguments concerning causal claims.
The first criterion, that of spatial and/or temporal contiguity, is often disputed. Effects do not need to follow immediately after a cause, nor do they need to be spatially close to the cause. E.g. strikes and demonstrations may be the (efficient) cause of a change of government, but the latter need not follow immediately after the demonstrations, nor need there be any spatial contiguity. Even though criterion a) cannot generally serve as a necessary condition for causation in the social sciences, a formalism of cause–effect relationships should be able to distinguish between ‘close’ and ‘remote’ causes of effects. Otherwise it would be difficult to express the ideas of ‘spurious cause’ and ‘causal chain’, both considered useful in the construction and criticism of causal explanations.

Condition b) cannot be expected to hold strictly with respect to social interactions. People tend to react differently upon others actions, even in otherwise similar situations. And, as Suppes (1970: 92) notes:

> Empirical studies of the sort done in psychology, sociology, and medicine hold little hope of establishing complete deterministic chains for the causes of actions. This is true whether we are analysing the sex habits of Eskimos, recidivism among parolees over forty, or church attendance by illiterates.

The ‘constant conjunction’ condition therefore must be reformulated. Here we will simply replace the ‘constant conjunction’ condition by a probabilistic relation: that the cause changes the probability of the effect.⁵ Note that “[i]t is not a matter of presenting evidence for causality by offering probabilistic considerations but it is part of the concept itself to claim relative frequency of co-occurrence of cause and effect” (Suppes 1970: 45). The incorporation of probabilistic aspects is thus not only of an epistemic nature. It is not incorporated because the sociologist does not (yet) know, but because ‘constant conjunction’ cannot reasonably

---

⁵ Suppes’ (1970) analysis starts from the assumption that causes should increase the probability of the effect. The subsequent discussion in the philosophical literature has shown that the concept of ‘increase of probability’ may not suffice for the analysis of probabilistic causality (e.g. Eells 1991). But since the problem is not central to our discussion we will content ourselves with the simple minded principle of a change of probability.
be claimed in sociology. Consequently, probability statements in this context should generally refer either to frequencies or to propensities, as also fits the needs of an empiricist program.

The condition c) requires a temporal framework for causal arguments. This is not normally included in the formulation of ‘causal models’ in the structural model literature. It is also not included in the formal representation of observational studies\(^6\) in e.g. biometry nor in more recent counterfactual or interventionist accounts. On the other hand, the temporal dimension of causal connections has often been recognised in sociology. Tuma and Hannan, introducing event–history analysis into sociology, acknowledge the interplay between temporal and causal analysis (1984: xi–xii, their italics):

Any attempt at forging a systematic framework for the empirical study of social change must confront two issues. One involves the development of *dynamic* models—models that describe the time paths of change in phenomena. The other involves the development of *causal* models—models that describe how change in some properties induces change in still other properties. …we rely heavily on the use of formal models to guide attempts at testing hypotheses about the processes and causes of change.

And Blossfeld and Rohwer (1995: 20) state:

…the important task of event history modelling is …to establish relevant empirical evidence that can serve as a link in a chain of reasoning about causal mechanisms. In this respect, event history models might be particularly helpful instruments because they allow a time-related empirical representation of the structure of causal arguments.

\(^6\) The definition of an observational study is a rather narrow one in this context. Rosenbaum (1995: 1) states: “An observational study concerns treatments, interventions, or policies and the effect they cause, and in this respect it resembles an experiment. A study without a treatment is neither an experiment nor an observational study.” Following his definition, many empirical sociological studies will not count as observational studies.
Suppes (1970) proposed one of the best known formalisations of probabilistic causality. His account includes a condition of temporal precedence of cause over effect, in contrast to many later attempts to clarify the concept (e.g. Eells, 1991). His starting point is the theory of probability based on systems of sets, called ‘events’. He adjoins a temporal indicator to these sets to indicate temporal sequences. But as witnessed by later contributions (e.g. Davis 1988), this strategy turned out to be rather limited. Savage (1972: 10) remarked with respect to the use of the term ‘event’ in probability theory: “...the concept of event as here formulated is timeless, though temporal ideas may be employed in the description of particular events.” Arjas and Eerola (1993: 384) note that in these formulations, “time is present (if at all) only as an index, distinguishing between what comes ‘before’ and what comes ‘after’.” Consequently, a more flexible representation of time order is needed. In the absence of condition a), it should at least be possible to formulate the timing of effects with respect to causes. Recent contributions therefore seek to enrich the formulations by borrowing heavily from dynamic theories of stochastic processes.

The program then is clear: to combine probabilistic models — re-expressing the constant conjunction (condition b) — with the idea of ‘the cause precedes the effect’ (condition c) to facilitate an empirical assessment of claims of causality. Causal statements are translated into the mathematical language of stochastic processes: \( Y = \{ Y_t, t \in T \} \) is a stochastic process with values in a finite set \( \mathcal{Y} \). The values of \( Y_t \) are interpreted as properties of units under study and events are changes of properties at time points \( t \). At any given point in time \( t \), the description of the evolution of the process may depend on conditions and events that occurred in the past, i.e. before \( t \), but not on what is the case at \( t \) or in the future, after \( t \). Causes acting on \( Y \) may then be introduced by considering them as changes in a further process \( X = \{ X_t, t \in T \} \). The process \( X \) may be treated as a time-dependent covariate and causal statements are therefore formulated as probabilistic relations between two (or more) stochastic processes. A time-dependent covariate records when a causal factor has changed its state. It signifies that an event of kind \( \mathcal{X} \) has taken place. Consequently, we would not say that a process \( X \) is a cause of a process \( Y \), but that a change in \( X \) at time \( t \) (an event at time \( t \) ) could be a
cause of a change in $Y$ at time $t'$, $t' > t$ (another event at a later time).

Often a canonical dynamic description of the stochastic processes can be given, one relating the ‘past’ of the processes to their ‘future’, and encapsulating their relevant probabilistic features. In this case the classical formulation of probabilistic causality (e.g. Suppes 1970) can be enhanced by allowing for an explicit representation of the timing of effects, generalising the Humean requirement of contiguity in time. Approaches along this line were advocated by Granger (1969) in the context of time series analysis and by Schweder (1970, 1986) for general Markov processes. These ideas were taken up more recently by Aalen (1987), Arjas/Eerola (1993), Parner/Arjas (1999), and Keiding (1999), providing a formal framework for probabilistic causality with a clear relation to time order. Many of these articles use genuinely epistemic notions for the interpretation of probabilities. This partly reflects the naming conventions used in the technical literature on stochastic processes and we will follow the convention here. But the mathematical formulation does not force us to accept an epistemic interpretation, and the possibility of interpretations in terms of propensities or frequencies should be kept in mind.

5.4. Mathematical Models

A dynamic description of stochastic processes fitting the above program can be outlined in the case of processes with discrete time parameters: \(^7\)

\(^7\) Much of the discrete-time theory extends directly to the continuous-time setting. The reason is that a continuous time martingale with respect to a right continuous filtration can be modified to have nice sample path properties, i.e. right-continuous paths with left hand limits. A thorough treatment presupposes a formidable technical machinery without adding much insight to the present discussion. But it should be noted that the continuous-time theory makes heavy use of continuity and of the denseness of the rationals within the real numbers. Many mathematical models of time try to avoid such strong assumptions (see Whitrow 1963: Chap. III for an early review). Moreover, the combination of a continuous-time theory using all the properties of the reals may collide with a concept of causality that is based both on the time ordering of cause and effect and on the distinction between direct and indirect causes (see e.g. Suppes 1970: 72).
Let \( Y = \{ Y_t, t = 0, 1, 2, \ldots \} \) be a stochastic process with values in a finite set \( \mathcal{Y} \).\(^8\) Suppose one is interested in what happens just after time \( t - 1 \). A good prediction is the conditional expectation of the change of \( Y_t \) from \( Y_{t-1} \), conditional on the previous history of the process, that is the random variable

\[
V_t = \mathbb{E}(Y_t - Y_{t-1} | Y_0, \ldots, Y_{t-1})
\]  

(5.1)

Putting

\[
U_t = \sum_{s=1}^{t} V_s = \sum_{s=1}^{t} \mathbb{E}(Y_s - Y_{s-1} | Y_0, \ldots, Y_{s-1})
\]

for the sum of the predicted values, one may write the original process \( Y_t \) as a sum of predictions in time and a remainder, the Doob decomposition:\(^9\)

\[
Y_t = Y_0 + U_t + M_t
\]

(5.2)

The prediction part \( U_t \) is a function of the previous history up to and including time \( t - 1 \) only, while \( M_t \) is a martingale, satisfying

\[
\mathbb{E}(M_t | Y_0, \ldots, Y_{t-1}) = M_{t-1}
\]

(5.3)

In fact, for the martingale difference \( M_t - M_{t-1} \) one finds

\[
\mathbb{E}(M_t - M_{t-1} | Y_0, \ldots, Y_{t-1}) = \mathbb{E}((Y_t - Y_{t-1}) - (U_t - U_{t-1}) | Y_0, \ldots, Y_{t-1}) = \mathbb{E}(Y_t - Y_{t-1} | Y_0, \ldots, Y_{t-1}) - V_t = V_t - V_{t-1} = 0
\]

The decomposition (5.2) generalises the additive regression decomposition into a ‘structural’ part \( U_t \) and a ‘random’ part \( M_t \) that is used in much of empirical social research. Accordingly, but somewhat ambiguously, the differences \( M_t - M_{t-1} \) are sometimes called the innovations

---

\(^8\) The assumption of a finite state space is not essential for the mathematical formulations used here. But subsequent discussions of the appropriateness of the formal model often presuppose a finite number of states.

\(^9\) Williams (1991) provides a thorough and vivid exposition for the discrete time case.
of the process. The predictions $U_t$, depending only on $Y_{t-1}, \ldots, Y_0$, are called predictable. Any other process $Z_t$ that depends only on the values of $Y_{t-1}, \ldots, Y_0$, the history of $Y$ strictly before $t$, will also be called predictable with respect to the process $Y$.

It may be instructive to see how a duration variable $T$, featuring prominently in event-history analysis, fits into the present framework. In that case one may put $Y_t = 1$ if the event happened at time $t$ or before, 0 otherwise. That is, $Y_t = \mathbb{I}[T \leq t]$, where $\mathbb{I}[A]$ is the indicator variable of the event $A$. For simplicity, one may also assume $Y_0 = 0$. The prediction process is then given by $V_t = \mathbb{E}(Y_t - Y_{t-1} \mid Y_0, \ldots, Y_{t-1})$. This quantity is 0 if $Y_{t-1}$ takes the value 1, since then both $Y_t$ and $Y_{t-1}$ must be 1. In other words: If the event happened before time $t$, there will be no change in the prediction of $Y_t$, because the one possible change in state is known to have occurred.

On the other hand, if there was no event at $t - 1$ or before, $Y_{t-1}$ as well as $Y_{t-2}, \ldots, Y_0$ are 0 and $V_t$ reduces to the probability $Pr(Y_t = 1 \mid Y_0 = 0, \ldots, Y_{t-1} = 0) = Pr(T = t \mid T \geq t)$. Thus $V_t$ reduces to a random function of the well known hazard function of $T$. If $\lambda(t) = Pr(T = t \mid T \geq t)$ denotes the hazard function of $T$, $V_t = \mathbb{I}[T \geq t] \lambda(t) = (1 - Y_{t-1}) \lambda(t)$. Note that $V_t$ in the present discussion is a random variable, depending on $Y_{t-1}, \ldots, Y_0$. In contrast, the hazard function $\lambda(t)$ treated in most texts on event–history analysis is a non-random transform of the distribution function. Furthermore, the accumulated prediction $U_t$ is a random sum of hazard functions, the sum extending over all $s \leq \min(T, t)$, the times $s$ before the event time $T$ or the observation time $t$, whatever comes first.

Following Aalen (1987), a dynamic statistical model is then defined as a parameterisation of the prediction increments $V_t$. Since $V_t$ depends on the history of the process up to and including $t - 1$ only, it can be interpreted as a description of the future, the likely events at time $t$, depending only on the knowledge of all past events. Alternatively, it is (an approximation of) the relative frequency of events at $t$ among all sequences with this history.

To introduce concepts of interdependence between several processes, it seems natural to embed the above concepts into a multivariate extension.
Basicallly, the history of the single process is replaced by one based on all relevant information available before time $t$. In the case of two processes $(Y_t, X_t)$ one might therefore define the conditionally expected increments as

$$V_t^Y = \mathbb{E}(Y_t - Y_{t-1} \mid Y_0, X_0, \ldots, Y_{t-1}, X_{t-1})$$  \hspace{1cm} (5.4)

and

$$V_t^X = \mathbb{E}(X_t - X_{t-1} \mid Y_0, X_0, \ldots, Y_{t-1}, X_{t-1}).$$  \hspace{1cm} (5.5)

Interpreting the conditional expectations above as an increase in knowledge, $V_t^Y$ will be based not only on the pre-$t$ history of $Y$ itself, but also on the knowledge of the development of $X$ up to and including $t - 1$. Thus, the expectation will change depending on the information provided by $X$. Symmetrically, $V_t^X$, the prediction of $X$ based on the common history of $X$ and $Y$ before $t$, will depend on the changes in $Y$ up to $t$. One may represent the two processes using the Doob decomposition with respect to the joint history of the processes as:

$$Y_t = Y_0 + \sum_{s=1}^{t} V_s^Y + M_t^Y \quad \text{and} \quad X_t = X_0 + \sum_{s=1}^{t} V_s^X + M_t^X$$  \hspace{1cm} (5.6)

Then $X_t$ is defined not to be causal for $Y_t$ in Aalen’s sense if and only if, first, the prediction errors $M_t^Y$ and $M_t^X$ are uncorrelated, and second, $U_t^Y$, the prediction of $Y_t$, may depend on $Y_0, \ldots, Y_{t-1}$ but not on $X_0, \ldots, X_{t-1}$. We will say that $X$ and $Y$ are locally autonomous if the first condition is satisfied.\(^{10}\) If the second condition holds, $Y_t$ is said to be locally independent of $X_t$. Otherwise, $Y_t$ is said to be locally dependent of $X_t$.\(^{11}\)

---

\(^{10}\) The notion of ‘autonomy’ has a long tradition in econometrics, especially in the context of simultaneous equation models. Aldrich (1989) provides a review.

\(^{11}\) A related concept is Granger non-causality, a concept often used within econometric time series analysis. A process $X$ is said not to cause $Y$ in Granger’s sense at $t$ if $Y_t \perp (X_{t-1}, \ldots, X_0) \mid Y_{t-1}, \ldots, Y_0$, i.e. where $Y_t$ and the pre-$t$ history of $X$, $(X_{t-1}, \ldots, X_0)$ are conditionally independent given the pre-$t$ history of $Y_t$ alone. Note that in the context of time series analysis the processes $Y_t$ and $X_t$ need not refer to ‘events’. The concept has been explored and extended in a series of papers by Florens and Mouchart (1982, 1985).
The condition of local (in-)dependence is asymmetric in the two processes. Indeed, in the example of two duration variables $Y_t = \mathbb{I}[T_1 \leq t]$ and $X_t = \mathbb{I}[T_2 \leq t]$, the process $X_{t-1}$ may enter as a time dependent covariate in the prediction (stochastic hazard) of the other, but not the other way around. This is in accord with the basic asymmetry of causal relations and also respects the notion of ‘cause precedes effect’.

The condition of local autonomy is introduced to ensure a certain autonomy of the two processes. When satisfied it is possible to envisage a change in the behaviour of one process after time $t$ without a change in its local relation to the other process or a corresponding immediate change in the other process. This should rule out processes that are merely related by definitions or ‘rules of the game’. Consider for example two gamblers throwing dice. Denote by $X^1$ ($X^2$) the result of a throw of the first (second) player. If the throws are independent, then also the prediction errors are independent and $X^1$ and $X^2$ are autonomous. But one may also look at the result of the play, a win, a draw, or a loss, for each player. Let $Y^1 = \mathbb{I}[X^1 > X^2] - \mathbb{I}[X^2 > X^1] \in \{-1, 0, 1\}$. Then $Y^2 = -Y^1$ and the processes are clearly not autonomous.
5.5. Statistical Methods

An empirical strategy to show local dependence of $Y$ on $X$ at $t$ is then to show that the prediction process $U_t^Y$ with respect to the joint history of the process is a non-constant function of $X_{t-1}, \ldots, X_0$. In the case of simple duration models, $V_t^Y = \mathbb{I}[Y \geq t] \lambda_\theta(t; X_{t-1}, \ldots, X_0)$. That is, the process $X$ appears as a time-dependent covariate in the hazard function for $Y$ at $t$, which is assumed to be parameterised by some $\theta \in \Theta$. One therefore needs to show that the pre-$t$ history of $X$ changes the hazard of $Y$ at $t$. Often it is possible to factor the likelihood of $(X, Y)_t$ for $\theta$ in such a way that only the part $U_t^Y(Y_t \mid Y_{t-1}, X_{t-1}, \ldots, Y_0, Y_0; \theta)$ figures in the computation of statistics. This is called a partial likelihood for $\theta$ since it does not depend on the specification of the joint distribution of $(Y, X)_t$. In particular, a model for the covariate process $X$ need not be specified. This allows for an attractive strategy to demonstrate local dependence since one can concentrate on the model for the conditional prediction of $Y$ given $X$ without worrying about the possibly complicated nature of $X$. In the context of counting process methods, Slud noted (1992: 97):

... that inferences could be made successfully without parametric specification of any probabilistic objects other than the failure counting process intensities. The latter point of view is especially liberating in problems with randomly time-varying covariates, where one is usually interested only in the effect of the covariates on the hazard of failure and where one can usually not provide convincing models of the stochastic variation of the covariates over time.\(^{12}\)

5.6. Events and their Descriptions

The formal frame of causal reasoning developed above has many merits with regard to dynamic formulations of causality. Despite some conceptual shortcomings in the reformulation of the ‘constant conjunction’\(^{12}\)

\(^{12}\) Arjas/Haara (1984), Slud (1992) and Greenwood/Wefelmeyer (1998) have studied the statistical properties of factorisations with time-dependent covariates.
condition it may readily be used to formulate claims of (non-) causality. As it stands, it relates to stochastic processes, i.e. collections of random variables. Often, statisticians are satisfied with formal references to variables as causes and effects, especially when arguing in the tradition of structural equation models. E.g. Pearl and Verma (1992: 91) speak of “stable causal mechanisms, which on a microscopic level, are deterministic functional relationships between variables, some of which are unobservable.” But neither variables nor what they stand for are generally admitted as causes or effects by social scientists. Background variables like sex or religion are not considered to be (representations of) possible causes.

One further prerequisite for the applicability of the above formalism in the social sciences is therefore a restriction on the entities that may be causes or effects. As Bunge (1963: 72, his italics) puts it, “there can be no causal links among states, nor among any other systems of qualities. States are not causes, but simply antecedents of later states. To regard states as causes amounts consequently to committing the fallacy of the post hoc, propter hoc.” Thus neither states nor things nor qualities of things can be causes or effects, only events can.¹³

But what are events? Hacker (1982: 17) says that “[e]vents, unlike objects, are directly related to time. They occur before, after, or simultaneously with other events. They may be sudden, brief or prolonged…None of these temporal predicates apply in the same way to objects.” But this special connection with time implies a difficulty for probabilistic theories of causality: “it is manifest that no event ever happens more than once, so that the causes and effects cannot be the same in all respects.”¹⁴ Therefore one cannot speak about the constant conjunction of cause and effect unless it is possible to also speak of kinds of events. While an event is something unique, events of the same kind can occur several times. But how can one define kinds of events? One possibility would be to delete some temporal descriptions from propositions about events. Such propositions might then be said to be about kinds of events.¹⁵ The

¹³ But see e.g. Mellor (1995: 129), who argues that “causation mostly links facts…. So no causation would be lost even if there were no particular events.”

¹⁴ Maxwell, cited in Bunge 1963: 50

¹⁵ See e.g. Scheffler 1994.
obvious shortcoming of this approach is that it concerns propositions, not events, and propositions do not qualify as causes, at least not in a realist account of causality.

Events presumably are not linguistic entities; like trees and molecules, events can be talked about, referred to, and described but they are not themselves statements, sentences, descriptions, or any other kind of linguistic units. Nor are events propositions; propositions are supposed to be abstract entities, whereas events are spatio-temporally bounded particulars. (Kim 1969: 198.)

An alternative is to relate events to changes in things. An event is a ‘movement’ by an object from the having of one to the having of another property, where those properties belong to the same quality space, and where those properties are such that the object’s successive havings of them implies that the object changes non-relationally. (Lombart 1986: 114)\(^{16}\)

This fits nicely with the proposed mathematical formalisation: The random variables \(Y_t\) refer to properties of things. These properties are represented by a definite set \(\mathcal{Y}\), the ‘property space’. And events are changes in things represented by variables, \(\{Y_t - Y_{t-1} \neq 0\}\). Events occur at a definite point in time.\(^{17}\) Kinds of events may than be described by certain transitions between elements of the set \(\mathcal{Y}\), or as classes of such transitions.

But this approach may be at once too general and too specific to serve its purpose as a general guideline in the social sciences. It may be too specific because the translation of ‘event’ into ‘change of property’ does

\(^{16}\) He argues that all events should be treated in this way.

\(^{17}\) Since an event in general takes some time, it seems inappropriate to say that they occur at a point in time. This creates considerable problems for a formalisation of time sequences, especially if it is based on the continuous-time theory of stochastic process. Hamblin argues that “the time-continuum, modelled on the real numbers, is richer than we need for the modelling of empirical reality.” (cited in Galton 1984: 19)
not capture the most general idea of event playing the role of an efficient cause. Parties, strikes, wars etc. are certainly events, and they generally are considered to have causal efficiency. Still, the notion of an event as a change of a property can only be adapted to such events at the price of some distortion of the event under study.

On the other hand, the notion of events and their probabilistic interdependence in time does still not capture the realist notion that causes should reflect mechanisms (Sørensen 1998), capacities (Cartwright 1989) or productivity (Bunge 1963). Events as changes of state may not involve mechanisms. They may only be sequences like sunset after sunrise, so that the concept of events as changes of properties of things is too general.

5.7. Agents, Actions, and Events

In the words of Bunge (1963: 46, his italics), “the reduction of causation to regular association, as propounded by Humeans, amounts to mistaking causation for one of its tests … . What we need is a statement expressing the idea—common to both the ordinary and the scientific usage of the word—that causation, far more than a relation, is a category of genetic connection, hence of change, that is a way of producing things, new if only in number, out of other things.” In statistical discussions, the exhibition of productivity of proposed causes is often side stepped. Instead, many accounts view causality through an analogy with planned, isolated experiments. Experiments are seen as a deliberate manipulation of causes with the goal to provide a magnitude of their effects. This magnitude is perceived as the difference between the value of a measurement on a subject in the presence of the cause and the value of the measurement on the same subject in the absence of the cause. The difference can never be observed and so relates to a counterfactual question. The theory therefore involves constructions of ‘similar worlds’ to identify such magnitudes.  

---

18 See Holland (1986), Pratt/Schlaifer (1984), Dempster (1990), Rubin (1990), Galles/Pearl (1998), Pearl (1999) and Robins (1999) for discussions and refinements. These approaches are closer in spirit to J.S. Mill’s attempts to codify methods of causal inquiry (Holland 1986: 950) than to elaborations of Humean criteria for the existence of causal links.
Since all the criteria for deriving magnitudes of effects rest on empirically untestable assumptions, they are met with scepticism from statisticians (e.g. Dawid 2000) and sociologists alike. Furthermore, counterfactual accounts are deterministic in that they refer to what would necessarily happen in the presence or absence of the cause. But such a deterministic outlook cannot easily adapt to the variability generally observed in the social sciences.

On the other hand, the insistence on the experimental analogy points to the importance of action based interpretations of causality. In fact, it is sometimes suggested that causality should be defined in terms of human actions and their impact on other humans. Von Wright (1972: Chap. 2.9) distinguishes between doing something and bringing about something and goes on to define \( P \) as a cause relative to \( Q \), and \( Q \) as an effect relative to \( P \), if and only if by doing \( P \) one could bring about \( Q \), or by suppressing \( P \) one could prevent \( Q \) from happening. Such a view partly reconciles Bunge’s search for productivity with counterfactual analysis: The capacity of a human agent to act and thereby to bring about certain events can hardly be denied. And this capacity includes the possibility of deliberately abstaining from that action. But the power to act and thereby to bring about an event is normally understood to mean that, counterfactually, if the agent would not have acted as, in fact, he did, then the event would not have happened.19

Many social phenomena are directly based on actions of individuals or organisations (see e.g. Blossfeld/Prein 1998). As far as sociology is concerned with these phenomena, there is no need to refer to an omnipotent experimenter or to seek rescue in designs that—always imperfectly—mimic the experimental setup of other sciences. Within these fields, sociology does not deal with associations among variables per se, but with events brought about or done by acting people. And claims for causal connection among events brought about by agents can be based on the causal capacities of the agents themselves.

19 Kelsen (1982) argues that the notion of cause and effect originated from idealised human action and reaction in society, that its origins lie in the projection of crime and punishment, guilt and retaliation, onto nature. An action based reasoning about causal connection would therefore be close to the ancient origins of the idea, but without projecting human capacities on God or nature.
It is tempting to seek the causal connection directly in sequences of actions. But actions should not be treated like events that enter into causal relations as causes and effects. There cannot be a similar connection between actions. Otherwise, as Alvarez/Hyman (1998) point out, one would be led to the idea that agents cause their actions, that actions are events caused by agents. But then “an agent who performs one action performs an infinite series of actions: he causes his action; he causes the causing of his action; he causes the causing of the causing of his action; and so on.” (p. 222) We will therefore say that agents cause the result of their action, that they bring about events and that causal connection exist, not between actions, but between events done or brought about by actions.

As Bach (1980) points out, the distinction between actions and events also relieves us from specifying times and places for actions. “Once we have specified all the relevant events in the act sequence and have described them as stemming from a mental episode in the way appropriate to action, we have said all we need to say about which actions were performed and what the agent did.” (p. 118)

It is sometimes argued that since human actors act intentionally and behaviour is goal-oriented, the intentions or motives of actors to bring about some effect in the future causes the actor to behave in a specific way in the present. Marini and Singer (1988: 377) say that

[a] major problem with use of the criterion of temporal order in which behavior occurs, or in which events resulting from behavior occur, is often not a good indication of causal priority. Because human beings can anticipate and plan the future, much human behavior follows from goals, intentions, and motives; i.e., it is teleologically determined. As a result, causal priority is established in the mind in a way that is not reflected in the temporal sequence of behavior or even in the temporal sequence of the formation of behavioral intentions.

