3 Generalisierte lineare Modelle (I)

Aufgabe 1

Der Datensatz `fakesoep` (Download von der Veranstaltungshomepage) ist dem sozioökonomischen Panel nachempfunden und soll nun mit Hilfe von R analysiert werden. Wir betrachten folgende Variablen (erhoben an $n = 3000$ befragten Personen):

`beink`	Bruttoverdienst im letzten Monat
`groesse`	Körpergröße
`alter`	Alter
`dauer`	Dauer der Betriebszugehörigkeit
`verh`	verheiratet (1: ja, 0: nein)
`geschl`	Geschlecht (1: weiblich, 0: männlich)
`deutsch`	Deutsche Staatsangehörigkeit (1: ja, 0: nein)
`abitur`	Abitur (1: ja, 0: nein)

(a) Untersuchen Sie die Variable `beink` hinsichtlich der Gestalt ihrer Verteilung. Was fällt Ihnen dabei auf?

(b) Im folgenden soll die Variable `beink` als Responsevariable betrachtet werden. Begründen Sie, weshalb die Gammaverteilung als Verteilungsannahme zur Modellierung dieser Responsevariablen geeignet sein könnte. Zeigen Sie nun, dass diese Verteilung zu einer Exponentialfamilie gehört. Bestimmen Sie die Größen θ, $b(\theta)$, ϕ, Erwartungswert und Varianz, sowie den natürlichen Link.

(c) Zeichnen Sie die Dichte einer gammaverteilten Zufallsgröße y für verschiedene Werte des `shape` Parameters, wobei für den Erwartungswert immer $E(y) = 1$ gelten soll.

(d) Fitten Sie ein GLM mit gammaverteiltem Response, allen Kovariablen (Haupteffekte) und natürlichem Link. Verwenden Sie anschließend den log-Link. Interpretieren Sie die Modelle. Welche Strukturannahme würden Sie hier bevorzugen?

(e) Erstellen Sie einen Plot, in dem die beobachteten und die durch Ihr Modell geschätzten Werte abgebildet werden. Betrachten Sie außerdem das Ergebnis der generischen R-Funktion `plot(glm.object)`, wobei für `glm.object` das von Ihnen soeben erstellte GLM-Objekt einzusetzen ist (und setzen Sie `which = 1:5`).

(f) Führen Sie für beide Modelle aus (c) eine *Backward*-Variablenselektion bezüglich des AIC-Kriteriums durch. Was passiert? Welches Modell würden Sie nun bevorzugen? (*Hinweis:* Verwenden Sie die Funktion `step()`.)

Aufgabe 2

In der vorangegangenen Aufgabe wäre auch ein GLM mit Invers-Gauß-verteiltem Response denkbar gewesen. Eine Invers-Gauß-verteilte Zufallsgröße y hat folgende Dichte:

$$f(y|\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right)$$

(a) Zeigen Sie zunächst, dass diese Verteilung zu einer Exponentialfamilie gehört. Bestimmen Sie die Größen θ, $b(\theta)$, ϕ, Erwartungswert und Varianz, sowie den natürlichen Link.

(b) Versuchen Sie nun, für die SOEP-Daten ein GLM mit natürlichem Link und allen Kovariablen zu fitten (Response: `beink`). Verwenden Sie anschließend den log-Link.