3 Generalisierte lineare Modelle (II)

Aufgabe 1

Gegeben seien Beobachtungen $(y_i, \mathbf{x}_i), i = 1, \ldots, n$. Dabei sei $y_i \in \{0, 1\}$ eine binäre Responsevariable und $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^T$ ein Vektor mit p Kovariablen. Der lineare Prädiktor ist durch $\eta = \mathbf{x}^T\boldsymbol{\beta}$ gegeben.

(a) Zunächst soll ein Probit-Modell gefittet werden. Leiten Sie die Score-Funktion $s(\boldsymbol{\beta})$ her.

(b) Die erwartete Fisher-Informationsmatrix $F(\boldsymbol{\beta})$ ist für GLMs allgemein gegeben durch

$$F(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \left(\frac{\partial h(\eta_i)}{\partial \eta} \right)^2 / \sigma_i^2,$$

mit $\sigma_i^2 = \text{var}(y_i)$.

Berechnen Sie mit Hilfe dieser Formel $F(\boldsymbol{\beta})$ für das Logit- und das Probit-Modell.

(c) In der Rechnung zu Aufgabe 1b) wurde festgestellt, dass im Logit-Modell $\frac{\partial h(\eta)}{\partial \eta} = \text{var}(y)$ gilt. Zeigen Sie, dass in allen GLMs mit kanonischem Link $\frac{\partial h(\eta)}{\partial \eta} = \frac{\text{var}(y)}{\phi}$ gilt. Welche weiteren Vorteile bietet die Verwendung des kanonischen Links im Vergleich zu anderen Linkfunktionen?

Aufgabe 2

Hypothesentests über den (p-dim.) Koeffizientenvektor $\boldsymbol{\beta}$ eines GLM sind meist linear und können in der Form

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \quad \text{gegen} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d}$$

dargestellt werden ($\mathbf{C} : r \times p$, $r \leq p$, $\text{rg}(\mathbf{C}) = r$, $d : r \times 1$).

(a) Sei nun $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$. Wie lauten $\mathbf{C}$ und $\mathbf{d}$, wenn folgende Hypothesen getestet werden sollen?

(i) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
(ii) $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$
(iii) $H_0 : \beta_1 = \beta_2,\ \beta_3 = \beta_4$
(iv) $H_0 : 2\beta_1 = \beta_2,\ 3\beta_1 = \beta_3,\ 4\beta_1 = \beta_4$

(b) Nehmen Sie an bei $\beta_1, \ldots, \beta_4$ handelt es sich um die Koeffizienten einer dummy-kodierten ordinalen Einflussgröße mit Levels $k = 0, 1, \ldots, 4$, wobei Kategorie 0 als Referenz-Kategorie verwendet wird. Welcher Modellierungsansatz wird dann in Teilaufgabe (a) (iv) überprüft?

(c) Welche Arten von Teststatistiken werden in der GLM-Theorie üblicherweise verwendet? In welcher Situation empfiehlt sich die Verwendung jeweils welcher Statistik?