Probabilistische Selektionsmodelle*

Ulrich Pötter

Zusammenfassung: Fehlende oder unvollständige Angaben in Umfragedaten sind unvermeidlich. Befragte geben manchmal keine Auskunft oder verweigern das Interview. Der Anteil der vollständigen Antworten ist selten größer als 40%. Was weiß man über die anderen 60%? Kann man problemlos die Angaben der 40% so behandeln als hätte man Angaben von allen ausgewählten Befragten? Probabilistische Selektionsmodelle sind ein Hilfsmittel, über Zusammenhänge zwischen erhaltenen und nicht erhaltenen Angaben nachzudenken. Ihre Anwendungsmöglichkeiten beschränken sich nicht auf die Analyse fehlender Angaben in Umfragen. Sie werden auch für die Untersuchung gruppierter, zensierter oder sonstiger unvollständiger Daten sowie für die Modellierung von Prozessen mit Selbstselektionen und die Evaluation von Maßnahmen benutzt. Nach einem kurzen Überblick über verschiedene Anwendungen wird zunächst diskutiert, welche Annahmen Selektionsmodellen zugrunde liegen. Es zeigt sich, dass die Verwendung probabilistischer Modelle auf Voraussetzungen verweist, die über empirisch zugängliche Annahmen über soziale Sachverhalte weit hinausgehen. Der Anteil modellimmanenter Spekulation kann allerdings verringert werden, wenn auch partielle Angaben der Befragten mit einbezogen werden. Die Grundlagen von Selektionsmodellen werden daher nicht nur für fehlende, sondern auch für gruppierte und vergröberte Angaben formuliert. Abschließend werden verschiedene Möglichkeiten der Sensitivitätsanalyse für probabilistische Selektionsmodelle dargestellt und die Beschränkungen enger Formulierungen demonstriert. Fehlende und gruppierte Einkommensangaben im ALLBUS 1996 dienen zur Illustration.

I. Einleitung

Fehlende oder unvollständige Angaben in Umfragedaten sind unvermeidlich. Schließlich steht es jedem Befragten frei, die Zumutung eines Interviews zurückzuweisen oder auf einzelne Fragen nicht zu antworten. Zudem mag ein Befragter nicht willens oder in der Lage sein, Auskünfte in der gewünschten Präzision zu geben. In Umfragen ist der Anteil vollständiger Angaben zu einer bestimmten Frage selten höher als 40%. Dennoch sollen Umfrageergebnisse Aufschluss über soziale Sachverhalte geben. Wenn dies nachvollziehbar gelingen soll, muss ein Zusammenhang zwischen den erhaltenen Angaben und den potentiellen Angaben aller ausgewählten Befragten hergestellt werden. Denn allein aufgrund der vollständigen Angaben soll es möglich sein, Aussagen

*Ich danke Andreas Behr, Eva Berns, Andreas Diekmann, Götz Rohwer und einem anonymen Gutachter für kritische Kommentare und hilfreiche Diskussionen.

über statistische Sachverhalte zu treffen, die sich auf alle Befragten beziehen. Wenn aber über 60% der Befragten nichts oder wenig bekannt ist, dann sind zur Rechtfertigung solcher Aussagen offenbar starke Annahmen erforderlich. Probabilistische Selektionsmodelle sind ein Hilfsmittel, solche Annahmen explizit zu formulieren und über ihre Konsequenzen nachzudenken.

Das Grundschema probabilistischer Selektionsmodelle lässt sich leicht angeben: Man möchte Aussagen über Aspekte der Verteilung einer statistischen Variablen Y machen. Nur wurden die Werte dieser Variablen nicht vollständig beobachtet. Dagegen ist die Verteilung einer Variablen Y^* bekannt, von der angenommen wird, sie stände in einer Beziehung zu der interessierenden Variablen Y. Zum Beispiel mag ein Befragter auf die Frage nach dem Einkommen mit einer genauen Angabe, mit einer ungefähren Angabe etwa in Form eines Intervalls, oder gar nicht antworten. Der Wert der Variablen Y^* ist dann entweder der Wert von Y, oder ein Intervall, in welches Y fällt, oder das Intervall $(0, \infty)$, das „keine Angabe" repräsentiert. Wäre bekannt, unter welchen Umständen jemand bei gegebenem Wert von Y eine mehr oder weniger genaue Auskunft Y^* gibt, dann könnte aus der Verteilung von Y^* und den Umständen auf Aspekte der Verteilung von Y geschlossen werden. Wird für die Antwortmöglichkeiten Y^* bei gegebenem Y eine bedingte Wahrscheinlichkeit angegeben, dann soll dieses probabilistische Modell im Folgenden ein allgemeines *Selektionsmodell* für (Y, Y^*) heißen.

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 nonresponse 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.

Dagegen sind sowohl in der mathematischen Statistik wie auch in der Biometrie und Ökonometrie seit etwa 30 Jahren eine Fülle von probabilistischen Selektionsmodellen entwickelt worden. Einerseits ist geklärt worden, unter welchen Modellvorstellungen relativ einfache Verfahren des Umgangs mit fehlenden Daten gerechtfertigt werden können. In solchen Modellen können die Einzelheiten des Zustandekommens unvollständiger Daten weitgehend ignoriert werden. Einen guten Überblick darüber geben die Bücher von Schafer (1997) und Little und Rubin (2002). Andererseits sind in der ökonometrischen Tradition hauptsächlich nicht ignorierbare Selektionsmodelle und entsprechende Schätzverfahren vorgeschlagen worden. Insbesondere die frühen Arbeiten von Heckman (1976, 1979) haben einen kaum zu überschätzenden Einfluss auf viele Bereiche der Sozialwissenschaften ausgeübt. Neuere Überblicke geben Nicoletti (2002) und Vella (1998). Obwohl die Abhängigkeit dieser Modelle von einer unübersichtlichen Mischung von Annahmen über Verteilungen, funktionale Formen von Regressionen, latente Variablen und Ausschlussrestriktionen früh kritisiert wurde (z.B. in Wainer 1986, 1989), haben sie sich in einigen Bereichen der Sozialwissenschaften als dominante analytische Methode durchgesetzt.

Dieser Artikel gibt einen Überblick über neuere Entwicklungen in der statistischen Literatur. Insbesondere werden die Form von Annahmen, die Einbeziehung unvollständiger Angaben und die Verwendung von Sensitivitätsanalysen diskutiert. Im nächsten Abschnitt wird zunächst ein kurzer Überblick über häufig verwandte probabilistische Selektionsmodelle und ihre Anwendungen in den Sozialwissenschaften gegeben. Anschließend wird am Beispiel der Einkommensangaben im ALLBUS 1996 die Form der Annahmen diskutiert, die in probabilistische Modelle einfließen. Die Annahmen verweisen nicht nur auf potentiell empirisch zugängliche soziale Sachverhalte, sondern enthalten immer auch modellimmanente Anteile, die nicht reifiziert werden sollten. Das mag die Zurückhaltung gegenüber solchen Modellen in der Stichprobentheorie rechtfertigen. Allerdings kann der Anteil modellimmanenter Spekulation verringert werden, wenn auch gruppierte, zensierte und andere partielle Angaben in die Analyse einbezogen werden. Abschnitt V. beschreibt die Vorgehensweise für den Fall ignorierbarer Selektionsmodelle. Im Abschnitt VI. wird die Ignorierbarkeit von Selektionsmodellen auch für gruppierte und vergröberte Angaben in einem wahrscheinlichkeitstheoretischen Rahmen definiert. Anschließend werden zwei Schätzverfahren vorgestellt, die die Einbeziehung von Kovariablen in Selektionsmodelle erlauben. Einige nicht ignorierbare Modelle werden im folgenden Abschnitt kurz vorgestellt. Abschließend werden Techniken der Sensitivitätsanalyse dargestellt. Sie erlauben eine Abschätzung der Abhängigkeit von Schlussfolgerungen von einigen zentralen modellimmanenten Annahmen und sind daher ein wesentliches Hilfsmittel für die Beurteilung von Selektionsmodellen.

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 Abschätzungen, 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.

II. Selektionsmodelle in den Sozialwissenschaften

Umfragedaten bilden eine wesentliche empirische Grundlage aller Sozialwissenschaften. Aber schon das Verfahren, mit dem Befragte ausgewählt werden, ist häufig mit dem Hinweis hinterfragt worden, die Auswahl sei selektiv gewesen. Ein oft und gern zitiertes Beispiel ist der spektakuläre Misserfolg der Wahlvorhersage der Zeitschrift Literary Digest für die Präsidentenwahl 1936 in den USA. Das Literary Digest hatte 60% der Stimmen für den Republikaner Landon vorhergesagt, aber Roosevelt gewann die Wahl mit 62%. Das Literary Digest hatte eine Stichprobe von Telefon- und Autobesitzern befragt. Eine klassische Erklärung des Fehlschlags besagt, Telefon- bzw. Autobesitz sei damals ein Anzeichen von Reichtum gewesen und reichere Personen hätten eher republikanisch gestimmt. Das Selektionsargument bezieht sich auf den gewählten Rahmen der Stichprobe, die Basis für die Auswahl von Befragten. Aber eine einfache Überlegung zeigt, dass diese Selektion nicht allein für den Misserfolg verantwortlich sein kann. 1936 besaßen ca. 40% der Haushalte ein Telefon. Hätten die Telefon- und Autobesitzer in der Tat zu 60% für Landon gestimmt, dann hätten von allen Haushalten, die weder Auto noch Telefon besaßen, über 75% für Roosevelt stimmen müssen. Betrachtet man die abgegebenen Wählerstimmen, so hätte der entsprechende Anteil sogar größer als 90% sein müssen (Bryson 1976). Das Verhältnis der Odds für Roosevelt in den beiden Gruppen der Telefonbesitzer und derjenigen ohne Telefon müsste also mehr als 1:20 betragen. Das Literary Digest hatte 10 Millionen Fragebogen verschickt, aber nur 2.3 Millionen zurückerhalten. Es liegt nahe zu vermuten, dass Landon-Anhänger eher als Roosevelt-Anhänger auf die Umfrage geantwortet haben. Wird angenommen, die 10 Millionen Befragten hätten tatsächlich zu 62% für Roosevelt gestimmt, muss das Verhältnis der Odds für Roosevelt in den beiden Gruppen der Antwortenden und der nicht Antwortenden nur 1:3 betragen, um zu der Diskrepanz zwischen Vorhersage und Wahlergebnis zu führen. Ein Verhältnis der Odds von 1:3 ist eine weit realistischere Größenordnung als die 1:20, die für die These einer Selektion zwischen Telefonbesitzern und Nichtbesitzern angenommen werden müsste. Daten aus Nachbefragungen haben dann auch die Bedeutung der Unterschiede im Antwortwortverhalten der Landon- bzw. Rooseveltanhänger bestätigt (Squire 1988; Cahalan 1989). Squire (1988: 132) schließt:

> „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.“

Brysons Überlegungen über den Misserfolg der Wahlvorhersage des Literary Digest verweisen zwar auf die möglicherweise gravierenden Folgen unvollständiger Angaben in Umfragen, benutzen aber keine probabilistischen Modelle etwa über das Antwortverhalten von Landon- bzw. Rooseveltanhängern. Seit der Mitte der 70er Jahre sind Verfahren entwickelt worden, die auf der Basis probabilistischer

Modelle für Teilnahme- und Antwortentscheidungen der Befragten versuchen, die Folgen unvollständiger Angaben abzuschätzen. Eine Variante, die auf Arbeiten von Heckman (1976, 1979, 1990) zurückgeht, ist von Engelhardt (1999) vorgestellt worden. Sie untersucht die Einkommensangaben in der Berliner Altersstudie (BASE), einer nach Alter und Geschlecht geschichteten Zufallsstichprobe von Berlinern und Berlinerinnen über 70 Jahren, die auf der Basis des Einwohnermelderegisters gezogen wurde (Mayer und Baltes 1996). An der Erstbefragung haben 928 Personen teilgenommen, das entspricht ca. 49% der Ausgangsstichprobe. Engelhardt verwendet neben dem Einkommen die Variablen Geschlecht, Alter, Familienstand, Schul- und Berufsausbildung, Wohnform, Interviewform sowie einen Demenzindikator. Vollständige Angaben zu diesen Variablen liegen für 842 Personen vor, zusätzliche Angaben zum Einkommen nur für 716 oder 77% der Personen, die an der Befragung teilgenommen haben. Engelhardt (1999: 716f) unterstellt eine Wahrscheinlichkeit für die 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,

> „die Heckman-Korrektur [bietet] aber auch Möglichkeiten, die in der explorativen Analyse liegen. Wenn—wie im Beispiel—die selektionskorrigierte Regressionsanalyse zu demselben Resultat kommt wie die unkorrigierte Schätzung, erhöht dies das Vertrauen in die OLS-Regression“ (1999: 721).

Sie betont aber die Abhängigkeit der Ergebnisse von den unterstellten Annahmen (1999: 713f) und zeigt, dass zumindest diejenigen Annahmen, die sich überprüfen lassen, wohl nicht gelten (1999: 719). Um tatsächlich ein erhöhtes „Vertrauen in die unkorrigierte Schätzung“ zu haben, müsste an Stelle eines einzigen alternativen Selektionsmodell, das zudem auf zweifelhaften Annahmen beruht, mehrere Selektionsmodelle verglichen werden. Heckmans Modell bietet aber keinen systematischen Ansatzpunkt, Modellannahmen zu variieren bzw. die Auswirkungen verletzter Annahmen quantitativ abzuschätzen. Die Möglichkeit, mit Hilfe von probabilistischen Selektionsmodellen über die Folgen unvollständiger Angaben nachzudenken, kann im Rahmen dieses Modells nur begrenzt genutzt werden.

In anderen Bereichen, etwa der historischen Demographie, werden dagegen manchmal probabilistische Selektionsmodelle eingesetzt, um über die Aussagekraft von Angaben zu spekulieren. So existieren oft nur unvollständige Angaben über die Lebensdauer von Menschen. Z.B. gibt es zu Geburts- bzw. Taufangaben in Kirchenregistern in vielen Fällen keine Angaben über das Todesdatum. Allerdings gibt es manchmal weitere Ereignisse wie Heiraten oder Kindergeburten, die aufgezeichnet 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 $(T_{\max}, \infty)$, wobei $T_{\max}$ der Zeitpunkt des letzten Ereignisses ist, das registriert wurde. Wenn angenommen wird, diese Angaben sagten nichts anderes als das jemand älter als $T_{\max}$ 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 $T_{\max}$ 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:

> „Die Regressionsschätzung basiert allerdings nur auf der Stichprobe der 99 erfolgreichen Auktionen, da nur für diese ein Verkaufspreis vorliegt. Nun könnte es sich hierbei um ein selektives Sample handeln.... Die Zwei-Stufen-Schätzmethode von Heckman ist eine Alternative, um einen eventuellen Stichprobenauswahlfehler zu kontrollieren“ (2002: 687).

Ein „Stichprobenauswahlfehler“ könnte aber nur vorliegen, wenn auch den Auktionen ohne Gebote ein „Preis“ zukäme. Diekmann und Wyder unterstellen wohl eine Größe, die mit „Zahlungsbereitschaft“ umschrieben werden kann. Ein Gebot wird abgegeben, falls die Zahlungsbereitschaft eines potentiellen Auktionsteilnehmers größer als das Mindestgebot der Auktion ist. Also fehlen Angaben über die Zahlungsbereitschaft, wenn der Startpreis der Auktion höher als diese Zahlungsbereitschaft ist. Daher könnte das Problem ähnlich wie fehlende Angaben in Umfragen oder Registern behandelt werden. Zwar bleibt unklar, was „Zahlungsbereitschaft“ unabhängig von einem konkreten Gebot in einer gegebenen Auktion bedeuten könnte, sogar, welcher Gruppe von Personen diese Größe zugeschrieben werden soll. Aber selbst wenn dem Konzept eine gewisse Plausibilität zugestanden wird, dann greift der Versuch, ein einfaches Selektionsmodell für fehlende Angaben zu verwenden, zu kurz. Denn zum einen ist bei Auktionen ohne Gebote die Zahlungsbereitschaft kleiner als der Startpreis, so dass die Zahlungsbereitschaft nicht vollständig unbekannt ist. Zum anderen ist der erzielte

Preis auch nicht gleich der Zahlungsbereitschaft des Höchstbietenden, sondern (bei mehr als einem Gebot) gleich der Zahlungsbereitschaft des Bieters mit dem zweit höchsten Gebot.[1] Ein Selektionsmodell für „Zahlungsbereitschaft“ müsste beide Aspekte, den der zusätzlichen Information aus den Mindestgeboten und den der erzielten Preise als zweit höchste Zahlungsbereitschaft berücksichtigen. Die Konzentration auf eine Schätzmethode behindert aber oft die Formulierung probabilistischer Modelle für Konzepte, die sich auf fehlende, abgeschnittene oder zensierte und nach Größe selektierte Beobachtungen stützen.

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

[1] Genauer: Bietet Person A zunächst maximal 20 Euro und später B 10 Euro, so beträgt der Auktionspreis 10 Euro. Der Auktionspreis gibt also die Zahlungsbereitschaft von B wieder. Bietet andererseits zunächst B 10 Euro, A später 20 Euro, so ist der Auktionspreis 10 Euro plus Mindesterhöhung der Auktion.

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 coneptually ..., 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).

Rubins Punkt (b) verweist zurück auf den Status unvollständiger Daten in den beiden Fällen. Verweigert jemand die Auskunft im Interview, so liegt das in seinem oder ihrem Ermessen, hängt aber nicht von den Interessen und Vorstellungen des Forschers ab. Dagegen sind Daten im kontrafaktischen Modell der Evaluation ebenso wie Überlegungen zur „Zahlungsbereitschaft" nur unvollständig aufgrund der Modellvorstellungen des Forschers. Erzielte Preise in Auktionen sind ebenso wie Ergebnisse von Arbeitsmarktmaßnahmen zumindest prinzipiell beobachtbar. „Zahlungsbereitschaft" im Rahmen von Auktionen oder der Erfolg einer Maßnahme (definiert als $Y_1 - Y_0$) ist dagegen nie beobachtbar. Auch wenn diese Unterscheidung eine graduelle ist, so muss doch genau angegeben werden, welche Aussagen getroffen werden sollen. Rubins Punkt (a) soll im Folgenden für unvollständige Angaben in Umfragen diskutiert werden.