But the connection of goals, intentions, and motives to acts and events seems to be much looser than Marini and Singer suggest. In fact, von
Wright (1972: Chap. 3) argues that in the ‘practical syllogism’ of the form: a) person \( P \) wants to achieve \( Y \); b) \( P \) believes that \( Y \) can be brought about when he is doing \( X \); c) Therefore \( P \) tries to do \( X \); the antecedences a) and b) cannot be understood as causes of the person’s doing \( X \). Based on observations of goals, intentions, and motives one may try to give a teleological explanation of behaviour. But such explanations are not causal, and they can coexist with causal connections between events that are done and brought about by agents. The fact that social agents can behave intentionally, based on expectations, does not reverse the time order underlying causal statements. The explanandum envisaged by Marini and Singer—why a certain person acts as she chooses to act—is simply different from the explanandum of causal analysis.

5.8. Conceptual Problems

In summary, we propose to investigate causal relations among events employing the concepts of local independence and local autonomy in those cases where one is concerned with events brought about or done by agents. However, the execution of such a program is hampered as well by technical as by philosophical problems. The latter concern the basic concepts of independence and autonomy themselves and we will exhibit some of the more disturbing aspects below.

5.8.1. Local Independence

In his “Foundations of the Theory of Probability” Kolmogorov (1950: 9) says that

one of the most important problems in the philosophy of the natural sciences is—in addition to the well-known one regarding the essence of the concept of probability itself—to make precise the premises which would make it possible to regard any given real events as independent.
The same applies, we think, to the concept of local independence or similar attempts to provide models for causal independence. Local independence cannot be demonstrated from observations alone. Even though conditional independence as well as local independence are not empirical concepts, it does not follow that they cannot be used in an empiricist program for the assessment of causality. They are needed as regulative ideas in modelling. The role of local and conditional independence is to suggest the kind of relations one needs to take into account, but not to describe the likely results of an investigation.

Second, observed relations between stochastic processes generally depend on the number of processes that are considered. If a further process is included, the local dependence between all processes may change. The theoretical background on which an analysis is grounded will to a certain extent determine the variables and histories to be considered in an analysis. In the words of Suppes

> It is important to emphasise … that what is to be taken as background or field will always be relative to the conceptual framework under discussion. In one theoretical approach to the causal analysis of phenomena, the field will include only the consideration of macroscopic bodies and their characteristics, but in another, it will go deeper and consider as well atomic objects and their properties. (1970: 74)

In this sense, there may exist several valid causal analyses based on different sets of stochastic processes. Arguments for the exclusion of certain processes will partly rely on ideas of causal non-dependence, which can be translated into local independence within the mathematical model. On the other hand, the theoretical background will rarely be specific enough to determine exactly what processes are to be considered. In consequence, results cannot be expected to be unique.

Third, and perhaps most disturbingly, the probabilistic concept of local independence does not fully conform with most notions of ‘explanation’.

---

20 A similar point has often been made in connection with the role of conditional independence in structural equation models. See e.g. Sobel (1997) and Holland in his discussion of Clogg (1992: 199).
There may be two different histories $H_1 = \{Y_{t-1}, A_{t-1}, \ldots \}$ and $H_2 = \{Y_{t-1}, B_{t-1}, \ldots \}$ that make the respective predictions for $Y_t$ locally independent of $\{X_{t-1}, \ldots, X_0\}$, showing that $X_t$ is at most an indirect cause of $Y_t$. But neither $H_1 \subseteq H_2$ nor $H_2 \subseteq H_1$ need to hold so that explanations (of spuriousness) are not unique.

Perhaps even worse, it may happen that $Y_t$ is locally independent of $X_t$ with respect to a history $H_1 = \{Y_{t-1}, A_{t-1}, \ldots \}$, but that it ceases to be so with respect to a larger history, say $H_2 = \{Y_{t-1}, A_{t-1}, B_{t-1}, \ldots \}$. Including more information for the prediction of $Y_t$ might destroy local independence. The work of Clogg and Haritou (1997) contains some valuable examples.

Finally, local dependence is not antisymmetric. Both processes may be locally dependent on each other. In this respect, the concept is weaker than many accounts of causality would require. On the other hand, if $X$ and $Y$ are mutually locally independent, then under slight regularity conditions (including uncorrelated innovations) $X$ and $Y$ are stochastically independent (e.g. Schweder 1979: 404). But stochastic independence is a much stronger concept than causal unrelatedness would seem to require.

5.8.2. Autonomy

The principle of local autonomy was introduced to ensure that the processes under consideration are not just expressions of a sole underlying process, so that it is meaningful to assess the properties of one process without regard to the other. The condition is formulated in terms of the uncorrelatedness of the martingales $M_t^X$ and $M_t^Y$, expressing the intuitive notion that what happens next to $X$, say, should not be directly related to what happens to $Y$ at the same time. The condition excludes two stochastic processes that are functionally related. In that case it may well be that the first process is locally independent of the second.

---

21 See Dawid 1979a, 1979b, 1980 for some examples concerning the ‘unexpected’ behaviour of conditional independence relations.
while the second is locally dependent on the first, but it would contradict common sense to say that the first process is a cause of the second.\footnote{See Aalen (1987: 188) for an example.}

But the condition fails in deterministic situations: Granger (1969: 430) used two deterministic processes as an example: If $Y_t = bt$ and $X_t = c(t + 1)$, then $V_t^Y = b$ (independent of $X_0, \ldots, X_{t-1}$). But one can as well write $V_t^Y = (b/c)(X_{t-1} - X_{t-2})$, which is dependent on $X_0, \ldots, X_{t-1}$. Certainly one would not like to call $X_t$ a cause for $Y_t$ even though the martingales corresponding to the two processes are trivially uncorrelated.

Second, in the context of counting process models, the assumption of autonomy is often replaced by the assumption of no common jumps of the two processes. This in turn implies that the martingales $M_t^Y$ and $M_t^X$ are uncorrelated and that $M_t^X M_t^Y$ is again a martingale (Fleming/Harrington 1991: 75). The condition of no common jumps is often easy to handle, but it may obscure somewhat the role of the condition. Schweder (1970), who starts with a common (Markov–) process with state space $\mathcal{X} \times \mathcal{Y}$, uses the assumption of no common jumps in $X$ and $Y$ explicitly as a condition for the existence of processes that can formally be partitioned into processes $X$ and $Y$.

On the other hand, the condition of no common jumps is often used for quite a different purpose: It can justify the construction of partial likelihoods. But the statistical considerations that lead to the use of only $U_t^Y(Y_t \mid Y_{t-1}, X_{t-1}, \ldots, Y_0, X_0; \theta)$ for likelihood construction and statistical inference should be carefully distinguished from considerations of the role of the two processes within a causal connection. If one is only willing to specify $U_t^Y(Y_t \mid Y_{t-1}, X_{t-1}, \ldots, Y_0, X_0; \theta)$ for statistical purposes one cannot analyse or simulate the dynamics of the compound process even if one might be willing to impute values for all $X_t$. This point was stressed both by Strotz and Wold (1960) and Cox (1992). Solving the estimation problem by concentrating on only a part of the system, even if justified, need not suffice to answer causal questions.
5.9. Conclusions

The discussions on causality, whether originating from a statistical perspective or from the methodology of the social sciences, have only rarely reflected the philosophical insight that a causal connection is a relation between events but not between variables, things, or qualities. Similarly, an agent based theory of causes that suggests itself in many parts of sociology was largely ignored in favour of counterfactual or system theoretic accounts. We have argued here that an agent based idea of causal connections between events can be supplemented by dynamic descriptions of series of events. When series of events of different kinds are represented by autonomous stochastic processes, the absence of a causal connection can be explicated by the concept of local independence. These concepts should be useful at least in those areas of empirical social research that are directly concerned with events brought about by agents.

We have not treated here a problem of central importance: the problem of spurious causes and of confounding. While the dynamic characterisation of series of events seems to allow for a better understanding and a more flexible formulation of these rather intuitive concepts (e.g. Parner and Arjas 1999), the variety of background conditions and situations generally encountered in social research may well preclude a comprehensive theoretical treatment of confounding. An examination of earlier attempts of demonstrating non-spuriousness in sociology, similar to Goldenberg’s (1998) article, will certainly enrich further theoretical developments.

5.10. Postscriptum

The preceding paper argues for an approach to causality that characterises causal relations as non-necessary relations between events brought about by agents. The non-necessity which has been advocated forcefully by Anscombe (1971) and others, is here embodied in a probabilistic and dynamic relation between events. And the notion of capacity (Cartwright 1989) or productivity (Bunge 1963) is in a social science context situated in the causal capacities of agents.
Although it was thought that this notion of causality, while rather weak, would nevertheless provide a useful conceptual framework for at least some parts of empirical social research, progress in the clarification of the open problems indicated at the end of the paper has been very slow. Most of the progress, however, is intimately connected with the main theme of the present work, namely the understanding and modelling of incomplete data. This is not too surprising given that the main technical ingredient of the approach depends on dynamic and probabilistic descriptions of events that did not yet occur.

But before I discuss some of the newer results it is necessary to give at least a brief account of a very different notion of causality that not only dominates much of the discussion in philosophy, the social sciences and in statistics, but is also strongly connected to the problem of incomplete data, though in quite a different way. The approach is commonly termed the counterfactual approach to causality. It aims to make precise formulations like: Suppose both $X$ and $Y$ happen. Then $X$ is a cause for $Y$ if, counterfactually, had $X$ not have happened, then $Y$ would not have happened. One may similarly insert “is the case that” or “obtain” or “occur” for “happen” in the above formulation and treat the $X$ and $Y$ as facts or propositions or states of affairs instead of events. The philosophical discussion of the counterfactual approach has explored these and several other possible causal relata. But within the statistical formulation (which is sometimes called the potential response approach) a causal relation is a relation between variables and not one between events or facts or propositions.

The literature on this notion of causality has grown tremendously during the past few years. To provide a reasonable overview of all aspects of the discussions would require not only a departure from the main arguments of the previous paper but also the introduction of further technical machinery. It should suffice here to mention a few books and review articles that take up most of the current topics discussed in conjunction with this notion of causality. The potential response approach was presented by Pearl (2000). His book features several possible approaches within the counterfactual tradition in econometrics and the social sciences. A much more concise discussion is given in his

The form of the counterfactual approach most often invoked in statistical analyses directly posits (statistical or random) variables as causal relata. It therefore allows a direct connexion with classical statistical methodology and in particular to methods developed for incomplete data. The connexion, very briefly, is this: Let $X$ denote a ‘cause’, i.e. a random variable with values in \{0, 1\}. Also, let $(Y_0, Y_1)$ denote a pair of random variables defined on a common probability space with $X$, say $\Omega$. They will serve to represent ‘effects’ of the ‘cause’ $X$. Thus, there is a function

$$(X, Y_0, Y_1): \Omega \rightarrow \{0, 1\} \times \mathcal{Y} \times \mathcal{Y}$$

The interpretation is that $Y_0$ is ‘observed’ if the ‘event’ \{X = 0\} ‘occurs',

268
while $Y_1$ is ‘observed’ if the ‘event’ $\{X = 1\}$ ‘occurs’.\(^{23}\) Note that the distribution of $Y_0$ is not the conditional distribution of some variable $Y$ given the ‘event’ $\{X = 0\}$. $Y_0$ is defined (‘exists’) whether or not $\{X = 0\}$ ‘occurs’, and the same applies to $Y_1$. The pair $(Y_0, Y_1)$ always ‘exists’ jointly.\(^{24}\)

Identifying causal relata with (statistical or random) variables and adopting the language of counterfactuals (including closest possible worlds etc.), the statistician can proceed using statistical techniques developed for incomplete data problems. It is stipulated that observations are coarsened versions of the triple $(X, Y_0, Y_1)$ so that ‘observations’ are given by the coarsened variables

$$(X, Y^*): \Omega \longrightarrow \{0, 1\} \times ((\mathcal{Y} \times \{Y\}) \cup (\{Y\} \times \mathcal{Y}))$$

I.e., one either observes $Y_0$ (if $X = 0$) or $Y_1$ (if $X = 1$) but never both. Now the ‘effect’ of the ‘cause’ $X$ (or $X(\omega)$ or $X(\omega, u)$ etc.) is defined to be some function of the tupel $(Y_0, Y_1)$. Hence, the “fundamental problem of causal inference” (Holland 1986: 947) arises: If the function depends

\(^{23}\) The literature on probability theory generally calls the set $\{X = 0\} = \{\omega \in \Omega \mid X(\omega) = 0\}$ an ‘event’. But the classical notion of sets is a static notion. As such, it neither captures dynamics nor does it easily lend itself to notions of agency or autonomy. The rather simplistic equivocation of sets with ‘events’ is used far too often to make the statistical version of the counterfactual approach palatable. But even when texts do not conflate mathematical objects with real events and changes, the obvious challenge for statistical analysis remains. It was the main difficulty that prompted the current article.

\(^{24}\) An explication often used refers to possible worlds such that if in fact $\{Y_0 = y, X = 0\}$, then some $y'$ is the value of $Y_1$ in the possible world closest to the present one, i.e. the possible world that is exactly like the present one except that $X = 1$. How ‘closest parallel worlds’ are to be interpreted, what criteria there are for ‘close’, and in particular in what sense these worlds ‘exist’, has always been a subject of intense philosophical discussion. See Grayling (1982: 68–77) for a succinct review. I have to mention here only that ‘existence’ in the context of possible worlds is no straightforward concept. And the exposition of the concept becomes even more involved when the entities of reference are random variables within a probabilistic model. In fact, most texts in the statistical literature simply bypass these problems altogether. Nevertheless, they sometimes seem to suggest that a rhetoric based on parallel or possible worlds and reasoning about counterfactual outcomes is the only ‘scientific’ way to answer causal questions (e.g. Phillips and Goodman 2006).
on both $Y_0$ and $Y_1$ (as it should as a measure of departure from $Y_0$ from $Y_1$), then its value is never ‘observed’.

There is, in this setup, only one obvious solution: The function representing ‘causal effects’ is linear and the CAR condition holds:

$$(Y_0, Y_1) \perp \!
\!
\!
\perp X$$

In that case,

$$\mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y_1 | X = 1) - \mathbb{E}(Y_0 | X = 0)$$

using both linearity of expectations and independence. But the right hand side can be ‘estimated’ by replacing expectations (means across $\omega \in \Omega$) by means across the ‘population’ (means across $u \in U$).\(^{25}\)

It turns out that the choice of any type of (statistical or random) variable as causal relata creates a major problem with this view. This is particularly visible in some sociological ‘applications’.\(^{26}\) Two recent examples may illustrate the ambiguities thus created. Harding (2003) published a “counterfactual model of neighborhood effects” on dropping out of school and on teenage pregnancy. A second example is the very detailed and much appraised study of Epstein et al. (2005) on Supreme Court decisions during ‘crisis’. But neither neighbourhoods nor crises do something either to young women or to Supreme Court judges. It is plain that young women’s living conditions do depend on their neighbourhoods

\(^{25}\) Even if the counterfactual account of causality is accepted, neither the restriction to linear ‘effects’ nor replacing differences between $Y_1(\omega, u)$ and $Y_0(\omega, u)$ by some sort of mean is necessitated by a counterfactual conception of causality. In an attempt to capture different ‘effects’ within different subpopulations (or subsets of $\Omega$) the econometric literature introduced ‘heterogeneous’ effects (Heckman 2001, 2006, 2008, Heckman and Smith 1997, Heckman et al. 2006). The extension is, however, just obfuscating the original problem. Furthermore, see Fragankakis et al. (2007) for a defence of the standard statistical view.

\(^{26}\) Shafer (1995: 556-558) also warns against mistaking (statistical or random) variables for causes within his predictive probability tree approach. He gives examples where variables, even if taken only as descriptions of causes, fail to be unique or are not even defined within a part of a probability tree. Moreover, random variables and statistical variables are conflated quite differently across the probability tree. Rather surprisingly, the statistical and applied literature does not pay much attention to these technical difficulties.
and the resources available from the neighbourhood. Similarly, Supreme Court judges certainly are well aware of the political situation and will react to it. But neither the neighbourhood nor ‘political crisis’ would qualify as a cause in the analysis proposed in the present article: That a cause ought to be identified with the event brought about by an act of an actor. Now, even with a liberal interpretation of an ‘event brought about by …’ probably neither ‘crises’ nor ‘neighbourhoods’ would qualify as causes. There is more to the examples than just a demonstration of a difference in definition (or personal preferences): If ‘cause’ should refer to more than just some ‘living conditions’ or ‘political climate’, i.e. background information; if, in other words, ‘causes’ are not just an assembly of arbitrary sets of (pre-) conditions but ensembles of states of affairs that necessitate the outcome, then neither a ‘political crisis’ nor the neighbourhood qualify as causes: Judges do not necessarily judge or vote ‘because’ of a ‘state of emergency’ (however defined by Epstein et al., and whether felt or real). And young women act and behave as they do, but not in an explicable way ‘because’ they grew up in a certain neighbourhood (however ‘neighbourhood’ is defined by Harding).

There is, in fact, some resistance against the indiscriminant use of any (statistical or random) variable as a ‘cause’. Heckman in his review (2006) distances himself from some aspects of the statistical literature on the ground that it relies decisively on the analogy with experimental designs:

One theme developed in my paper is that major limitations hamper the statistical treatment effect literature in answering important social science questions. These limitations are not surprising since the statistical treatment effect literature is an offshoot of the experimental design literature in biostatistics. My essay shows that “technical” assumptions invoked in the statistical treatment effect literature have unappealing implications for social science. (2006: 138)

However, Heckman’s review does not depart from the rest of the statistical literature in that the questions and models he proposes refer to counterfactual measures based on the joint distributions of a set of random variables plus the idea that the values of some of the variables can be manipulated. In the contrast to the all embracing approaches
of Harding and Epstein et al. (where nearly everything can become a ‘cause’), the ‘manipulations’ envisaged by Heckman are basically the implementations of certain policies. Furthermore, he advocates that the relevance of CAR-type ‘assumptions’ ought to be judged in the context of (economically or sociologically) reasonable, relevant, and explicit models of self-selection, a requirement that echoes the requirement of Neyman cited in Chapter 3.

Another important difference of this literature to the ideas presented in the present article is that dynamic aspects are central to the article but basically absent in practically all suggestions derived within the counterfactual approach.\(^{27}\) In contrast, much of the classical philosophical literature always concentrated on events as the relata of causes and therefore relied on dynamic concepts. Even though a few philosophers have taken up the clarifications achieved within the statistical discussion, the impact of philosophers and philosophically informed discourses on the developments within statistics has been minor.\(^{28}\)

Sociologists often claim another distinctive feature of causes from mere conditions, namely ‘mechanisms’. The ‘mechanisms’ envisaged are however as abstract (or elusive) as the ‘selection processes’ stressed in the economic literature. To take a recent example, Ní Bhrolcháin and Dyson (2007), discussing the role of causation in demographic research, cite as diverse ‘mechanisms’ as astrological conjunctions and famines as qualifying as ‘mechanisms’ ‘causing’ changes in total period birth rates and other aggregates.\(^{29}\) The tendency to extend the notion of ‘mechanism’ beyond its normal use in order to make it a distinguishing feature of causality within the social sciences seems to be at least as far fetched as

\(^{27}\) Whether authors embrace counterfactual concepts or not, the statistical literature is rather silent with regard to dynamics and models of change. The few recent exceptions are Lok et al. (2004), Bray et al. (2006), Aalen and Frigessi (2007), Neugebauer and van der Laan (2006a,b), Neugebauer et al. (2007), and Moodie et al. (2007).


the similar strategy based on self-selection. In any case, the aspirations of most proponents of such a distinction to create a condition as general as possible (sometimes even to comprehensively cover all problems of sociology simultaneously) makes all such efforts rather pointless. In fact, the tendency to embrace nearly everything that can be described at all as a possible mechanism runs counter to any reasonable effort to distinguish coincidences, circumstances, and conditions from actions and events.

A careful review of the recent literature would need to distinguish the ‘mechanism’ idea from an argument based on ‘manipulation’, where the latter is mainly based on similarities to ‘experiments’. The latter view has been advocated mainly by Pearl (2000, 2003). It is pursued in much of the current literature on ‘evidence based medicine’, ‘effects’ of labour market regulations and programs, etc. Since not even the much cited recent works of Heckman helped to draw the distinction between ‘experiments’ and a reasonable analysis of self-selection processes (conjectured or known, observed or unobserved, random or fixed), it may be permissible to ignore this rather small differences.

However, the approach based on analogues of ‘experiments’ or ‘manipulations’ or respective relations between inputs and outcomes mediated by ‘mechanisms’ might seem to be close to the agent based dynamic concept proposed in this paper. Albeit apparent similarities, the consequences of the different formulations are profound. A manipulation/mechanism etc. might simply involve the daily application of a certain amount of drug \( a \). But this may be accomplished in a variety of ways: Allowing a certain group of people to swallow a pill of a certain makeup, injecting a certain amount of the same substance to every one in a group (with or without his or her consent), or designing an ‘experiment’ that chooses the ‘treated’ patients by some randomisation device. Advocates of this form of ‘experiment’ (the ‘gold standard’ in epidemiology and in studies of ‘risks’) probably assume that people assigned to either treatment or control will

---

30 See also the detailed discussion in Halpern and Pearl (2005).
31 See Lindley (2002) for critical remarks from a Bayesian point of view and Hacking (1988) for illuminating background information on the origins of statistical ‘experiments’.
follow henceforth the prescription given to them without hesitation, or thought, or consideration of intermediate results of the ‘experiment’, or their own considerations, their personal situation, discussions with relatives and friends, or even perceived well-being. That this is wrong in most circumstances of practical importance just reiterates Heckman’s critique in that an evaluation of ‘policy interventions’ must take into account a reasonable (probabilistic) model of self-selection.

In consequence, it seems unhelpful to insist on the pivotal role of ‘controlled experiments’ as a gold standard after ‘adjusting’ results somehow based on models of self-selection and similar processes. Instead, the whole idea of a fixed, though unknown ‘effect’ discernable from idealised ‘experiments’ should to be discredited. \(^{32}\)

Even though most of the statistical literature focuses on the counterfactual approach to causality, there has been some progress along the lines of problems indicated in the final section of my article. In particular, Didelez (2007, 2008) discusses the definition of local independence and Aalen and Frigessi (2007) make a few helpful comments on the concept of autonomy. Geneletti’s contribution (2007) provides a refreshing view of a non-counterfactual view on causality.

---

\(^{32}\) The literature on ‘causal effects’ pursuing the pretended ‘gold standard’ of randomised experiments is burgeoning. But only a tiny fraction of examples provided by that literature in the social sciences stands up against its own proclaimed aims. Even twin-studies (often invoked to provide a basis for judging ‘causality’ unambiguously) are probably tainted by self-selection since identical twins are more frequent with older couples (and higher education, and …).
A Non-parametric Mean Residual Life Estimator. An Example from Market Research

6.1. Summary

The mean residual life (mrl) function dynamically describes the average time to an event, depending on the time since the previous event. It provides a forecast in parallel with the development of the underlying process. From a theoretical point of view, the mrl characterises the distribution of the process completely, but in contrast to other characterisations like the hazard rate, it has a direct interpretation in terms of average behaviour.

We use Kaplan–Meier integrals (weighted averages of residual times) to construct a nonparametric estimator of the mrl. We use results from Stute (1995) and Yang (1994) to describe the asymptotic behaviour of this estimator and derive an approximate variance formula.

We present a small simulation study and apply the estimator (and the variance formula) to data pertaining to purchase time behaviour from the Homescan Panel™, A. C. Nielsen, Germany.
6. A Non-parametric Mean Residual Life Estimator

6.2. Introduction

The mean residual life (mrl) function dynamically describes the average time to an event, depending on the time since the previous event. As an important example for functionals of the Kaplan–Meier estimator it has been studied by many authors, e.g., Gill (1983), Gijbels and Veraverbeke (1991), Yang (1994) and Stute (1995). From a theoretical point of view, the mrl characterises the distribution of the process completely, see e.g. Shaked and Shanthikumar (1991). The mrl function can therefore be used in model formulation just as densities or hazard functions are. The mrl function is defined as a conditional expectation of the time to an event given that that time is larger than a given value. Its computation thus involves integrals over unbounded intervals of the real line. While conditional expectations are easily interpreted and often the object of immediate interest in applications, the fact that integrals over unbounded domains are involved severely hampers the analysis of estimators in the presence of censoring. This issue has been discussed extensively by Stute and Wang (1993), where a strong law of large numbers is given for such functionals. Instead of integrating with respect to the cumulative hazard rate – as was done in previous work – the authors use integrals with respect to the distribution function of the underlying random variable. Along the same lines, Stute (1995) provides a central limit theorem in this situation. A crucial role is played by the Kaplan–Meier weights, see Stute and Wang (1993: 1593) and Stute (1995: 423), respectively. These quantities generalise the well-known weights for non-parametric estimation in sampling theory, where the inverse inclusion probabilities are used as weighting factors. In the presence of independent censoring, however, the suitable weights are stochastic and vary between the observations.

6.3. An Example from Market Research

A central goal of market research is to describe the market performance of ”fast moving” consumer goods (fmcg). Those are products which are perishable or quickly used up, like food or detergents in contrast to
cars, washing machines etc. Throughout this paper, we primarily have fmcg in mind when we talk about products. Both manufacturers and retailers have a strong interest in identifying the more or less successful items of a product class (pc). One way of doing this is to collect data on the purchase behaviour of households. Specifically, we may observe the purchase acts of the participating consumers for a pc of interest during a fixed period, say one year. We call the duration between two consecutive purchase acts by the same household an interpurchase time (ipt). As a pc consists of several items, different types of ipt occur: On the aggregate level, we have the time between two purchases in the pc, while for each product, there is also an item-specific ipt starting and ending with a purchase of the specific article.

In order to reduce problems of dependencies between observations we use only one ipt of each type per household in the sample. When interpreting the data, we also have to be aware of possible censorings. They occur at the end of the observation period, in case no repurchase has taken place by then.

