4 Mehrkategoriale Regressionsmodelle

Aufgabe 4

Der Datensatz `knie` (vgl. Tutz, 2000; Download von der Veranstaltungshomepage) entstammt einer klinischen Studie, die die Wirkung eines Sprays auf die Verringerung von Druckschmerzen im Knie nach zehntägiger Behandlung von Sportverletzungen untersucht. Folgende Variablen stehen zur Verfügung:

`TH`	Therapie (Behandlung=1/Placebo=0)
`AGE`	Alter (metrisch)
`GEN`	Geschlecht (männlich=1/weiblich=0)
`PAIN`	Schmerzen unter normierter Belastung (ordinal; geringe Schmerzen=1, ... , starke Schmerzen=4)

(a) Lesen Sie den Datensatz `knie` in R ein. Modifizieren Sie die kategorialen Variablen so, dass diese von R mit dem korrekten Skalenniveau identifiziert werden, und zentrieren sie die metrische Variable `AGE` um den Wert 30.

(b) Geben Sie allgemein die Definition des kumulativen Logit-Modells für eine ordinale Zielgröße $Y \in \{1, \dots, k\}$ an. Weshalb wird dieses Modell auch *Proportional-Odds-Modell* genannt?

(c) Motivieren Sie Ihr Modell aus (b) über die latente (metrische) Variable U, für die $U = -\mathbf{x}'\boldsymbol{\gamma} + \epsilon$ angenommen werde (bei gegebenem Prädiktor-Vektor $\mathbf{x}$). Außerdem gelte $E(\epsilon) = 0$, sowie

$$Y = r \;\Leftrightarrow\; \theta_{r-1} < U \leq \theta_r,$$

mit Schwellenwerten $-\infty = \theta_0 < \theta_1 < \cdots < \theta_{k-1} < \theta_k = \infty$. Die Zufallsgröße ϵ habe allgemein die Verteilungsfunktion F. Welche Funktion muss für F angenommen werden, damit sich Ihr Modell aus (b) ergibt?

(d) Machen Sie sich mit der Funktion `polr()` aus dem Package `MASS` zum Fitten eines kumulativen Logit-Modells vertraut. Was ist hier anders im Vergleich zu der aus der Vorlesung (bzw. (b)) bekannten Definition des Modells? Fitten Sie nun ein kumulatives Logit-Modell mit der Variable `PAIN` als Response und den Variablen 'Therapie', 'Alter' und 'Geschlecht' als Kovariablen (Haupteffekte).

(e) Geben Sie für das Modell aus (d) die konkrete Modellformel an und interpretieren Sie die einzelnen Parameter.

(f) Nehmen Sie in das Modell aus (d) nun die Variable `AGE` noch quadratisch mit auf. Was fällt dabei auf? Welches Modell würden Sie bevorzugen? Kann man dieses Modell reduzieren?

(g) Prognostizieren Sie für das in (f) gewählte Modell für eine 22-jährige Frau die Verteilung der Stärke des Schmerzes, sowohl für den Fall, dass das Spray eingesetzt wurde, als auch für das Placebo.

Aufgabe 5

Betrachten Sie weiterhin den Datensatz `knie` aus Aufgabe 4. Sehr flexible und vielfältige multivariate Regressionsmodelle lassen sich in `R` mit dem Paket `VGAM` (Yee, 2010) fitten.

(a) Installieren Sie das Package `VGAM` und verschaffen Sie sich mit Hilfe der Dokumentation einen ersten Überblick.

(b) Fitten Sie Ihr Modell aus Aufgabe 4. Verwenden Sie hierzu die Funktion `vglm()` mit geeigneter `family`.

(c) Probieren Sie noch andere Link-Funktionen als den Logit-Link, z.B. das *Proportional-Hazards-Modell*. Letzteres ist über

$$P(Y \leq r|\mathbf{x}) = 1 - \exp(-\exp(\theta_r + \mathbf{x}'\boldsymbol{\gamma})), \quad r = 1, \ldots, k-1,$$

definiert.

(d) Gehen Sie bei der Modellierung der Knie-Daten nun zu einem sequentiellen Modell über.
Hinweis: Wählen Sie hier als `family 'sratio'` *oder* `'cratio'`.

(e) Lässt sich Ihr (kumulatives oder sequentielles) Modell (für den vorliegenden Datensatz) noch flexibler gestalten?