III. Beispiel: Einkommensangaben im ALLBUS

Das Haushaltsnettoeinkommen ist von zentraler Bedeutung in vielen Bereichen der Sozialforschung, etwa der Armutsforschung und der Haushaltstheorie. Dennoch ist selbst über das mittlere Hausaltsnettoeinkommen empirisch wenig bekannt. Das Statistische Jahrbuch 2001 weist auf der Basis der Einkommens- und Verbrauchsstichprobe (EVS) 1998 einen monatlichen Mittelwert von 5115 DM aus, 5346 DM in Westdeutschland, 4059 DM in Ostdeutschland (Statistisches Bundesamt 2001: 566ff). Dagegen weist der Datenreport 1999 auf der Basis des SOEP für das Jahr 1996 ein mittleres Haushaltsnettoeinkommen von 1978 DM (2061 DM West, 1644 DM Ost) aus (Statistisches Bundesamt 2000: 584). Die beiden Datensätze (EVS und SOEP) unterscheiden sich zwar deutlich: Die EVS ist eine Quotenstichprobe mit über 60000 beteiligten Haushalten, das SOEP ist eine weit kleinere Zufallsstichprobe.[2] Aber wegen der großen Abweichungen wäre

[2] Der Materialband zum ersten Armuts- und Reichtumsbericht der Bundesregierung (Bundesregierung 2001) enthält einen hilfreichen Vergleich verschiedener verfügbarer Datenquellen. Fehlende und unvollständige Einkommensangaben im SOEP und ihre Behandlung sind im Datenreport nicht angegeben. Die Arbeiten von Frick und Grapka (2003), Riphahn und Serfling (2002) und Schräpler (2004) geben einen Überblick.

es sicherlich wünschenswert, unabhängigen Aufschluss über die Verteilung des Haushaltseinkommens in der BRD zu erhalten. Von den großen regelmäßigen sozialwissenschaftlichen Umfragen enthält auch der ALLBUS eine Frage nach dem Haushaltsnettoeinkommen sowie nach dem persönlichen Einkommen. Der ALLBUS 1996 wurde als Melderegisterstichprobe durchgeführt. Grundgesamtheit waren Personen ab 18 Jahren in Privathaushalten (einschließlich Deutsch sprechender Ausländer) in West- und Ostdeutschland. Dabei wurden 3518 Interviews realisiert, 2402 davon in den alten und 1116 in den neuen Bundesländern. Personen in den neuen Bundesländern sind also überproportional befragt worden. Ihr Anteil in der Nettostichprobe beträgt 31,7%, der Bevölkerungsanteil betrug ca. 19%. Die berichtete Ausschöpfungsquote betrug 54,2%, d.h. nur 54,2% der angestrebten Interviews wurden realisiert (Koch 2002: 33).[3] Von den 3518 Befragten, die einem Interview zustimmten, antworteten gerade einmal 1772 oder 50,4% auf die Frage nach dem monatlichen Haushaltsnettoeinkommen. Weitere 906 Personen oder 25,8% machten Angaben in gruppierter Form. Insgesamt hat man also nur von etwa 41% der ausgewählten Personen eine valide Antwort auf die Frage nach dem Haushaltseinkommen erhalten, darunter 14% in gruppierter Form. Darüber hinaus gibt es aber noch 293 Personen, die zwar Angaben zu ihrem persönlichen Einkommen machten, aber keine Angaben zum Haushaltseinkommen. Diese Angaben könnten als untere Grenzen für das Haushaltseinkommen benutzt werden. Die Fallzahlen sind in der folgenden Tabelle zusammengestellt.

s_1	Bruttostichprobe	$\|s_1\| = n_1 = 6491$
s_2	Nettostichprobe	$\|s_2\| = n_2 = 3518$
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	$\|s_6\| = n_6 = 213$

Abbildung 1 zeigt die Dichte der genauen Angaben zum Haushaltsnettoeinkommen. Diese Angaben sind insbesondere in den höheren Einkommensbereichen zu einem großen Teil auf volle 1000 DM Beträge gerundet. So sind von den 122 Angaben von mehr als 7000 DM 120 in vollen 100 DM Beträgen und 107 in vollen 500 DM Beträgen angegeben, dagegen 85 Angaben in vollen 1000 DM Beträgen. Auf der anderen Seite sind von den 802 Angaben von 3000 DM oder weniger nur 277 in vollen 500 DM Beträgen angegeben. Die Rundungsregeln der Befragten sind unbekannt, hängen aber offenbar von der Einkommenshöhe ab. Auch die „genauen“ Einkommensangaben können nur als grobe Näherung des tatsächlichen

[3]Untersuchungen über den möglicherweise selektiven Ausfall von geplanten Interviews im ALLBUS, dem so genannten Unit-Nonresponse, sind von Koch (1997) und Schneekloth und Leven (2003) vorgelegt worden.

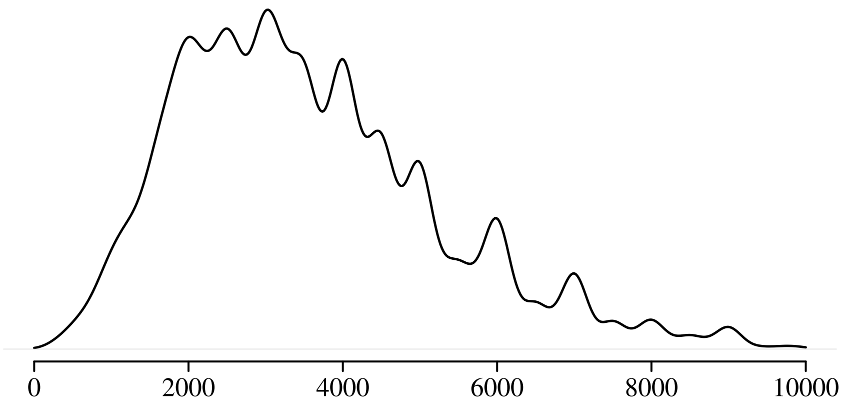

Abbildung 1: Kern-Dichte-Schätzung der genauen Angaben zum Haushaltseinkommen in DM, eingeschränkt auf das Intervall [1,9999]. Als Kern wurde eine Normalverteilung mit Standardabweichung 166 benutzt. Der Dichte-Schätzer ist etwas unterglättet.

verfügbaren Haushaltseinkommen betrachtet werden. Die Rundung in den Einkommensangaben führt zu einer weiteren Unsicherheit 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 Haushaltsnettoeinkommen 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

[4]Die Beschränkung des betrachteten Einkommenbereichs ist sowohl für nicht-probabilistische wie für probabilistische Modelle notwendig. Für nicht-probabilistische Überlegungen ist das unmittelbar einsichtig, weil Haushalten ohne Einkommensangaben jedes beliebige Einkommen zukommen könnte. Aber auch in probabilistischen Modellen sind Einschränkungen notwendig. Denn ohne solche Einschränkungen ist der Erwartungswert nicht einmal ein stetiges Funktional bezüglich der Kolmogorov-Metrik $d(F, G) := \sup_y |F(y) - G(y)|$ zwischen Verteilungsfunktionen F und G (Lehmann 1999: 391). Ohne Einschränkungen existieren also keine gleichmäßig konsistenten Tests, der Bootstrap funktioniert nicht etc. Wird das Verhalten von Schätzern nicht einfach unter einem als 'wahr' angenommenen Modell untersucht, sondern werden alle Modelle zugelassen, die mit den gegebenen Beobachtungen statistisch verträglich sind, dann ist der Bereich möglicher Erwartungswerte unendlich groß, selbst wenn die Existenz endlicher vierter Momente unterstellt wird (Davies 1995: 205). Im Zusammenhang mit Selektionsmodellen ist das Problem von Robins und Ritov (1997) untersucht worden (vgl. Abschnitt VII.).

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 von Stichprobe zu Stichprobe ändern und erlaubten keine statistischen Aussagen. Man erhält also einen Mittelwert, der wegen der sehr unterschiedlichen Auswahlsätze in Ost- und Westdeutschland wohl nichts über das durchschnittliche Haushaltseinkommen in Deutschland sagt, und ein Konfidenzintervall, dessen Überdeckungseigenschaften schlicht unbekannt sind.[6]

Soll der Rahmen der klassischen Stichprobentheorie nicht vollständig verlassen werden, dann muss die gesamte (realisierte) Stichprobe einschließlich der gruppierten Angaben betrachtet werden. Dies erfordert Verfahren zur Behandlung unvollständiger Angaben. Das einfachste Verfahren hält an der Idee fest, die Angaben der Befragten als fixes Datum zu behandeln. Unvollständige Angaben werden als Bereiche aufgefasst, in denen der tatsächliche Wert liegt. Ist die Auskunft eines Befragten, das Einkommen liege zwischen 2000 DM und 2500 DM, dann wird unterstellt, das exakte Einkommen sei eine der Zahlen 2000, 2001, 2002,...,2499. Bei einer Statistik wie dem Mittelwert wird die Berechnung für jeden der möglichen Werte in diesem Bereich durchgeführt. Das Ergebnis ist ein Bereich von Mittelwerten. Handelt es sich um zusammenhängende Intervalle, dann reicht es, die Statistik für die Extremwerte der Antwortintervalle auszuwerten.[7] Das Verfahren kann nur dann sinnvoll angewandt werden, wenn die Bereiche unvollständiger Daten beschränkt sind. Daher können nur Einkommensangaben etwa unter 10000 DM behandelt werden. Betrachtet man zunächst s'_3, so ergibt sich für die möglichen Mittelwerte das Intervall $[3589, 3785)$. Das Intervall der möglichen Mittelwerte ist deutlich länger als das naive 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.[8] 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.

[5] In der Teilstichprobe s'_3 ergibt sich sogar ein Anteil von 34,1%, ein „signifikanter" Unterschied zu 31,7%.

