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

Ruhr-Universität Bochum
Sommersemester 2007

Arbeitsblatt 4

Einige lineare Modelle

1) Gesundheitsausgaben: Die Variable `hc049xx` erfasst zusätzliche Zahlungen für Medikamente. 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` (vgl. Arbeitsblatt 3).

Lineare Modelle werden in `R` durch die Funktion `lm` berechnet. Ihr einziges erforderliches Argument ist eine Formel der Form `y ~ x1 + x2 + x3`. Die linke Seite gibt den Namen der abhängigen Variablen an. Die durch `+` getrennten Terme sind die Namen der unabhängigen Variablen. Die Funktion `lm(Mausgaben ~ ALTER)` würde also eine lineare Regression von Zuzahlungen auf das Alter berechnen. Als Daten werden dabei ohne weitere Angaben die zuletzt definierten Variablen benutzt. Um Unklarheiten zu beseitigen, sollte daher immer auch angegeben werden, auf welche Daten man sich gerade bezieht. Das geschieht durch die Form `lm(Mausgaben ~ ALTER + SEX, data=dat)`, wobei das Argument von `data=` ein `data frame` sein muss.

Aufgaben:

a) Definieren Sie nach dem Einlesen der Datensätze `dn`, `cv` und `hc` entsprechende Variable `Mausgaben`, `ALTER`, und `SEX`. Beachten Sie die Kodierungen und fehlende Werte. Berechnen Sie dann ein lineares Modell `lm(Mausgaben ~ ALTER + SEX, data=dat)`.

b) Das Ergebnis eines Aufrufs von `lm()` ist eine *Liste* von Ergebnissen. Eine Zusammenfassung der Ergebnisse liefert die `summary()` Funktion.

```
l <- lm(Mausgaben ~ ALTER + SEX, data=dat)
summary(l)
```

liefert also einen ersten Überblick über die Ergebnisse der linearen Regression. Elemente der Ergebnisliste können durch ihre Namen angesprochen werden. So liefert `l$coefficients` nur den Koeffizientenvektor der Regression. Zudem können viele Teilergebnisse auch durch spezielle Funktionen extrahiert werden: `coef(l)` ist äquivalent zu `l$coefficients`. `vcov()` liefert die Varianz-Kovarianzmatrix der geschätzten Koeffizienten. Weitere solche Funktionen werden später erwähnt.

Sind die Koeffizienten des Modells `l` auf dem 5% Niveau signifikant?

c) Berechnen Sie eine lineare Regression von `Mausgaben` auf `ALTER, SEX` nur für diejenigen, die tatsächlich Zuzahlungen geleistet haben. Nennen Sie das Ergebnis `l2`. Verändern sich die Koeffizienten von `Alter` bzw. `SEX` im Vergleich zum ersten Modell?

d) Benutzen Sie die Variable `dn010.` (höchster Schulabschluss), um eine Variable `SCHULE` zu definieren. Berechnen Sie eine lineare Regression von `Mausgaben` auf `ALTER, SEX, SCHULE`. Nennen Sie das Ergebnis `l3`. Vergleichen Sie die Ergebnisse mit denen aus Aufgabenteil b).

e) Interaktionen von Variablen können in dem Formel Argument von `lm()` etwa durch `lm(Mausgaben ~ SEX+ALTER*SCHULE)` angegeben werden. Dies ergibt die Haupteffekte von `SEX, ALTER, SCHULE` sowie die Interaktion von `ALTER` mit `SCHULE`. Berechnen Sie diese Regression. Nennen Sie das Ergebnis `l4`. Welche Koeffizienten sind auf dem 5% Niveau signifikant? Sollte man die nicht signifikanten Variablen aus der Regression ausschließen?

2) Diagnostik: Diagnostiken der Regressionsmodelle lassen sich durch den Aufruf entsprechender Funktionen berechnen, deren Argument die Ergebnisliste des Regressionsmodells ist. Im folgenden soll das Modell mit den Ergebnissen `l3` benutzt werden.

a) Probieren Sie den Befehl `plot(l3)`. Für lineare Modelle werden vier Plots erzeugt. Welche Informationen erhält man? Ergeben sich Anzeichen für Ausreißer?

b) Die Hutmatrix kann durch `hatvalues(l3)` berechnet werden. Plotten Sie die Werte der Hutmatrix. Vergleichen Sie die Werte mit 3*Anzahl Kovariable/Anzahl Beobachtungen. Gibt es Hinweise auf Ausreißer in den Kovariablen?

c) Berechnen Sie studentisierte Residuen (`rstudent(l3)`) und plotten Sie sie. Gibt es Hinweise auf Ausreißer?

d) Identifizieren Sie mindestens einen Ausreißer.

e) Plotten Sie die studentisierten Residuen gegen die vorhergesagten Werte. Sie erhalten die vorhergesagten Werte durch `fitted(l3)`. Gibt es Hinweise auf Heteroskedastie in Richtung der vorhergesagten Werte?

f) Gibt es Hinweise auf Heteroskedastie in Abhängigkeit von `ALTER`?

g) Die empirische Einflussfunktion `DFBETA`, also $\hat{\beta}_{-i} - \hat{\beta}$, wird durch `dfbeta(l3)` berechnet. Welche Dimension hat das Ergebnis? Zeigen die Einzelergebnisse der Einflussfunktion Hinweise auf Ausreißer? Welche Variable ist besonders betroffen?

h) Wiederholen Sie die Diagnostik für das Modell `l3`, wenn nur Beobachtungen mit tatsächlichen Zuzahlungen betrachtet werden und nur Zuzahlungen unter 1000 Euro berücksichtigt werden. Gibt es immer noch Anzeichen für Ausreißer? Für Heteroskedastie?