3 Generalisierte lineare Modelle (V)

Aufgabe 1

Der Datensatz `foodstamp` (Künsch, Stefanski & Carroll, 1989, *JASA*; Download von der Veranstaltungshomepage) enthält spaltenweise in folgender Reihenfolge die Variablen

`y`	Teilnahme am US-Essensmarkenprogramm (ja=1/nein=0)
`TEN`	Mietverhältnis (ja=1/nein=0)
`SUP`	Ergänzungseinkommen (ja=1/nein=0)
`INC`	Monatseinkommen

(a) Schätzen Sie das logistische Regressionsmodell

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \beta_1 \text{TEN} + \beta_2 \text{SUP} + \beta_3 \log(\text{INC}+1).$$

Interpretieren Sie die Parameterschätzung.

(b) Über die Diagonalelemente h_{ii} der generalisierten Hat-Matrix H, werden so genannte High-Leverage-Punkte (Punkte mit extremer Lage im Designraum) identifiziert. Berechnen Sie die h_{ii} mit der R-Funktion `hatvalues()` und plotten Sie die für obiges Modell erhaltenen Werte gegen die Indizes i.

(c) Die studentisierten Pearson-Residuen, $r_{i,s}^P = r_i^P/(\sqrt{1-h_{ii}})$, sollten für gruppierte Daten mit genügend großen n_i approximativ normalverteilt sein. Berechnen Sie $r_{i,s}^P$ für den vorliegenden (ungruppierten) Datensatz, plotten Sie die erhaltenen $r_{i,s}^P$ gegen i und untersuchen die Verteilung an Hand eines Normal-Quantil-Plots.

(d) Eine weitere Möglichkeit zur Bestimmung von einflussreichen Beobachtungen ist Cook's Distance,

$$c_i = (\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}})^T \text{cov}(\hat{\boldsymbol{\beta}})^{-1} (\hat{\boldsymbol{\beta}}_{-i} - \hat{\boldsymbol{\beta}}),$$

dabei bezeichnet $\hat{\boldsymbol{\beta}}_{-i}$ den ML-Schätzer bei Entfernen der i-ten Beobachtung ($\hat{\boldsymbol{\beta}}$ ist der Schätzer bei Verwendung aller Beobachtungen). Schreiben Sie eine Funktion zur Berechnung der c_i und plotten Sie die sich für Ihr Modell aus (a) ergebenden Werte gegen die Indizes.

Eine sog. *one-step* Approximation $\hat{\boldsymbol{\beta}}_{-i,1}$ von $\hat{\boldsymbol{\beta}}_{-i}$ erhält man, indem man mit dem Wert $\hat{\boldsymbol{\beta}}$ startet und lediglich eine Fisher-Scoring Iteration durchführt. Man kann zeigen, dass in einem (univariaten) GLM

$$\hat{\boldsymbol{\beta}}_{-i,1} = \hat{\boldsymbol{\beta}} - w_i^{1/2}(1-h_{ii})^{-1/2} r_{i,s}^P (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{x}_i$$

gilt, wobei $\mathbf{W} = \text{diag}(w_1,\ldots,w_n) = \mathbf{W}(\hat{\boldsymbol{\beta}}) = \text{diag}(w_1(\hat{\boldsymbol{\beta}}),\ldots,w_n(\hat{\boldsymbol{\beta}}))$ die Gewichtungsmatrix und $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)^T$ die Designmatrix darstellt. Bei der Berechnung von $r_{i,s}^P$ werde $\sigma_i^2 = \phi v(\mu_i)$ verwendet.

(e) Bestimmen Sie allgemein die Werte $c_{i,1}$, die sich ergeben, falls man in (d) $\hat{\boldsymbol{\beta}}_{-i}$ durch $\hat{\boldsymbol{\beta}}_{-i,1}$ ersetzt.

(f) Berechnen und zeichnen Sie für Ihr Modell aus (a) die Werte $c_{i,1}$. Vergleichen Sie das Ergebnis mit Ihrem Resultat aus (d).

4 Mehrkategoriale Regressionsmodelle (I)

Aufgabe 1

Für das multinomiale Logit-Modell mit Referenzkategorie k gibt es die alternativen Modelldarstellungen

(1)
$$\log\left(\frac{P(Y=r|\boldsymbol{x})}{P(Y=k|\boldsymbol{x})}\right) = \boldsymbol{x}'\boldsymbol{\beta}_r$$

sowie

(2)
$$P(Y=r|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta}_r)}{1+\sum_{s=1}^{k-1}\exp(\boldsymbol{x}'\boldsymbol{\beta}_s)} \qquad \text{und} \qquad P(Y=k|\boldsymbol{x}) = \frac{1}{1+\sum_{s=1}^{k-1}\exp(\boldsymbol{x}'\boldsymbol{\beta}_s)}$$

($r = 1, \ldots, k-1$). Zeigen Sie die Äquivalenz der beiden Darstellungsformen!