[6] Zur Berechnung eines Gesamtmittelwerts kann auch einfach der bekannte Bevölkerungsanteil in Ost- und Westdeutschland zur Kombination der Ergebnisse in Ost- und Westdeutschland benutzt werden. Dann werden Auswahlsätze einfach ignoriert. Für komplexere Fragen eignet sich eine solche naive Randanpassung allerdings nicht. Zusammenhänge zwischen Stichprobengewichten und Unit-nonresponse werden von Kalton (2002), Kalton und Flores-Cervantes (2003) diskutiert. Little und Vartivarian (2003) kritisieren einige klassische Verfahren.

[7] Die Grundidee ist recht alt (Cochran 1977: Kap. 13.2). Aber schon bei Statistiken wie der Varianz ergeben sich Probleme, effiziente Algorithmen für die Berechnung der Intervalle zu finden (Fishman und Rubin 1998; Rohwer und Pötter 2001: Kap. 19; Ferson et al. 2002). Manski (2003) gibt einen guten Überblick über neuere Ergebnisse. Für Kreuztabellen werden neuere Verfahren von Dobra und Fienberg (2000) beschrieben.

[8] Ich rechne die $n_3 - n'_3 = 45$ Angaben außerhalb von $[1, 9999]$ dazu.

IV. Stichproben und probabilistische Auswahlmodelle

Da beide Ansätze selten weiterhelfen, wurde versucht, die Fragestellung umzuformulieren. Der hierbei zumeist eingeschlagene Weg opfert einen wesentlichen Ausgangspunkt der Stichprobentheorie, der von den Berichten der Befragten als fixem Datum ausgeht. Stattdessen werden die Angaben der Befragten als Realisationen von Zufallsvariablen im Sinne der Wahrscheinlichkeitstheorie aufgefasst. Für probabilistische Modelle existieren bereits Methoden zur Analyse unvollständiger Daten. Zudem können in diesem Rahmen auch Modelle der Entstehung unvollständiger Angaben entwickelt werden.

Die Durchführung eines solchen Ansatzes ist konzeptionell weit schwieriger und konsequenzenreicher als oft angenommen wird. Einen Teil des Weges gehen *Superpopulationsmodelle*: Sie unterstellen, dass die interessierenden Größen in einer Gesamtheit $\mathcal{U}$ durch einen Zufallsprozess zustande gekommen seien, der sich durch eine Wahrscheinlichkeitsverteilung F_θ beschreiben lässt. Etwa: Das Einkommen der Bevölkerung der BRD wird als Realisation von $N := |\mathcal{U}|$ unabhängigen und identisch log-normalverteilten Zufallsvariablen erzeugt. Das ist offenbar keine realistische Annahme über das Zustandekommen von Einkommen. Die Metaphorik des „als ob durch einen Zufallsprozess zustandegekommen“ erlaubt aber relativ kompakte Beschreibungen von empirischen Verteilungen durch die Parameter θ 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 θ 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 Haushaltseinkommen gar nicht antwortet oder nur in gruppierter 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 $(\Omega, \mathcal{B}, \lambda)$ ein hinreichend großer Wahrscheinlichkeitsraum, mit dem alle diese Variablen beschrieben werden können, dann lässt sich etwa das Haushaltseinkommen als Funktion von $u \in \mathcal{U}$ und $\omega \in \Omega$ auffassen:

$$Y\colon \mathcal{U} \times \Omega \longrightarrow \{1, 2, 3, \ldots\} =: \mathcal{Y}$$

$Y(u, \omega)$ ist also das Haushaltseinkommen, das einer Person $u \in \mathcal{U}$ bei Realisierung von $\omega \in \Omega$ zukommt. Eine Person u hat in Abhängigkeit von ω verschiedene Einkommen. Aber es gibt ein $\omega_0 \in \Omega$, für das $(Y(u, \omega_0), u \in \mathcal{U})$ den Haushaltseinkommen $(y(u), u \in \mathcal{U})$ in der BRD entspricht.

Der Zusammenhang mit Aussagen über Durchschnitte der $Y(., \omega)$ über alle $u \in \mathcal{U}$ wird hergestellt, indem allen u gleiche Wahrscheinlichkeitsverteilungen zugeschrieben werden und die Unabhängigkeit von $Y(u, .)$ und $Y(u', .)$ für verschieden u, u' angenommen wird.

Tatsächlich in der Stichprobe beobachtet wird aber nur eine mengenwertige 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

$$S\colon \Omega \longrightarrow \mathcal{P}(\mathcal{U}) \setminus \{\emptyset\}$$

eine Stichprobe, wobei $\mathcal{P}(\mathcal{U})$ die Potenzmenge von $\mathcal{U}$ bezeichnet und S bzgl. $(\Omega, \mathcal{B})$ 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

$$Y^*\colon \mathcal{U} \times \Omega \longrightarrow \{\{y\} \,|\, y \in \mathcal{Y}\} \cup \{[1, 400), \ldots, [15000, \infty)\} \cup \{\mathcal{Y}\} \cup \{*\}$$

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 $\{\omega \,|\, S(\omega) = s\}$, als auch auf ganz Ω sind $Y(u, .)$ und $Y^*(u, .)$ Zufallsvariablen im Sinn der Wahrscheinlichkeitstheorie. 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 Vytlacil (2002: 332).

Durchschnitte über $\mathcal{U}$ und Durchschnitte über den Wahrscheinlichkeitsraum $(\Omega, \mathcal{B}, \lambda)$ führen nicht nur zu unterschiedlichen numerischen Ergebnissen, sie sind nicht einmal konzeptionell verbunden.[9] Zwar garantieren asymptotische Aussagen wie starke Gesetze oder Ergodensätze, dass die beiden Durchschnitte im Grenzwert ($|\mathcal{U}| \to \infty$ oder $\mathbb{E}(|S|)/|\mathcal{U}| \to c \notin \{0, 1\}, |\mathcal{U}| \to \infty$, 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 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

[9] Eine weitere Konsequenz probabilistischer Ansätze betrifft Designvariablen wie die Ost/West-Differenzierung im ALLBUS. Wird eine solche Variable als Konstante bzw. als degenerierte Zufallsvariable aufgefasst, kann sie wie in der klassischen Stichprobentheorie verwandt werden. Insbesondere können Gewichtungsverfahren benutzt werden, um Designaspekte zu berücksichtigen. Werden Designvariablen dagegen als Zufallsvariablen wie alle anderen behandelt, dann hängen diese Variablen von allen anderen Variablen eines Modells ab und Gewichtungsverfahren verlieren ihre Gültigkeit. Die beiden Ansätze führen z.B. bei Regressionsmodellen zu unterschiedlichen Ergebnissen.

> 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 $u \in \mathcal{U}$ 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. $\{F_\theta \,|\, \theta \in \Theta\}$ 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.

V. 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.[10] 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

[10] 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.

Verteilung F_θ von $Y(u,.)$ durch einen endlich-dimensionalen Vektor $\theta \in \mathbb{R}^k$ parametrisiert, etwa $Y(u,.){=}_d\mathrm{N}(\mu,\sigma^2)$, also normalverteilt mit Erwartungswert μ und Varianz σ^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 μ und Varianz σ^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[11]

$$L(\theta; y^*(u), u \in s) = \prod_{u\in s} \int_{v\in y^*(u)} dF_\theta(v)$$

Im Fall der Angaben zum Haushaltseinkommen im ALLBUS ergibt sich

$$\begin{aligned} &L(\theta; y^*(u), u \in s_2) \\ &\quad= \prod_{u\in s_4} \phi(y(u);\theta) \prod_{u\in s_3\setminus s_4} \int_{v\in y^*(u)} \phi(v;\theta)\,dv \prod_{u\in s_2\setminus s_3} \int_{v\in\mathbb{R}} \phi(v;\theta)\,dv \end{aligned}$$

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 Terme 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 μ 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 (μ,σ^2). Die Maximierung der entsprechenden Likelihood führt zu $\hat\mu = 8,1113$ und $\hat\sigma = 0,5371$. Da $\mathbb{E}_\theta(Y) = \exp(\mu + \sigma^2/2)$

[11] Die Formulierung ist bei absolut stetigen Verteilungen wie der Normalverteilung offenbar nicht korrekt. Denn falls es mindestens eine exakte Beobachtung gibt, wird der entsprechende Term 0. Es gibt verschiedene Vorschläge, wie auch absolut stetige Verteilungen in dieser allgemeinen Form behandelt werden können, vgl. Jacobsen und Keiding (1995), Gill et al. (1997) und Nielsen (2000).

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 $\hat{\theta} = 2,9879 * 10^{-4}$. Der Erwartungswert unter diesem Modell ist allerdings ∞. Man würde auch einen Erwartungswert von ∞ 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 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 $\mathcal{U}$ ist dagegen sicher immer endlich. Hier zeigt sich der bereits angedeutete Perspektivenwechsel: An Stelle eines Durchschnitts über alle $u \in \mathcal{U}$ interessiert der Durchschnitt über die $\omega \in \Omega$, durch die der Parameter θ erst seine Bedeutung erhält. Es gibt aber zwischen dem Mittelwert einer endlichen Menge $\mathcal{U}$ 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 $\{Y^*(u,.) \,|\, u \in s\}$

$$L(F; y^*(u), u \in s) = \prod_{u \in s} \int_{v \in y^*(u)} dF(v)$$

Im Fall von s_2 ergibt sich

$$\begin{aligned} &L(F; y^*(u), u \in s_2) \\ &\quad = \prod_{u \in s_4} F(y(u)) - F(y(u)_-) \prod_{u \in s_3 \setminus s_4} \int_{v \in y^*(u)} dF(v) \prod_{u \in s_2 \setminus s_3} \int_{v \in \mathbb{R}} dF(v) \end{aligned}$$

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 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

[12] Die Verteilung der geschätzten Mittelwerte für die Replikationen ist multimodal. Daher ergibt sich ein sehr asymmetrisches Konfidenzintervall, wenn wie hier die Perzentile der Verteilung zur Konstruktion benutzt werden.

