Angewandte Regression — Serie 11

1. Wir fahren mit der Aufgabe 1 aus Serie 10 fort. Wir betrachten das folgende kumulative Logit-Modell (*proportional-odds model*, vgl. Skript, Kap. 14.2.h):

$$\text{logit}\langle\gamma_k\langle\underline{x}_i\rangle\rangle = \beta_1\,\texttt{SES}_i + \beta_2\,\texttt{LE}_i - \alpha_k$$

Die Daten sind im Data Frame `mental.dat` gegeben, wobei jede Zeile einer Beobachtung entspricht.

a) Passen Sie das lineare Modell mit `regr()` an. **R-Hinweise:**

```
> d.mental$Y <- ordered(d.mental$Y, levels=c("Well",...))
> library(MASS)
> r.mental <- regr(Y ~ SES+LE, family="ordered", data=d.mental)
```

b) Machen Sie eine Residuenanalyse mit `plot()`.

Quelle: A. Agresti, *Categorical Data Analysis*, Wiley, 1990, p.325.

2. In einem Experiment wurde die Reissfestigkeit von Kevlar-49-Fasern gemessen. Aus 8 verschiedenen Rollen (`Spool`) und unter vier verschiedenen Zugskräften (`Stress` in MPa) wurde die Zeit (`Failure`) in Stunden bis zum Reissen der Fasern gestoppt. Der Datensatz `kevlar49.dat` beinhaltet das Ergebnis des Experiments.

Quelle: M.J. Crowder, A.C. Kimber, R.L. Smith, T.J. Sweeting, *Statistical Analysis of Reliability Data*, Chapman&Hall, p90.

Lesen Sie die Daten ein und schränken Sie die Daten auf `Stress`≥ 24 ein, da einige Messdaten für `Stress`< 24 zensuriert sind und wir die Regressionstechniken für zensurierte Daten noch nicht kennen. Achtung, `Spool` muss ein Faktor sein.

a) Passen Sie ein lineares Modell

$$\log(\texttt{Failure}) = \beta_o + \beta_1\texttt{Stress} + \beta_2\texttt{Spool} + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

an und machen Sie eine Residuenanalyse.

In den folgenden Teilaufgaben möchten wir die Weilbull-Regression machen. Dazu ein kleiner Zusatz. Wir haben in der Vorlesung gelernt, dass die Dichte der Weilbull-Funktion gegeben ist durch

$$f(x) = \frac{\alpha}{\sigma}(x/\sigma)^2 \exp(-(x/\sigma)^\alpha).$$

Die Weilbull-Verteilung hat somit die Form (wenn man die obige Dichte integriert)

$$F(x) = 1 - \exp(-(x/\sigma)^\alpha) = 1 - S(x).$$

Die $S(x) = \exp(-(x/\sigma)^\alpha)$ nennt man die Survival-Funktion. Sie beschreibt die Überlebenswahrscheinlichkeit als Funktion der Zeit x. Logarithmiert man $S(x)$, so erhalten wir

$$\log(S(x)) = -(x/\sigma)^\alpha.$$

Bringt man das Vorzeichen auf die Seite von $S(x)$ und logarithmiert man nochmals, so erhalten wir

$$\log(-\log(S(x))) = \alpha \log(x) - \alpha \log(\sigma).$$

Dh, wenn unsere Daten Weibull verteilt sind, so muss die empirische Survival-Funktion $\hat{S}(x)$, bestimmt durch

$$\hat{S}(x) = \frac{\text{Anzahl Beobachtungen} \geq x}{\text{Anzahl Beobachtungen}},$$

nach der Transformation $\log(-\log(\hat{S}(x)))$ im groben linear in $\log(x)$ sein.

b) Bestimmen Sie jeweils für die Gruppen mit gleichem `Stress`-Wert die empirische Survival-Funktion $\hat{S}(x)$ zu jedem $x =$`Failure` und machen Sie einen Plot von $(\log(x), \log(-\log(\hat{S}(x))))$. Kommentieren Sie den Output. Hinweis: Im Datensatz `kevlar49.dat` sind die `Failure` bezüglich `Stress` schon geordnet. `split()` und `cumsum()` können hilfreich sein.

c) Für die Weibull-Regression

$$\log(\sigma_i) = \underline{x}_i^T \underline{\beta}$$

müssen Sie das Package survival laden und die Funktion `survreg()` benützen.

```
library(survival)
rr <- survreg(Surv(Failure,rep(1,nrow(dd)))~Stress+Spool, data=dd)
```

d) Wie hängen diese Ergebnisse mit Teilaufgabe a) zusammen?

e) Machen Sie eine Residuenanalyse. Die Residuen erhalten Sie mit `residuals(...)[,"median"]`. Tragen Sie diese gegen die Werte des linearen Prädiktors (`rr$linear.predictors`) und gegen die Eingangsvariablen auf. Hinweis: Benützen (`scatter.smooth()`) für den Plot.

3. Wir betrachten den Datensatz

```
d.ertrag <- c(35.6, 34.9, 36.0, 30.2, 36.2, 35.6, 35.8, 35.9, 36.1)
```

a) Berechnen Sie für den Datensatz den M-Schätzer mit Hubers ψ-Funktion (c=1.345) und das asymptotische 95%-Konfidenzintervall. Verwenden Sie dazu die R-Funktion `rlm(...)` aus dem Package `MASS`. Vergleichen Sie dieses Konfidenzintervall mit dem klassischen Konfidenzintervall.

b) Stellen Sie die Daten graphisch dar und zeichnen Sie im gleichen Plot die geschätzten Lageparameter ein.