Prof. Dr. J. Althammer
Dr. U. Pötter
Empirische Gesundheitsforschung

Ruhr-Universität Bochum
Sommersemester 2007

Arbeitsblatt 3

1) Gesundheitsausgaben: Gesundheitsausgaben sowohl für Zuzahlungen wie für private oder zusätzliche Versicherungen wurden im Modul `hc` (Health Care) abgefragt (`hc045xx, ..., hc061xx`). Insbesondere gibt `hc049xx` zusätzliche Zahlungen für Medikamente an. Beachten Sie die Kodierung für Deutschland sowie die Kodierung der fehlenden Angaben. Falls es keine Angabe auf die direkte Frage (`hc049e`) gab, wurden ungefähre Werte abgefragt. Diese Angaben (und die Höhe der abgefragten Grenzen) finden sich in den Variablen `hc049v1, hc049v2, hc049v3`. Die Zusammenfassung dieser Angaben finden Sie in der Variablen `hc049ub`.

Aufgaben: a) Lesen Sie den Datensatz `hc` ein:

```
library(foreign)
### hc
hc <- read.dta("./data/sharerel1_hcD.dta", convert.factors=F)
attach(hc)
```

b) Rekodieren Sie die unvollständigen Angaben in `hc049e` und `hc049ub` (setzen Sie die Werte für „refusal“ und „don’t know“ auf `NA`).

c) Berechnen Sie den Mittelwert der Ausgaben aus der Variablen `hc049e`. Berechnen Sie den Anteil der Personen ohne Zuzahlungen.

d) Berechnen Sie den Mittelwert für die Werte in `hc049e` und denen in `hc049ub`, wenn der Mittelwert der Angaben in `hc049ub` benutzt wird. Berechnen Sie die Mittelwerte für die zusätzlichen Ausgaben, wenn jeweils entweder das Minimum der Angaben in `hc049ub` oder das Maximum benutzt wird. Stellen Sie das Ergebnis in einer Grafik dar (etwa mit Boxplots). Wie verändern sich die Ergebnisse, wenn die Personen ohne Zuzahlungen ausgeschlossen werden?

Hinweis: Die Rekodierung der Werte in `hc049ub` geht am einfachsten mit folgenden Trick:

```
ausgmub <- hc049ub
ausgmub[ausgmub > 7] <- NA
cod <- c(75,150,225,300,450,600,800)
ausgmubm <- cod[ausgmub]
```

2) Generierte Variable: Für einige Fragen wurden aus den Angaben der Befragten neue Variable generiert. Insbesondere wurden Angaben über Einkommen, Vermögen, Gesundheit, Haushaltszusammensetzung und Ausbildung zusammengeführt. Informationen zu den generierten Variablen finden sich in kurzen, zusätzlichen Dokumentationen mit den Namen `*_gv_*.pdf`. Die generierten Angaben über Ausbildungsabschlüsse orientieren sich an der OECD ISCED-1997 (International Standard Classification of Education). Dies erlaubt vergleichende Analysen mit Daten verschiedener Länder. Aus der ISCED wurde zudem eine Variable generiert, die die äquivalente Ausbildungsdauer enthält.

Aufgaben: a) Lesen Sie die beiden Datensätze `dn` (Demographie) und `gv_edu` ein:

```
### dn
dn <- read.dta("./data/sharerel1_dnD.dta", convert.factors=F)
### gv_edu
gve <- read.dta("./data/sharerel1_gv_eduD.dta", convert.factors=F)
### Zusammenfuehren:
dat <- merge(dn, gve, by=c("sampid2","cvid"))
### Aufraeumen
rm(dn,gve)
attach(dat)
```

b) Vergleichen Sie die Angaben in den Variablen `dn010.` und `dn012.` mit den Angaben in `edu` und `yedu`. Beachten Sie die Kodierung fehlender Werte sowie sonstiger (ausländischer) Abschlüsse. Beachten Sie auch die Kodierung der ISCED Kategorien 2A, 2B etc.

c) Matchen Sie zu dem Datensatz `dat` die Angaben aus `cv`, die insbesondere die Angabe zum Geschlecht enthält. Überlegen Sie sich, wie man den Zusammenhang zwischen Ausbildungsdauer (`yedu`) und Geschlecht und Alter graphisch darstellen kann. Betrachten Sie insbesondere Darstellungen der bedingten Verteilung der Ausbildungsdauer gegeben Alter und Geschlecht. Beachten Sie auch die Kodierung fehlender Werte für die Variable Alter.

Hinweis: Benutzen Sie Varianten von `barplot`, `cdplot`, `coplot`, `mosaicplot` oder `spineplot`. Einige dieser Graphik-Befehle können durch *Formeln* definiert werden. Symbolische Formeln dieser Art werden auch zur Spezifikation von Regressionsmodellen benutzt. Sie haben die Form `y ~ x | a`. Dabei ist `y` der Name der abhängigen Variable, `x` der der unabhängigen Variable, und `a` der Name einer Variable, auf deren Werte konditioniert wird.