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 ∞ verschoben werden. Das geschieht in den Bootstrap-Replikationen so oft, dass die obere Schranke des Konfidenzintervalls ehrlicherweise durch ∞ 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 $\omega \in \Omega$.

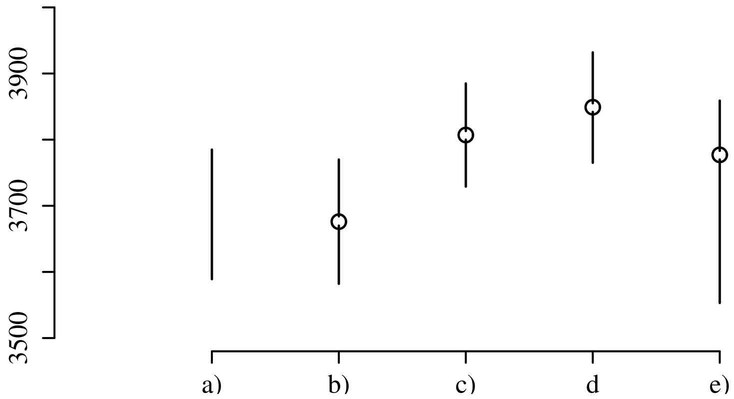

Abbildung 2: Geschätztes mittleres Haushaltseinkommen und 95% Konfidenzintervalle. a) Nichtparametrische Schranken auf s_3', b) Normalverteilung ohne gruppierte Angaben, c) Normalverteilung unter Einschluss gruppierter Angaben, d) Lognormalverteilung unter Einschluss gruppierter Angaben, e) Nichtparametrischer Mittelwert unter Einschluss gruppierter Angaben

VI. 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.

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 $y^*(u) = \{1000, 1500\}$ falls $y(u) = 1000$, sonst aber immer den genauen Betrag, dann ergibt sich die folgende Verteilung auf $\mathcal{Y}^*$:

$$\begin{aligned}
&\Pr(Y^*(u,.) = \{1000, 1500\}) = 1/3\\
&\Pr(Y^*(u,.) = \{1000\}) = 0\\
&\Pr(Y^*(u,.) = \{1500\}) = 1/3\\
&\Pr(Y^*(u,.) = \{2000\}) = 1/3
\end{aligned}$$

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 $Y(u,.)$: $\Pr(Y(u,.) = 1000) = 0$, $\Pr(Y(u,.) = 1500) = 2/3$, $\Pr(Y(u,.) = 2000) = 1/3$. Für den Zusammenhang zwischen Y und Y^* wird $\Pr(Y^*(u,.) = \{1000, 1500\} \,|\, Y(u,.) = 1000) = 1/2 = \Pr(Y^*(u,.) = \{1000, 1500\} \,|\, Y(u,.) = 1500)$ 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 $\Pr(Y^*(u,.) = \{1000, 1500\} \,|\, Y(u,.) = 1000)$ und $\Pr(Y^*(u,.) = \{1000, 1500\} \,|\, Y(u,.) = 1500)$. 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 Vergröberung* (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:

$$\Pr(Y^*(u,.) = y^* \,|\, Y(u,.) = y) \text{ ist konstant auf } y \in y^* \tag{1}$$

$$\Pr(Y^*(u,.) = y^* \,|\, Y(u,.) = y) = \Pr(Y^*(u,.) = y^* \,|\, Y(u,.) \in y^*) \tag{2}$$

$$\{Y^*(u,.) = y^*\} \perp\!\!\!\perp \{Y(u,.) = y\} \;|\; \{Y(u,.) \in y^*\} \tag{3}$$

$$\Pr(Y(u,.) = y \,|\, Y^*(u,.) = y^*) = \Pr(Y(u,.) = y \,|\, Y(u,.) \in y^*) \tag{4}$$

Dabei bedeutet $A \perp\!\!\!\perp B \,|\, C$ die bedingte stochastische Unabhängigkeit der Ereignisse A und B gegeben C. $\{\Pr(Y^*(u,.) = y^* \,|\, Y(u,.) = y) \,|\, y* \in \mathcal{Y}^*, y \in y^*\}$ soll im Folgenden *Selektionsmodell* heißen.

Die Bedingungen (1) und (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 $\Pr(Y^*(u,.) = y^* \,|\, Y(u,.) = 1234)$. Ist dagegen $y(u) = 1123$, so erzwingt (1) eine Auswahl zwischen den drei Antwortmöglichkeiten 1123, [1000,1250) und $[1,\ldots,\infty)$ 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 (2) ist nur eine Umformulierung dieser Beschreibung. Denn (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 (3) die Unabhängigkeit des Ereignisses „gruppierte Angabe“ von den tatsächlichen Einkommen innerhalb des Intervalls. Ist $y^* = \mathcal{Y} = [1,2,\ldots,\infty)$, dann ist die Bedingung $Y(u,.) \in [1,2,\ldots,\infty)$ immer erfüllt und (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 (3) ist automatisch erfüllt.

Die Bedingung (4) ist besonders hilfreich, weil sie nicht von den zugrunde liegenden Werten $Y(u,.)$ sondern von den beobachteten Angaben $Y^*(u,.)$ ausgeht. Ist etwa $y^*(u) = [1000, 1250)$, dann verlangt (4), dass die Verteilung der tatsächlichen Werte $Y(u,.)$ sich nicht von der bedingten Verteilung der $Y(u,.)$ unterscheidet, wenn bekannt ist, dass $Y(u,.)$ in dem Intervall [1000,1250) liegt. Genau dies ist in der Konstruktion der Likelihoodfunktionen im letzten Abschnitt verwandt worden.

Schreibt man $\Pr_\theta(Y(u,.) = y)$, um die Abhängigkeit der Verteilung von einem Parameter $\theta \in \Theta$ anzugeben, und entsprechend $\Pr_\gamma(Y^*(u,.) = y^* \,|\, Y(u,.) = y)$ mit $\gamma \in \Gamma$ für das Selektionsmodell, dann kann die Verteilung der Beobachtungen $Y^*(u,.)$ wie folgt aufgespalten werden:

$$
\begin{aligned}
&\Pr{}_{\theta,\gamma}(Y^*(u,.) = y^*) \\
&\quad = \sum_{y \in y^*} \Pr{}_{\theta,\gamma}(Y^*(u,.) = y^*, Y(u,.) = y) \\
&\quad = \sum_{y \in y^*} \Pr{}_{\theta}(Y(u,.) = y) \Pr{}_{\gamma}(Y^*(u,.) = y^* \,|\, Y(u,.) = y)
\end{aligned}
$$

$$= \quad \Pr{}_\gamma(Y^*(u,.) = y^* \,|\, Y(u,.) = y, y \in y^*) \sum_{y \in y^*} \Pr{}_\theta(Y(u,.) = y)$$
$$= \quad \Pr{}_\theta(Y(u,.) \in y^*) \Pr{}_\gamma(Y^*(u,.) = y^* \,|\, Y(u,.) \in y^*)$$

Die dritte Gleichung folgt aus der CAR-Bedingung (1), die letzte Gleichung aus (2). Sind die Parameter θ und γ variationsunabhängig, gibt es also zu jedem Element $(\theta, \gamma) \in \Theta \times \Gamma$ eine Wahrscheinlichkeitsverteilung $\Pr_{\theta,\gamma}$, dann kann bei Likelihoodbetrachtungen für θ der Selektionsteil $\Pr_\gamma(Y^*(u,.) = y^* \,|\, Y(u,.) \in y^*)$ 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 genauen Angabe oder gar nicht, dann ist $\mathcal{Y}^* = \{\{y\} \,|\, y \in \mathcal{Y}\} \cup \{\mathcal{Y}\}$. 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 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.[13] 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 $\Pr_\gamma(Y^*(u,.) = y^* \,|\, Y(u,.) \in y^*)$ 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).

Zusammenfassend lässt sich sagen, dass erstens probabilistische Überlegungen zu einer nicht trivialen Charakterisierung von Bedingungen führen, unter denen klassische Methoden für unvollständige Beobachtungen korrekt sind: die CAR-Bedingungen. Zweitens zeigt sich, dass Selektionsmodelle empirisch nicht identifiziert sind: Es kann immer ein CAR-Modell konstruiert werden, das die Daten exakt reproduziert, ganz unabhängig davon, wie die Daten „tatsächlich" entstanden sind. Die CAR-Annahme und damit die Verwendung klassischer Likelihood-Methoden lässt sich empirisch nicht hinterfragen. Drittens ist es selbst unter der CAR-Bedingung möglich, einige (semi-) parametrische Selektionsmodelle empirisch zurückzuweisen, wenn neben exakten und vollständig fehlenden Angaben auch gruppierte oder andere partielle Angaben vorliegen und entsprechend modelliert werden. Man darf aber bei all dem Fortschritt nicht vergessen, dass probabilistische Modelle untersucht werden. Die Modelle können nicht umstandslos mit dem realen Verhalten von Befragten gleichgesetzt werden. Denn die Modelle unterstellen u.a., alle Befragten würden ihr Antwortverhalten nach einer Wahrscheinlichkeitsverteilung auswürfeln, die zudem für alle gleich wäre.