Statistical inference concerning durations falls into the realm of survival analysis, where the most common quantity is the hazard rate. In this paper, however, we will focus on the mean residual life function (mrl) instead: Denote by $T$ the random variable describing the length of the ipt and by $F$ its cumulative distribution function. Then the mrl $m(t)$ at some time $t$ equals the average remaining time to repurchase, given this event has not yet taken place:

$$m(t) := \mathbb{E}_T(T - t \mid T > t) = \int_t^\infty u - t \, dF(u) \frac{1}{1 - F(t)}$$

(6.1)

We will discuss the mrl in more detail in Section 6.4. For the moment, we only note that the mrl is a conditional expectation and thus describes the mean behaviour of consumers. This feature makes it an interesting quantity for traditional marketing, where it is impossible to focus on specific, single consumers and one has to deal with the whole group of households instead. As an illustration, we present data pertaining to purchase time behaviour from the Homescan Panel™ of A. C. Nielsen, Germany. It is based on approximately 5,800 households surveyed in
2001. We focus on two items of an anonymised pc. Central facts of this pc and the two selected items, 'focus' and 'competitor', are given in the next table:

<table>
<thead>
<tr>
<th>Class</th>
<th>penetration</th>
<th>market share</th>
<th>proportion of repurchasing HH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus</td>
<td>0.58</td>
<td>-</td>
<td>0.78</td>
</tr>
<tr>
<td>Comp.</td>
<td>0.03</td>
<td>0.014</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>0.14</td>
<td>0.077</td>
<td>0.61</td>
</tr>
</tbody>
</table>

In words, 58% of the German households purchased the pc in 2001, 3% bought the focus item and 14% bought the competitor item at least once. With market shares of 1.4% and 7.7%, neither of the two products has an overwhelming influence on the development of the pc. On the aggregate level, censoring occurred in 22% of the observations. On the item level, censoring is rather heavy, 50% and 39%, respectively. In Section 6.5, we present a non-parametric estimator for the mrl in this situation, see (6.6). Applied to our data truncated at \( \tau^* = 365 \) days\(^1\), the resulting curves are given in Figure 6.1.

On a qualitative level, the estimated mrls suggest the following interpretation: The mrl of the pc serves as a reference. To some extent, it provides information on how the average consumer purchases the typical item of this pc. The mrl of the competitor stays rather close to the pc’s mrl. This implies that its consumers perceive this specific item to 'represent' the pc: whenever the average consumer repurchases the pc, the customer of the competitor's item repurchases this specific product. Clearly, this makes the competitor a successful article within the pc.

---

\(^1\) Truncation to \([0, 365]\) means that we used only repurchase times that happened within a year. We are thus looking at the conditional mrl given that repurchase happened within a year. Purchase data from the panel households is available for the following months as well, if we neglect the slight but inevitable panel mortality. Consequently, one might aim at extending the truncation time \( \tau^* \), thus diminishing the loss of information. On the other hand, for a household to contribute to the actually used purchase data, it is required that its overall reporting quality exceeds a certain level over the whole period in question. As a result, extending the relevant observation period leads to a decreased sample size presumably offsetting any benefits.
From the focus item's point of view, the situation is considerably less comfortable: Approximately two months (56 days) after the last purchase act, the average remaining time to repurchase of the focus item is 40 days longer than it is for the pc.

![Figure 6.1: Estimated mrl functions according to (6.6) for the data presented in Section 6.3. Solid: Product class. Dashed: Focus item. Dotted: Competitive item.](image)

To determine whether the differences between the estimated mrl functions really support these arguments or merely occurred 'by chance', we have to introduce a stochastic model of our situation and estimate variances of the mrl estimates as well. We suggest several variance formulae in Section 6.6 and evaluate their performance in section 6.7. Concerning the real life data, it turns out that the corresponding estimations do not render any differences between the mrls significant.
6.4. Mean Residual Life Function

The mrl function defined in (6.1) determines the distribution uniquely. If $F$ is absolutely continuous one can compute the distribution function from the mrl function by

$$1 - F(t) = \frac{m(0)}{m(t)} \exp \left\{ - \int_0^t \frac{1}{m(u)} \, du \right\},$$  \hspace{2cm} (6.2)

see Shaked and Shanthikumar (1991: 614). The mrl function can thus be used as a characterisation of the distribution in the same way as the density or the hazard function can.

As an example, consider a mrl function which is linear on some interval $I = [t_1, t_2]$. For $t \in I$, the mrl thus has the form

$$m(t) = a \cdot (t - t_1) + m(t_1),$$

where $m(t_1) > 0$ and $a \geq -1$. In case $a < 0$, we also have to ensure that $t_2 \leq t_1 - \frac{m(t_1)}{a}$.

For $a \neq 0$, the distribution function on $I$ then equals a Pareto distribution scaled by $F(t_1)$:

$$F(t) = F(t_1) \cdot \left( \frac{a \cdot (t - t_1)}{m(t_1)} + 1 \right)^{-\frac{a+1}{a}}, \hspace{1cm} t \in I$$

Specifically, if $a = -0.5$, the distribution function is linear and the distribution on $I$ is consequently uniform.

For $a = 0$, the distribution function on $I$ is exponential:

$$F(t) = F(t_1) \cdot \exp \left( - \frac{t - t_1}{m(t_1)} \right)$$

Thus, if the mrl function is given on the interval $I$ we get the behaviour of $F$ restricted to $I$, and complete knowledge of $F$ on $I$ if $F(t_1)$ was known. Note that both formulae simplify considerably in terms of ‘local coordinates’ $F(t)/F(t_1)$ and $(t - t_1)/m(t_1)$. 

280
In the presence of censoring there may be a point in time after which it is impossible to gather further information. Let $\tau$ be the least upper bound on the period of observation. Then the best one can hope for is to estimate integrals up to that point in time. The definition of the mrl function, however, requires an evaluation of the integrals to $\infty$, not only up to $\tau$. Moreover, the truncation point $\tau$ is generally unknown (and not easily estimable).

We will therefore consider to estimate

$$m^* (t, \tau^*) := \mathbb{E}(T - t \mid T \in (t, \tau^*)) = \frac{\int_t^{\tau^*} u - t \, dF(u)}{\int_t^{\tau^*} dF(u)}$$

for some $\tau^* \leq \tau$. This truncated mrl function $m^*$ need not be related to the mrl function $m$ in any obvious way. However, $m^*$ is still a conditional expectation with an easy interpretation and important applications. Moreover, as a generalisation of (6.2), the truncated mrl function uniquely identifies the distribution function up to $\tau^*$ through

$$F(\tau^*) - F(t) = \frac{m^*(0, \tau^*)}{m^*(t, \tau^*)} \exp \left( - \int_0^t \frac{du}{m^*(u, \tau^*)} \right)$$

To see this, note that for continuously differentiable (in $t$) $m^*$

$$\frac{\partial m^*(t, \tau^*)}{\partial t} = \frac{\partial}{\partial t} \frac{\int_t^{\tau^*} u - t \, dF(u)}{\int_t^{\tau^*} dF(u)}$$

$$= \left( -tf(t) - \int_t^{\tau^*} f(u) \, du + tf(t) \right) \frac{1}{F(\tau^*) - F(t)}$$

$$+ \frac{f(t)}{(F(\tau^*) - F(t))^2} \int_t^{\tau^*} (u - t)f(u) \, du$$

$$= -1 + m^*(t, \tau^*) \frac{f(t)}{F(\tau^*) - F(t)}$$

Thus

$$-\frac{\partial}{\partial t} \ln (F(\tau^*) - F(t)) = \frac{\partial m^*(t, \tau^*)/\partial t + 1}{m^*(t, \tau^*)}$$
so that
\[ F(\tau^*) - F(t) = \exp \left( - \int_0^t \frac{\partial m^*(v, \tau^*)}{\partial v} \bigg|_{v=u} m^*(u, \tau^*) \, du \right) + \frac{1}{m^*(u, \tau^*)} \, du \]
from which (6.4) follows by noting that the first fraction under the integral is the logarithmic derivative of \( m^*(u, \tau^*) \).

### 6.5. Kaplan–Meier Integrals

With complete data an estimator of the mrl function is
\[
\hat{m}(t) := \frac{\int_t^\infty u - t \, dF_n(u)}{\int_t^\infty dF_n(u)} = \frac{\sum_{i=1}^n (t_i - t) \, \mathbb{I}[t_i > t]}{\sum_{i=1}^n \mathbb{I}[t_i > t]} \quad (6.5)
\]
With censored observations it is natural to replace \( F_n \) by the Kaplan-Meier estimator
\[
1 - \hat{F}_n(t) = \prod_{i=1}^n \left( 1 - \frac{\delta(i)}{n - i + 1} \right) \mathbb{I}[z(i) \leq t]
\]
where \( z_i = \min\{t_i, u_i\} \) are the censored observations, \( \delta_i = \mathbb{I}[t_i \leq u_i] \) are the censoring indicators and \( z(1) \leq \ldots \leq z(n) \) are the ordered values of the observed times while the \( \delta(i) \) are the corresponding censoring indicators. For a given upper limit \( \tau^* \), the equation (6.5) becomes
\[
\hat{m}^*(t, \tau^*) := \frac{\int_t^\tau^* u - t \, d\hat{F}_n(u)}{\int_t^\tau^* d\hat{F}_n(u)} = \frac{\sum_{i=1}^n (t_i - t) \, \mathbb{I}[t_i > t] w_i}{\sum_{i=1}^n \mathbb{I}[t_i > t] w_i} \quad (6.6)
\]
where the \( w_i \) are the jump sizes \( \hat{F}_n(t_i) - \hat{F}_n(t_i-) \) of the Kaplan–Meier estimator. Note that if the largest observation is censored we will not force \( 1 - \hat{F}_n \) to be zero after that observation.

In order to analyse the estimator (6.6) we propose a simple stochastic model that will allow the calculation of approximate variances (within that model, of course) and that might be used to gauge the performance
of the estimator on real data sets. We will assume that the observations arise from independent and identically distributed copies of the random variables \((T, U)\), where \(T\) and \(U\) are independent and \(T\) has distribution function \(F\) while \(U\) has distribution function \(G\). Then \(1 - H(t) = (1 - F(t))(1 - G(t))\) is the survivor function of the random variable \(Z := \min\{T, U\}\). We set \(\delta := \mathbb{1}(T \leq U)\) and let \(\tau := \inf\{t \mid H(t) = 1\}\) be the least upper bound of the support of \(Z\). In the following we will assume that \(F\) is absolutely continuous while we allow for an arbitrary distribution of \(G\). Note, however, that Stute’s (1995) results are valid for general \(F\) and general \(G\), while Yang’s results (1994) allow \(G\) to vary with the observations but requires absolute continuous \(F\). Our restriction to absolutely continuous \(F\) and identically distributed censoring times is however often applicable and simplifies the formulations considerably. Extending the integrals with respect to \(\hat{F}_n\) up to \(\tau\) requires delicate considerations on the behaviour of \(\hat{F}_n\) near \(\tau\). The point is well discussed by Stute and Wang (1993) and by Gill (1994). We will here avoid an explicit discussion by either truncating to \(\tau^*\) or by assuming \(\tau = \infty\). In the simulations we will use a distribution function \(F\) with a compact support strictly included in the support of \(G\).

6.6. Variance Formulae

The weights in the formula for \(\hat{m}^*\) will in the presence of censoring depend on all the censored observations preceding a given event time. The weights \(w_i\) are therefore not independent random variables in our model. There is, however, a general representation of Kaplan–Meier integrals in terms of sums of independent random variables given by Stute (1995: 425). Using that representation, standard results on sums of independent random variables can be used to derive variances of the estimator \(\hat{m}^*\). In the case of absolutely continuous \(F\), the representation simplifies considerably and can be written as

\[
\int_0^\infty \phi_k(u) \, d\hat{F}_n(u) - \int_0^\tau \phi_k(u) \, dF(u) =
\]
6. A Non-parametric Mean Residual Life Estimator

\[
\frac{1}{n} \sum_{i=1}^{n} \int_{0}^{\infty} \frac{\phi_k(u) - \mathbb{E}(\phi_k(T) \mid T > u)}{1 - G(u_-)} \, dM_i(u) + o_P(n^{-1/2}) \quad (6.7)
\]

for \(k = 1, 2\) and where

\[
\phi_1(u) := \mathbb{I}[u > t](u - t) \quad , \quad \phi_2(u) := \mathbb{I}[u > t]
\]

and where

\[
M_i(t) := \mathbb{I}[t_i \leq t, \delta_i = 1] - \int_{0}^{t} \frac{\mathbb{I}[z_i \geq u]}{1 - F(u)} \, dF(u)
\]

is the martingale for the counting process \(\mathbb{I}[t_i \leq t, \delta_i = 1]\) with respect to the standard filtration (see the Appendix for a derivation).

The representation gives among other results a variance formula via a standard martingale argument (details are deferred to the Appendix):

\[
\text{Var} \left( \int_{0}^{\infty} \frac{\phi_k(u) - \mathbb{E}(\phi_k(T) \mid T > u)}{1 - G(u_-)} \, dM_i(u) \right) = \\
\int_{0}^{\infty} \frac{(\phi_k(u) - \mathbb{E}(\phi_k(T) \mid T > u))^2}{1 - G(u_-)} \, dF(u) =: \sigma_k^2 , \quad k = 1, 2 \quad (6.8)
\]

This formula was also derived via a direct argument by Yang (1994).

While the above formulae are useful in theoretical work, practical computations will also have to rely on a direct empirical counterpart of Stute’s representation. We start from the ordered values \(z_{(i)}\), \(i = 1, \ldots, n\) and the corresponding values \(\delta_{(i)}\) and \(\phi_{ki} := \phi_k(z_{(i)})\). We then define (assuming no ties):

\[
\gamma_i := \exp \left( \sum_{j=1}^{i-1} \frac{1 - \delta_{(j)}}{n - j} \right) , \quad i = 2, \ldots, n , \quad \gamma_1 := 1
\]

\[
a_{ki} := \delta_{(i)} \phi_{ki} \gamma_i , \quad i = 1, \ldots, n
\]

\[
b_{ki} := \frac{1 - \delta_{(i)}}{n - i} \sum_{j=i+1}^{n} a_{kj} , \quad i = 1, \ldots, n - 1 , \quad b_{kn} := 0
\]
\[ c_{ki} := \sum_{j=1}^{i-1} \frac{b_{kj}}{n-j}, \quad i = 2, \ldots, n, \quad c_{k1} := 0 \]

Finally, we set
\[ A_{ki} := a_{ki} + b_{ki} - c_{ki}, \quad i = 1, \ldots, n, \quad k = 1, 2 \] (6.9)

Writing
\[ \hat{m}^*(t, \tau^*) = \frac{\int_{\tau^*}^{\tau} u - t \, d\hat{F}_n(u)}{\int_{t}^{\tau^*} d\hat{F}_n(u)} =: \frac{\hat{A}_1}{\hat{A}_2} \] (6.10)

we get an expression for the variance of the fraction using the delta method:
\[ \text{Var}(\hat{m}^*(t, \tau^*)) \approx \frac{\hat{A}_1^2 \hat{\sigma}_2^2}{n\hat{A}_2^4} + \frac{\hat{\sigma}_1^2}{n\hat{A}_2^2} - 2\frac{\hat{A}_1}{n\hat{A}_2^3} \text{Cov}(\hat{A}_1n, \hat{A}_2n) \] (6.11)

where \( \hat{\sigma}_1^2 \) is the estimated variance of \( \hat{A}_1 \) and \( \hat{\sigma}_2^2 \) is the estimated variance of \( \hat{A}_2 \).

6.7. A Small Simulation Study

The variance approximations of the previous section can be implemented in various ways. In this section we will use simulations to evaluate some of the possible choices. To keep things simple, we use only the exponential distribution with expectation 1 truncated to \([0, 2]\) for \( F \), i.e. \( F(t) = (1 - \exp(-t))/(1 - \exp(-2)) \) on \( t \in [0, 2] \). The truncation assures finiteness of all moments with respect to \( F \) and thus all regularity requirements for the validity of the variance formulae are fulfilled. We evaluate the mrl at 0.2, 0.5, 1.0 and 1.5. The mrl at these points is 0.6435, 0.5691, 0.4180 and 0.2293. The survival probabilities \( 1 - F(t) \) at the evaluation points are 0.79, 0.55, 0.27, and 0.10. As censoring distributions we use the exponential distribution with expectations 5 and 1 corresponding to censoring probabilities of 0.12 and 0.43, respectively.
6. A Non-parametric Mean Residual Life Estimator

The corresponding expected proportions at risk are 0.65, 0.33, 0.10, and 0.02 using a censoring distribution with expectation 1, and 0.76, 0.49, 0.22, and 0.08 using a censoring distribution with expectation 5. Sample sizes are 200 and 1000. We use 1000 simulations for each of the combinations.

Table 6.1.: Simulation results for the mrl estimator

<table>
<thead>
<tr>
<th>Experiment</th>
<th>true mrl</th>
<th>t=0.2</th>
<th>t=0.5</th>
<th>t=1.0</th>
<th>t=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>n = 200, (U \simeq \text{exp}(1))</td>
<td>Min 0.4637, Mean 0.6376, Max 0.8173, NA 0, Var ((\times 10^2)) 0.3130, as. Var 0.2769</td>
<td>0.4637</td>
<td>0.3561</td>
<td>0.1457</td>
<td>0.0020</td>
</tr>
<tr>
<td></td>
<td>Min 0.5315, Mean 0.6442, Max 0.7486, NA 0, Var ((\times 10^2)) 0.1534, as. Var 0.1647</td>
<td>0.5315</td>
<td>0.4427</td>
<td>0.2859</td>
<td>0.0747</td>
</tr>
<tr>
<td></td>
<td>Min 0.5718, Mean 0.6427, Max 0.7180, NA 0, Var ((\times 10^3)) 0.5558, as. Var 0.5538</td>
<td>0.5718</td>
<td>0.4926</td>
<td>0.3319</td>
<td>0.1178</td>
</tr>
<tr>
<td></td>
<td>Min 0.5837, Mean 0.6434, Max 0.7070, NA 0, Var ((\times 10^3)) 0.3353, as. Var 0.3295</td>
<td>0.5837</td>
<td>0.5144</td>
<td>0.3590</td>
<td>0.1891</td>
</tr>
</tbody>
</table>

Table 6.1 gives the mrl at the evaluation points together with summary statistics for the estimated mrl from the simulations. The mrl estimator is slightly downward biased for smaller numbers of observations and for
later times. This was to be expected from the supermartingal structure of Kaplan-Meier integrals as discussed by Stute and Wang (1993) and the bias of the Kaplan-Meier estimator itself as given in Fleming/Harrington (1991: 99). Possible consequences and remedies are described by Miller (1983) and Gill (1994: section 8).

In our situation, the asymptotic variances from (6.8) and (6.11) can be computed explicitly. They are also given in Table 6.1. For \( n = 1000 \), the agreement with the variances estimated in the simulation experiment is excellent. For \( n = 200 \), however, the variances of the mrl estimates tend to be larger than the asymptotic variances suggest. Note that the variances are multiplied by 100 in the case of \( n = 200 \) and by 1000 in the case of \( n = 1000 \).

Turning to the empirical variance estimators, a simple approach would be to use the empirical variance of the terms \( A_{1i}/B_{2i} \), thus avoiding the use of (6.11). Even when using the empirical variances of the \( A_{ki} \) separately, one might try simpler versions of (6.11). Yang (1994: 342) proposed to use just the variance of the numerator. Since the denominator is consistent, an appeal to Slutsky’s lemma shows this to be a consistent variance estimator. But this would work well only if the expectation of the numerator would be 0. A further possibility is to use both empirical variances, but to ignore the covariance. But the covariance is by definition far from 0. In fact, in our simulation setup, the results using these three estimators had no clear relation with the observed variance of the mrl. We will therefore not present the results in our simulation reports.

A second approach uses the Kaplan–Meier estimator (and its estimated variance) in the denominator of the mrl and in its variance estimator. We will not pursue this possibility here, since we are mainly interested in the performance of Kaplan–Meier integrals. We will therefore also use only the Kaplan–Meier integrals in empirical versions of (6.8).

A third possibility is a direct application of the discrete representation of Stute. This means to compute the terms in (6.9) and use the empirical means, variances, and covariances of these terms in (6.11). We denote this estimator by \( \hat{\sigma}^2_S \).
In \( \hat{\sigma}_S^2 \), one might replace the terms \( \hat{A}_{kn} \) by the respective terms used in the mrl estimator. This should give estimators that are better centered and possibly be less variable. We denote this estimator by \( \hat{\sigma}_{mrl}^2 \).

Lastly, we use an empirical version of Yang’s formula (6.8) where we use the estimated mrl for the inner expectation and the empirical Kaplan–Meier integrals for the outer integral. The denominator in the integrand is estimated by the Kaplan–Meier estimator. Moreover, in (6.11), we use the mrl estimators for the \( \hat{A}_{kn} \) terms. This version will be denoted by \( \sigma_Y^2 \).

Note that, from a computational point of view, \( \hat{\sigma}_Y^2 \) requires the computation of the mrl at all (unique) points \( z_i \). In contrast, both \( \hat{\sigma}_{mrl}^2 \) and \( \hat{\sigma}_S^2 \) can easily be computed point-wise but are relatively more expensive when the variances are required at all observed points. All three variants, however, are computationally much cheaper than a bootstrap approach.

<table>
<thead>
<tr>
<th>Table 6.2.: Simulation results for variance estimators, ( n = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U \simeq \exp(1) )</td>
</tr>
<tr>
<td>sim. Var (( \times 10^2 ))</td>
</tr>
<tr>
<td>( \hat{\sigma}_S^2 )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( \hat{\sigma}_{mrl}^2 )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>( \hat{\sigma}_Y^2 )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6.3.: Simulation results for variance estimators, ( n = 200 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( U = \exp(0.2) )</td>
</tr>
<tr>
<td>sim. Var (( \times 10^2 ))</td>
</tr>
<tr>
<td>( \hat{\sigma}_S^2 )</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Looking first at the results for the case $n = 200$ and $U \simeq \exp(0.2)$ (Table 6.2), all three variance estimators are in rather close agreement with the simulated variances, even at $t = 1.5$. Moreover, $\hat{\sigma}_S^2 \leq \hat{\sigma}_{mrl}^2 \leq \hat{\sigma}_Y^2$ in the mean over all simulations, where also the variability of the estimators increases in this order.

The situation is less favourable in the case of heavy censoring and $n = 200$, given in Table 6.3. Here, all estimators show a large variability. For $t \leq 1$, $\hat{\sigma}_S^2$ seems to work best. However, at $t = 1.5$, none of the estimators is even close to the variability of the mrl estimator. But note that in this case the asymptotic variance is also not close to the observed variability of the mrl estimator.

### Table 6.3.: Simulation results for variance estimators, $n = 200$

<table>
<thead>
<tr>
<th>$\hat{\sigma}_{mrl}^2$</th>
<th>Min</th>
<th>Mean ($\times 10^2$)</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0011</td>
<td>0.0012</td>
<td>0.0010</td>
<td>0.0003</td>
</tr>
<tr>
<td>0.1682</td>
<td>0.1863</td>
<td>0.1917</td>
<td>0.1474</td>
</tr>
<tr>
<td>0.0044</td>
<td>0.0053</td>
<td>0.0080</td>
<td>0.0097</td>
</tr>
</tbody>
</table>

### Table 6.4.: Simulation results for variance estimators, $n = 1000$

<table>
<thead>
<tr>
<th>$U \simeq \exp(1)$</th>
<th>sim. Var ($\times 10^3$)</th>
<th>t=0.2</th>
<th>t=0.5</th>
<th>t=1.0</th>
<th>t=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\sigma}_S^2$</td>
<td>0.5558</td>
<td>0.7852</td>
<td>1.0237</td>
<td>1.1719</td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0006</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>Mean ($\times 10^3$)</td>
<td>0.6155</td>
<td>0.8133</td>
<td>1.1273</td>
<td>1.1074</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>0.0039</td>
<td>0.0052</td>
<td>0.0069</td>
<td>0.0041</td>
<td></td>
</tr>
</tbody>
</table>

### Table 6.5.: Simulation results for variance estimators, $n = 1000$

<table>
<thead>
<tr>
<th>$U \simeq \exp(1)$</th>
<th>sim. Var ($\times 10^3$)</th>
<th>t=0.2</th>
<th>t=0.5</th>
<th>t=1.0</th>
<th>t=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\sigma}_{mrl}^2$</td>
<td>0.6233</td>
<td>0.8281</td>
<td>1.1671</td>
<td>1.1989</td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0006</td>
<td>0.0004</td>
<td></td>
</tr>
<tr>
<td>Mean ($\times 10^3$)</td>
<td>0.0043</td>
<td>0.0061</td>
<td>0.0098</td>
<td>0.0092</td>
<td></td>
</tr>
<tr>
<td>Max</td>
<td>0.5579</td>
<td>0.7436</td>
<td>1.0644</td>
<td>1.2491</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\hat{\sigma}_Y^2$</th>
<th>Min</th>
<th>Mean ($\times 10^3$)</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0006</td>
<td>0.0003</td>
</tr>
<tr>
<td>0.5579</td>
<td>0.7436</td>
<td>1.0644</td>
<td>1.2491</td>
</tr>
<tr>
<td>0.0007</td>
<td>0.0010</td>
<td>0.0017</td>
<td>0.0036</td>
</tr>
</tbody>
</table>
6. A Non-parametric Mean Residual Life Estimator

<table>
<thead>
<tr>
<th>$U \simeq \exp(0.2)$</th>
<th>$t=0.2$</th>
<th>$t=0.5$</th>
<th>$t=1.0$</th>
<th>$t=1.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sim. Var ($\times 10^3$)</td>
<td>0.3353</td>
<td>0.3854</td>
<td>0.3766</td>
<td>0.2688</td>
</tr>
<tr>
<td>$\hat{\sigma}^2_S$</td>
<td>Min 0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0002</td>
</tr>
<tr>
<td>Mean ($\times 10^3$)</td>
<td>0.3302</td>
<td>0.3658</td>
<td>0.3781</td>
<td>0.2826</td>
</tr>
<tr>
<td>Max 0.0004</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td>$\hat{\sigma}^2_{mrl}$</td>
<td>Min 0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0002</td>
</tr>
<tr>
<td>Mean ($\times 10^3$)</td>
<td>0.3305</td>
<td>0.3661</td>
<td>0.3787</td>
<td>0.2834</td>
</tr>
<tr>
<td>Max 0.0004</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0005</td>
<td></td>
</tr>
<tr>
<td>$\hat{\sigma}^2_Y$</td>
<td>Min 0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0002</td>
</tr>
<tr>
<td>Mean ($\times 10^3$)</td>
<td>0.3311</td>
<td>0.3676</td>
<td>0.3837</td>
<td>0.2978</td>
</tr>
<tr>
<td>Max 0.0004</td>
<td>0.0004</td>
<td>0.0005</td>
<td>0.0005</td>
<td></td>
</tr>
</tbody>
</table>

Turning to the case $n = 1000$ with light censoring (Table 6.4), all three variance estimators are close together and close to the simulated variances. Once again, we find $\hat{\sigma}^2_S \leq \hat{\sigma}^2_{mrl} \leq \hat{\sigma}^2_Y$ in the mean over all simulations. We had expected to see $\hat{\sigma}^2_Y$ perform less satisfactorily than the other two estimators since with a small amount of censoring the explicit use of the Kaplan–Meier estimator $1 - \hat{G}$ as a weight in (6.8) might result in unstable behaviour. But this seems not to be the case here.

