

Verallg. lineare Modelle (Rg-2a) — Serie 3 (Geordnete diskrete Zielgrößen)

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, ...), die in den letzten 3 Jahren stattfanden, beschreibt.

Wir betrachten das folgende kumulative Logit-Modell (*proportional-odds model*, vgl. Skript, Kap. 3.1.e):

$$\text{logit}\langle\gamma_k(\underline{x}_i)\rangle = \beta_1 \text{SES}_i + \beta_2 \text{LE}_i - \alpha_k \quad (1)$$

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 Wett-Verhältnisse (*odds*) für $\text{SES} = 1$ und $\text{SES} = 0$ bei gegebenem “life events score” **LE** und bei festem Level k ?
- d) Berechnen und zeichnen Sie die Wahrscheinlichkeiten $P\langle Y \geq k \mid \text{LE}, \text{SES} \rangle$ für die Stufen $k = 0, 1, 2, 3$ und Werte $\text{SES} = 0, 1$ und $\text{LE} = 0, 1, \dots, 9$.
Tun Sie dasselbe auch für die einzelnen Klassen, $P\langle Y = k \mid \text{LE}, \text{SES} \rangle$ für die Stufen $k = 0, 1, 2, 3$ und Werte $\text{SES} = 0, 1$ und $\text{LE} = 0, 1, \dots, 9$.

R-Hinweise:

Damit Sie nicht zuviel Zeit für's 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>. Copy-pasten 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. a) Wir betrachten in der folgenden Aufgabe Daten über eine Lungenkrankheit (Pneumocoinosis) bei Arbeitern in Kohlebergwerken. Insgesamt 371 Personen wurden je nach Grad der Krankheit einer von drei Kategorien “normal”, “leicht” und “schwer” ($k = 0, 1, 2$) zugeordnet. Zudem wurde die Dauer der Tätigkeit im Kohlebergbau erfasst:

Zeit t_i (Jahre)	Kat. $Y_i^{(0)}$ (normal)	Kat. $Y_i^{(1)}$ (leicht)	Kat. $Y_i^{(2)}$ (schwer)
5.8	98	0	0
15.0	51	2	1
21.5	34	6	3
27.5	35	5	8
33.5	32	10	9
39.5	23	7	8
46.0	12	6	10
51.5	4	2	5

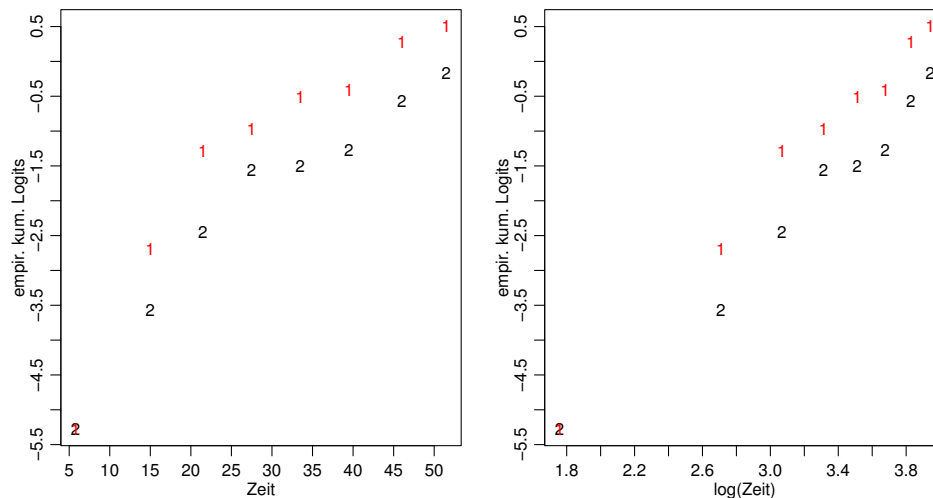
In der folgenden Abbildung sind die empirischen kumulativen Logits

$$\log\left(\frac{Y_i^{(2)}+0.5}{m_i-Y_i^{(2)}+0.5}\right), \text{ bezeichnet mit } \mathbf{2}$$

und

$$\log\left(\frac{Y_i^{(2)}+Y_i^{(1)}+0.5}{m_i-Y_i^{(2)}-Y_i^{(1)}+0.5}\right), \text{ bezeichnet mit } \mathbf{1}$$

gegen die Zeit t_i in (a) und gegen $\log(t_i)$ in (b) aufgetragen. Welches Modell scheint Ihnen aufgrund der Grafik sinnvoller? Begründen Sie Ihre Antwort.



- b) Schätzen Sie das Modell mit R und kommentieren Sie das Resultat.

R-Hinweise:

Die Daten sind im Dataframe `coalmin.dat` gegeben. Die Variablen heißen `Category` (Kategorien 1, 2 und 3 (entsprechen 0, 1, 2 in unserer Notation)) und `Period` (Zeit in Jahren). Die Zielvariable muss wieder zuerst in einen geordneten Faktor umgewandelt werden (vgl. Aufgabe 1b). Benützen Sie dann wieder die Funktion `polr()`.

- c) Berechnen Sie die Wahrscheinlichkeiten für einen Arbeiter, nach 10 resp. 30 Jahren schwer an Pneumoconiosis zu erkranken.

Quelle; P. McCullagh and J.A. Nelder, *Generalized Linear Models*, 1989, p. 178.