VII. Ignorierbare Ausfälle: Konditionale CAR-Modelle

In vielen Fällen gibt es neben den Angaben $(y^*(u), u \in s)$ zu den interessierenden Größen $(y(u), u \in \mathcal{U})$ 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: $X : \mathcal{U} \times \Omega \to \mathcal{X}$. Die CAR-Bedingungen können konditional auf die Werte dieser Kovariablen formuliert werden:

$$\Pr(Y^*(u,.) = y^* \,|\, Y(u,.) = y, X(u,.) = x) \text{ ist konstant auf } y \in y^* \quad (5)$$

$$\begin{aligned} &\Pr(Y^*(u,.) = y^* \,|\, Y(u,.) = y, X(u,.) = x) \\ &\quad = \Pr(Y^*(u,.) = y^* \,|\, Y(u,.) \in y^*, X(u,.) = x) \end{aligned} \quad (6)$$

$$\{Y^*(u,.) = y^*\} \perp\!\!\!\perp \{Y(u,.) = y\} \;|\; \{Y(u,.) \in y^*\}, X(u,.) \quad (7)$$

$$\begin{aligned} &\Pr(Y(u,.) = y \,|\, Y^*(u,.) = y^*, X(u,.) = x) \\ &\quad = \Pr(Y(u,.) = y \,|\, Y(u,.) \in y^*, X(u,.) = x) \end{aligned} \quad (8)$$

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 Kovariabler in den Bedingungen (5) – (8).[14]

Eine Idee, die schon von Pearson Anfang des letzten Jahrhunderts formuliert wurde, geht davon aus, dass Y eine lineare Regression auf den 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 $\mathcal{U}$ mit dem jeweils gleichen β 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 $Y \mid X, I[R = 1]$ 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 β 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.[15] Der naive Mittelwert, der nur vollständige Angaben berücksichtigt, ist um 90 DM kleiner und befindet sich am unteren Rand des Konfidenzintervalls.

---

[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: $\mathrm{Cov}(Y_1, Y_2 \mid U = u) = \mathbb{E}(Y_1 Y_2 \mid U = u) - \mathbb{E}(Y_1 \mid U = u)\mathbb{E}(Y_2 \mid U = u) = \mathbb{E}\left((u + V)(u - V)\right) - u^2 = \mathrm{Var}(V) - u^2 = 1 - u^2$. $Y_1 \perp\!\!\!\perp Y_2 | U = u$ 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 (7) kann durch die Einbeziehung weiterer Kovariabler also auch verletzt werden. Weitere Probleme bei der Interpretation von „Information“ bei bedingten Modellen diskutieren Dubra und Echenique (2004).

[15] Die 6 Beobachtungen ohne Altersangabe wurden ausgeschlossen, die Schätzung von β 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.

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 bezogen auf die Regressionsgleichung. Diese ist in der Regel nicht direkt von Interesse. Im Fall des ALLBUS soll eine Aussage über das mittlere Haushaltseinkommen getroffen werden, nicht aber über einen Regressionszusammenhang. Zudem sind die Kovariablen X, die für alle Befragten zur Verfügung stehen, hauptsächlich durch das Stichprobendesign und die Fragebogenkonstruktion bestimmt, nicht durch inhaltliche Überlegungen. Daher wird versucht, den Regressionszusammenhang $\mathbb{E}(Y \mid X)$ möglichst allgemein, d.h. ohne parametrische Annahmen, zu modellieren. Das Haushaltseinkommen ist sicherlich keine lineare Funktion des Alters. Theoretisch gibt es keinen sinnvollen Zusammenhang zwischen dem Alter eines Befragten und dem Haushaltseinkommen, es sei denn, es handelt sich um Ein-Personen-Haushalte. Der lineare Term für das Alter sollte daher flexibler, etwa durch Spline-Funktionen dargestellt werden. Zudem sollten möglichst alle Interaktionen zwischen den Kovariablen berücksichtigt werden. Wenn aber keine parametrischen Annahmen oder wenigstens Annahmen über die Glätte der Regressionsbeziehung getroffen werden können, dann gibt es innerhalb des Ansatzes nicht einmal ein Schätzverfahren, das gleichmäßig konsistent ist (Robins und Ritov 1997: 294f). Selbst wenn diese theoretischen Schwierigkeiten ignoriert werden, bleibt das praktische Problem, ein relativ stabiles und gleichzeitig allgemein akzeptierbares Regressionsmodell für Y gegeben X zu formulieren.

Eine Möglichkeit, zumindest einige der theoretischen Probleme zu umgehen, ergibt sich aus einem Rückgriff auf eine Idee der Stichprobentheorie 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 $\pi(u,x) := \Pr(R(u,.) = 1 \mid X(u,.) = x) = \Pr(R(u,.) = 1 \mid X(u,.) = x, Y(u,.) = y)$ 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 $R(u,.) \perp\!\!\!\perp Y(u,.) \mid X(u,.)$. Ist die Auswahlwahrscheinlichkeit für alle Kovariablenwerte größer als eine positive Schranke, $\pi(u,X) > \sigma > 0$ für alle $X(u,.)$, dann gilt

$$
\begin{aligned}
&\mathbb{E}\left(\frac{I[R(u,.) = 1]Y(u,.)}{\pi(u,X)}\right) \\
&\quad = \quad \mathbb{E}\left(\mathbb{E}\left(\frac{I[R(u,.) = 1]Y(u,.)}{\pi(u,X)} \,\middle|\, Y(u,.) = y, X(u,.) = x\right)\right)
\end{aligned}
$$

$$= \mathbb{E}\left(\frac{Y(u,.)}{\pi(u,X)}\mathbb{E}\left(I[R(u,.)=1]\,\middle|\,Y(u,.)=y,X(u,.)=x\right)\right)$$
$$= \mathbb{E}\left(\frac{\pi(u,X)Y(u,.)}{\pi(u,X)}\right)=\mathbb{E}\left(Y(u,.)\right)$$

Daher ist

$$\hat{\mu}=\sum_{u\in s}\frac{I[R(u,.)=1]Y(u,.)}{\pi(u,X)}/\sum_{u\in s}\frac{I[R(u,.)=1]}{\pi(u,X)}$$

ein erwartungstreuer Schätzer des Erwartungswerts der $Y(u,.)$, und zwar ganz unabhängig von der Form des bedingten Erwartungswerts $\mathbb{E}(Y\,|\,X)$. 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 $\{u\in s\,|\,R(u,.)=1\}$ für die Schätzung. Ein Teil der Information kann durch die Addition eines weiteren, von Funktionen der $X(u,.)$ und $I[R(u,.)=1]Y(u,.)$ abhängigen Terms zur Schätzgleichung zurückerhalten werden.[16] Allerdings können im Rahmen gewichteter Schätzer partielle Angaben nur sehr rudimentär Berücksichtigung finden. Die Schätzgleichung beruht zunächst nur auf vollständigen Angaben, partielle Angaben werden im Gegensatz zum Pearson-Lawley Ansatz oder zu Likelihood-Methoden nur über weitere additive Terme in die Schätzgleichung eingeführt.

VIII. Nicht ignorierbare Ausfälle

Bei der Diskussion von Selektionsmodellen wird überlegt werden müssen, was geschieht, wenn sie nicht die CAR-Bedingung erfüllen. In diesem Fall hängt $\Pr(Y^*(u,.)=y^*\,|\,Y(u,.)=y)$ aufgrund von (1) von $y\in y^*$ ab und diese bedingten Wahrscheinlichkeiten könnten modelliert werden. Andersherum kann von (4) ausgegangen und entsprechend $\Pr(Y(u,.)=y\,|\,Y^*(u,.)=y^*)$ modelliert werden. Eine Reihe anderer Möglichkeiten der Konstruktion nicht ignorierbarer Selektionsmodelle sind denkbar. Das wohl bekannteste Modell ist von Heck-

[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 $Y(u,.)$ gegeben $X(u,.)$ bzw. gegeben $X(u,.), R(u,.)=1$ ab (Rotnizky und Robins 1997; Scharfstein und Irizarry 2003; van der Laan und Robins 2003). Beide Verteilungen sind aber nicht von direktem Interesse und ihre nichtparametrische Schätzung, die zusätzliche probabilistische Annahmen vermeidet, ist bei den üblichen Stichprobengrößen wie im ALLBUS sehr instabil.

man (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 Zusammmenhang 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 $Y(u,.) \perp\!\!\!\perp R^*(u,.)$ gilt. Denn der Ausdruck $\Pr(Y^*(u,.) = \mathcal{Y} \mid Y(u,.) = y) = \Pr(R(u,.) = 2 \mid Y(u,.) = y) = \int_{-\infty}^{0} f(v \mid Y(u,.) = y)\, dv$ müsste aufgrund von (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 $\mathbb{E}(Y(u,.)) := \mu$, $\mathbb{E}(R^*(u,.)) := \mu^*$, $\mathrm{Var}(R^*(u,.)) := 1$, $\mathrm{Var}(Y(u,.)) := \sigma^2$ und $\mathrm{Corr}(Y(u,.), R^*(u,.)) := \rho$, dann ist die bedingte Dichte von $Y(u,.)$ gegeben $R(u,.) = 1$

$$\phi(y \mid R^*(u,.) \geq 0) = \frac{1}{\Pr(R(u,.) = 1)} \frac{1}{\sigma} \phi((y-\mu)/\sigma) \Phi \left(\frac{1}{\sqrt{1-\rho^2}} (\mu^* + \frac{\rho}{\sigma}(y - \mu)) \right)$$

wobei ϕ bzw. Φ 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 (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 „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“.