In the case with heavy censoring ($U \simeq \exp(1)$) we see that $\hat{\sigma}^2_Y$ compares favourably with the other two estimators: It is somewhat closer to the simulated variances and has less variability. To look closer at the problem with the other two estimators, we sampled one of the $A_{ki}$ from each of the 1000 simulation runs. Density estimates of the numerator and denominator variables are given in Figures 6.2 and 6.3. The distributions of the empirical $A_k$ terms are far from normal. They are multimodal with one of the modes close to 0 and they have rather heavy tails. Moreover, the empirical versions of the $A_{ki}$ are dependent so that the variances of the $A_{ki}$ are difficult to estimate accurately. This might explain the larger variability of $\hat{\sigma}^2_S$ and $\hat{\sigma}^2_{mrl}$ compared to $\hat{\sigma}^2_Y$. Moreover, with a larger proportion of censored observations the estimator of the distribution of the censoring variable stabilises and thus also Yang’s estimator stabilises.

In conclusion, our limited experience suggests that the variance estimator $\hat{\sigma}^2_Y$ is to be preferred in cases of heavy censoring, while with light censoring all estimators behave similarly.
Lastly, we look at a borderline case where the expectation in (6.8) is finite but where the bias condition (1.6) of Stute (1995: 425) is violated. Suppose $F$ is exponential with expectation 1 and $G$ is exponential with expectation 5. In this case, the mrl of $F$ is constant 1. The function $C(x)$ is given by

$$C(x) = \int_0^x \frac{1}{(1 - H(u))(1 - G(u))} dG(u) = \frac{0.2}{1.2} (e^{1.2x} - 1)$$

and thus

$$\int \phi_k(x) \sqrt{C(x)} dF(x)$$

diverges for $k = 1, 2$. With a sample size of $n = 1000$ (Table 6.6), the variance estimators $\hat{\sigma}^2_{mrl}$ and $\hat{\sigma}^2_S$ are rather larger than the simulated variances while $\hat{\sigma}^2_Y$ is still quite close. Moreover, the latter is much less variable than the other two. In fact, looking at the behaviour of the $A_{ki}$, they show very heavy tails with occasional huge values.

Table 6.6.: Simulation results, $F \simeq \exp(1), G \simeq \exp(0.2), n = 1000$

<table>
<thead>
<tr>
<th></th>
<th>t=0.2</th>
<th>t=0.5</th>
<th>t=1.0</th>
<th>t=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>mrl</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.8742</td>
<td>0.8424</td>
<td>0.7749</td>
<td>0.7495</td>
</tr>
<tr>
<td>Mean</td>
<td>0.9950</td>
<td>0.9936</td>
<td>0.9906</td>
<td>0.9836</td>
</tr>
<tr>
<td>Max</td>
<td>1.1390</td>
<td>1.1382</td>
<td>1.2160</td>
<td>1.2711</td>
</tr>
<tr>
<td>Var ($\times 10^3$)</td>
<td>1.5778</td>
<td>2.2273</td>
<td>4.0604</td>
<td>7.5099</td>
</tr>
<tr>
<td>$\hat{\sigma}^2_Y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.0010</td>
<td>0.0013</td>
<td>0.0019</td>
<td>0.0025</td>
</tr>
<tr>
<td>Mean ($\times 10^3$)</td>
<td>1.8755</td>
<td>2.7261</td>
<td>5.0650</td>
<td>9.3061</td>
</tr>
<tr>
<td>Max</td>
<td>0.0148</td>
<td>0.0233</td>
<td>0.0551</td>
<td>0.1101</td>
</tr>
<tr>
<td>$\hat{\sigma}^2_{mrl}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.0010</td>
<td>0.0013</td>
<td>0.0019</td>
<td>0.0025</td>
</tr>
<tr>
<td>Mean ($\times 10^3$)</td>
<td>1.8863</td>
<td>2.7482</td>
<td>5.1377</td>
<td>9.5365</td>
</tr>
<tr>
<td>Max</td>
<td>0.0152</td>
<td>0.0242</td>
<td>0.0587</td>
<td>0.1216</td>
</tr>
<tr>
<td>$\hat{\sigma}^2_Y$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Min</td>
<td>0.0009</td>
<td>0.0012</td>
<td>0.0019</td>
<td>0.0025</td>
</tr>
<tr>
<td>Mean ($\times 10^3$)</td>
<td>1.5599</td>
<td>2.2317</td>
<td>4.0548</td>
<td>7.3789</td>
</tr>
<tr>
<td>Max</td>
<td>0.0047</td>
<td>0.0075</td>
<td>0.0196</td>
<td>0.0442</td>
</tr>
</tbody>
</table>
6. A Non-parametric Mean Residual Life Estimator

Figure 6.2.: Density estimates for $A_{1i}$. Solid line: $t = 0.2$, short dashed line: $t = 0.5$, dashed and dotted line: $t = 1.0$, long dashed line: $t = 1.5$.

Table 6.7.: Simulation results, $F \simeq \exp(1)$, $G \simeq \exp(0.2)$, $n = 200$

<table>
<thead>
<tr>
<th></th>
<th>mrl</th>
<th>t=0.2</th>
<th>t=0.5</th>
<th>t=1.0</th>
<th>t=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td></td>
<td>0.7533</td>
<td>0.6608</td>
<td>0.6291</td>
<td>0.4774</td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td>0.9857</td>
<td>0.9819</td>
<td>0.9772</td>
<td>0.9663</td>
</tr>
<tr>
<td>Max</td>
<td></td>
<td>1.2871</td>
<td>1.3645</td>
<td>1.4910</td>
<td>1.7176</td>
</tr>
<tr>
<td>Var ($\times 10^2$)</td>
<td></td>
<td>0.8320</td>
<td>1.2447</td>
<td>2.1862</td>
<td>4.1630</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$\hat{\sigma}_S^2$</th>
<th>t=0.2</th>
<th>t=0.5</th>
<th>t=1.0</th>
<th>t=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>$\times 10^2$</td>
<td>0.0032</td>
<td>0.0036</td>
<td>0.0040</td>
<td>0.0043</td>
</tr>
<tr>
<td>Mean</td>
<td>$\times 10^2$</td>
<td>1.0050</td>
<td>1.4446</td>
<td>2.6152</td>
<td>4.5671</td>
</tr>
<tr>
<td>Max</td>
<td>$\times 10^2$</td>
<td>0.1212</td>
<td>0.1998</td>
<td>0.3797</td>
<td>0.6479</td>
</tr>
</tbody>
</table>
Table 6.7: Simulation results, $F \simeq \exp(1)$, $G \simeq \exp(0.2)$, $n = 200$

<table>
<thead>
<tr>
<th>$\hat{\sigma}^2_{mrl}$</th>
<th>t=0.2</th>
<th>t=0.5</th>
<th>t=1.0</th>
<th>t=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0032</td>
<td>0.0036</td>
<td>0.0041</td>
<td>0.0043</td>
</tr>
<tr>
<td>Mean ($\times 10^2$)</td>
<td>1.0286</td>
<td>1.4921</td>
<td>2.7655</td>
<td>5.0157</td>
</tr>
<tr>
<td>Max</td>
<td>0.1368</td>
<td>0.2372</td>
<td>0.5054</td>
<td>0.8394</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\hat{\sigma}^2_Y$</th>
<th>t=0.2</th>
<th>t=0.5</th>
<th>t=1.0</th>
<th>t=1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.0033</td>
<td>0.0037</td>
<td>0.0042</td>
<td>0.0044</td>
</tr>
<tr>
<td>Mean ($\times 10^2$)</td>
<td>0.7962</td>
<td>1.1502</td>
<td>2.1417</td>
<td>4.0573</td>
</tr>
<tr>
<td>Max</td>
<td>0.0429</td>
<td>0.0661</td>
<td>0.1197</td>
<td>0.4599</td>
</tr>
</tbody>
</table>

Looking at a sample size of $n = 200$ (Table 6.7), the mrl has a somewhat larger downward bias. The variance estimators based on the $A_{ki}$ are now rather far from the simulated variances especially at larger $t$. The estimator $\hat{\sigma}^2_Y$ is closer to the simulated variances. In conclusion, there seems to be some leeway to improve on variance estimators based on the $A_{ki}$, possibly also in the case of light censoring.

6.8. Appendix

For a function $\phi$ with $\mathbb{E}(|\phi(T)|) < \infty$ which also meets the appropriately modified moment conditions (1.5) and (1.6) in Stute (1995: 425), Stute (1995) gives a representation of the Kaplan–Meier integral $\int_0^\tau \phi(u) d\hat{F}_n(u)$ in terms of sums of independent random variables up to $o_p(n^{-1/2})$. We specialise to absolutely continuous distributions $F$ and $G$ and assume $\tau = \infty$. This rather special case leads to a transparent derivation of the main variance formula and allows to compare the results of Stute (1995) with those of Yang (1994). In particular, Stute’s moment condition (1.5) and Yang’s condition (ii) (Yang 1994: 339) simply reads

$$\int_0^\infty \frac{\phi(u)^2}{1 - G(u)} dF(u) < \infty$$

Stute’s representations in the special case can be written

$$\frac{\delta \phi(Z)}{1 - G(Z)} + (1 - \delta) \frac{\mathbb{E}(\phi(T) | T > Z)}{1 - G(Z)}$$
Figure 6.3.: Density estimates for $A_{2i}$. Solid line: $t = 0.2$, short dashed line: $t = 0.5$, dashed and dotted line: $t = 1.0$, long dashed line: $t = 1.5$.

\[
\int \int \frac{\phi(w) \, 1(v < Z) \, 1(v < w)}{1 - H(v)} \, dF(w) \, d\Lambda^G(v) \quad (6.12)
\]

where $\Lambda^G(t) := \int_0^t \frac{dG(u)}{1 - G(u)}$ is the integrated hazard function of $G$. The expectation of the above expression with respect to $(\delta, Z)$ is easily seen to be $E(\phi(T))$: Compute the conditional expectation of the first term given $\{T = t\}$ to see that the expectation of that term is $E(\phi(T))$. For the last two terms, simply write out the expectation with respect to $(\delta, Z)$.

Subtracting the expectation $E(\phi(T))$ and evaluating (6.12) at a fixed
argument \( t \) we get

\[
\delta \left[ \frac{\phi(t)}{1 - G(t)} - \frac{\mathbb{E}(\phi(T) \mid T > t)}{1 - G(t)} \right] + \frac{\mathbb{E}(\phi(T) \mid T > t)}{1 - G(t)}
- \int \int \frac{\phi(w) \mathbb{1}(v < t) \mathbb{1}(v < w)}{1 - H(v)} dF(w) d\Lambda^G(v)
- \mathbb{E}(\phi(T))
\]

(6.13)

We will start by re-expressing the double integral. For this, note that \( \Lambda^H = \Lambda^F + \Lambda^G \) from the definition of the distribution function of \( Z \). Thus,

\[
\int \int \frac{\phi(w) \mathbb{1}(v < t) \mathbb{1}(v < w)}{1 - H(v)} dF(w) d\Lambda^G(v)
= \int \int \frac{\phi(w) \mathbb{1}(v < t) \mathbb{1}(v < w)}{1 - H(v)} dF(w) d(\Lambda^H(v) - \Lambda^F(v))
= \int \int \frac{\phi(w) \mathbb{1}(v < t) \mathbb{1}(v < w)}{(1 - H(v))^2} dH(v) dF(w)
- \int \int \frac{\phi(w) \mathbb{1}(v < t) \mathbb{1}(v < w)}{(1 - H(v))(1 - F(v))} dF(v) dF(w)
= \int \left( \frac{1}{1 - H(\min(t, w))} - 1 \right) \phi(w) dF(w)
- \int \int \frac{\phi(w) \mathbb{1}(v < t) \mathbb{1}(v < w)}{(1 - H(v))(1 - F(v))} dF(v) dF(w)
\]

where we used Fubini’s theorem in the second equation, and where the first term in the third equation results from a transform of variables.

The last three terms in (6.13) can thus be written as

\[
\frac{\mathbb{E}(\phi(T) \mid T > t)}{1 - G(t)} - \int \int \frac{\phi(w) \mathbb{1}(v < t) \mathbb{1}(v < w)}{1 - H(v)} dF(w) d\Lambda^G(v)
\]
\[ - \mathbb{E}(\phi(T)) \]
\[ = \int_t^\infty \frac{\phi(w)}{1 - H(t)} dF(w) \]
\[ - \iint \frac{\phi(w) \mathbb{I}(v < t) \mathbb{I}(v < w)}{1 - H(v)} \, dF(w) \, d\Lambda^G(v) \]
\[ - \int_0^\infty \phi(w) \, dF(w) \]
\[ = \int_t^\infty \frac{\phi(w)}{1 - H(t)} \, dF(w) - \int \left( \frac{1}{1 - H(\min(t, w))} - 1 \right) \phi(w) \, dF(w) + \iint \frac{\phi(w) \mathbb{I}(v < t) \, dF(v) \, d\Lambda^G(v)}{(1 - H(v))(1 - F(v))} - \int_0^\infty \phi(w) \, dF(w) \]
\[ = - \int_0^t \frac{\phi(w)}{1 - H(w)} \, dF(w) + \int_0^t \frac{\mathbb{E}(\phi(T) | T > v)}{1 - H(v)} \, dF(v) \]
\[ = - \int_0^t \frac{\phi(w)}{1 - G(w)} \, d\Lambda^F(w) + \int_0^t \frac{\mathbb{E}(\phi(T) | T > v)}{1 - G(v)} \, d\Lambda^F(v) \]
\[ = - \int \frac{\phi(w) - \mathbb{E}(\phi(T) | T > w)}{1 - G(w)} \, \mathbb{I}(w < t) \, d\Lambda^F(w) \]

Defining
\[ M_i(t) := \mathbb{I}[t_i \leq t, \delta_i = 1] - \int_0^t \frac{\mathbb{I}[z_i \geq u]}{1 - F(u)} \, dF(u) \]

to be the martingale for the counting process \( \mathbb{I}[t_i \leq t, \delta_i = 1] \) with respect to the standard filtration, we can finally write (6.12) for the \( i \)-th observation as
\[ \int \frac{\phi(u) - \mathbb{E}(\phi(T) | T > u)}{1 - G(u)} \, dM_i(u) \]  

(6.14)

In sum, we have the representation
\[ \int_0^\infty \phi(u) \, d\hat{F}_n(u) - \int_0^\infty \phi(u) \, dF(u) = \]
\[
\frac{1}{n} \sum_{i=1}^{n} \int_{0}^{\infty} \frac{\phi(u) - \mathbb{E}(\phi(T) \mid T > u)}{1 - G(u)} \, dM_i(u) + o_P(n^{-1/2})
\]  

(6.15)

which is just (6.7).

With the last representation at hand it is easy to derive a variance expression using standard martingale arguments:

\[
\text{Var}_{Z, \delta} \left( \int_{0}^{\infty} \phi(u) \, d\hat{F}_n(u) - \int_{0}^{\infty} \phi(u) \, dF(u) \right)
\]

\[
= \mathbb{E}_{Z, \delta} \left( \left( \int_{0}^{\infty} \phi(u) \, d\hat{F}_n(u) - \int_{0}^{\infty} \phi(u) \, dF(u) \right)^2 \right)
\]

\[
\approx \mathbb{E}_{Z, \delta} \left( \left( \int_{0}^{\infty} \frac{\phi(u) - \mathbb{E}(\phi(T) \mid T > u)}{1 - G(u)} \, dM(u) \right)^2 \right)
\]

\[
= \mathbb{E}_{Z, \delta} \left( \int_{0}^{\infty} \left( \frac{\phi(u) - \mathbb{E}(\phi(T) \mid T > u)}{1 - G(u)} \right)^2 \, d\langle M, M \rangle(u) \right)
\]

\[
= \mathbb{E}_{Z, \delta} \left( \int_{0}^{\infty} \mathbb{1}(Z > u) \left( \frac{\phi(u) - \mathbb{E}(\phi(T) \mid T > u)}{1 - G(u)} \right)^2 \, d\Lambda^F(u) \right)
\]

\[
= \int \int \mathbb{1}(z > u) \left( \frac{\phi(u) - \mathbb{E}(\phi(T) \mid T > u)}{1 - G(u)} \right)^2 \frac{1}{1 - F(u)} \, dF(u) \, dH(z)
\]

\[
= \int (1 - H(u)) \frac{\left( \phi(u) - \mathbb{E}(\phi(T) \mid T > u) \right)^2}{(1 - H(u))(1 - G(u))} \, dF(u)
\]

\[
= \int \frac{(\phi(u) - \mathbb{E}(\phi(T) \mid T > u))^2}{1 - G(u)} \, dF(u)
\]

This is also Yang’s (1994) variance formula valid for arbitrary \( G \). We will also need the covariances of the representations for \( \phi_1 \) and \( \phi_2 \). Shortening \( \phi_k(u) - \mathbb{E}(\phi_k(T) \mid T > u) \) to \( R(\phi_k)(u) \) we have by a similar reasoning as
above:

\[
\text{Cov} \left( \int_0^\infty \frac{R(\phi_1(u))}{1 - G(u)} \, dM(u), \int_0^\infty \frac{R(\phi_2(u))}{1 - G(u)} \, dM(u) \right) \\
= \mathbb{E}_{Z, \delta} \left( \int_0^\infty \frac{R(\phi_1(u))R(\phi_2(u))}{(1 - G(u))^2} \, d \langle M, M \rangle(u) \right) \\
= \mathbb{E}_{Z, \delta} \left( \int_0^\infty \mathbb{1}(Z > u) \frac{R(\phi_1(u))R(\phi_2(u))}{(1 - G(u))^2} \, d\Lambda^F(u) \right) \\
= \int \int \mathbb{1}(z > u) \frac{R(\phi_1(u))R(\phi_2(u))}{(1 - G(u))^2} \, d\Lambda^F(u) \, dH(z) \\
= \int_0^\infty \frac{R(\phi_1(u))R(\phi_2(u))}{1 - G(u)} \, dF(u)
\]

6.9. Postscriptum

The suggested estimator in this paper is of the form

\[
\hat{m}(t) = \int_t^\infty \phi(u) \, dF_n(u)
\]

where \(F_n(.)\) is the Kaplan-Meier estimator of the distribution function. The analysis then proceeds by replacing the dependent weights of the Kaplan-Meier estimator by a representation using independent contributions. This was first suggested by Yang (1994) and Stute (1995).

These results have recently been extended to cases where in the integral depends on further parameters or when covariates are present. In particular, Sellero et al. (2005) consider estimates of integrals of the form

\[
\hat{m}(t) = \int_t^\infty \int \phi(u, v) \, dF_n(u, v)
\]

where \(F_n(., .)\) is the joint distribution function to possibly censored variable \(T\) and a covariate \(Y\). Uña-Álvarez and Rodríguez-Campos (2004) use a smoothed version of the Kaplan-Meier estimator to provide
estimators of expectations of bivariate functions of a possibly censored variable and a covariate.

Wang and Jing (2001) and Qin and Zhao (2007) as well as Zhao and Qin (2006) study the application of empirical likelihood methods to Kaplan-Meier integrals.

Delecroix et al. (2008) propose a weighting procedure similar to the ones proposed by Koul et al. (1981) and Leurgans (1987) in order to estimate conditional expectations of censored variables given a covariate.

While there was not much progress in the analysis of strong representations for more general Kaplan-Meier processes there are now some interesting suggestions for estimators of the mean residual life functions and regression versions of it. In particular, Chen and Cheng (2005) and Chen et al. (2005) use inverse probability of censoring weighted extensions of a model first proposed by Maguluri and Zhang (1994). Instead of a proportional mean residual life function, Chen and Cheng (2006) suggested a linear function of the covariates. Jeong et al. (2008) consider estimators of the median residual life based on inverting functions of the Kaplan-Meier estimator.
7

Covariate Effects in Periodic Hazard Rate Models

7.1. Summary

Labour market participation, consumer behaviour, and many other phenomena exhibit strong periodic patterns that result from cyclic behaviour, constraints on the timing of events, or seasonal variation. While these periodicities can generally be neglected when dealing with small data sets or coarsely grouped event times, they pose challenges to the analysis of large data sets with precise recordings. It seems natural to require that statistical models used in the analysis of such data sets reproduce any underlying periodicities. In particular, the conditional hazard rate given covariates should be periodic for all possible values of the covariates. We show that this requirement severely restricts the class of covariate effects models.

We define periodicities by points of zero crossings of the derivative of the hazard rate. We then develop the concepts of hazard envelope and essential extrema. These allow the construction of classes of covariate effect models with time varying coefficients that respect the underlying periodic structure.
7.2. Introduction

Labour market participation, consumer behaviour, and many other phenomena exhibit strong periodic patterns that result from cyclic behaviour, constraints on the timing of events, or seasonal variation. These phenomena become apparent when large data sets with precise recordings of the timing of events become available. Figure 7.1 exhibits the hazard rate of the inter-purchase time of an 1-litre ice-cream package. The estimate is based on data provided by the German Homescan Panel of A.C. Nielsen. The data contain information on the day of purchases for some 8,400 households over a period of three years.

![Hazard rate of inter-purchase times (days) of ice-cream packages.](image)

The (discrete) daily hazard rate oscillates with maxima at 7, 14, 21 days and so on. It is plausible to assume that the reason for this behaviour is the weekly purchase schedule of most households. This argument is supported by the fact that these patterns occur across sociodemographic subgroups, e.g. regardless of whether there are children present in the household or not, cf. Figure 7.2.

As a second example, Figure 7.3 presents the estimated hazard rate of
Figure 7.2.: Conditional hazard rate of inter-purchase times (days) of ice-cream packages. Households without children: solid line, Households with children: dotted line.

job durations in Germany for the years 1975 to 1990. The estimate is based on a subsample of records of the social security administration (see Bender et al. 1996) covering some 400,000 job spells. The hazard rate shows large annual peaks, somewhat smaller quarterly peaks and also monthly peaks. However, the number of job durations away from these peaks is still considerable. Again, the findings are similar across subgroups (e.g. men and women). An obvious reason for this pattern is that institutional and juridical regulations in general restrict ending a job to the end of a quarter or to the end of a (calendar) year. Of course, such regulations ought to be the same for all socio-demographic subgroups. In fact, the impact of the regulations is extremely strong: Figure 7.4 depicts the total number of job exits per calendar day for the period 1990–1999. The numbers are based on the complete data of the social security administration. The number of job exits not coincident with a month’s end is generally below 100.
7. Covariate Effects in Periodic Hazard Rate Models

Figure 7.3.: Hazard rate of job durations (days) in Germany 1975–1990.

Figure 7.4.: Total number (in million) of job exits by day in Germany 1990–1999. Number for 31.12.1992 truncated.

7.3. Marginal and Conditional Hazard Rates

In both examples, strong external influences cause both the conditional and the marginal hazard rates to oscillate. These influences are common in the sense that local maxima and minima appear at the same times for the
7.3. Marginal and Conditional Hazard Rates

marginal as well as for the conditional hazard rates. Regression models should account for this periodic behaviour: The conditional hazard rates implied by the models should exhibit the same periodicities for all values of the covariates. Moreover, the periodicities of the conditional hazard rates should be the same as those of the implied marginal hazard rates. Surprisingly, however, no regression model strictly satisfies these requirements.

Let $T > 0$ and $X$ be random variables on a common probability space representing duration and covariate information. Denote by $\lambda(t) = \frac{f(t)}{1 - F(t)}$ and $\lambda(t \mid x) = \frac{f(t \mid x)}{1 - F(t \mid x)}$ the marginal and conditional hazard rates. Here, the conditioning is on the events $\{X = x\}$, $f(t)$ and $f(t \mid x)$ are the marginal and conditional densities, and $F(t)$ and $F(t \mid x)$ are the marginal and conditional cumulative distribution functions.

For simplicity, we assume that $\lambda(t \mid x)$ is twice continuously differentiable with respect to $t$. If $t_1, t_2, \ldots$ are the locations of minima and maxima of $\lambda(t \mid x)$, then $\dot{\lambda}(t_i \mid x) = 0$, $i = 1, 2, \ldots$, where $\ddot{\lambda}(t \mid x)$ is the derivative of the conditional hazard rate with respect to $t$. A possible though rather strict formulation of the above requirements becomes: There is a sequence $0 < t_1 < t_2 \ldots$ such that

$$\dot{\lambda}(t_i) = 0 = \dot{\lambda}(t_i \mid x), \text{ for } i = 1, 2, \ldots \text{ and for all } x. \quad (7.1)$$

To see that in fact no non-trivial model can satisfy this condition, we need to consider the relation between marginal and conditional hazard rates and their derivatives. The marginal hazard rate is given by a time-dependent “convex combination” of the conditional hazard rates:

$$\lambda(t) = \mathbb{E}(\lambda(t \mid X) \mid T \geq t) \quad (7.2)$$

where the expectation is with respect to the distribution of $X$ conditional on the event $\{T \geq t\}$.

Differentiating this relation leads to

$$\dot{\lambda}(t) = \mathbb{E}(\dot{\lambda}(t \mid X) \mid T \geq t) + [\ddot{\lambda}(t)^2 - \mathbb{E}((\lambda(t \mid X)^2 \mid T \geq t)] \quad (7.3)$$
When the derivatives $\dot{\lambda}(t \mid x)$ vanish for all $x$, the first term becomes 0. By Jensen’s inequality, the second term is negative unless $\lambda(t \mid x)$ is constant in $x$. Thus, if all derivatives of conditional hazard rates vanish at a point $t_i$, then the derivative of the marginal hazard rate has to be negative.

To illustrate, consider two subgroups distinguished by the covariate $X \in \{0, 1\}$ and assume

$$\Pr := \Pr(X = 0) = \frac{1}{2} = \Pr(X = 1) := \Pr,$$

$$\lambda_0(t) := \lambda(t \mid X = 0) = \frac{5}{4} + \sin t,$$

$$\lambda_1(t) := \lambda(t \mid X = 1) = 2 \cdot \lambda_0(t),$$

Then

$$\lambda(t) = \frac{\lambda_0(t) \cdot \exp(-\Lambda_0(t)) + \lambda_1(t) \cdot \exp(-\Lambda_1(t))}{\exp(-\Lambda_0(t)) + \exp(-\Lambda_1(t))},$$

where

$$\Lambda_i(t) := \int_0^t \lambda_i(s) \, ds, \quad i = 0, 1,$$

are the cumulative hazard rates of the subgroups. Figure 7.5 displays the different extrema of the marginal hazard rate $\lambda(t)$ and the conditional hazard rates $\lambda_0(t)$ and $\lambda_1(t)$. Although $\lambda(t) \in [\lambda_0(t), \lambda_1(t)]$ holds due to (7.2), the time-dependence of the convex combination causes the derivative $\dot{\lambda}$ to vanish at different times than $\dot{\lambda}_0$ and $\dot{\lambda}_1$.

