Thilo Ernst
Ulrich Pötter
Rechentechniken der Datenanalyse

Ruhr-Universität Bochum
Sommersemester 2006

Arbeitsblatt 7

1) Die Daten vom letzten Arbeitsblatt sollen wieder eingelesen werden. Als Variablen wurden betrachtet: individuelles (Arbeits-) Einkommen (`I1111099`), Alter (`D1110199`), Geschlecht (`D11102LL`), Schuljahre (`D1110999`) und Arbeitszeit (Stunden je Jahr) (`E1110199`) der 1999 Welle der PSID-Studie. Es sollen nur Beschäftigte `E1110299==1` der Hauptstichprobe (`X11104LL==11`) betrachtet werden, die auch zur Stichprobe gerechnet werden (`X1110399==1`).

Allerdings werden extreme Fälle ausgeschlossen: Einkommen über 150000 $, Arbeitsstunden über 4000/Jahr, Stundenlöhne kleiner als 1 $ und größer als 100 $:

```
X <- read.csv("../../cnef/psid/pequiv99.csv",head=T)

X <- subset(X, E1110299==1 & X11104LL==11 & X1110399==1 &
            E1110199 < 4000 & I1111099 < 150000 &
            I1111099/E1110199 < 100 & I1111099/E1110199 > 1,
            c(X11101LL, X1110299, D11102LL, D1110199,
              D1110999, E1110199, I1111099)
            )

X <- na.omit(X)
attach(X)
```

Außerdem soll wieder die lineare Regression von log(Stundenlohn) auf Alter, Geschlecht und Schuljahre betrachtet werden.

```
erg <- lm(I(log(I1111099/E1110199)) ~ D1110199 +
            D11102LL + D1110999 + E1110199)
summary(erg)
```

2) Nicht-lineare Effekte: Nicht-lineare Effekte von Kovariablen lassen sich am einfachsten und effektivsten durch Splines darstellen. Splines werden in der library `splines` definiert. Man kann dabei vorgeben, wie glatt die Funktion der Kovariablen sein soll. Am Beispiel des Alters etwa:

```
library(splines)
erg2 <- lm(I(log(I1111099/E1110199)) ~ ns(D1110199,df=3) +
            D11102LL + D1110999)
summary(erg2)
```

Die Koeffizienten der Spline-Terme lassen sich nicht direkt interpretieren. Man kann sich aber die Effekte malen lassen:

```
termplot(erg2,terms=1,rug=T,se=T)
```

Für Vorhersagen muss man nun die verwendete Spline-Basis berücksichtigen:

```
datneu <- data.frame(D1110199=18:70,D11102LL=1,D1110999=10)
plot(18:70,predict(erg2,newdata=datneu),type="l")
```

3) Additive Modelle: Man spricht von additiven Modellen, wenn Modelle der Form

$$\mathbb{E}(Y \mid X_1 = x_1, \ldots X_p = x_p) = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$$

angepasst werden. Die Funktionen sollen möglichst glatt sein. Wieder bieten sich Spline-Funktionen als Grundlage für die Wahl der Funktionen f_j an. Allerdings sollte nun der Grad der Glattheit möglichst günstig gewählt werden. Das Paket `mgcv` erlaubt (nicht nur) dies:

```
library(mgcv)
erg3 <- gam(I(log(I1111099/E1110199)) ~ s(D1110199) +
            D11102LL + D1110999)
summary(erg3)
```

Die `summary`-Methode für diese Modelle gibt nicht einmal mehr die Koeffizienten der Spline-Basis explizit an. Man braucht wieder eine Möglichkeit, den geschätzten Effekt zu plotten:

```
plot.gam(erg3,select=1,rug=T,se=T,jit=T)
```