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 μ^* durch $\Phi^{-1}(\mathrm{P}(R = 1))$ konstruiert werden, wobei $\mathrm{P}(R = 1)$ den Anteil vollständiger Beobachtungen in der Stichprobe s angibt. Dagegen müssen μ, σ und ρ auf der Basis der vollständigen Beobachtungen geschätzt werden. Insbesondere der Parameter ρ, 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 μ, σ, ρ 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 III.). Nach diesem Ergebnis hätte ein erheblicher Teil der Population ein stark negatives Haushaltseinkommen. Zudem wäre die „Antwortwortbereitschaft“ $R^*(u,.)$ 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 $\Pr(R(u,.) = 1 \,|\, Y(u,.) = y, X(u,.) = x)$ 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 ρ erfordert eine parametrische Einschränkung der Wirkung der Kovariablen. Aber die Auswirkungen etwa von Linearitätsannahmen in $\Pr(R(u,.) = 1 \,|\, Y(u,.) = y, X(u,.) = x)$ sind ebenso wenig wie die Auswirkungen willkürlicher Verteilungsannahmen einfach zu überblicken.

Immerhin ergibt das Heckmansche Modell einen ersten Ansatzpunkt zur Modellierung nicht ignorierbarer Selektionsprozesse. Zudem können auch partielle Angaben relativ leicht einbezogen werden. In der Tat kann das Modell so erweitert werden, dass viele häufig auftretende Probleme mit unvollständigen Angaben in diesem Rahmen formuliert werden können (z.B. Crouchley und Ganjali 2002). Weiterhin wurden Verteilungsannahmen abgeschwächt (Das et al. 2003) und asymptotische Analysen verfeinert (Rotnitzky et al. 2000). Ausgehend von Heckmanschen Selektionsmodellen können die Auswirkungen „lokaler“ Abweichungen von den Annahmen approximiert werden (Copas und Li 1997).

IX. Sensitivität

In der sozialwissenschaftlichen Praxis reicht es zumeist nicht, sich für ein Selektionsmodell zu entscheiden und sodann Ergebnisse nur unter dieser Modellannahme zu präsentieren. Wird mit einer CAR-Annahme begonnen, so muss dennoch Rechenschaft über die Konsequenzen der Annahme abgelegt werden.

Das ist umso wichtiger, als die CAR-Annahme dazu verführen könnte, nicht über Selektionsprozesse nachzudenken. Zwar können „lokale“ Abweichungen von der CAR-Annahme allgemein beschrieben werden. Damit ist es auch möglich, die Konsequenzen der CAR-Annahme für statistische Aussagen zu approxmieren (Copas und Eguchi 2001). Aber probabilistische CAR-Modelle können kaum als realistische Modelle für Stichprobenausfälle angesehen werden. Daher wird eine lokale Approximation, so nützlich sie theoretisch ist, zumeist nicht befriedigen können. Die Approximation läuft Gefahr, die Konsequenzen von Annahmen über den Selektionsprozess nur in den engen Grenzen probabilistisch ähnlicher Modelle zu untersuchen. Rosenbaums Sensitivitätsanalyse (2002: Chap. 4) knüpft an Überlegungen über die Variable $R^*(u,.)$ an, wie sie auch in Heckmans Modell verwandt wird. Rosenbaums Methode geht von einer größeren Menge von Selektionsmodellen aus und erlaubt damit eine umfassendere Abschätzung der Effekte von nicht ignorierbaren Selektionsprozessen. Die Methode verbleibt aber in der unterstellten Modellwelt und ist zudem bisher nur für spezielle Statistiken formuliert worden.

Robins und Mitarbeiter haben vorgeschlagen, von den gewichteten Schätzfunktionen

$$\sum_{u \in s} \frac{I[R(u,.)=1]}{\pi(u,X)}(Y(u,.)-\mu)$$

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

$$\begin{aligned}
&\mathbb{E}\left(1-\frac{I[R(u,.)=1]}{\pi(u,X,Y)}\right) \\
&\quad = \quad 1-\mathbb{E}\left(\frac{1}{\pi(u,X,Y)}\mathbb{E}\left(I[R(u,.)=1] \mid X(u,.),Y(u,.)\right)\right)=0
\end{aligned}$$

ist, kann die Schätzfunktion erweitert werden:

$$\sum_{u \in s} \frac{I[R(u,.)=1]}{\pi(u,X,Y)}(Y(u,.)-\mu)+\left(1-\frac{I[R(u,.)=1]}{\pi(u,X,Y)}\right)\phi(X;\mu)$$

Wird etwa logit $\pi(u,x,y)=x(u)\beta+\alpha y(u)$ gesetzt und

$$\phi(X;\mu,\alpha)=\frac{\mathbb{E}(Y(u,.)\exp(\alpha Y(u,.)) \mid R(u,.)=1,X(u,.))}{\mathbb{E}(\exp(\alpha Y(u,.)) \mid R(u,.)=1,X(u,.))}-\mu$$

gewählt, dann ergibt sich ein *doppelt robuster* Schätzer, der auch dann konsistent ist, wenn die Auswahlgleichung fehlspezifiziert ist. Dabei kann der Koeffizient α 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 α auch in sehr großen Datensätzen kaum stabil schätzbar. Robins und Mitarbeiter haben daher vorgeschlagen, eine Reihe von Werten des Koeffizienten α fest zu wählen und die Auswirkungen auf die Schätzung von μ zu notieren. Wie ihre Simulationen und Beispiele aber zeigen, sind auch diese Schätzungen sehr instabil. Zudem erscheint noch unklar, wie partielle Angaben in diesem Ansatz direkt berücksichtigt werden können (Robins 1997; Robins und Gill 1997).

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 III. und verzichten auf probabilistische Annahmen über den Selektionsprozess. Sie führen nur dann zu relativ engen Intervallen, wenn $Y(u, .)$ beschränkt ist. Im Fall des Haushaltseinkommens ergeben sich daher Intervalle für die möglichen Werte der Durchschnitte wie schon im Abschnitt III.. Diese Intervalle sind auch die theoretischen Grenzen der Methode von Robins et al., wenn ihr Koeffizient α alle Werte zwischen $-\infty$ und ∞ durchläuft (Scharfstein et al. 1999: 1108). Raghhunathan 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.“ (Manski und Horowitz 2000: 87f).

X. 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 Ausschöpfungsraten 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 angwiesen. Existieren aber gruppierte, zensierte oder fehlklassifizierte Berichte der Befragten, dann müssen diese Angaben ernst genommen und in die Berechnung von Statistiken einbezogen werden. Es werden Verfahren benötigt, die auch diese partiellen Angaben berücksichtigen. Selektionsmodelle erlauben es, über die Voraussetzungen 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.

Literatur

ALLBUS, 1998: Codebuch des kumulierten ALLBUS 1980–96, ZA–Nummer 1795, Release 98.01. Köln: Zentralarchiv für Empirische Sozialforschung.

Berk, Richard A., und *Subash C. Ray,* 1982: Selection bias in sociological data. Social Science Research 11: 352–398.

Bryson, Maurice C., 1976: The Literary Digest poll: Making of a statistical myth. American Statistician 30: 184–185.

Bundesregierung, 2001: Lebenslagen in Deutschland. Daten und Fakten. Materialband zum ersten Armuts- und Reichtumsbericht der Bundesregierung. BundestagsdrucksachenNr. 14 / 5990.

Cahalan, Don, 1989: The Digest poll rides again. Public Opinion Quarterly 53: 129–133.

Cochran, William G., 1977^3: Sampling Techniques. New York: Wiley.

Copas, John B., und *H. G. Li,* 1997: Inference for non-random samples (with discussion). Journal of the Royal Statistical Society, B, 59: 55–95.

Copas, John B., und *Shinto Eguchi,* 2001: Local sensitivity for selectivity bias. Journal of the Royal Statistical Society, B, 63: 871–895.

Crouchley, Rob, und *Mojtaba Ganjali,* 2002: The common structure of several models for non-ignorable dropout. Statistical Modeling 2: 39–62.

Das, Mitali, Whitney K. Newey und *Francis Vella,* 2003: Nonparametric estimation of sample selection models. Review of Economic Studies 70: 33–58.

Davies, Laurie, 1995: Data features. Statistica Neerlandica 49: 185–245.

Dawid, A. Philip, 2000: Causal inference without counterfactuals (with discussion). Journal of the American Statistical Association 95: 407–448.

Diekmann, Andreas, und *David Wyder,* 2002: Vertrauen und Reputationseffekte bei Internet-Auktionen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 674–693.

Dobra, Adrian, und *Stephen E. Fienberg,* 2000: Bounds for cell entries in contingency tables given marginal totals and decomposable graphs. Proceedings of the National Academy of Sciences 97: 11885–11892.

Dubra, Juan, und *Federico Echenique,* 2004: Information is not about measurability. Mathematical Social Sciences 47: 177–185.

Engelhardt, Henriette, 1999: Lineare Regression mit Selektion: Möglichkeiten und Grenzen der Heckman-Korrektur. Kölner Zeitschrift für Soziologie und Sozialpsychologie 51: 706–723.

Ferson, Scott, Lev Ginzburg, Vladik Kreinovich, Luc Longpré und *Monica Aviles,* 2002: Computing variance for interval data is NP-hard. ACM SIGACT News 33: 108–118.