7.4. Hazard Envelopes

On the other hand, if the conditional hazard rates $\lambda_i(t), \ i = 0, 1$, oscillate strongly and commonly in the sense that (w.l.o.g.)

$$\ddot{\lambda}_0(t_{2i}) < 0, \quad \ddot{\lambda}_1(t_{2i}) < 0,$$

then
Figure 7.5.: Due to the time-dependence of (7.2), the marginal hazard rate (dashed line) does not have the same extrema as the conditional hazard rates (solid lines).

\[
\begin{align*}
\dot{\lambda}_0(t_{2i-1}) &> 0, & \dot{\lambda}_1(t_{2i-1}) &> 0, \\
\lambda_0(t_{2i}) &> \lambda_1(t_{2i-1}), & \lambda_0(t_{2i}) &> \lambda_1(t_{2i+1}), \ \forall i \geq 1,
\end{align*}
\]

hold, then the marginal hazard rate \( \lambda(t) \), being a pointwise convex combination of \( \lambda_0(t) \) and \( \lambda_1(t) \) as well as a differentiable function of \( t \), must have a local maximum in each interval \( (t_{2i-1}, t_{2i+1}) \), \( i \geq 1 \), and a local minimum in each interval interval \( (t_{2i}, t_{2i+2}) \), \( i \geq 1 \). Qualitatively speaking, \( \lambda(t) \) oscillates as well, cf. Figure 7.6.

Thus the marginal hazard rate will oscillate in a similar way as the conditional hazard rates if the latter have common minima and maxima and if they oscillate strongly enough. To bound the behaviour of the marginal hazard rate, we introduce the concept of the hazard envelope \((\underline{\lambda}(t), \overline{\lambda}(t)):\n\]

\[
\underline{\lambda}(t) := \inf_x \{ \lambda(t \mid x) \} , \quad \overline{\lambda}(t) := \sup_x \{ \lambda(t \mid x) \}
\]

(7.4)

Note that in general neither \( \underline{\lambda}(t) \) nor \( \overline{\lambda}(t) \) need to correspond to any member of the family of conditional hazard rates. Nevertheless, it follows
from (7.2) that the marginal hazard rate at a given $t$ is bounded by the extreme points of the conditional hazard rates. Thus

$$\lambda(t) \leq \lambda(t) \leq \lambda(t) \quad \forall t$$ (7.5)

Suppose next that maxima of the hazard envelope occur at even numbered times $t_{2i}$ while minima occur at odd numbered times $t_{2i-1}$. Suppose further that the conditional hazard rates have common minima and maxima at $t_{2i}$ and $t_{2i-1}$, respectively. We say that the conditional hazard rates have an *essential* maximum at $t_{2i}$ if

$$\lambda(t_{2i-1}) < \lambda(t_{2i}) > \lambda(t_{2i+1})$$ (7.6)

We say that the conditional hazard rates have an *essential* minimum at $t_{2i+1}$ if

$$\lambda(t_{2i}) > \lambda(t_{2i+1}) < \lambda(t_{2i+2})$$ (7.7)

An essential maximum implies at least one maximum of the marginal hazard rate in the interval $(t_{2i-1}, t_{2i+1})$, while an essential minimum implies at least one minimum of the marginal hazard rate in the interval $(t_{2i}, t_{2i+2})$. 

Figure 7.6.: The marginal hazard rate necessarily oscillates between the two dashed lines.
7.5. Consequences for Model Choice

We are now in the position to formulate more reasonable requirements for regression models in situations with strong periodicities: Suppose there is a sequence of time points \( t_{2i}, i \geq 1 \) at which maxima of the hazards are to occur. Think of the weekly maxima in the hazard rate of inter-purchase times or the quarterly maxima in the hazard rate of job durations. Then one might want to restrict attention to models of covariate effects that, firstly, admit maxima of the conditional hazard rates at the \( t_{2i} \) for all values of the covariates, that, secondly, admit the existence of common minima of the conditional hazard rates at some sequence of times \( t_{2i-1} \), and that thirdly, admit essential maxima at all the \( t_{2i} \) even in the presence of non-trivial covariate effects.

Consider the class of proportional hazards models with

\[
\lambda(t \mid x) = \lambda_0(t)\psi(x\beta), \quad \psi(x\beta) > 0
\]  

(7.8)

For this class the envelope hazard coincides with certain conditional hazard rates if the support of the distribution of the covariates is compact. The situation is then very similar to that depicted in Figures 7.5 and 7.6. The first and second requirements are easily met. In fact, they simply depend on the choice of an appropriate baseline hazard rate \( \lambda_0(t) \). Whether or not a local maximum is essential will depend both on the extent of covariate effects and the amplitude of the baseline hazard rate. Thus proportional hazard rate models are certainly feasible candidate models.

But what happens if one wants, for some good reason, use non-proportional hazards models? Consider the accelerated failure time model. This model posits a scaling effect of covariates: Suppose that \( T_x \) is a random variable representing duration conditional on the covariate value \( x \). Suppose further that there is a random variable \( T_0 \) on the same probability space as \( T_x \) such that

\[
T_x = T_0/\psi(x\beta), \quad \psi(x\beta) > 0
\]  

(7.9)
and such that the $T_0$ have identical distributions for all values of the
covariates. The conditional hazard rates are then of the form:

$$
\lambda(t \mid x) = \psi(x\beta) \cdot \lambda(\psi(x\beta) \cdot t), \ \psi(x\beta) > 0
$$

(7.10)

But in such a model, maxima and minima of conditional hazard rates for
different values of the covariates cannot coincide, except in the trivial
case of no covariate effect, $\psi(x\beta) \equiv 1$.

Does this rule out the use of accelerated failure time models and many
other non-proportional hazards models? Not if one is prepared to allow
the effects of covariates to change with time. But how does one allow for
time-varying covariate effects without destroying the defining features of
the accelerated failure time model? After all, if one allows for general
time dependent effects $\beta(t)$ and plugs this into (7.10), then the “scaling
the time axis” property is destroyed, while there is no obvious way of
plugging a time indexed $\beta(t)$ into (7.9).

There is, however, a quite natural way to define changing covariate effects
that respects the scaling interpretation of accelerated failure time models.
One has to change the global “change of scale” interpretation of covariate
effects into a local property at a point in time. That can be done by using
derivatives. Starting with the “scale change” interpretations in terms
of random variables as in (7.9), one can consider the derivative of the
baseline duration with respect to duration with covariate value $x$. That
derivative should be influenced, at a point in time, by the covariate effect
at that same point in time. One might thus write

$$
\frac{\partial t_0}{\partial t_x} \bigg|_{t_x = u} = \psi(x\beta(u))
$$

(7.11)

But then

$$
t_0 = \int_0^{t_x} \psi(x\beta(u)) \, du =: \Psi(t_x; \tilde{\beta})
$$

where $\tilde{\beta}$ contains the covariate information and the changes in covariate
effects. Therefore

$$
T_x = \Psi^{-1}(T_0; \tilde{\beta})
$$
The hazard rate corresponding to this model of covariate effects is

$$\lambda(t \mid x) = \psi(x\beta(t)) \cdot \lambda_0(\Psi(t; \hat{\beta})) \quad (7.12)$$

Note that this differs from the naive idea to plug in some $\beta(t)$ into (7.10) while it preserves the interpretation of the effects of covariates as a (local) scaling of the time axis.\(^1\)

With this definition of varying covariate effects it is now easy to exhibit versions of the accelerated failure time model that do respect the requirements formulated at the beginning of this section. If we choose

$$\Psi(t_{2i}; \hat{\beta}) = t_{2i}$$
$$\dot{\Psi}(t; \hat{\beta}) > 0$$
$$\ddot{\Psi}(t_{2i}; \hat{\beta}) = 0 \text{ and } \dot{\lambda}_0(t_{2i}) = 0$$

then

$$\dot{\lambda}(t_{2i} \mid x) = \dot{\psi}(x\beta(t_{2i})) \cdot \lambda_0(\Psi(t_{2i}; \hat{\beta})) + \psi(x\beta(t_{2i}))^2 \cdot \dot{\lambda}_0(\Psi(t_{2i}; \hat{\beta})) = 0$$

With this choice of $\Psi(t)$, $\beta(t)$, and $\lambda_0(t)$, all conditional hazard rates will have the same points of maxima as the baseline $\lambda_0(t)$. As in the case of proportional hazards models, whether the maxima are essential depends on the amplitude of $\lambda_0(t)$ and on the covariate effects process $\psi(x\beta(t))$. But since the envelope hazard will in general not coincide with any of the conditional hazard rates, one needs to compute the envelope hazard explicitly.

7.6. Postscriptum

As far as I am aware of the literature, there is no newer contribution to the analysis of covariate effects in models with periodic behaviour. There have been contributions to the analysis of Poisson point processes

\(^1\) A version of this model for time-dependent covariates has been proposed by Cox and Oakes (1984: 67). Robins and Tsiatis (1992) developed an estimator for this model.
7. Covariate Effects in Periodic Hazard Rate Models

with periodic intensities (see Helmers et al. (2007) for a recent review). However, this literature is silent on the consequences on the form of regression models or the general form of regression models.
8

A Multivariate Buckley-James Estimator

8.1. Summary

Buckley and James (1979) extended the least-squares estimator to cover the case of censored dependent variables. I consider a generalisation of their estimator to the multivariate case based on a non-parametric estimator of the joint distribution of the residuals.

8.2. Introduction

Buckley and James (1979) introduced a regression technique suitable for censored dependent variables. Their estimator uses the least-squares estimating equations and an updating mechanism based on a non-parametric estimator of the residual distribution to deal with the censoring. The procedure is attractive because the use of the least-squares technique allows for an easy interpretation of results and the use of residual analysis, while the updating scheme is general enough to accommodate various forms of censoring and grouping. Consequently, many generalisations of the basic technique have been proposed.
In this paper I explore a possible extension to multivariate dependent variables. Related work, especially that of Lin and Wei (1992), Lee, Wei and Ying (1993), Pan and Kooperberg (1999), and Hornsteiner and collaborators in a series of papers (1996, 1997, 1998), is mainly inspired by the literature on generalised estimating equations. It concentrates on the estimation of the marginal effects of covariates on each of the dependent variables. Accordingly, the least-squares estimating equations are modified to accommodate the multivariate character of the dependent variables. Less emphasis is put on the updating scheme that deals with the censoring problem. The authors suggest to use non-parametric estimators of the marginal distributions of the residuals only. I propose to incorporate the multivariate information from the residual distribution into the updating scheme.

In the next section I introduce Buckley and James’ approach to regression estimation with censored observations and, in section 8.4, indicate why it works. Next I consider the multivariate case. Generalisations of the missing information principle are treated in section 8.6. This leads to a multivariate extension of Buckley and James’ approach that uses the multivariate information also for the updating scheme. In the final section I examine the performance of the estimator through examples.

8.3. Buckley-James Estimators

Suppose that conditionally on some covariates $x$, the random variable $Y$ follows a linear regression

$$Y = x\beta + \epsilon$$  \hspace{1cm} (8.1)

where $x$ is a $1 \times p$ vector of covariates including a constant, $\beta$ is a $p \times 1$ vector of unknown regression coefficients, and $\epsilon$ is a random variable with mean zero and finite variance. If $Y$ is the logarithm of a positive random variable representing a duration or time to an event, this model is sometimes called accelerated failure time model (Cox/Oakes 1984: Chap. 5.2).
In many applications only censored observations from $Y$ are available. More precisely, suppose that the observations are given by the censored variable $Z$ and censoring indicator $\delta$:

$$Z := \min(C, Y), \quad \delta := \mathbb{I}[C \geq Y],$$

where $\mathbb{I}$. is the indicator function and the censoring variable $C$ is (conditionally) independent of $Y$. The observations are $n$ independent and identically distributed realizations from $(x, \epsilon, C)$. The $n \times (p + 2)$ data matrix is given by $(z_i, \delta_i, x_i)_{i=1,\ldots,n}$.

In the absence of censoring one can estimate $\beta$ by minimising the least-squares criterion

$$\sum_{i=1}^{n} (y_i - x_i\beta)^2 = n \int e^2 \hat{F}_n(e) = \sum_{i=1}^{n} \int (y - x_i\beta)^2 d\hat{F}_{ni}(y) \quad (8.2)$$

where $\hat{F}_n(e)$ is the empirical distribution function of the residuals $e_i = y_i - x_i\beta$, and $\hat{F}_{ni}(y) = \mathbb{I}[y_i < y]$ is the empirical distribution of just one observation $y_i$.

Miller (1976) and Leurgans (1987), using the second and third representation respectively, proposed replacing the empirical distributions by versions appropriate for censored data. Instead of taking the least-squares criterium (8.2) as their starting point, Buckley and James (1979) suggested to modify the least-squares estimating equations

$$\sum_{i=1}^{n} x'_i(y_i - x_i\hat{\beta}) = 0 \quad \text{or} \quad \sum_{i=1}^{n} x'_iy_i = \left(\sum_{i=1}^{n} x'_ix_i\right) \hat{\beta} \quad (8.3)$$

In the presence of censoring they proposed to replace the censored observations $Z$ by the conditional expectation of $Y$ given the observed (censored) data $Z$ and the covariates:

$$Y^* = \mathbb{E}_\beta(Y \mid z, \delta, x) = \delta z + (1 - \delta)\mathbb{E}_\beta(Y \mid Y \geq z, x) \quad (8.4)$$

Note the dependence of the conditional expectation on the unknown parameter $\beta$. Replacing $Y$ in expression (8.3) by its conditional expectation
gives
\[
\frac{1}{n} \sum_{i} x_i' \hat{E}_\beta(Y \mid z_i, \delta_i, x_i) = \frac{1}{n} \left( \sum_{i=1}^{n} x_i' x_i \right)^{\hat{\beta}}
\]  
(8.5)

In other words, the Buckley-James estimator \( \hat{\beta} \) solves the normal score function for \( \beta \) when the expectation on the left hand side is computed using \( \hat{\beta} \).

Using the model formula (8.1) and a fixed \( \beta \), an empirical version of the conditional expectation can be evaluated:

\[
\hat{E}_\beta(Y \mid z_i, \delta_i, x_i) = \hat{y}_i(\beta) = \delta_i z_i + (1 - \delta_i) \left( x_i \beta + \int_{e_i}^{\infty} e d\hat{F}_\beta(e) \right) \frac{\hat{S}_\beta(e_i)}{\hat{S}_\beta(e_i)}
\]

\[
= \delta_i z_i + (1 - \delta_i) \left( x_i \beta + \sum_{k=i}^{n} v_{ik}(\beta)(z_k - x_k \beta) + x_i \beta \right)
\]

(8.6)

where \( \hat{F}_\beta \) is an estimator of the distribution function of the residuals (e.g. the Kaplan-Meier estimator), \( \hat{S}_\beta \) is the estimated survivor function \( 1 - \hat{F}_\beta \), and I have put

\[
v_{ik}(\beta) = \begin{cases} 
    \frac{w_k(\beta)}{\hat{S}_\beta(e_i)} & \text{if } e_i < e_k \\
    0 & \text{otherwise}
\end{cases}
\]

and

\[
w_k(\beta) = \hat{P}_\beta(\epsilon = e_k)
\]

so that \( w_i(\beta) \) is the height of the jump of the estimated distribution at the \( i \)-th residual.\(^1\) A solution \( \hat{\beta} \) of the estimating equation (8.3) therefore

\(^1\) For ease of notation it is assumed here that the observations are ordered according to the magnitude of the corresponding residuals.
satisfies:

\[
\hat{\beta} = \left( \sum_{i=1}^{n} x'_i x_i \right)^{-1} \left( \sum_{i=1}^{n} \delta_i x'_i z_i + \sum_{i=1}^{n} (1 - \delta_i) x'_i \hat{y}_i(\hat{\beta}) \right)
\]  

(8.7)

This leads to a straightforward iterative procedure for the computation of \( \hat{\beta} \):

1. Assign starting values \( \hat{\beta}^0 \).
2. Compute \( \hat{y}_i(\hat{\beta}^j) \) according to (8.6) using the Kaplan-Meier procedure as an estimator for the distribution of the residuals.
3. Compute \( \hat{\beta}^{j+1} \) using the right hand side from (8.7).
4. Go back to step 2 unless some convergence criterion is met.

To be numerically efficient, this simple iterative strategy needs elaboration. Following the steps of the algorithm, the basic choices are:

1. Starting values may be obtained using the least-squares estimator treating all observations as uncensored. This was suggested by Buckley and James (1979). Other choices, e.g. using only uncensored observations, are of course possible, but do not seem to have a decisive influence on the procedure.

2. The Kaplan-Meier estimator is not uniquely defined on the whole real line if the largest residual is censored. Buckley and James suggest to always treat the largest residual as uncensored. This will lead to an underestimation of the regression constant, but should scarcely affect the other regression estimators. Other choices are discussed by Efron (1988), while Lai and Ying (1991) propose to smooth the risk sets.

4. The iteration may not converge to a unique value. This is due to the fact that the right hand side of (8.7) is a piecewise linear function in \( \beta \). Changing \( \beta \) does not change the weights \( v_{ik}(\beta) \) unless the ranks of the residuals change. Therefore, the iterations may oscillate between several values \( \hat{\beta} \). The discontinuity of (8.7) hampers the analytic treatment of the estimator. Moreover, the
number of limiting values in finite samples is not predictable, but may potentially be rather large (Currie, 1996). Fortunately, the phenomenon seems to be of practical interest only in rather small samples, in situations where the effect of covariates is small, or when the convergence criterion is very strict (Wu/Zubovic, 1995).²

8.4. Score Functions and Censoring

To appreciate why the Buckley-James procedure is a “good” generalisation of estimating equations to censored variables it is helpful to consider it from a more general point of view. Especially the relation between score functions with and without censoring is revealing. Write \( \ell(\beta) = \ell(\beta; Y, x) = x'(Y - x\beta) \) for the score function from the normal linear regression model (8.1). The expectation satisfies

\[
\mathbb{E}_\beta(\ell(\beta; Y, x)) = 0 \tag{8.8}
\]

Moreover, the root \( \hat{\beta} \) of the empirical version of the expectation (8.8),

\[
\frac{1}{n} \sum_i \ell(\hat{\beta}; y_i, x_i) = 0
\]

is the maximum likelihood estimator. Even if the distribution is not normal — so that the root of the score function need no longer be a maximum likelihood estimator — \( \hat{\beta} \) is consistent and often highly efficient. In the presence of censoring, the censored normal score function \( \ell^* \) can be expressed as

\[
\ell^*(\beta; Z, \delta, x) = \mathbb{E}(\ell(\beta; Y, x) \mid Z, \delta, x) \tag{8.9}
\]

the conditional expectation of the score function with complete observations given the incomplete observations (see e.g. Ibragimov/Has’minskii

² Wu and Zubovic (1995) suggested to use the arithmetic mean of all limit values of the algorithm as estimator. This suggestion may be useful in situations where a unique estimator is required (e.g. simulations, using the procedure as building block for more complicated models, etc.). Otherwise, the different values of the limiting cycle of estimators are often very close and it may suffice to report just one of them.
1981: Chap. 1.7). This relation between score functions for complete and incomplete observations makes the score function an attractive starting point for the construction of estimators.

It remains to consider the computation of the conditional expectation. From the perspective of the normal linear regression model one might try to use the normal distribution. This was proposed by Schmee and Hahn (1979) and Aitkin (1981). However, one can only expect the good properties of the estimators even outside the normal distribution to extend to censored data situations if the conditional expectation is computed from a non-parametric estimator. In the case of right censored observations, the Kaplan-Meier estimator, being a non-parametric maximum likelihood estimator solving a self-consistency equation, seems to be an appropriate choice. In fact, Lai and Ying (1994), following Ritov (1990) and Severini and Wong (1992), provide a general argument for the use of self-consistent estimators in the computation of conditional expectations for censored and truncated observations. To outline the reasoning it is best to regard the estimation problem as one involving both $\beta$ and the distribution of $\epsilon, F$, as unknown parameters. Here, $\beta$ is the parameter of interest and $F$ is treated as a nuisance parameter. In such a context, one may consider the score function corresponding to the profile likelihood. The profile log-likelihood is derived from the log-likelihood $\ell(\beta, F)$ by replacing $F$ with an estimator $\hat{F}_\beta$ treating $\beta$ as known. It is thus a function of $\beta$ only. Symbolically, then, one may write

$$
\frac{d}{d\beta} \ell(\beta, \hat{F}_\beta) = \frac{\partial}{\partial \beta} \ell(\beta, F)|_{(\beta, \hat{F}_\beta)} + \frac{\partial}{\partial F} \ell(\beta, F)|_{(\beta, \hat{F}_\beta)} \frac{\partial}{\partial \beta} \hat{F}_\beta \tag{8.10}
$$

for its score function. If $\hat{F}_\beta$ is of maximum likelihood type, the sample mean of the second term vanishes. One needs only to consider the score function for $\beta$ that would result if $F$ was known.

This holds for all unbiased estimating equations for $F_\beta$. But an estimator $\hat{F}_\beta$ that maximises the likelihood $\ell(\beta, F)$ in $F$ for $\beta$ fixed automatically provides an estimator of the least favourable submodel $\beta \mapsto (\beta, \hat{F}_\beta)$ for

---

3 The argument extends to estimating equations that are not derived from likelihood functions. See Bickel et al. (1998: Chap. 7.7) for a general discussion of the construction of estimators along these lines.
the estimation of $\beta$ and therefore (8.10) approximates the efficient score function, making efficient estimation of $\beta$ feasible. Thus one would like to use an estimator $\hat{F}_\beta$ that simultaneously solves an estimating equation and maximises a non-parametric likelihood.

Taking $F_\beta(u) - \mathbb{1}[Y - x\beta \leq u]$ as a score function for $F_\beta$ in the uncensored case, one is led via the projection of scores (8.9) to an estimator of $F_\beta$ that satisfies the corresponding self-consistency equation, namely

$$0 = \mathbb{E}_n \hat{\ell}^*(\hat{F}_\beta) = \mathbb{E}_n(\mathbb{E}_{F_\beta} \hat{\ell}(\hat{F}_\beta) \mid Z, \delta, x) = \hat{F}_\beta(u) - \frac{1}{n} \sum_{i=1}^n \hat{F}_\beta(u \mid z_i, \delta_i, x_i)$$

(8.11)

But the estimator $\hat{F}_\beta$ that solves the self-consistency equations and maximises the non-parametric likelihood is the Kaplan-Meier estimator. On the other hand, considering $\mathbb{E}(\partial \ell(\beta, \hat{F}_\beta; Y, x) / \partial \beta \mid Z, \delta, x)$ as the profile score function in the presence of censoring, one is led to the estimating equations

$$0 = \mathbb{E}_n \mathbb{E}_{F_\beta}(\hat{\ell}(\hat{\beta}; Y, x) \mid Z, \delta, x)$$

$$= \frac{1}{n} \sum_{i=1}^n x'_i(x_i \hat{\beta} - \mathbb{E}_{F_\beta}(Y \mid z_i, \delta_i, x_i))$$

$$= \frac{1}{n} \sum_{i=1}^n x'_i(x_i \hat{\beta} - \hat{y}_i(\hat{\beta}))$$

leading back to (8.5). From this perspective, then, both the choice of the normal score function $\hat{\ell}(\beta) = x'(Y - x\beta)$ as a starting point and the use of the Kaplan-Meier estimator are the appropriate extension of an estimating equation technique to censored data.

8.5. Multivariate Extensions

To consider the multivariate situation I write $Y = (Y_1, \ldots, Y_k)'$ for the column vector of $k$ dependent variables. The covariates are given by a
A $k \times kp$ matrix $\mathbf{x}$ where the $j$-th row corresponds to the $p$ covariates $x_j$ of the $j$-th dependent variable $Y_j$ with zeros padded in the appropriate places. The regression coefficients are given by a column vector $\beta$ of dimension $kp \times 1$. The multivariate linear model can then be presented as

$$Y = \mathbf{x}\beta + \epsilon = \begin{pmatrix} x_1 & 0 & \ldots & \ldots & 0 \\ 0 & x_2 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & x_k \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \epsilon \quad (8.12)$$

with residual vector $\epsilon$. The mean of the residuals is $\mathbb{E}(\epsilon) = 0$ and the covariances are given by $\mathbb{E}(\epsilon\epsilon') = \Omega$. As before, the covariate vectors $x_j$ are assumed to contain a constant. Note that in the case of equal effects $\beta_1 = \beta_2 = \ldots = \beta_k$ the $\mathbf{x}$ matrix can be reduced to a $k \times p$ matrix.

Now suppose that the data are censored by a $k$-dimensional variable $\mathbf{C} = (C_1, \ldots, C_k)'$. Instead of $Y$ only the vectors $\mathbf{Z} = (Z_1, \ldots, Z_k)' = (\min(Y_1, C_1), \ldots, \min(Y_k, C_k))' = \min(\mathbf{Y}, \mathbf{C})$ and $\delta = (\mathbb{I}[C_1 \geq Y_1], \ldots, \mathbb{I}[C_k \geq Y_k])' = \mathbb{I}[\mathbf{C} \geq \mathbf{Y}]$ are observed. Note that here and in the sequel minima, indicator functions, and (in-)equalities are interpreted component-wise.

To render the conditional distribution of $Y$ identifiable from the censored version $(\mathbf{Z}, \delta)$ I will assume that the censoring vector $\mathbf{C}$ and the vector $\mathbf{Y}$ are (conditionally on $\mathbf{x}$) independent. Moreover, the support of $\mathbf{Y}$ is assumed to be contained in the support of $\mathbf{C}$.\(^4\)

Using this model, Lin and Wei (1992), Lee, Wei and Ying (1993), and Hornsteiner and collaborators (1996, 1997, 1998) proposed extensions to the one-dimensional Buckley-James estimator. In these papers, a solution to an equation similar to (8.7) is used. Both Lin and Wei (1992), and Lee, Wei and Ying (1993) use $k$ least-squares estimating equations

\(^4\) Some of the censoring patterns of interest in event history analysis, e.g. censoring of the recurrence times in a semi-Markov process by a fixed observation interval, are not easily represented in this setup. Reference to an underlying process would be necessary to line up censorings and durations according to their timing on a common time scale. See Dabrowska and Lee (1996), Li and Lagakos (1997), and Tsai and Crowley (1998) for some discussion.

$$\sum_{i=1}^{n} x'_i V(\hat{\alpha})^{-1}(y_i - x_i \hat{\beta}) = 0$$

(8.13)

in an attempt to gain efficiency. To deal with the censoring, all these proposals use an updating scheme parallel to the one-dimensional case, namely the conditional expectations

$$Y_{j}^{**} = \mathbb{E}_{\beta_j}(Y_j \mid z_j, \delta_j, x_j)
= \delta_j z_j + (1 - \delta_j) \mathbb{E}_{\beta_j}(Y_j \mid Y_j \geq z_j, x_j), \quad j \in \{1, \ldots, k\}$$

(8.14)

from the $j$-th model equation. This leads to the correct mean structure while using only the marginal distributions of the residuals. The conditional expectations are then computed from the marginal Kaplan-Meier estimators of the distribution of the residuals. While Lin and Wei and Lee, Wei and Ying simply use the marginal Kaplan-Meier estimators, Hornsteiner (1998) also considers pooled and weighted versions to increase efficiency in certain situations. In addition to the contributions of $Y_{j}^{**}$ in the updating scheme, both Hornsteiner et al. and Pan and Kooperberg (1999) also base their estimating equations for $\alpha$ on the values of $Y_{j}^{**}$. As Hornsteiner (1998: 49) notes, this approach is approximately valid only if the amount of censoring is small.

