3 Generalisierte lineare Modelle

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.

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

(d) Zeichnen Sie die Dichte einer gammaverteilten Zufallsgröße y für verschiedene Werte des `shape` Parameters, wobei für den Erwartungswert immer $E(y) = 1$ gelten soll.
Hinweis: Nutzen Sie die Funktion `manipulate()` aus dem Paket `manipulate`.

(e) 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?

(f) Erstellen Sie zur Diagnose Ihres GLMs mit log-Link

- einen Plot, in dem die beobachteten Werte (y_i) gegen die durch das Modell geschätzten Werte ($\hat{y}_i$) abgetragen werden.
- einen Plot, in dem die durch das Modell geschätzten linearen Prädiktoren ($\hat{\eta}_i$) gegen die Residuen des Modells (r_i) abgetragen werden.

Beurteilen Sie beide Darstellungen jeweils kurz.

Aufgabe 2

In Aufgabe 1 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) Zeichnen Sie mithilfe der Funktion `dinvgauss()` aus dem Paket `statmod` die Dichte einer Invers-Gauß-verteilten Zufallsgröße y für verschiedene Werte von λ, wobei für den Erwartungswert immer $E(y) = 1$ gelten soll.
Hinweis: Nutzen Sie die Funktion `manipulate()` aus dem Paket `manipulate`.

(c) Versuchen Sie nun, für den Datensatz `fakesoep` ein GLM für die Responsevariable `beink` mit natürlichem Link und allen Kovariablen zu fitten. Verwenden Sie anschließend den log-Link.