Thilo Ernst
Ulrich Pötter
Ruhr-Universität Bochum
Sommersemester 2006
Rechentechniken der Datenanalyse

Arbeitsblatt 8

1) Um die Daten aus verschiedenen Wellen eines Panels als einen Datensatz behandeln zu können, ist es meist erforderlich, die Daten der verschiedenen Wellen einzeln einzulesen und dann anhand der in jeder Welle vorhandenen eindeutigen Identifikationsvariablen zusammenzufügen. Der entsprechende Befehl in R ist `merge()`. Zunächst einige Beispieldaten:

```
X <- data.frame(id=c(1,2,3,4,5),
  Vorname=c("Egon", "Franz", "Thorsten", "Katrin", "Susi"))

Y <- data.frame(id=c(1,2,3,4,5),
  Alter=c(18,53,22,34,21),
  Einkommen=c(1000,2500,7385,990,956))

X; Y
Z <- merge(X,Y)
Z
```

Ohne weitere Optionen benutzt `merge()` die in beiden Datensätzen gleich benannten Variablen, um die Daten zeilenweise zuzuordnen (in diesem Beispiel also `id`). Die zu verwendenden Variablen lassen sich mit `by.x` und `by.y` auch explizit angeben. Im Beispiel entspricht das Standardverhalten folgender Angabe: `merge(X, Y, by.x="id", by.y="id")`.

Zudem sind im resultierenden Datensatz ohne weitere Einstellungen nur diejenigen Daten zusammengeführt, die gleichzeitig in beiden ursprünglichen Datensätzen enthalten sind:

```
Y <- data.frame(id=c(1,2,3,4,6),
  Alter=c(18,53,22,34,21),
  Einkommen=c(1000,2500,7385,990,956))

Z <- merge(X,Y)
X; Y; Z
```

Sollen alle Zeilen bzw. Fälle des einen und/oder anderen Datensatzes im `merge`-Ergebnis enthalten sein, lässt sich dies durch die Optionen `all=TRUE`, `all.x=TRUE` und/oder `all.y=TRUE` erreichen.

2) Da tatsächliche Panel-Datensätze meist deutlich größer als das Beispiel sind (und ihre Verarbeitung entsprechend speicherintensiv ist), ist es ratsam, aus jeder Panel-Welle nur ausgewählte Variablen einzulesen und diese Daten dann zusammenzuführen.

In R bietet es sich an, die Daten einer Welle erst wie gewohnt einzulesen, dann mit `subset` einige Variablen auszuwählen und den Originaldatensatz vor dem Einlesen der nächsten Welle aus dem Speicher zu entfernen.

Das lässt sich mit einer selbstgeschriebenen Funktion teilweise automatisieren:

```
getsubset <- function(year) {
  print("reading data ...")
  dat <- read.csv(paste("D:/.../pequiv", year, ".csv", sep=''),
    header=TRUE)

  print("creating subset ...")
  dats <- subset(dat, select=c(
    "X11101LL",                      # id-variablen
    paste("X11102", year, sep=''),
    "X11104LL",
    paste("D11101", year, sep=''),  # demografisches
    "D11102LL",
    paste("D11103", year, sep=''),
    paste("D11104", year, sep=''),
    paste("E11101", year, sep=''),  # Erwerbsstatus
    paste("E11102", year, sep=''),
    paste("I11110", year, sep=''))  # Einkommen
  )
  rm(dat)
  print("done.")
  return(dats)
}
```

Mit `X <- getsubset(80)` werden also die ausgewählten Variablen der 1980er-Welle eingelesen.

```
X <- getsubset(80)
Y <- getsubset(85)
dim(X); dim(Y)
X <- merge(X,Y)
dim(X)
```

3) Eine mögliche Verwendung von Paneldaten betrifft folgende Fragestellung: Wenn der Anteil 'relativ armer' Haushalte pro Welle mehr oder weniger konstant ist, bleibt noch die Frage, ob es sich über die Zeit stets um die gleichen derart klassifizierten Haushalte handelt.

'Relativ arme' Haushalte lassen sich pro Welle bspw. als diejenigen Haushalte definieren, die über weniger als 60% des durchschnittlichen Haushaltseinkommens der jeweiligen Welle verfügen.

Um eine solche Fragestellung bearbeiten zu können, müssen die vorliegenden Individualdaten zunächst in einen Datensatz auf Haushaltsebene überführt werden. Pro Welle werden die Haushalte dann mit einer Indikatorvariable 'relative Armut' versehen. Diese Operationen lassen sich gut in einer Funktion zusammenfassen:

```
gethhsubset <- function(year) {
  print("reading data ...")
  dat <- read.csv(paste("D:/.../pequiv", year, ".csv", sep=''),
    header=TRUE)

  print("selecting vars ...")
  index <- !duplicated( dat[[ paste("X11102", year, sep='') ]] )
  dats <- subset(dat, index,
    select=c(
      paste("X11102", year, sep=''),  # hh id number
      paste("I11113", year, sep=''),  # hh postgov income
      paste("D11106", year, sep=''),  # no. of persons in hh
      paste("D11107", year, sep='')  # no. of children in hh
    )
  )
  rm(dat, index)
  names(dats)[1] <- "HHID"

  print("creating poverty indicator ...")
  minc <- mean(dats[,2], na.rm=TRUE)
  dats[[ paste("ARM", year, sep='') ]] <- dats[,2] < 0.6 * minc
  print("done.")
  return(dats)
}

dat80 <- gethhsubset(80)
```

Schließlich müssen die Wellen in einem einzigen Datensatz zusammengeführt werden (`merge()`); dann lässt sich z.B. die maximale Zahl aufeinanderfolgender Jahre bestimmen, in denen sich jeder Haushalt im Zustand 'relativ arm' befunden hat (`rle()`).