8.6. The Missing Information Principle and Non-parametric Estimation of Censored Multivariate Observations

Starting with a score function $\dot{\ell}(\beta)$ derived from a likelihood $\ell(\beta)$, the missing information principle suggests to use the conditional expectation of $\dot{\ell}(\beta)$ based on all the available information, not just the information
from the marginal distributions. Thus one may consider the conditional expectations

\[ Y^*_j = \mathbb{E}_\beta(Y_j \mid z, \delta, x) = \delta_j z_j + (1 - \delta_j) \mathbb{E}_i(Y_j \mid Y_j \geq z_j, (z, \delta, x)), \quad j \in \{1, \ldots, k\} \]

(8.15)

instead of (8.14). This conditional expectation is based on all the information on \( Y \) available from the data while (8.14) uses only the information from the distribution in the \( j \)-th dimension. Extending the argument from section 8.4 one would expect (8.15) to give an appropriate generalisation of one-dimensional censored regression if it was possible to exhibit a self-consistent estimator of the multivariate distribution of the censored residuals. Also, from a more practical point of view, it seems advantageous to use as much information as possible in dealing with the censoring process without imposing strong extraneous assumptions. If the degree of censoring is high and if there is considerable correlation within \( Y \) or \( C \), one might expect (8.15) to perform better than (8.14).

In the context of multivariate proportional hazards models this approach was implicitly suggested by Prentice and Hsu (1997) and Cai and Prentice (1995). On the other hand, this extension has not been discussed in the context of the Buckley-James approach. This is not by accident: in the computation of \( \mathbb{E}(Y \mid Z, \delta, x) \) one would need a non-parametric estimator of the joint distribution of \( \epsilon \) from censored data that additionally should solve a self-consistency equation, maximise a non-parametric likelihood, and, for practical reasons, should allow for easy computation of conditional expectations along half-lines or orthants.

In dimension 2 or higher, there is no unique self-consistent non-parametric maximum likelihood estimator (NPMLE) of the distribution function of \( \epsilon \). In fact, the EM type argument leading to (8.11) will not even result in a consistent estimate. To fix ideas, consider the two-dimensional problem, \( k = 2 \), disregarding covariates for the moment. Suppose one observes \((z_1, z_2, 0, 1)\), censored in the first component, but exactly observed in the second. This says that the underlying tuple \((y_1, y_2)\) is located on the ray \(\{(y_1, y_2) \mid y_1 > z_1, y_2 = z_2\}\) parallel to the first axis. But if the distribution
of \((Y_1, Y_2)\) is absolutely continuous, the probability of obtaining another uncensored observation lying on this ray is 0.

Without uncensored observations on the ray there is no empirical support for the computation of the distribution function along this ray. To compute a self-consistent estimator, one needs an expression for 
\[
\Pr((Y_1, Y_2) \leq (u_1, u_2) \mid (Y_1, Y_2) \in \{(y_1, y_2) \mid y_1 > z_1, y_2 = z_2\})
\]
the last term in the self-consistency equation (8.11) based on a current estimate of the joint distribution. If there are no uncensored observations on the ray, the conditioning event has probability 0 for all sensible starting estimates. Therefore the conditional probability can be defined arbitrarily. But updates of the estimator based on the self-consistency equation will not change due to probability mass transferred from the censored observation to uncensored observations, thus leading to inconsistent estimators.

In response to these difficulties several alternative estimators of the joint distribution of multivariate censored observations have been developed. Pruitt (1993) describes six estimators, summarises their known properties, and compares their small sample behaviour in a limited Monte Carlo experiment. Further comparisons are contained in van der Laan (1997). Some of these estimators are based on a decomposition of the joint distribution into conditional times marginal distributions. The approaches then proceed using the one-dimensional Kaplan-Meier estimator. But the resulting estimators will generally depend on the ordering of the decomposition. Other approaches use smoothing techniques for singly censored observations, thus depending on the choice of a smoothing parameter. The proposals of Dabrowska (1988, 1989) and Prentice and Cai (1992) use special representations of the multivariate survivor function, both representations giving rise to explicit estimators of the distribution function. Gill (1992) provides a lucid introduction to these methods, and both are discussed in Pruitt’s 1993 article. Though computationally attractive, both estimators are neither solutions to some self-consistency equation nor are they of maximum likelihood type.

All these approaches may yield negative mass for the increments of the estimated distribution function (Pruitt 1991). This property is especially disturbing when one is interested in computing conditional expectations.
\[ \mathbb{E}_f(Y_1 \mid Y_1 > z_1, Y_2 = y_2) \] which may result in values \( \leq z_1 \) for these estimators. Moreover, the implied computation of conditional expectations used in (8.15) are indetermined in general and cannot directly be used in a generalisation of the Buckley-James procedure.

In contrast, there is an essentially unique non-parametric estimator for discrete censored data maximising a likelihood. It was first considered by Campbell (1981a,b). This let van der Laan (1995, 1996, 1997) to consider a non-parametric MLE based on discretised censored observations. In the two-dimensional case, let \( D = D_1 \times D_2 \) be a rectangle covering the observations so that \((z_{1i}, z_{2i}) \in D\) for all observations. Partition the side of \( D_1 \) into \( q_1 \) intervals of equal length, \( i_{1,l}, l = 1, \ldots, q_1 \). Partition \( D_2 \) into \( q_2 \) intervals \( i_{2,l}, l = 1, \ldots, q_2 \), also of equal length. This partitions \( D \) into \( q_1 q_2 \) congruent rectangular boxes \( i_{1,l} \times i_{2,m} \). Now coarsen the observations as follows: if the observation is uncensored \(((\delta_1, \delta_2) = (1, 1))\) or censored in both dimensions \(((\delta_1, \delta_2) = (0, 0))\), keep the data as \((z_1, z_2, \delta_1, \delta_2)\). If the observation is censored in only one dimension \(((\delta_1, \delta_2) = (0, 1) \text{ or } (\delta_1, \delta_2) = (1, 0))\), replace the uncensored dimension by the interval it falls into. That is, if the observation is censored in the first dimension, \((z_1, z_2, 0, 1)\), replace \( z_2 \) with the interval \( i_{2,l} \) to which \( z_2 \) belongs. The corresponding \((y_1, y_2)\) are therefore assumed to lie in the strip \( \{(y_1, y_2) \mid y_1 > z_1, y_2 \in i_{2l}\} \). Moreover, the strip is restricted to the domain \( D \), \( \{(y_1, y_2) \mid y_1 > z_1, y_2 \in i_{2l}\} \cap D \). Similarly, observations only censored in the second dimension, \((z_1, z_2, 1, 0)\), are grouped into \((i_{1,l}, z_2, 1, 0) \cap D \).

In figure 8.1 five observations are depicted. The filled circles represent uncensored observations while the hollow ones represent singly and doubly censored observations. Feasible values of \((y_1, y_2)\) in the case of singly censored observations lie on the rays indicated by solid lines, while values corresponding to the doubly censored observation lie in the orthant indicated by the broken line. The box around the figure indicates the domain \( D \) which is partitioned by intervals of equal length along its two sides. The resulting grid is shown by light lines. The coarsening of the observations does not change the uncensored or doubly censored observations. However, the values of \((y_1, y_2)\) corresponding to the two singly censored observations are now assumed to lie in the shaded strips. While the rays
do not contain any uncensored observations, the strip corresponding to the observation censored in the second dimension now contains an uncensored observation. For the reduced data the self-consistency equa-

\[ \text{Figure 8.1.: Coarsening censored observations} \]

tions contain the term \( \Pr((Y_1, Y_2) \leq (u_1, u_2) \mid (Y_1, Y_2) \in \{(y_1, y_2) \mid y_1 > z_1, y_2 \in i_2, l(z_2)\}) \) for observations singly censored in the first dimension, where \( l(z_2) = \{l \in \{1, \ldots, k\} \mid z_2 \in i_2\} \). In general, there will be uncensored observations in the strips corresponding to the conditioning event. Thus, changes in the mass attributed to the uncensored observations will be reflected in the updating scheme for the singly censored observations.
One may hope that this recaptures the properties of self-consistent estimators in the discrete multivariate and the one-dimensional case, albeit at the cost of throwing away some data.

In fact, van der Laan (1996) showed that the self-consistent MLE based on the reduced data is uniformly consistent and asymptotically normal\(^5\). To achieve asymptotic efficiency of the reduced data MLE, he shows that the length of the coarsening intervals \(i\) in the two-dimensional case have to shrink to 0 at a rate slower than \(n^{-1/18}\) (van der Laan 1996: Theorem 5.1). This does not provide much guidance for sample sizes practically encountered. His simulations (1997) suggest that a small interval length of 0.02 for the square \([0, 1] \times [0, 1]\) and \(n = 200\) works well. Our limited experience indicates that in order to attain stable estimates of conditional expectations for the use in Buckley-James iterations it is expedient to use rather larger coarsening intervals.

The procedure is easily generalised to \(k\) dimensions. All observations with \(0 < \sum_{i=1}^{k} \delta_i < k\) are coarsened to a lattice in \(D = D_1 \times \ldots \times D_k\) induced by a partition of the \(D_i\) into intervals of equal length. This will ensure that the conditioning events in the self-consistency equations will have positive \(k\)-dimensional contents. The estimation procedure for the non-parametric self-consistent MLE of the reduced \(k\)-dimensional data can be summarised as follows:

---

\(^{5}\) In his simulations and the proofs van der Laan uses a slightly more complicated method of data reduction than the one proposed above. It involves a simultaneous coarsening of the censoring variables \(C\) in addition to the coarsening of the uncensored dimensions. If \(Y\) is independent of \(C\) this is no longer true for the coarsened data version, since \(\Pr(Y_1 \in i_{1,i}, \delta_i = 1) = \Pr(Y_1 \in i_{1,i}, C_1 \geq Y_1) = \int_{i_{1,i}} 1 - G_1(u) dF_1(u),\) where \(F_1\) and \(G_1\) are the (marginal) distributions of \(Y_1\) and \(C_1\), respectively. Thus the likelihood no longer factors into a term only containing \(F\) and another only depending on the censoring distribution \(G\). Van der Laan’s proposal retains the orthogonality between \(C\) and \(Y\) and thus allows asymptotic arguments based on a sequence of identical models. From a practical point of view and considering that the independence of the censoring scheme cannot be ascertained from the observations one may as well assume that the non-parametric likelihood in the coarsened model factors. One should then bear in mind that different models for the original and coarsened experiment are used, and that one changes models when changing the coarsening grid.
1. Choose a region $D = D_1 \times \ldots \times D_k$. I use $D_l = ] \min_i z_{li} - \sigma, \max_i z_{li} + \sigma [$. Note that the choice $\sigma = 0$ will exclude observations that are either right censored in this component at the maximum, or are uncensored in this component at the minimum of the observations.

2. Choose the number $q_l$ of intervals $i_l$ for each dimension $l$. Partition each side $D_l$ into $q_l$ intervals $i_{l,1}, \ldots, i_{l,q_l}$. I use left open and right closed intervals. Partition $D$ accordingly in $\prod_{l=1}^k q_l$ boxes $i_{1,m_1} \times \ldots \times i_{k,m_k}$.

3. Choose starting values. The NPMLE is discrete. It suffices to specify point masses for $\Pr(Y = y)$. We choose to put mass $1/n$ on all uncensored observations. The mass of $1/n$ of censored observations is equally spread over the strips implied by the censoring pattern of that observation. To all uncensored observations in the strip and to all intersections of the strip with other strips or with the boundary of $D$ the appropriate part of $1/n$ is added. This will produce a superset of the support points of the NPMLE. Pruitt (1993), Betensky and Finkelstein (1999), and Prentice (1999) discuss the exact determination of the support points of the NPMLE in the two-dimensional case, but the formulation does not easily generalise to higher dimensions.

4. Iterate the self-consistency equations: For each support point $y$ compute the new value $\Pr^{j+1}(Y = y)$ as the mean of the conditional probabilities given the observed information, $1/n \sum_i \Pr^j(Y = y | Z_i, \delta_i)$, where the probability of the conditioning event is the sum over the probabilities $\Pr^j(Y = y)$ lying in the strip determined by $(Z, \delta)_i$.

5. Stop the iteration using some convergence criterion. I use the maximum of $|\Pr^{j+1}(Y = y) - \Pr^j(Y = y)|$ over all support points as convergence criterion.

This EM algorithm generally converges very slowly. Especially the mass of points not in the support of the MLE, but given positive mass by our determination of starting values, decreases only slowly to 0. Prentice
(1999) and Betensky and Finkelstein (1999) proposed to use a direct constraint maximisation algorithm based on the likelihood function. But the approach will fail if the maximum of the likelihood is not unique. This happens if there are strips (or orthants) corresponding to censored observations that intersect $D$ without intersecting other strips or uncensored observations. Region of non-uniqueness can be ascertained in the two-dimensional case, though the procedure is quite tedious. Excluding these region from the maximisation problem would make direct maximisation algorithms very appealing. Unfortunately, we did not find a feasible formulation for the regions of non-uniqueness in the $k$-dimensional case. In contrast to the direct maximisation approaches the EM algorithm is not hampered by the possible non-uniqueness of the NPMLE. It simply does not change estimates in the regions of non-uniqueness. Since the estimator is to be used repeatedly based on changing data in the Buckley-James procedure, it seems appropriate to use the slow but reliable EM algorithm.

## 8.7. The Multivariate Buckley-James Estimator

With a NPMLE for the distribution of multivariate censored data at hand, an algorithm for the computation of multivariate regression estimators in the model (8.12) using the Buckley-James approach can be described as follows:

1. Compute starting values for $\beta$. I use the least-squares estimator treating all observations as uncensored.

2. For the $j + 1$-th iteration, compute the NPMLE of the residuals based on the data $(z_i - x_i \hat{\beta}^j, \delta_i)$.

3. Compute new values of the dependent variable as $Y^{*(j+1)} = \hat{y}(\hat{\beta}^j)$ according to (8.15). The conditional expectations of the censored residuals $e_i$ are evaluated as the weighted means of the residuals $e_k$. The estimates from step 2 are used as weights and the summation is over the regions determined by the censoring pattern.
4. Compute new regression coefficients $\hat{\beta}^{j+1}$ using a least squares regression of $Y^{*(j+1)}$ on $x$.

5. Go back to step 2 unless some convergence criterion is met. I use the maximum of $|\hat{\beta}_m^j - \hat{\beta}_m^{j+1}| / \max(|\hat{\beta}_m^j|, 1)$, where $m$ indexes the elements of $\beta$.

The distinctive feature of the estimator is the use of the joint distribution of the residuals to compute expected values in step 3. To illustrate the effectiveness of the computations, I generate $n = 300$ bivariate normal observations with $Y_1 \sim N(0, 1)$, $Y_2 \sim N(0, 1)$ and corr$(Y_1, Y_2) = 0.8$. These are censored in the second dimension only by $C_2 \sim N(2.4, 1)$. Figure 8.2 compares the estimated expected values of the censored observations (circles) based on the joint distribution (diamonds) with those based on the marginal distribution only (crosses). The estimates based on the joint distribution are clearly better in mimicking the underlying distribution than are the estimates based on the marginal distribution only. Figure 8.3 compares the two approaches in the case of independent components $Y_1 \sim N(0, 1)$, $Y_2 \sim N(0, 1)$, once again with $n = 300$ and $C_2 \sim N(2.4, 1)$. In this situation the estimates based on the joint distribution may be thought to fare less well. While the joint distribution cannot supply any additional information over the marginal distribution, the estimator based on the joint distribution looses information due to the coarsening. In this (and the previous) example I partitioned the first dimension into 10 intervals. It seems apparent from figure 8.3 that the estimates based on the joint distribution do not suffer strongly from the coarsening.

As an example for the effect of joint versus marginal estimation on the regression coefficients I use data from Wei, Lin and Weissfeld (1989: Table 1). The data give natural logarithm of the number of days, $z_{li}$, to virus positivity in the $l$-th serum sample of the $i$-th patient, $l = 1, 2, 3; i = 1, \ldots, 36$. There are thus three time dimensions. Patients were treated with ribavirin. There are three treatment groups: placebo, low dose, and high dose. This covariate information is coded in two dummy variables indicating low dose group and high dose group, respectively. There are six observations with missing values in one of the $z_{li}$. These were excluded from the analysis. Table 8.1 compares the estimated regression
coefficients from a model using a marginal Kaplan-Meier estimator with the proposed method using the joint distribution estimator. The latter was computed using a coarsening to five intervals of equal length in each of the three dimensions. The procedure converged after four Buckley-James iterations in each of which the computation of the NPMLE took four to five iterations. The resulting estimated coefficients are all slightly smaller than the coefficients from the marginal estimator.
Table 8.1: Dependent variable: natural logarithm of days to virus positivity

<table>
<thead>
<tr>
<th></th>
<th>marginal</th>
<th>joint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant 1</td>
<td>1.893</td>
<td>1.893</td>
</tr>
<tr>
<td>Constant 2</td>
<td>2.170</td>
<td>2.166</td>
</tr>
<tr>
<td>Constant 3</td>
<td>2.179</td>
<td>2.149</td>
</tr>
<tr>
<td>low dose 1</td>
<td>0.692</td>
<td>0.674</td>
</tr>
<tr>
<td>high dose 1</td>
<td>0.542</td>
<td>0.530</td>
</tr>
<tr>
<td>low dose 2</td>
<td>0.168</td>
<td>0.128</td>
</tr>
<tr>
<td>high dose 2</td>
<td>0.028</td>
<td>0.021</td>
</tr>
<tr>
<td>low dose 3</td>
<td>0.596</td>
<td>0.530</td>
</tr>
<tr>
<td>high dose 3</td>
<td>0.252</td>
<td>0.229</td>
</tr>
</tbody>
</table>

8.8. Discussion

The suggested multivariate Buckley-James estimator seems to be a feasible alternative to approaches based on the marginal distribution of the residuals. I have tried it with real and simulated datasets with up to 4000 observations and up to 10 dimensions. The most time consuming part of its computation is the estimation of the joint distribution of the residuals, which may often take 20 to 30 iterations. It would therefore be of interest to develop reliable direct maximisation procedures for the NPMLE.

An obvious obstacle to the use of the estimator is the lack of a variance estimator for regression coefficients. This is due to the fact that there is no variance expression for the NPMLE. Nevertheless, it might be possible to obtain variance estimators from a numerical approximation of the score function.
8.8. Discussion

Figure 8.3.: Conditional expectations: Correlation 0
8.9. Postscriptum


8.9.1. Multivariate Approaches

The monograph by Martinussen and Scheike (2006: Chap. 9) discusses recent approaches to the estimation in multivariate censored data. More

9

ON PROPORTIONALITY OF REGRESSION COEFFICIENTS IN MIS-SPECIFIED GENERAL LINEAR REGRESSION MODELS

9.1. Summary

It is shown that estimated regression coefficients in mis-specified general linear regression models (the conditional distribution of the variable of interest given the regressors is a function of a linear combination of the regressors) are approximately proportional to the “true” regression coefficients. The constant of proportionality is computed explicitly and a second order approximation is given. The connexion of this result with similar but stronger findings under the additional assumptions of a convex loss function and normally distributed regressors (Li and Duan 1989: 1009–1052) is clarified.

9.2. Introduction

Over recent years a large amount of empirical evidence—both obtained from real data sets and simulations—has been accumulated for a peculiar
stability property of regression models. Using different and possibly mis-specified models results in estimates of regression parameters that are roughly proportional. This is true also if certain kinds of incomplete data are ignored in the analysis. It has long been known that Probit and Logit models often give similar results (Chambers and Cox 1967). D’Agostino et al. (1990) show that a pooled logistic regression model and two variants of Cox’ proportional hazards model give nearly identical results for their data sets and offer some reason for the identity. Based on simulations a similar comparison is made by Ingram and Kleinman (1989). Doksum and Gasko (1990) discuss the general relationship between models in binary regression and survival analysis. Addison and Portugal (1987) compare several parametric regressions for a data set on the duration of unemployment and find proportional regression coefficients in all models. Bergström and Edin (1992) give similar results also, using data on unemployment duration. In addition they show that grouping of the dependent variable will attenuate the regression coefficients by a constant factor. Petersen (1991) and Petersen and Koput (1992) treat the effects of grouping in censored regression.

In contrast to the abundance and generality of the empirical findings, theoretical explanations are rare and cover only special cases. Attenuation effects were shown by Bretagnolle and Huber-Carol (1985) for Cox’ proportional hazards model if covariates are left out. Ruud (1986) reviews earlier results on proportionality of regression parameters in “limited dependent variable” models with normal regressors. Results for mis-specified correlated Logits were obtained by Neuhaus et al. (1992). Consistency results for linear location-scale models were given by Gould and Lawless (1988). In the case of least squares estimates in a grouped normal regression Stewart (1983) shows that proportionality is asymptotically exact for jointly normal regressors. Chung and Goldberger (1984) generalise his findings to arbitrary information loss under a linearity condition on the reverse regression. More recently, Goldenshluger and Polyak (1993) showed consistency of the least squares estimator in linear models posing very weak conditions on the error term (the models allow for auto-regression and other forms of dependence as well as nearly arbitrary distributions) when the regressors are i.i.d. with a symmetric joint distribution, with expectation 0 and finite fourth moment.
However, there are two general theoretical approaches towards an explanation that seem to cover most of the empirical findings. The first approach places restrictions on the joint distribution of the covariates. It originated from observations on the behaviour of ordinary least squares estimators in nonlinear situations when the regressors are normal (Brillinger 1983). These results were later generalised for models of binary outcomes and other “limited dependent variable” models (see Ruud 1986). Li and Duan (1989) offered the most general proof to date for the class of general linear regression models. These include models for binary outcomes, transformation models, censored regression and generalised linear models.

The second approach is based on a Taylor approximation argument that requires no conditions on the distribution of the regressors. Solomon (1984, 1986) was apparently the first to use it for the comparison of a proportional hazards model when an accelerated life model is true and for an accelerated life model when a proportional hazards model is true. Struthers and Kalbfleisch (1986) gave more general conditions for the behaviour of the partial likelihood estimator (similar to the ones given later by Hjort (1992)), but also used a Taylor approximation argument to get the proportionality result in the partial likelihood case. In a discussion of a paper by Cox and Reid (1987), Skinner (1987) hinted at the possibility of extending the argument of Solomon to the general linear regression case. Later he expanded on the interpretability of proportional regression coefficients (Skinner 1989). In that paper he used the same arguments as Li and Duan for the general linear regression case but did not give a proof via the alternative Taylor approximation.

In this paper I will make Skinner’s remarks explicit. The constant of proportionality will be computed and the second order approximation will be given. Then I try to elucidate the relation between Skinner’s and Li and Duan’s result.
9.3. Notation

The conditional distribution of an outcome variable $Y$ given the values of covariates $x$ describes a general linear regression model if the distribution of $Y$ depends on $x$ only through a linear combination of the $x$. The conditional density (or probability) of $Y$ may be written as

\[ f(y \mid \alpha_0 + x\alpha) \]  

(9.1)

where $f(\cdot \mid \cdot)$ belongs to some class $\mathcal{F}$ of regular conditional densities, $x$ is a $(1 \times p)$ row vector, $\alpha_0$ is a scalar, and $\alpha$ is the $(p \times 1)$ vector of regression coefficients. It will be assumed that the parameter $\alpha$ is identified at least up to a scalar multiple:

\[ f(y \mid \alpha_0 + x\alpha) = g(y \mid \alpha_0^* + x\alpha^*) \quad \forall x, y; f, g \in \mathcal{F} \]

\[ \implies \alpha^* = \gamma \alpha, \quad \gamma \in \mathbb{R} \setminus \{0\} \]  

(9.2)

The class of general regression models retains the interpretability of relative magnitudes of regression coefficients, a point stressed by Skinner (1989): Let $e_i$ be the $i$-th unit vector, $\delta$ a real scalar. Then

\[ f(y \mid (x + \delta e_i)\alpha) = f(y \mid x\alpha + \delta \alpha_i) \]

\[ = f\left(y \mid x\alpha + \delta \frac{\alpha_i}{\alpha_j}\right) \]

\[ = f\left(y \mid (x + \delta \frac{\alpha_i}{\alpha_j}e_j)\alpha\right) \]

so that $\alpha_i/\alpha_j, \alpha_j \neq 0$ is the amount of change in $x_j$ required to achieve an effect on the density equivalent to a unit change in $x_i$. By the identifiability assumption (9.2) these “equivalent effects” ratios (Skinner 1989) do not depend on the chosen representation, $f$.

It is sometimes useful to describe the model given by (9.1) in terms of random variables instead of densities. Let us write

\[ Y = T(h(x\alpha, \epsilon)), \quad \epsilon \sim G(.) \]  

(9.3)
where $\epsilon$ is a real valued random variable with distribution function $G(.)$ independent of $x$, $h(., .)$ is a strictly monotonic function in its second argument and $T(.)$ is a (not necessarily strictly) monotonic function. A transformation from (9.1) to the representation (9.3) may be accomplished by choosing $T(h(., .))$ to be the generalised inverse of the conditional distribution function in (9.1) and $G(.)$ to be the uniform distribution on $(0, 1)$. If (9.3) holds, (9.1) follows because the independence of $\epsilon$ and $x$ implies that the conditional distribution of $Y$ given $X = x$ only depends on $x \alpha$. Note that the representation (9.3) is not unique: The functions $T(.)$ and $h(., .)$ can always be merged into a single $h^*(., .)$ and $G(.)$ may be chosen as the uniform distribution by accommodating the function $h^*(., .)$ accordingly. However, (9.3) allows an interpretation in terms of a latent variable $Y^* = h(x\alpha, \epsilon)$ that is observed only after some form of transformation or information loss represented by the function $Y = T(Y^*)$. Most of the “limited dependent variable” models are motivated by such a construction. Moreover, if a continuous covariate $x_i$ with non-zero regression coefficient $\alpha_i$ is present, and $h(., .)$ is also strictly monotonic in its first argument, then a conceptually simple estimator consistent for $\alpha/\|\alpha\|$ is easily constructed: take the normalised value $\hat{\alpha}$ that maximises the rank correlation between $x\alpha$ and $Y$ (see Han (1987) for details). Obviously this works even if $T(.)$ and $h(., .)$ are unknown. Note that the existence of consistent estimators implies identifiability as in (9.2) for that class of models (see e.g. Deistler and Seifert 1978).