Fishman, George S., und *David S. Rubin,* 1998: Best- and worst-case variances when bounds are available for the distribution function. Computational Statistics & Data Analysis 29: 35–53.

Frick, Joachim R., und *Markus M. Grabka,* 2003: Missing data in the German SOEP: Incidence, imputation and its impact on the income distribution. Discussion Papers 376. Berlin: DIW.

Gill, Richard D., 1997: Nonparametric estimation under censoring and passive registration. Satistica Neerlandica 51: 35–54.

Gill, Richard D., *Mark J. van der Laan* und *James M. Robins*, 1997: Coarsening at random: Characterizations, conjectures, counter-examples. S. 255–294 in: *D.Y. Lin* und *Thomas R. Fleming* (HG.), 1997: Proceedings of the 1st Seattle Symposium in Biostatistics: Survival Analysis. Berlin: Springer.

Grünwald, Peter D., und *Joseph Y. Halpern,* 2003: Updating probabilities. Journal of Artificial Intelligence Research 19: 243–278.

Heitjan, Daniel F., 1994: Ignorability in general incomplete-data models. Biometrika 81: 701–708.

Heitjan, Daniel F., und *Donald B. Rubin* 1991: Ignorability and coarse data. The Annals of Statistics 19: 2244–2253.

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.

Heckman, James J., 1979: Sample selection bias as a specification error. Econometrica 47: 153–161.

Heckman, James J., 1990: Varieties of selection bias. American Economic Review 80: 313–318.

Holland, Paul W., 1986: Statistics and causal inference (with discussion). Journal of the American Statistical Association 81: 945–970.

Jacobsen, Martin, und *Niels Keiding*, 1995: Coarsening at random in general sample spaces and random censoring in continuous time. The Annals of Statistics 23: 774–786.

Jonker, Marianne A., 2003: Estimation of life expectancy in the Middle Ages. Journal of the Royal Statistical Society, A, 166: 105–117.

Kalton, Graham, 2002: Models in the practice of survey sampling (revisited) (with discussion). Journal of Official Statistics 18: 129–161.

Kalton, Graham, und *Ismael Flores-Cervantes*, 2003: Weighting methods. Journal of Official Statistics 19: 81–97.

Koch, Achim, 1997: Teilnahmeverhalten beim ALLBUS 1994. Soziodemographische Determinanten von Erreichbarkeit, Befragungsfähigkeit und Kooperationsbereitschaft. Kölner Zeitschrift für Soziologie und Sozialpsychologie 49: 98–122.

Koch, Achim, 2002: 20 Jahre Feldarbeit im ALLBUS: Ein Blick in die Blackbox. ZUMA Nachrichten 51: 9–37.

Lawley, D.N., 1943: A note on Karl Pearson's selection formulae. Proceedings of the Royal Society of Edinburgh A, 62, 28–30.

Le Cam, Lucien, und *Grace Lo Yang,* 1990: Asymptotics in Statistics. Some Basic Concepts. Berlin: Springer.

Lehmann, Erich Leo, 1999: Elements of Large-Sample Theory. Berlin: Springer.

Little, Roderick J., und *Donald B. Rubin*, 2002^2: Statistical Analysis with Missing Data. New York: Wiley.

Little, Roderick J., und *Sonya Vartivarian*, 2003: On weighting the rates in non-response weights. Statistics in Medicine 22: 1589–1599.

Madow, William G. und *Ingram Olkin* (Hg.), 1983: Incomplete Data in Sample Surveys. 3 Bände. New York: Academic Press

Manski, Charles F., 1993: The selection problem in econometrics and statistics. S. 73–84 in *G.S. Maddala, C.R. Rao* und *H.D. Vinod* (Hg.), 1993: Handbook of Statistics, Vol. 11 Amsterdam: Elsevier.

Manski, Charles F., 2003: Partial Identifiability of Probability Distributions. Berlin: Springer.

Manski, Charles F., und *Joel L. Horowitz,* 2000: Nonparametric analysis of randomized experiments with missing covariate and outcome data (with discussion). Journal of the American Statistical Association 95: 77–88.

Manski, Charles F., und *Elie Tamer*, 2002: Inference on regressions with interval data on a regressor or outcome. Econometrica 70: 519–546.

Mayer, Karl Ulrich, und *Paul B. Baltes* (Hg.), 1996: Die Berliner Altersstudie. Berlin: Akademie-Verlag.

Mohler, Peter, Achim Koch und *Siegfried Gabler,* 2003: Alles Zufall oder? Ein Diskussionsbeitrag zur Qualität von face to face Umfragen in Deutschland. ZUMA Nachrichten53: 10–15.

Nicoletti, Cheti, 2002: Correcting for sample selection bias: Alternative estimators compared. `http://www.iser.essex.ac.uk/activities/seminars/Monday_Afternoons/archive/papers/poverty3.pdf`

Nielsen, Søren Feodor, 2000: Relative coarsening at random. Statistica Neerlandica 54: 79–99.

Qin, Jing, Denis Leung und *Jun Shao,* 2002: Estimation with survey data under nonignorable nonresponse or informative sampling. Journal of the American Statistical Association 97: 193–200.

Riphahn, Regina T., und *Oliver Serfling*, 2002: Item non-response on income and wealth questions. Forschungsinstitut zur Zukunft der Arbeit. Discussion Paper No. 573.

Robins, James M., 1997: Non-response models for the analysis of non-monotone non-ignorable missing data. Statistics in Medicine 16: 21–37.

Robins, James M., und *Richard D. Gill,* 1997: Non-response models for the analysis of non-monotone ignorable missing data. Statistics in Medicine 16: 39–56.

Robins, James M., und *Ya'acov Ritov,* 1997: Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Statistics in Medicine 16: 285–319.

Rohwer, Götz, und *Ulrich Pötter*, 2001: Grundzüge der sozialwissenschaftlichen Statistik. Weinheim: Juventa.

Rosenbaum, Paul R., 2002^2: Observational Studies. Berlin: Springer.

Rotnitzky, Andrea, und *James M. Robins*, 1997: Analysis of semi-parametric regression models with non-ignorable non-response. Statistics in Medicine 16: 81–102.

Rotnitzky, Andrea, James M. Robins und *Daniel O. Scharfstein,* 1998: Semiparametric regression for repeated outcomes with nonignorable nonresponse. Journal of the American Statistical Association 93: 1321–1339.

Rotnitzky, Andrea, David R. Cox, Matteo Bottai und *James M. Robins*, 2000: Likelihood-based inference with singular information matrix. Bernoulli 6: 243–284.

Schafer, Joseph L., 1997: Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Scharfstein, Daniel O., Andrea Rotnitzky und *James M. Robins*, 1999: Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). Journal of the American Statistical Association 94: 1096–1146.

Scharfstein, Daniel O., und *Rafael A. Irizarry,* 2003: Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. Biometrics 59: 601–613.

Schneekloth, Ulrich, und *Ingo Leven*, 2003: Woran bemisst sich eine „gute" allgemeine Bevölkerungsumfrage? Analysen zu Ausmaß, Bedeutung und zu den Hintergründen von Nonresponse in zufallsbasierten Stichprobenerhebungen am Beispiel des ALLBUS. ZUMA Nachrichten 53: 16–57.

Schräpler, Jörg-Peter, 2004: Respondent behavior in panel studies — A case study for income-nonresponse by means of the SOEP. Erscheint in: Sociological Methods & Research.

Squire, Peverill, 1988: Why the 1936 Literary Digest poll failed. Public Opinion Quarterly 52: 125–133.

Statistisches Bundesamt (Hg.), 2000: Datenreport 1999. Zahlen und Fakten über die Bundesrepublik Deutschland. Bundeszentrale für politische Bildung, Schriftenreihe Band 365, Bonn.

Statistisches Bundesamt, 2001: Statistisches Jahrbuch. Wiesbaden.

Stolzenberg, Ross M., und *Daniel A. Relles,* 1997: Tools for intuition about sample selection bias and its correction. American Sociological Review 62: 494–507.

van der Laan, Mark J., und *James M. Robins,* 2003: Unified Methods for Censored Longitudinal Data and Causality. Berlin: Springer.

Vella, Francis, 1998: Estimating models with sample selection bias: A survey. The Journal of Human Resources 33: 127–169.

Vytlacil, Edward, 2002: Independence, monotonicity, and latent index models: An equivalence result. Econometrica 70: 331–341.

Wainer, Howard (Hg.), 1986: Drawing Inferences from Self-Selected Samples. Mahwah: Lawrence Erlbaum.

Wainer, Howard, 1989, 1992: Eeelworms, bullet holes, and Geraldine Ferraro: Some problems with statistical adjustment and some solutions (with discussion). Journal of Educational Statistics 14: 121–140. Wieder abgedruckt S. 129–207 in *Juliett Popper Shaffer* (Hg.), 1992: The Role of Models in Nonexperimental Social Science: Two Debates. Washington: American Educational Research Association and American Statistical Association.

Winship, Christopher, und *Robert D. Mare,* 1992: Models for sample selection bias. Annual Review of Sociology 18: 327–350.

Zaffalon, Marco, 2002: Exact credal treatment of missing data. Journal of Statistical Planning and Inference 105: 105–122.

Korrespondenzanschrift: Dr. Ulrich Pötter, Ruhr–Universität Bochum, Fakultät für Sozialwissenschaft, Sektion Methodenlehre und Statistik, Universitätsstr. 150, 44801 Bochum

E-Mail: ulrich.poetter@ruhr-uni-bochum.de