3 Generalisierte lineare Modelle

Aufgabe 8

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) Fitten Sie ein Logit-Modell mit Prädiktor $\eta = \beta_0 + \beta_1 \text{TEN} + \beta_2 \text{SUP} + \beta_3 \log(\text{INC}+1)$. Interpretieren Sie kurz die geschätzten Parameter.

(b) Bestimmen Sie (allgemein) die Pearson- und Devianz-Residuen für das Logit-Modell. Berechnen Sie diese für ihr Modell aus Teilaufgabe (a) und plotten Sie sie gegen die Indizes i.

(c) Über die Diagonalelemente h_{ii} der generalisierten Hat-Matrix $\boldsymbol{H}$, werden so genannte High-Leverage-Punkte (Punkte mit extremer Lage im Designraum) identifiziert. Geben Sie $\boldsymbol{H}$ in allgemeiner Form an und schreiben Sie in `R` eine Funktion, die h_{ii} berechnet. Berechnen Sie diese für ihr Modell aus Teilaufgabe (a) und plotten Sie sie gegen die Indizes i. Vergleichen Sie das Ergebnis mit dem Resultat der Funktion `hatvalues()`.

(d) Die studentisierten Pearson-Residuen, $r^P_{i,s} = r^P_i/(\sqrt{1-h_{ii}})$, sollten für gruppierte Daten mit genügend großen n_i approximativ normalverteilt sein. Berechnen Sie diese für ihr Modell aus Teilaufgabe (a) (mit ungruppierten Daten) und plotten Sie sie gegen die Indizes i. Untersuchen die Verteilung an Hand eines Normal-Quantil-Plots.

(e) 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. Berechnen Sie diese für ihr Modell aus Teilaufgabe (a) und plotten Sie sie gegen die Indizes i.

Aufgabe 9

Im Folgenden wird wieder der Datensatz `fakesoep` von Übungsblatt 4 betrachtet.

(a) Logarithmieren Sie die Responsevariable `beink`. Fitten Sie anschließend ein (unpenalisiertes) GLM mit gaußverteiltem Response und allen Kovariablen (nur Haupteffekte).

(b) Geben Sie an, wieviele verschiedene Modelle durch alle denkbaren Kombinationen dieser Kovariablen möglich sind.

(c) Führen Sie mit Hilfe der `R`-Funktion `step()` schrittweise Variablenselektionen für Ihr Modell aus Teilaufgabe (a) auf Basis des AIC

- ausgehend vom vollen Modell und
- ausgehend von einem reinen Intercept-Modell

durch. Lassen Sie jeweils Schritte in beide Richtungen zu (`direction=''both''`). Vergleichen Sie die Ergebnisse.

(d) Berechnen Sie mithilfe der Funktion `penalized()` aus dem `R`-Paket `penalized` den Lasso-Schätzer Ihres Modells aus Teilaufgabe (a).
Plotten Sie die Koeffizientenpfade des Modells und vergleichen Sie Anzahl der selektierten Variablen mit den Ergebnissen aus Teilaufgabe (c).

(e) Wiederholen Sie nun die Schritte der Teilaufgaben (a) - (d) für ein Modell, das neben allen Haupteffekten auch alle paarweisen Interaktionen beinhaltet. Vergleichen Sie insbesondere die Anzahl der selektierten Variablen zwischen Lasso und der schrittweisen Variablenselektion.