To complete the description of the stochastic setup let $(Y_i, X_i)_{i=1,..n}$ be a sequence of i.i.d. vectors where the conditional density of $Y$ given $X = x$ is given by (9.1), and the marginal distribution of $X$ is non-degenerate with $\mathbb{E}(X'X) = \Sigma > 0$ and—without loss of generality—$\mathbb{E}(X) = 0$. Sometimes further assumptions on the existence of moments, the differentiability of certain functions and the interchange of differentiation and integration will be needed. These will be clear from the context.

Only maximum likelihood type estimators will be considered explicitly. Denote by

$$
\sum_i \ell(y_i, x_i; \beta) = \sum_i \ln g(y_i \mid x_i\beta)
$$

(9.4)
the log-likelihood function under an assumed model $g(\cdot \mid \cdot)$ in the class of general regression models but not necessarily identical to the data generating model (9.1). The first order conditions for a maximum of the log-likelihood function are given by setting the pseudo score functions to 0:

$$\sum_i U(y_i, x_i; \hat{\beta}) = \sum_i \frac{\partial}{\partial \beta} \ell(y_i, x_i; \beta) \bigg|_{\beta = \hat{\beta}} = \sum_i x_i^T g^*(y_i, x_i \hat{\beta}) = 0 \quad (9.5)$$

where $g^*(\cdot, \cdot)$ denotes the derivative of $\ln g(\cdot \mid \cdot)$ with respect to its second argument. The formulation allows for nuisance parameters as long as the solution of (9.5) is unaffected by them. The sequence of estimators $\hat{\beta}_n, n \to \infty$ solving (9.5) will be Fisher consistent for the solution of

$$E \left( U(Y, X; \hat{\beta}) \right) = E \left( X' g^*(Y, X \hat{\beta}) \right)$$

$$= E_X \left( X' E_{Y|X} \left( g^*(Y, x \hat{\beta}) \mid X = x \right) \right) = 0 \quad (9.6)$$

provided such a solution exists. This generally implies strong consistency. Under some further smoothness assumptions on $\ell(\cdot)$, asymptotic normality of the estimators follows. See (Li and Duan 1989: 1029pp) and the literature on the behaviour of maximum likelihood type estimators in mis-specified models (see e.g. Huber 1967, White 1982, 1983, Gourieroux et al. 1984, Fahrmeir 1990). Hjort (1992) gave similar results for censored data using the counting process framework.

Since most of the arguments below use only the basic structure $E(X' g^*(\cdot, \cdot))$ or similar structures for the expected value of the log-likelihood $\ell(\cdot)$ under the true model, the results will also pertain to any estimators of analogous form. This includes e.g. Cox’ partial likelihood in the proportional hazards case, and any $M$-estimator derived from maximising a criterion function by setting its derivative with respect to $\beta$ to 0.
9.4. Two Special Cases: Linear Models and Least Squares under Information Loss

9.4.1. Linear Models

The “true” model is given by a location-scale model with a fixed density $f_0(.)$. That is

$$f(y \mid \alpha_0 + x\alpha) = \sigma^{-1}f_0\left(\frac{y - \alpha_0 - x\alpha}{\sigma}\right), \ Y = \alpha_0 + x\alpha + \sigma\epsilon, \ \epsilon \sim F_0$$

Estimation is carried out by assuming another location-scale model with density $g_0(.)$:

$$g(y \mid \beta_0 + x\beta) = \tau^{-1}g_0\left(\frac{y - \beta_0 - x\beta}{\tau}\right) \quad (9.7)$$

If maximum likelihood estimation based on the model (9.7) is used, the expected score function $U(.)$ for $\beta$ becomes

$$\mathbb{E}(U(Y, X; \beta)) = -\frac{1}{\tau}\mathbb{E}_X \left( X'\mathbb{E}_{Y \mid X} \left[ g_0^*\left(\frac{Y - \beta_0 - x\beta}{\tau}\right) \mid X = x \right] \right) = 0$$

Conditional on $X = x$

$$Y - \beta_0 - x\beta = x(\alpha - \beta) + (\alpha_0 - \beta_0) + \sigma\epsilon \quad (9.8)$$

Setting $\beta = \alpha$ makes the inner expectation independent of $X$ so that the expectation with respect to $X$ will be 0 (remember $\mathbb{E}(X') = 0$). Therefore $\beta = \alpha$ is a solution of the expected score function which implies consistency of the regression coefficients despite mis-specification of the disturbance distribution. This is the result of Gould and Lawless (1988): If a location-scale model is true, then using any other location-scale model for maximum likelihood estimation will lead to consistent estimators of the regression slopes, without any further restrictions on the covariates. Note that the nuisance parameters $\tau$ and $\alpha_0$ do not affect the solution of the expected score equation.
9.4.2. Least Squares under Transformations and Information Loss

Assume that a general regression model holds for a latent variable \( Y^* \) with fixed distribution of the disturbance \( \epsilon \). Let \( Y = T(Y^*) \) be observed, where \( T(\cdot) \) is some arbitrary function. Examples include binary regression models with \( T(z) = \mathbb{I}[z \geq 0] \) (where \( \mathbb{I}[] \) is the indicator function), grouped data situations, and transformation of the dependent variable as in Box-Cox transforms. The least squares normal equations with expectation

\[
\mathbb{E}(X'(Y - X\hat{\beta} - \hat{\beta}_0)) = 0
\]

are used as estimating equations. The solution for the slopes is

\[
\hat{\beta} = \Sigma^{-1}\mathbb{E}(X'Y)
\]

Now assume that the conditional expectation of \( X \) given \( Y^* \) is linear in \( Y^* \):

\[
\mathbb{E}(X \mid Y^* = y^*) = \delta(y^* - \mathbb{E}(Y^*))
\]

(9.9)

This places some restrictions on the distribution of the regressors \( X \). E.g., the condition is implied by joint normality of the regressors together with the regression structure given by (9.1).

Observe that \( \delta \) is the least squares projection of \( X \) on \( Y^* \),

\[
\delta = \mathbb{E}(X'Y^*)/\text{var}(Y^*)
\]

Then

\[
\mathbb{E}(X'Y) = \mathbb{E}_{Y^*} \left( Y\mathbb{E}_X (X \mid Y^*) \right)
\]

\[
= \mathbb{E}_{Y^*} \left( Y\delta(y^* - \mathbb{E}(Y^*)) \right) = \delta \text{cov}(Y, Y^*)
\]

By the preceding example, the regression slopes in the underlying model are given by \( \alpha = \Sigma^{-1}\mathbb{E}(X, Y^*) \). Therefore

\[
\beta = \Sigma^{-1}\mathbb{E}(X'Y) = \Sigma^{-1}\delta \text{cov}(Y, Y^*)
\]

\[
= \Sigma^{-1} \left( \mathbb{E}(X'Y^*)/\text{var}(Y^*) \right) \text{cov}(Y, Y^*)
\]
\[
\Sigma^{-1} (\Sigma \alpha / \text{var}(Y^*)) \text{cov}(Y, Y^*) = \alpha \left( \text{cov}(Y, Y^*) / \text{var}(Y^*) \right)
\]

(9.10)

The least squares estimator in the model under information loss is proportional to the underlying vector of regression slopes and the proportionality factor is given by the least squares regression slope of \( Y \) on \( Y^* \).

A special case of the result was given in Stewart (1983) for grouped normal regression. The present version is due to Chung and Goldberger (1984). They also point out that it is sufficient to treat the regression slopes of the least squares regression of \( Y^* \) on \( X \) as the “true” parameter \( \alpha \) without assuming an explicit model for the conditional distribution of \( Y^* \) given \( X \). Thus, assuming only the existence of moments, a regression result is compared with the one that would have been obtained if there had been no information loss. This operationalistic view may prove to be very helpful for judging empirical findings.

9.5. Skinner’s Approach

Returning to the general situation, assume that data are generated by a model of the form (9.1) for some \( f(., \), \( \alpha_0, \alpha \). Score functions from an assumed model \( g(., .) \) of the form (9.5) are used for estimation. These estimators will converge to the solution of the expected score equation (9.6). If a solution exists and is unique, Skinner (1987) proposed to use a Taylor approximation to \( \beta(\alpha) \) in terms of \( \alpha \) around \( \alpha = 0 \).

In fact, if \( \alpha = 0 \), setting \( \beta = 0 \) makes \( g^*(., .) \) independent of \( X \) so that the outer expectation in (9.6)—the covariance between \( X \) and \( \mathbb{E}_{Y|X}(g^*(Y, x\beta) \mid X = x) \)—vanishes. So a solution of (9.6) exists for \( \alpha = 0 \).

Taking derivatives with respect to \( \beta \) in the expected score equation (9.6) at \( \alpha = \beta = 0 \)—assuming the interchange of derivative and expectation is
justified—results in
\[ \frac{\partial}{\partial \beta} \mathbb{E}(X'g^*(Y, X\beta)) \bigg|_{\alpha=\beta=0} = \mathbb{E}(X'Xg^*_2(Y, 0)) \]

where \( g^*_2(y, \eta) \) is the derivative of \( g^*(y, \eta) \) with respect to \( \eta \). Since \( -g^*_2(y, \eta) \) is the “observed” information for \( \eta \) in the assumed model, it will be positive unless the model is unidentified. Therefore its expectation with respect to \( Y \) given \( X \) will also be positive so that the derivative reduces to
\[ \frac{\partial}{\partial \beta} \mathbb{E}(U(Y, X; \beta)) = \kappa \Sigma \] (9.11)

for some scalar \( \kappa < 0 \).

Taken together, the implicit function theorem may be invoked and the function \( \beta(\alpha) \) may be approximated by
\[ \beta(\alpha) \simeq \beta(0) + D\beta(0)\alpha = D\beta(0)\alpha \]

around \( \alpha = 0 \), where \( D \) is the differential operator. It remains to compute \( D\beta(0) \) explicitly. Set
\[ k(x\alpha, x\beta) := \mathbb{E}_{Y|X}(g^*(Y, x\beta) \mid X = x) \]

Then
\[
D\beta(0) = \frac{\partial \beta(\alpha)}{\partial \alpha} \bigg|_{\alpha=0} \\
= -\left[ \frac{\partial}{\partial \beta} \mathbb{E}_X(X'k(X\alpha, X\beta)) \bigg|_{\alpha=\beta=0} \right]^{-1} \times \\
\left[ \frac{\partial}{\partial \alpha} \mathbb{E}_X(X'k(X\alpha, X\beta)) \bigg|_{\alpha=\beta=0} \right] \\
= -\left[ k_2(0, 0)\Sigma \right]^{-1} \left[ k_1(0, 0)\Sigma \right] \\
= -\frac{k_1(0, 0)}{k_2(0, 0)} I = \gamma I \] (9.12)
where $k_1(\mu, \eta), k_2(\mu, \eta)$ are the partial derivatives of $k(\mu, \eta)$ with respect to $\mu$ resp. $\eta$ and interchanges of integral and derivative are once again assumed to be valid. Since the terms for the covariances of $X$ cancel, $D\beta(0)$ reduces to a scalar times the identity matrix. Thus the slopes of the regression under the assumed model are (to first order) proportional to the regression slopes under the data generating model.

It is instructive to compare this result to the example 9.4.2. When the linearity condition (9.9) does not hold, the conclusion of the example may not be applicable. Nevertheless, it is possible, with the help of the preceding formula to give an approximation to the least squares estimator under information loss. Specifically, let $T(z) = \mathbb{1}[z \geq 0]$. That is, least squares method is applied to binary data generated from $T(x\alpha + \alpha_0 + \varepsilon)$, $\varepsilon \sim F(.)$. Then:

$$\frac{\partial}{\partial \beta} \mathbb{E}_X(X'Y|X(T(x\alpha + \alpha_0 + \varepsilon) - x\beta - \beta_0)) = -\mathbb{E}_X(X'X) = -\Sigma$$

Moreover,

$$\frac{\partial}{\partial \alpha} \mathbb{E}_X(X'Y|X(T(x\alpha + \alpha_0 + \varepsilon) - x\beta - \beta_0)) \bigg|_{\alpha=0} = \frac{\partial}{\partial \alpha} \mathbb{E}_X(X'(1 - F(-X\alpha - \alpha_0))) \bigg|_{\alpha=0} = f(-\alpha_0)\Sigma \quad (9.13)$$

Therefore, the constant of proportionality in this case is given by

$$\gamma = f(-\alpha_0)$$

If $(X, \varepsilon)$ is jointly normal, this coincides with the solution (9.10). Further explicit expressions for the constant of proportionality in special cases are given in Solomon (1984) and Galler and Pötter (1992).

Higher orders of approximation may be obtained from the implicit function theorem together with a further application of the chain rule along the lines of (9.12). To second order, the Taylor approximation is given by

$$\beta(\alpha) \simeq D\beta(0)\alpha + \frac{1}{2}(\alpha'D_1^2\beta(0)\alpha)_{l=1...p}$$
with $D_1^2 \beta(0)$ the Hessian of $\beta_1(\alpha) \mid_{\alpha=0}$. Now

$$
D^2 \beta(0) = -\left[ \mathbb{E}(X'Xg_2^*(Y, 0)) \right]^{-1} \times \left( D_1^2 \mathbb{E}_X(X'k(X\alpha, X\beta)) + 2D\beta(0)D_{1,2}^2 \mathbb{E}_X(X'k(X\alpha, X\beta)) + (D\beta(0))^2 D_2^2 \mathbb{E}_X(X'k(X\alpha, X\beta)) \right)
$$

where $D_1^2$ is the second order differential operator with respect to $\alpha$, $D_2^2$ the one with respect to $\beta$ and $D_{1,2}^2$ the one with respect to $\alpha$ and $\beta$. At $\alpha = \beta = 0$ all terms except the first (inverse) term evaluate to $\mathbb{E}(X_i X_j X_l)_{i,j,l \in \{1,...,p\}}$ times a constant. Therefore, the second order term vanishes if all third order moments vanish. This happens e.g. if the joint distribution of $X$ is symmetric with respect to 0, and the third moments exist. Note that no assumption of absolute continuity of the distribution of the covariates needs to be imposed. Some explicit results on the second order terms were given by Solomon (1984) in the special case of the partial likelihood under accelerated failure models.

9.6. The Approach Based on Normal Regressors

The second general approach to explain the stability of regression ratios starts with placing restrictions on the distribution of regressors. Let the expected log-likelihood $\mathbb{E}(\ell(Y; \beta_0 + X\beta))$ be convex in $X\beta$. If $\mathbb{E}_X|X\alpha(X\beta \mid x\alpha)$ is linear in $x\alpha$ for all $\beta \in \mathbb{R}^p$, then the maximiser of the likelihood $\hat{\beta}$ converges to $\gamma \alpha$, where $\gamma$ is a scalar (Li and Duan 1989, Theorem 2.1). Since $\alpha$ is unknown in applications, the linearity condition will normally be required for all values of $\alpha$. In this case the proportionality of regression coefficients holds for all values of $\alpha$, not only approximately for small $\alpha$. But the restriction on the marginal distribution of the regressors is rather severe: It is e.g. implied by elliptical distributions, but discrete covariates or non-elliptical distributions are excluded.
The proof of Li and Duan’s result is very simple:

$$
\mathbb{E} (\ell(Y; b_0 + Xb)) \\
= \mathbb{E}_{X\alpha,\varepsilon} \left( \mathbb{E}_{X|X\alpha,\varepsilon} \left( \ell \left( T(h(X\alpha, \varepsilon)); b_0 + Xb \right) \mid X\alpha, \varepsilon \right) \right) \\
\leq \mathbb{E}_{X\alpha,\varepsilon} \left( \ell \left( T(h(X\alpha, \varepsilon)); b_0 + \mathbb{E}_{X|X\alpha,\varepsilon} \left( Xb \mid X\alpha, \varepsilon \right) \right) \right) \\
\quad \text{(by Jensen’s inequality)} \\
= \mathbb{E}_{X\alpha,\varepsilon} \left( \ell(T \left( h(X\alpha, \varepsilon) \right); b_0 + c + \gamma X\alpha) \right) \\
\quad \text{(by linearity)}
$$

Therefore the expected log-likelihood at some $b$ is always smaller than the expected log-likelihood at $\gamma \alpha$, and the result follows.

The convexity condition can be dispensed with at the price of strengthening the distributional assumptions. If $X$ is jointly normal, then the conclusion of the theorem holds without the convexity condition. Li and Duan (1989: Theorem 2.2) prove this by a direct appeal to properties of the multivariate normal distribution. However, it is possible to prove proportionality of regression results using Stein’s lemma, thus providing some insight into a possible connexion with the previous results. Recall Stein’s lemma: if $X$ is multivariate normal with covariance $I$ (for simplicity) and $g(.)$ a real, differentiable function with $\mathbb{E}(|Dg(X)|) < \infty$, then $\mathbb{E}(X'g(X)) = \mathbb{E}(Dg(X))$ (see e.g. Ibragimov and Has’minskii 1981: 25).

Now suppose the regressors are multivariate normal. Then

$$
0 = \mathbb{E}(U(X\beta, Y)) = \mathbb{E}_{X} (X'k(X\alpha, X\beta)) \\
= (\alpha, \beta) \mathbb{E}_{X} \left( \frac{k_{1}(X\alpha, X\beta)}{k_{2}(X\alpha, X\beta)} \right)
$$

by Stein’s lemma. Therefore,

$$
\beta = - \frac{\mathbb{E}_{X}(k_{1}(X\alpha, X\beta))}{\mathbb{E}_{X}(k_{2}(X\alpha, X\beta))} \alpha \quad (9.14)
$$

This proves proportionality without using the convexity of the log-likelihood, but requires normality. Note that the constant of proportionality agrees with the previous result (9.12) for $\alpha = \beta = 0$, so that a Taylor argument would lead to the result of Skinner, though using
much stronger assumptions. Note also that the proof starts with the expected score function, not with the expected log-likelihood. These similarities together with the second order result (9.13) prompt for generalisations either via generalisations of Stein’s lemma or via Skinner’s approach. But trying higher orders of the approximation seems to be less promising: although the accumulating moment conditions may lead to approximations to the normal distribution, it would require increasingly stronger differentiability assumptions for the score function, thus excluding many interesting estimators. Moreover, such assumptions are absent in the approach based on normal regressors.

Another strategy might be to apply generalisations of Stein’s lemma to non-normal distributions, perhaps on the lines of Cacoullos and Papathanasiou (1992). They give identities similar to Stein’s that hold for a larger class of densities, so that a standard integration by parts argument works. Unfortunately, they also show that the class of densities where this works also has to fulfil a moment condition: The conditional expectation of $X_i$ given all other variables is linear in these variables. This is rather close to the conditions imposed by Li and Duan.

9.7. Conclusions

Under a wide variety of circumstances regression models seem to possess a remarkable property of stability: the relative magnitude of the regression coefficients are stable across different models used for estimation and under many forms of information loss. Despite of the many empirical findings a solid theoretical understanding of this property is still missing. I have concentrated here on one specific approach towards an explanation: the use of Taylor approximations via expected estimating equations. This approach has been implicit in the work of Solomon (1984), Skinner (1987) and others. I developed a more explicit version, computed the constant of proportionality and gave a second order approximation. This, and a reanalysis of the other general approach via normality assumptions showed some similarities between them. But still neither could be reduced to the other. Moreover, as the examples in section
9.4 demonstrate, neither approach is able to incorporate sufficiently these special results: in the first example the constant of proportionality equals one throughout the parameter space without any conditions on the regressors, in the second linearity of the inverse regression is needed. This is implied by the assumption of normal errors together with the linear form of the models, but not vice versa. Moreover, it was possible to compute the constant of proportionality having a nice interpretation, while the computation from the approximation formula is generally awkward.

While it seems relatively straightforward to extend both approaches to include stochastic forms of data loss (censoring, selection etc.) and to give results for multivariate response models, general results on the set of models and situations in which a given model leads to consistent estimators (up to scale) seem not to be easily derived. The work of Lazrieva and Toronjadze (1991) gives some answers in this direction, though they deal with the partial likelihood scheme which lends itself easily to such questions.

9.8. Postscriptum

While there has been no progress in developing a general theory of misspecified regression models and in particular of misspecified regression incorporating incomplete data, there is a wealth of additional results concerning particular models.

Hattori (2006) investigates an additive hazards model and shows that tests of zero effects of covariates are consistent despite misspecification of the covariate model. Angrist et al. (2006) consider quantile regression estimators and use a representation in terms of a weighted mean squared error loss function to derive an omitted variable bias formula. Müller et al. (2008) investigate the effects of estimators of semi-Markov models when transition distributions are misspecified. Müller (2007) considers misspecified nonlinear regression models. O’Brien et al. (2006) consider the effect of misspecification on the power of tests of association of covariates with a dependent variable in the class of generalised linear


A more general approach has been advocated by Royall and Tsou (2003) who construct robust adjusted likelihood ratios in general parametric models (see also Kateri and Balakrishnan (2008) for an application to contingency tables). Patilea (2001) investigates the connection between convexly parametrised models and the behaviour of maximum likelihood estimators when the model is misspecified. Classes of convexly parametrised models are also used by van der Laan and Robins (2003) in their discussion of the double robustness of certain estimators in incomplete data models. Finally, Brown et al. (2006) survey recent results and generalisations of Stein's Lemma.
A

AN IMPLEMENTATION OF A CLASS OF DISTRIBUTION FUNCTION
ESTIMATORS

This section describes an implementation of a basic building block in
many incomplete data problems, namely the construction of estimators
of a cumulative distribution function from right censored data. The
building block has been used for the multivariate Buckley-James esti-
mator suggested in Chapter 8 and is closely related to the estimation
method used in Chapter 6. I will start by describing the general setup of
the software and will then discuss the particular algorithms used in a
variety of special cases.

The procedure is implemented in TDA (available from http://www.
stat.rub.de/tda.html). A command gdf is provided that can be
used to calculate marginal and joint distribution and survivor functions
based on possibly incomplete (censored) data. The syntax of the com-
mand is shown in Box 1. Most parameters are optional. Required is the
name of a variable (specified with the y1 parameter) and the name of an
output file to be given on the right-hand side.

Input data must be defined for $i = 1, \ldots, n$ units. Each unit can contribute
observations for (a subset of) $m$ dimensions. If available, $y_{ij}$ is the
observation for unit $i$ in dimension $j$. In general, each observation
A. An Implementation of a Class of Distribution Function Estimators

consists of two parts

\[(y_{ij}, \delta_{ij})\]

where \(y_{ij}\) is the observed value, and \(\delta_{ij}\) indicates whether the observation is uncensored \((\delta_{ij} = 1)\), or is right censored \((\delta_{ij} = 0)\). The corresponding data matrix variables can be specified with the \(y1\) and \(cen\) parameters, respectively. The following combinations are possible.

1. Only observed values are specified with the \(y1\) parameter. Then all observations are assumed to be exact with corresponding values.

2. Observed values are specified with the \(y1\) parameter and a censoring indicator is specified with the \(cen\) parameter. An observation is then interpreted as exact at \(y_{ij}\) if \(\delta_{ij} = 1\) and is interpreted as right censored if \(\delta_{ij} = 0\).

By default, the command assumes a single dimension \((m = 1)\) and \(n = NOC\) units, where \(NOC\) is the number of cases in the current data matrix. The \(grp\) parameter can be used to specify multivariate data. The syntax is

\[grp = ID, L1,\]

where \(ID\) and \(L1\) are names of data matrix variables. Each block of data matrix rows where \(ID\) has identical values is interpreted as data for one unit. Any values are possible and since the data matrix is always sorted with respect to \(ID\) and \(L1\), it is not required that blocks are contiguous. The \(L1\) variable must contain positive integers. The number of different integers found in this variable is interpreted as the number dimensions. Again, it is not required that these numbers are contiguous. Each data matrix row provides one observation for the unit given by the \(ID\) variable and dimension given by the corresponding \(L1\) variable.

To illustrate, consider the data in Box 2. In this example there are three units and six observations in all. The number of dimensions is \(m = 3\). Dimensions are mapped to the values of the \(L1\) variable in ascending order. Thus, dimension 1, 2, and 3 correspond, respectively, to \(L1 = 1\), \(L1 = 3\), and \(L1 = 4\). The first unit \((ID = 1)\) has observations for
Box 1: Syntax for gdf command.

```
gdf (opt=..., method, def. 1
1 = marginal calculation
2 = joint calculation, method 1
3 = joint calculation, method 2
prn=..., output option, def. 0
0 = distribution functions
1 = survivor functions
2 = expected values
yl=..., variable name for observed values
cen=..., variable name for censoring information
grp=ID,L1, specification of dimensions
sc=..., offset for domain (method 1 and 2), def. 0
n=..., number of boxes in grid (method 1), def. 100
mxit=..., maximal number of iterations (method 1), def. 20
tolf=..., tolerance for convergence (method 1), def. 0.001
d=..., delta specification (method 2), def. 0.1
fmt=..., print format for output file, def. 10.4
prot=..., protocol file with diagnostic information
) = fname;
```

Box 2: Example data to illustrate grp parameter.

```
<table>
<thead>
<tr>
<th>ID</th>
<th>L1</th>
<th>YL</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
```

all three dimensions. The observation is exact in the first and second dimension, and right censored in the third dimension. Unit 2 contributes an exact observation to the second dimension and a right censored
observation to the third dimension. The third unit (ID = 5) contributes a right censored observation to the first dimension.

A.1. Marginal Distribution Functions

If opt = 1 (default), the command calculates marginal distributions, separately for each dimension present in the input data. Assuming that there are \( n_j \) observations for the \( j \)th dimension, the data are:

\[(y_{ij}, \delta_{ij}), \ldots, (y_{njj}, \delta_{njj})\]

Depending on prn, these data are used to calculate a distribution function (prn=0), a survivor function (prn=1), or expected values (prn=2).

Marginal EDF: Exact Observations

If the data for one dimension contain only exact observations, the command calculates a standard empirical distribution function,

\[F_j(y_{ij}) = \sum_{i=1}^{n_j} \mathbb{I}(Y_j \leq y_{ij})\]

if prn = 0, or survivor function,

\[S_j(y_{ij}) = 1 - F_j(y_{ij})\]

if prn = 1. Here \( Y_j \) denotes the variable in the \( j \)th dimension and refers to the possible values \( y_{ij}, i = 1, \ldots, n_j \). \( \mathbb{I}() \) denotes the indicator function. In this case, if prn = 2, the expected values equal the observed values.

