Angewandte Regression — Serie 10

1. Wir betrachten in dieser Aufgabe Daten einer Untersuchung über Geistesschwäche bei 40 Testpersonen.
Die Zielgrösse `Y` bezeichnet den Grad der Geistesschwäche (`Well`, `Mild`, `Moderate` oder `Impaired`). `Well` ist die niedrigste Kategorie ($k = 0$), `Impaired` die höchste ($k = 3$) .
Als erklärende Faktoren wurden der sozialökonomische Status `SES` und ein sogenannter "life events index" `LE` verwendet. Die Variable `SES` besitzt zwei Stufen: 1 für einen hohen, 0 für einen tiefen sozialökonomischen Status. Der "life events index" ist ein Mass, das die Anzahl und Stärke von wichtigen Lebensereignissen (Geburt, neue Arbeitsstelle, Scheidung, Todesfall, ...) in den letzten 3 Jahren beschreibt.

Wir betrachten das folgende kumulative Logit-Modell (*proportional-odds model*, vgl. Skript, Kap. 14.2.h:

$$\text{logit}\langle\gamma_k\langle\underline{x}_i\rangle\rangle = \beta_1\,\texttt{SES}_i + \beta_2\,\texttt{LE}_i - \alpha_k \tag{1}$$

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

a) Schauen Sie sich die Daten zuerst grafisch an. Welche ersten Schlüsse ziehen Sie daraus, erwarten Sie positive oder negative Koeffizienten β_1 bzw. β_2?
R-Hinweise:
> `plot.design(d.mental[,1:2])`
> `plot.design(d.mental[,-2])`

b) Passen Sie nun das Modell (1) an. Wie lauten die geschätzten Koeffizienten und Schwellenwerte? Bestätigen die Resultate die Eindrücke aus Teilaufgabe a)?
R-Hinweise:
Zuerst muss die Zielvariable als geordneter Faktor gespeichert werden:
> `d.mental$Y <- ordered(d.mental$Y, levels=c("Well",...))`
Verwenden Sie dann die in den R-Hinweisen beschriebene Funktion `polr()` aus dem Package `MASS`:
> `library(MASS)`
> `r.mental <- polr(Y ~ SES+LE, data=d.mental)`

c) Um welchen Faktor unterscheiden sich die Wettverhältnisse (*odds*) für `SES` = 1 und `SES` = 0 bei gegebenem "life events score" `LE` und bei festem Level k?

d) Berechnen und zeichnen Sie die Wahrscheinlichkeiten $P\langle Y \geq k \mid \texttt{LE}, \texttt{SES}\rangle$ für die Stufen $k = 0, 1, 2, 3$ und Werte $\texttt{SES} = 0, 1$ und $\texttt{LE} = 0, 1, \ldots, 9$.
Tun Sie dasselbe auch für die einzelnen Klassen, $P\langle Y = k \mid \texttt{LE}, \texttt{SES}\rangle$ für die Stufen $k = 0, 1, 2, 3$ und Werte $\texttt{SES} = 0, 1$ und $\texttt{LE} = 0, 1, \ldots, 9$.
R-Hinweise:
Damit Sie nicht zuviel Zeit fürs Programmieren brauchen, finden Sie den ganzen Code, den Sie für die Berechnung und Zeichnung des Verlangten benötigen im Netz unter `ftp://.../R/polr.plots.R`. Kopieren Sie den Code in Ihr File. Versuchen Sie in erster Linie die Bilder zu verstehen und zu interpretieren.

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

2. Wir betrachten den Datensatz einer Umweltumfrage. Es wurde erhoben, wer die Verantwortung für die Umwelt tragen soll und ob die Leute sich durch die Umweltverschmutzung beeinträchtigt fühlen. Weiter wurden Variable wie Bildungsniveau, Geschlecht, Parteizugehörigkeit etc. aufgenommen. Hier soll der Zusammenhang zwischen der Zielgrösse Hauptverantwortung (`Hauptv`) und geeigneten anderen Variablen untersucht werden.
Laden Sie den Datensatz mit `source("ftp://stat.ethz.ch/NDK/Source-WBL-2/R/umwelt.R")`. Er lässt sich dann mit `t.d` ansprechen.

a) Schätzen Sie die Koeffizienten des multinomialen Modells mit den erklärenden Variablen `Schule` und `Beeintr`. Wie ist der Koeffizient mit der Bezeichnung `SchuleAbitur` für `Hauptv = Staat` zu interpretieren? Hinweis: log(odds(...)).
R-Hinweis:
```
library(nnet); t.rm <- multinom(Hauptv ~ Schule + Beeintr, data=t.d)
```

b) Bestimmen Sie die odds ratios für die Zuweisung der Verantwortung an die Einzelnen und Studierte, je gegen die übrigen Kategorien.
R-Hinweis: Verwenden Sie `predict(..., type="probs")`; genauer:
```
## data.frame mit allen Kombinationen der Levels von Schule und Beeintr
t.dt <- data.frame(table(t.d$Schule,t.d$Beeintr))
names(t.dt) <- c("Schule", "Beeintr", "Freq")
t.p <- predict(t.rm, newdata=t.dt, type="probs"); cbind(t.dt[,1:2], t.p)
# für Beeintr = nicht
log(t.p[5,1] / sum(t.p[5,2:3]))-log(sum(t.p[1:4,1]) / sum(t.p[1:4,2:3])) ...
```

c) Bestimmen Sie ein geeignetes Modell, ausgehend von
`Hauptv`∼`Alter+Geschlecht+Schule+Wohnlage+Ortsgroesse +Partei+Beeintr`.
Hinweis: Verwenden Sie `drop1()` und `anova()` (für Modellvergleiche), `step()` funktioniert hier leider nicht.

3. (fakultativ) Für die Kontrolle einer grossen Unternehmung wurden getätigte Transaktionen von zwei bestimmten Typen untersucht. Man ist an der aufgewendeten Zeit (resp. den Kosten) der Transaktionen interessiert. Das Data Frame `transaction.dat` enthält Daten von 261 Zweigstellen:

`Time`	Total-Zeit (in Minuten), die für die Transaktionen gebraucht wurde
`Type1`	Anzahl Transaktionen vom Typ 1
`Type2`	Anzahl Transaktionen vom Typ 2

a) Lesen Sie die Daten ein mit `read.table("http://.../transaction.dat", header=T)` und schauen Sie diese zuerst an, z.B. mit `pairs()`.

b) Passen Sie ein (gewöhnliches) lineares Modell mit normalverteilten Fehlern an. Überprüfen Sie die Residuen. Weshalb passt dieses Modell nicht?

c) Passen Sie nun mit `glm()` ein verallgemeinertes lineares Modell mit der Gamma-Verteilung an. Formulieren Sie das verallgemeinerte Modell. Welche Link-Funktion ist hier angebracht? Probieren Sie den kanonischen Link und den Identitätslink (`link=identity`) aus und begründen Sie, warum der zweite vorzuziehen ist.
Versuchen Sie dasselbe auch mit `regr()` zu rechnen.
R-Hinweise:
```
> glm(Time ~ Type1 + Type2, family=Gamma(link=...), data=d.trans)
```
Oder:
```
> regr(Time ~ Type1 + Type2, family=Gamma(link=...), data=d.trans)
```