3) Die generierten Variablen für Einkommen und Vermögen fassen die sehr detaillierten Angaben zu verschiedenen Komponenten von Einkommen und Vermögen zusammen, die in den Modulen `ep`, `hh`, `ho` und `as` erfasst wurden. Da hier besonders

viele Angaben fehlen, wurden jeweils 5 Varianten der Datensätze erzeugt, in denen fehlende Angaben *imputiert* wurden. Eine genaue Beschreibung des Ausmaßes unvollständiger Angaben und des Vorgehens bei der Imputation finden sich in Kapitel 10-12 von `http://www.share-project.org/new_sites/Documentation/TheSurvey.pdf`. Dort findet sich auch die Liste der Variablennamen und deren Kodierung.

Aufgaben: a) Geben Sie einen kurzen Überblick über das verwandte Imputationsverfahren für Einkommen (Kap. 10 der obigen Dokumentation).

b) Lesen Sie den ersten und zweiten Datensatz der generierten Einkommensvariablen ein und matchen Sie ihn zu den Angaben aus `cv`, etwa:

```
library(foreign)
cv <- read.dta("./data/sharerel1_cv_rD.dta", convert.factors=F)
inc1 <- read.dta("./data/sharerel1_gv_inc1D.dta", convert.factors=F)
inc2 <- read.dta("./data/sharerel1_gv_inc2D.dta", convert.factors=F)

dat <- merge(cv,inc1)
dat <- merge(dat,inc2,by=c("sampid2","cvid"))
### Beachten Sie die Namensgebung!
### Die Angaben aus inc1 erhalten das Suffix ".x", die aus inc2
### das Suffix ".y"
names(dat)
### Aufraeumen
rm(inc1,inc2,cv)

attach(dat)
```

c) Wieviele fehlende Werte gibt es für das jährliche persönliche Bruttoeinkommen in beiden Versionen (`yre.x` und `yre.y`)? Wieviele Einkommen liegen über 100.000,- Euro? Was sind die jeweiligen Maxima und die entsprechenden durchschnittlichen Einkommen?

d) Schätzen Sie die Dichte der persönlichen Bruttoeinkommen (`yre`) aus den beiden imputierten Datensätzen und zeichnen Sie sie in ein Fenster, etwa:

```
plot(density(yre.x), main="Pers\"onliches Bruttoeinkommen",
     xlab="Einkommen", xlim=c(0,70000), lwd=1.6, bty="l",
     col="red")
lines(density(yre.y), lwd=1.6, xlim=c(0,70000), col="blue")
legend(40000,0.00002,c("1. Imputation", "2. Imputation"),
       col=c("red","blue"), lty=c(1,1))
```

e) Benutzen Sie die Variablen `irf.x` und `irf.y`, um den Umfang der imputierten Angaben abzuschätzen. Sind die Angaben in den beiden Versionen gleich?

f) Vergleichen Sie die Werte direkt, indem Sie eine neue Variable aus der Differenz der Angaben in `yre.x` und `yre.y` bilden. Beschreiben Sie das Ergebnis mit einem Dichteschätzer. Würden Sie die einzelnen imputierten Angaben für zuverlässig halten? Beschreiben Sie die Variabilität der imputierten Werte für die Personen. Wie würden Sie die Angaben kombinieren?

4) Einige zusammenfassende Gesundheitsindizes werden in `sharerel1_gv_healthD.dta` bereitgestellt: `eurod` ist ein Depressionsindex, `spheu` fasst subjektive Gesundheitseinschätzung zusammen, `chronic` zählt die Anzahl angegebener chronischer Erkrankungen, `symptoms` zählt Symptome, `bmi` gibt den Body-Mass Index an, `adl` ist ein Index für Einschränkungen bei täglichen Aktivitäten etc.

Aufgaben: a) Lesen Sie den Datensatz ein. Matchen Sie den Datensatz mit dem Datensatz `sharerel1_phD.dta`. Die Variable `spheu` ist eine Rekodierung der Angaben in `ph002.` und `ph053.`. Beachten Sie die Kodierung fehlender Werte und versuchen Sie, die Konstruktion von `spheu` nachzuvollziehen.

```
library(foreign)
health <- read.dta("./data/sharerel1_gv_healthD.dta",
                   convert.factors=F)
ph <- read.dta("./data/sharerel1_phD.dta", convert.factors=F)
dat <- merge(health,ph)
attach(dat)
rm(health,ph)
```

b) Stellen Sie den Zusammenhang zwischen `spheu` sowie Alter und Geschlecht graphisch dar. Sie müssen dazu den Datensatz `cv` zu `dat` matchen:

```
cv <- read.dta("./data/sharerel1_cv_rD.dta", convert.factors=F)
detach(dat)
dat <- merge(dat,cv)
attach(dat)
```

Beachten Sie die Kodierung fehlender Werte in `yrbirth` und `gender`. Benutzen Sie Varianten von `barplot`, `cdplot`, `coplot`, `mosaicplot` oder `spineplot` für die graphische Darstellung.

c) Wieviele fehlende Werte gibt es bei der Variablen `bmi`? Was sind Minima und Maxima? Stellen Sie die Verteilung des Body-Mass Indexes graphisch dar. Benutzen Sie einen Dichte-Schätzer. Welchen Einfluss haben besonders große oder kleine Werte? Benutzen Sie einen Boxplot, um die Unterschiede zwischen den Geschlechtern darzustellen, etwa:

```
plot(density(bmi,na.rm=T), main="Verteilung des BMI",
     xlab="BMI", ylab="Dichte")
boxplot(bmi ~ gender, notch=T, col="lightblue",
        names=c("M\"anner","Frauen"), main="Verteilung des BMI")
```