Box 3 provides an illustration. Input data are given by the variable YL. The command

\[\text{gdf (yl=YL) = df0;}\]
Box 3: Illustration of calculations with exact observations.

<table>
<thead>
<tr>
<th>YL</th>
<th>D</th>
<th>Y</th>
<th>F</th>
<th>df0 (prn = 0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-1.0000</td>
<td>0.1667</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.0000</td>
<td>0.3333</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>5.0000</td>
<td>0.5000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7.0000</td>
<td>0.8333</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>9.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>YL</th>
<th>D</th>
<th>Y</th>
<th>S</th>
<th>ds0 (prn = 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>-1.0000</td>
<td>0.8333</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1.0000</td>
<td>0.6667</td>
<td></td>
</tr>
<tr>
<td>-1</td>
<td>1</td>
<td>5.0000</td>
<td>0.5000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>7.0000</td>
<td>0.1667</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>9.0000</td>
<td>0.0000</td>
<td></td>
</tr>
</tbody>
</table>

creates the output file df0, the command
gdf (yl=YL,prn=1) = ds0;
creates the output file ds0.\(^1\) The first column in the output files shows the dimension, then follow the value of the variable (sorted in ascending order) and the corresponding value of the distribution or survivor function. If the input data refer to more than one dimension, the output file will contain the same information separately for each dimension.

Marginal EDF: Right Censored Observations

A second situation occurs if the input data, for one dimension, contain both, exact and right censored observations. The command then uses the standard Kaplan-Meier procedure to calculate a marginal distribution, or survivor, function. Considering the jth dimension, the observations for \(Y_j\) are sorted in ascending order. If there are exact and right censored observations for the same value of the variable, exact observations come first, followed by right censored observations. The highest value of \(Y_j\) is always treated as uncensored. In consequence, the expectation computed from the estimated distribution function will always be finite. Then, in the order from lowest to highest values, the mass of each right censored observation is distributed to the right (over all observations with larger values, see Efron (1988)).

\(^1\) The data file for this and the following examples is gdf1.dat. The command file is gdf1.cf. Both are contained in the TDA example archive.
A. An Implementation of a Class of Distribution Function Estimators

Box 4: Marginal distributions with right censored observations.

<table>
<thead>
<tr>
<th>YL</th>
<th>DELTA</th>
<th>D</th>
<th>Y</th>
<th>F</th>
<th>D</th>
<th>Y</th>
<th>S</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1.0000</td>
<td>0.1667</td>
<td>1</td>
<td>1.0000</td>
<td>0.8333</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>3.0000</td>
<td>0.3750</td>
<td>1</td>
<td>3.0000</td>
<td>0.6250</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>4.0000</td>
<td>0.6875</td>
<td>1</td>
<td>4.0000</td>
<td>0.3125</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5.0000</td>
<td>1.0000</td>
<td>1</td>
<td>5.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

An example is given in Box 4. Input data are given by the variables YL and DELTA. The command

\[
gdf (yl=YL, cen=DELTA) = df1;\]

computes the distribution function. Adding the parameter prn = 1 computes the corresponding survivor function.

A.1.1. Marginal EDF: Expected Values

If prn = 2, the \( gdf \) command calculates expected values based on the marginal distributions. If an observation is exact its expected value equals its observed value. If the observation is right censored, the command calculates

\[
y_{ij}^* = \mathbb{E}_{F_j}(Y_j \mid Y_j > y_{ij}) = \int_{y_{ij}}^{\infty} y \, dF_j / \int_{y_{ij}}^{\infty} dF_j
\]

where \( F_j \) is the Kaplan-Meier estimate of the marginal distribution function in the \( j \)th dimension.

To illustrate, consider the data in Box 5 (same as in Box 4). The command is now

\[
gdf (y1=YL, cen=DELTA, prn=2) = de1;\]
Box 5: Expected values based on marginal distribution.

<table>
<thead>
<tr>
<th>YL</th>
<th>DELTA</th>
<th>D</th>
<th>Y</th>
<th>CEN</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3.0000</td>
<td>1</td>
<td>3.0000</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>3.0000</td>
<td>0</td>
<td>4.5000</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>2.0000</td>
<td>0</td>
<td>4.1250</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.0000</td>
<td>1</td>
<td>1.0000</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>4.0000</td>
<td>1</td>
<td>4.0000</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>5.0000</td>
<td>1</td>
<td>5.0000</td>
</tr>
</tbody>
</table>

The resulting output file, also shown in Box 5, contains four columns. The first column refers to the current dimension, the second column contains the observed values, and the third column shows the censoring status of the observation. The last column contains the expected value. This will equal the observed value if the observation is uncensored, otherwise the expected value as calculated from the Kaplan-Meier survivor function. Note that with this option, the input data are not sorted. Note also that the largest observation is always assumed to be uncensored.

A.2. Joint Distributions

We now discuss the calculation of joint distribution functions. This is done separately for each marginal pattern which is found in the input data. The word "marginal pattern" refers to a combination of dimensions,

\[(d_1, \ldots, d_{m_k})\] where \(1 \leq d_1 < d_2 < \cdots < d_{m_k} \leq m\)

\(m\) being the maximal number of dimensions as given by the input data. For ease of notation, the following discussion refers to a full marginal pattern, i.e. that \(m_k = m\). Now, let \(n\) denote the number of units for the marginal pattern. We then have \(m\) observations for each unit, as follows:

\[(y_{il}, \delta_{il}), \ldots, (y_{im}, \delta_{im})\] \(i = 1, \ldots, n\)
Based on these data, the command calculates marginal distributions (if opt = 1) or joint distributions (if opt = 2 or opt = 3). Calculations depend on whether the data contain right censored observations. The contents of the resulting output file depends on the prn parameter. The command calculates a distribution function if prn = 0, a survivor function if prn = 1, or expected values if prn = 2.

### A.2.1. Joint Distribution: Exact Observations

If all observations are exact, the command calculates a standard $m$-dimensional distribution function, defined as

$$F(y_1, \ldots, y_m) = \sum_{i=1}^{n} \prod_{j=1}^{m} \mathbb{1}(Y_j \leq y_i)$$

or the corresponding survivor function

$$S(y_1, \ldots, y_m) = \sum_{i=1}^{n} \prod_{j=1}^{m} \mathbb{1}(Y_j > y_i)$$

The functions are calculated and tabulated in the output file, for all data points in the input data.

To illustrate, consider the data shown in Box 6. There are four units, all having observations for two dimensions. The joint distribution function, shown in the upper half of the right part of the box, was calculated with the command

```
gdf (grp=ID,L1,yl=YL,opt=2,prn=0) = df3;
```

The corresponding survivor function was calculated with the command

```
gdf (grp=ID,L1,yl=YL,opt=2,prn=1) = ds3;
```

In both cases, the first column in the output file refers to the current marginal pattern, followed by the dimensions (values of L1 variable) that define this pattern. The next $m$ columns contain the observations,
Box 6: Joint distribution with exact observations

<table>
<thead>
<tr>
<th>ID</th>
<th>L1</th>
<th>YL</th>
<th>MP</th>
<th>D1</th>
<th>D2</th>
<th>F</th>
<th>Y1</th>
<th>Y2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.2500</td>
<td>-1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>2.0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.5000</td>
<td>1.0000</td>
<td>2.0000</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>-1.0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.5000</td>
<td>2.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1.0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>0.7500</td>
<td>3.0000</td>
<td>1.5000</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>2.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>1.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

sorted in ascending or descending order. The final column contains the corresponding value of the distribution or survivor function.

If there are only exact observations, and *prn* = 2, the output file will simply contain the observations, a censoring indicator that always has value 1, and expected values that equal the observed values. In this case, the input data are not sorted.

A.2.2. Censored Observations, Method 1

We now consider a situation where the multivariate data contain right censored observations. Unfortunately, there is no simple generalization of the 1-dimensional Kaplan-Meier procedure. Discussion in the literature has proposed several different approaches. The gdf command offers two methods. The first one (selected with *opt* = 2) can be used with observations which might be censored simultaneously in several dimensions. The second method (selected with *opt* = 3) can only be used with observations which are censored at most in one dimension. This section describes the first method (*opt* = 2).

This method uses an iterative EM-like procedure that tries to find a self-consistent estimate of the distribution function. In explaining the
A. An Implementation of a Class of Distribution Function Estimators

procedure we refer, again, to a full marginal pattern. Data are given by

\[(y_i^{(1)}, \delta_i^{(1)}), \ldots, (y_i^{(m)}, \delta_i^{(m)}), \quad i = 1, \ldots, n\]

In a first step, we calculate a domain as an \(m\)-dimensional interval

\[D = D_1 \times \cdots \times D_m\]

where

\[D_j = \left[ \min_i y_{ij} - \sigma, \max_i y_{ij} + \sigma \right]

By default, the offset is \(\sigma = 0\). This implies that observations which have an exact component on a left side of the domain, or a right censored component on a right side of the domain, will not be used.\(^2\) In order to include all observations one can specify a positive offset with the \texttt{sc} parameter, see Box 1.

The iterative procedure is based on a partition of the domain into a grid of boxes. The \(j\)th dimension is partitioned into \(q_j\) intervals

\[i_j(k) = \left[ l_j(k), u_j(k) \right], \quad k = 1, \ldots, q_j\]

These intervals, and the corresponding boxes, are treated as open on the left side and closed on the right side. Since the number of boxes rapidly increases in higher dimensions, we require the user to specify a total number of boxes for the whole grid with the \texttt{n} parameter, default is \(n = 100\). The command then tries to find an integer \(q\) such that \(q^m \approx n\) and sets

\[q_1 = \ldots = q_m = q\]

The minimum is \(q = 1\), that is, the grid consists of only a single box. Let now \(B_j = \{1, \ldots, q_j\}\). Then, each

\[(k_1, \ldots, k_m) \in B_1 \times \cdots \times B_m\]

\(^2\) Corresponding values of the distribution function, or survivor function, will then be -1, and expected values will then equal the observed (possibly censored) values.
A.2. Joint Distributions

refers to one box in the grid, namely

\[ B(k_1, \ldots, k_m) = i_1(k_1) \times \cdots \times i_m(k_m) \]

We now define for each observation \((y_{ij}, \delta_{ij})\) a subset

\[ b_j(y_{ij}, \delta_{ij}) \subseteq B_j \]

containing pointers to those boxes (in the \(j\)th dimension) where the observation possibly has values. Explicitly, if the observation is exact, the definition is

\[ b_j(y_{ij}, \delta_{ij}) = \{ k \in B_j \mid y_{ij} \in i_j(k) \} \]

and if the observation is right censored, the definition is

\[ b_j(y_{ij}, \delta_{ij}) = \{ k \in B_j \mid \exists \delta > 0 : y_{ij} + \delta \in i_j(k) \} \]

Then, for each unit \(i\),

\[ \hat{b}_i = b_1(y_{i1}, \delta_{i1}) \times \cdots \times b_m(y_{im}, \delta_{im}) \]

provides pointers to those boxes in the domain where unit \(i\) has, possibly, an \(m\)-dimensional value. Of course, \(\hat{b}_i\) will be empty, if unit \(i\) has a component not covered by the domain.

Using these notations, Box 7 shows the iterative algorithm.\(^3\) The maximal number of iterations can be specified with the \texttt{mxit} parameter, default is 20. Convergence is assumed if

\[ \max_{k_1, \ldots, k_m} \{ |f(k_1, \ldots, k_m) - f'(k_1, \ldots, k_m)| \} \leq TOLF \]

where \(f'()\) refers to the density from the previous iteration. By default, \(TOLF = 0.001\); other values can be specified with the \texttt{tolf} parameter.

Information about the number of iterations and the final value of the convergence criterion is given in the standard output. In any case,

\(^3\) The notation ‘+’ means that the expression on the right-hand side is added to the expression on the left-hand side.
A. An Implementation of a Class of Distribution Function Estimators

Box 7: Iterative algorithm for joint distribution (method 1)

1. \( \forall (k_1, \ldots, k_m) \in B_1 \times \cdots \times B_m : f^*(k_1, \ldots, k_m) = 0 \)
2. \( \forall i \forall (k_1, \ldots, k_m) \in \bar{b}_i : f^*(k_1, \ldots, k_m) + = \frac{1}{n | \bar{b}_i |} \)
3. \( \forall (k_1, \ldots, k_m) \in B_1 \times \cdots \times B_m : f(k_1, \ldots, k_m) = 0 \)
4. \( \forall i \forall (k_1, \ldots, k_m) \in \bar{b}_i : \)
   \[ f(k_1, \ldots, k_m) + = \frac{1}{n} \sum_{(l_1, \ldots, l_m) \in \bar{b}_i} f^*(l_1, \ldots, l_m) \]
5. end if convergence has been achieved, or the maximal number of iterations has been reached.
6. \( \forall (k_1, \ldots, k_m) \in B_1 \times \cdots \times B_m : \)
   \[ f^*(k_1, \ldots, k_m) = f(k_1, \ldots, k_m) \]
7. continue with (3)

depending on prn, the gdf command finally calculates a distribution function, a survivor function, or expected values. In order to explain the calculation, let

\[ k_{ij} = \min \{ k_j \mid k_j \in b_j(\delta_{ij}) \} \]

meaning that \( k_{ij} \) refers to the box where, in dimension \( j \), observation \( y_{ij} \) begins. The distribution function is then calculated using the formula

\[ F(y_{i1}, \ldots, y_{im}) = \sum_{k_1 = 1, \ldots, k_{i1}} \cdots \sum_{k_m = 1, \ldots, k_{im}} f(k_1, \ldots, k_m) \]

Correspondingly, calculation of the survivor function uses the formula

\[ S(y_{i1}, \ldots, y_{im}) = \sum_{k_1 = k_{i1}, \ldots, q_1} \cdots \sum_{k_m = k_{im}, \ldots, q_m} f(k_1, \ldots, k_m) \]
Box 8: Example data set (Pruitt [1993])

<table>
<thead>
<tr>
<th>ID</th>
<th>L1</th>
<th>Y</th>
<th>CEN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>

If prn = 2, the command calculates expected values. To explain this option, let \((y_{i1}, \ldots, y_{im})\) be one of the \(m\)-dimensional observations. The expected value, in the \(j\)th dimension, will equal \(y_{ij}\) if this component is not censored. Otherwise, it is calculated by

\[
\frac{\sum_{(k_1, \ldots, k_m) \in b_i} z_j(k_1, \ldots, k_m) f(k_1, \ldots, k_m)}{\sum_{(k_1, \ldots, k_m) \in b_i} f(k_1, \ldots, k_m)}
\]

\(z_j(k_1, \ldots, k_m)\) is the \(j\)th component of the mean value of all exact observations falling in box \((k_1, \ldots, k_m)\) or, if the box does not contain exact observations, equals the mean of \(i_j(k_j)\).

Example 1. For a first illustration we use some example data from Pruitt [1993], shown in Box 8.⁴ In order to estimate a distribution function, we use the command

\[
gdf\left(\text{opt}=2, \text{prn}=0, \text{yl}=Y, \text{cen}={\text{CEN}}, \text{grp}={\text{ID}}, {\text{L1}}, \text{n}=64, \text{sc}=0.5\right) = \text{df5};
\]

⁴ The data file is gdf2.dat, contained in the TDA example archive.
A. An Implementation of a Class of Distribution Function Estimators

Box 9: Estimated distribution function

<table>
<thead>
<tr>
<th>MP</th>
<th>D1</th>
<th>D2</th>
<th>Y1</th>
<th>Y2</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.0000</td>
<td>6.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2.0000</td>
<td>4.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3.0000</td>
<td>5.0000</td>
<td>0.0001</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4.0000</td>
<td>3.0000</td>
<td>0.1250</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5.0000</td>
<td>2.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6.0000</td>
<td>7.0000</td>
<td>0.6290</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>7.0000</td>
<td>1.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>8.0000</td>
<td>8.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Box 10: Estimated expected values

<table>
<thead>
<tr>
<th>MP</th>
<th>D1</th>
<th>D2</th>
<th>Y1 (obs)</th>
<th>Y2 (obs)</th>
<th>CEN</th>
<th>Y1 (est)</th>
<th>Y2 (est)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1.0000</td>
<td>6.0000</td>
<td>0</td>
<td>1</td>
<td>5.6197</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2.0000</td>
<td>4.0000</td>
<td>0</td>
<td>1</td>
<td>5.5839</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3.0000</td>
<td>5.0000</td>
<td>0</td>
<td>0</td>
<td>6.6101</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4.0000</td>
<td>3.0000</td>
<td>1</td>
<td>1</td>
<td>4.0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>5.0000</td>
<td>2.0000</td>
<td>1</td>
<td>0</td>
<td>5.0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>6.0000</td>
<td>7.0000</td>
<td>1</td>
<td>1</td>
<td>6.0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>7.0000</td>
<td>1.0000</td>
<td>0</td>
<td>0</td>
<td>7.8492</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>8.0000</td>
<td>8.0000</td>
<td>1</td>
<td>1</td>
<td>8.0000</td>
</tr>
</tbody>
</table>

In this example, we have used a total number of 64 boxes and added a small offset to the domain in order to cover all observations. \(^5\)

Box 9 shows the resulting output file. Estimated expected values, calculated with the prn=2 option are shown in Box 10.

Example 2. For a second illustration, we create 100 data points

\[(y_{i1}, y_{i2}) \quad i = 1, \ldots, 100\]

where

\[y_{i1} = i, \quad y_{i2} = i + r_i\]

\(^5\) The command file is gdf2.cf.

366
and $r_i$ are random numbers which are equally distributed in $[-10, 10]$. We then have randomly censored 13 of these data points in the second dimension. The resulting data points are shown in Figure 1.

We then estimated expected values, first using a marginal Kaplan-Meier procedure. The resulting estimated values are indicated in Figure 1 by cross (x) symbols. We then used the iterative procedure described above to estimate expected values. The resulting estimated values are indicated in Figure 1 by ♦ symbols. They obviously provide somewhat better estimates.

---

Figure A.1.: Illustration of observed data points and estimated expected values. Estimation with marginal Kaplan-Meier procedure (x) and with joint estimation (method 1, n = 100 boxes).

---

\[\text{The command file for data generation and estimation is gdf3.cf in the TDA example archive.}\]
A. An Implementation of a Class of Distribution Function Estimators

Box 11: Algorithm for joint distribution (method 2)

1. Sort observations \((y_{i1}, \ldots, y_{im})\) in ascending order, first wrt first component, then wrt second component, and so on; in case of ties, exact observations precede censored observations.
2. For \(i = 1, \ldots, n\) : \(f(i) = 1/n\)
3. \(i = 1\)
4. If \((y_{i1}, \ldots, y_{im})\) is exact in all components, continue with step (8).
5. Let \((y_{i1}, \ldots, y_{im})\) be censored in dimension \(j_0\). Calculate an index set \(B(i, j_0)\) containing indices of all data points \((y_{k1}, \ldots, y_{km})\) for which:
   a) \(y_{kj} > y_{ij}\)
   b) \(\forall j \neq j_0 : y_{kj}\) is exact
   c) \(\forall j \neq j_0 : |y_{kj} - y_{ij}| \leq \Delta_{j_0}/2\)

6. \(\forall k \in B(i, j_0) : f(k) + = \frac{f(i)}{|B(i, j_0)|}\)
7. \(f(i) = 0\)
8. \(i += 1\)
9. If \(i \leq n\) continue with (4).

A.2.3. Censored Observations, Method 2

We now describe an alternative approach (selected with \(\text{opt} = 3\)) that can be used when the observations are censored in at most a single dimension. The basic idea is quite simple: we use a local Kaplan-Meier procedure based on all observations that are available in one of the censored dimensions.

Box 11 explains the algorithm. It is controlled by parameters \(\Delta_j\) that define
the size of the subsets of the domain used for the local Kaplan-Meier procedure. These parameters are calculated by using the parameter $d$ that can be specified by the user, see Box 1. Then

$$\Delta_j = d W_j$$

where $W_j$ denotes the width of the domain in dimension $j$. Default is $d = 0.1$.

The algorithm results in densities, $f(i)$, for all data points $i$ that do not contain censored observations. Depending on $\text{prn}$, they are finally used to calculate a distribution function

$$F(y_{i1}, \ldots, y_{im}) = \sum_{(y_{k1}, \ldots, y_{km}) \leq (y_{i1}, \ldots, y_{im})} f(k)$$

if $\text{prn} = 0$, or a survivor function

$$S(y_{i1}, \ldots, y_{im}) = \sum_{(y_{k1}, \ldots, y_{km}) \geq (y_{i1}, \ldots, y_{im})} f(k)$$

if $\text{prn} = 1$. These values are tabulated in the output file for all data points. The data points are not sorted.

If $\text{prn} = 2$, the command calculates expected values. For each data point $(y_{i1}, \ldots, y_{im})$, if $y_{ij}$ is exact, this will equal the corresponding expected component. If $y_{ij}$ is censored, the expected value is calculated by

$$\hat{y}_{ij} = \frac{\sum_{k \in B(i,j)} y_{kj} f(k)}{\sum_{k \in B(i,j)} f(k)}$$

For the definition of $B(i,j)$ see Box 11. If this index set is empty, $\hat{y}_{ij}$ will equal the observed value, $y_{ij}$.

Example 3. For an illustration, we use the 2-dimensional data from example 2. There are 100 data points, 13 are censored in the second dimension. Since the data points are censored in only a single dimension, we can use both methods, 1 and 2. Box 12 shows the censored data points
Box 12: Estimated expected values (method 1 and method 2)

<table>
<thead>
<tr>
<th>Y1 (obs)</th>
<th>Y2 (obs)</th>
<th>CEN</th>
<th>Y1 (est)</th>
<th>Y2 (est)</th>
<th>Y2 (est)</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.0000</td>
<td>27.3464</td>
<td>1 0</td>
<td>34.0000</td>
<td>37.6603</td>
<td>37.3177</td>
</tr>
<tr>
<td>36.0000</td>
<td>30.3996</td>
<td>1 0</td>
<td>36.0000</td>
<td>37.6603</td>
<td>38.0035</td>
</tr>
<tr>
<td>43.0000</td>
<td>15.3593</td>
<td>1 0</td>
<td>43.0000</td>
<td>46.2104</td>
<td>46.2348</td>
</tr>
<tr>
<td>45.0000</td>
<td>43.5366</td>
<td>1 0</td>
<td>45.0000</td>
<td>48.9264</td>
<td>48.4735</td>
</tr>
<tr>
<td>49.0000</td>
<td>50.6783</td>
<td>1 0</td>
<td>49.0000</td>
<td>54.3749</td>
<td>51.9167</td>
</tr>
<tr>
<td>68.0000</td>
<td>50.5112</td>
<td>1 0</td>
<td>68.0000</td>
<td>61.4348</td>
<td>71.7563</td>
</tr>
<tr>
<td>69.0000</td>
<td>47.4082</td>
<td>1 0</td>
<td>69.0000</td>
<td>61.4348</td>
<td>71.7563</td>
</tr>
<tr>
<td>73.0000</td>
<td>49.9009</td>
<td>1 0</td>
<td>73.0000</td>
<td>74.5214</td>
<td>73.0277</td>
</tr>
<tr>
<td>75.0000</td>
<td>65.2591</td>
<td>1 0</td>
<td>75.0000</td>
<td>74.5214</td>
<td>73.7913</td>
</tr>
<tr>
<td>79.0000</td>
<td>38.7495</td>
<td>1 0</td>
<td>79.0000</td>
<td>74.5214</td>
<td>79.1623</td>
</tr>
<tr>
<td>84.0000</td>
<td>53.5625</td>
<td>1 0</td>
<td>84.0000</td>
<td>82.3746</td>
<td>83.4225</td>
</tr>
<tr>
<td>85.0000</td>
<td>62.9593</td>
<td>1 0</td>
<td>85.0000</td>
<td>82.3746</td>
<td>85.3742</td>
</tr>
<tr>
<td>93.0000</td>
<td>33.8836</td>
<td>1 0</td>
<td>93.0000</td>
<td>99.9475</td>
<td>103.4777</td>
</tr>
</tbody>
</table>

and estimated expected values for their censored component.\(^7\) It is seen that, in this example, both methods give quite similar estimates.

Example 4. If \(d = 1\), the algorithm uses all observations (in the censored dimension) and becomes identical with a standard marginal Kaplan-Meier procedure. For an illustration see command file gdf5.cf in the TDA example archive.

---

\(^7\) The command file is gdf4.cf.


Bach, K. 1980: Actions are not events. Mind 89: 114–120.


Bibliography


Bibliography


Bibliography


Chen, Hua Yun, Roderick J. Little 2001: A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. Lifetime Data Analysis 7: 207–224.


Bibliography


Bibliography


Bibliography


Bibliography


Hacker, P.M.S. 1982: Events and objects in space and time. Mind 91: 1–19.
Bibliography


Heckman, James J. 1976: The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. Annals of Economic and Social Measurement 5: 475–492.


Bibliography


Heckman, James J., Jeffrey Smith (with the assistance of Nancy Clements) 1997: Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. Review of Economic Studies 64: 487–535.


Bibliography


Bibliography


Horowitz, Joel L., Charles F. Manski, Maria Ponomareva, Jörg Stoye 2003: Computation of bounds on population parameters when the data are incomplete. Reliable Computing 9: 419–440.


Huang, Jie, David Harrington 2004: Dimension reduction in the linear model for right-censored data: Predicting the change of HIV-I RNA levels using clinical and protease gene mutation data. Lifetime Data Analysis 10: 425–443.
Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Bibliography


Phillips, Carl V., Karen J. Goodman 2006: Causal criteria and counterfactuals; nothing more (or less) than scientific common sense. Emerging Themes in Epidemiology 3: 5.


Bibliography


Riphahn, Regina T., Oliver Serfling 2002: Item non-response on income and wealth questions. Forschungsinstitut zur Zukunft der Arbeit. Discussion Paper No. 573:


Bibliography


Schafer, Joseph L. 2003: Multiple imputation in multivariate problems when the imputation and analysis models differ. Statistica Neerlandica 57: 19–35.


Bibliography


Song, Xiao, Shuangge Ma, Jian Huang, Xiao-Hua Zhou 2007: A semiparametric approach for the nonparametric transformation survival model with multiple covariates. Biostatistics 8: 197–211.


Bibliography


Bibliography


Bibliography


Wang, Qihua 2008: Probability density estimation with data missing at random when covariables are present. Journal of Statistical Planning and Inference 138: 568–587.


Bibliography


Bibliography


Bibliography


