2 Regression mit binärem Response

Aufgabe 16

Die Zusammenhangsstruktur dreier binärer Variablen X_A, X_B und X_C lasse sich durch das Modell $X_AX_B/X_AX_C/X_BX_C$ beschreiben. Welcher Zusammenhang besteht dann zwischen dem zugehörigen loglinearen Modell und einem Logit-Modell mit X_B als abhängiger Variable und X_A und X_C als Einflussgrößen?

Aufgabe 17

Für die verschiedenen Ausprägungen der mehrkategorialen Einflussgröße Ausbildungsstand x sei die Wahrscheinlichkeit, dass die betreffende (arbeitslose) Person weniger als 6 Monate arbeitslos ist, $\pi(x) = P(y = 1|x)$ bekannt; y ist eine dichotome Variable (1: Arbeitslosigkeit $\leq$ 6 Monate, 0: Arbeitslosigkeit $>$ 6 Monate). Es gilt

x	$\pi(x)$
1 : Keine Ausbildung	0.68
2 : Lehre	0.65
3 : Fachspezifische Ausbildung	0.57
4 : Hochschulabschluss	0.71

(a) Wählen Sie eine (0-1)-Kodierung für die kategoriale Einflussgröße und geben Sie das entsprechende Logit-Modell (mit Nebenbedingung) an.

(b) Berechnen Sie die relativen Chancen von Kategorie 3 gegenüber Kategorie 4. Was sagen diese relativen Chancen aus?

(c) Bestimmen Sie den Parametervektor $\boldsymbol{\beta}$ des Logit-Modells, wobei $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T$.

Aufgabe 18

Zur Untersuchung der Erklärungskraft von Kovariablen auf einen binären Response mit den Werten 0 bzw. 1 dient u.a. folgendes nach Gini benanntes Maß:

$$G = \frac{\sum_{i=1}^{n} \hat{\pi}_i^2 - n\bar{p}^2}{n\bar{p}(1-\bar{p})}$$

Dabei steht $\bar{p}$ für den Anteil an Einsern bei der Zielgröße in der Stichprobe. Zeigen Sie, dass G dem Bestimmtheitsmaß R^2 im Falle einer binären Zielgröße entspricht.

Aufgabe 19

Zur Modellierung von Überdispersion in einem Binomial-Modell besteht ein sinnvoller Ansatz darin, die zusätzliche Variabilität durch eine latente Variable zu erklären. Anstatt einer Binomialverteilung für die Anzahl der Erfolge in einer Gruppe $i = 1, \ldots, N$

$$y_i \sim B(n_i, \pi_i)$$

wird nun folgendes Modell angenommen:

- Für die latente Variable $D_i \in [0; 1]$ gelte:

$$\begin{aligned} E(D_i) &= \pi_i \\ Var(D_i) &= \delta\pi_i(1-\pi_i) \quad \text{mit} \quad \delta > 0 \end{aligned}$$

- Gegeben $D_i = \vartheta_i$ gelte:

$$y_i | D_i = \vartheta_i \sim B(n_i, \vartheta_i)$$

(a) Zeigen Sie, dass dann für die marginale Verteilung von y_i gilt:

$$\begin{aligned} E(y_i) &= n_i\pi_i \\ Var(y_i) &= n_i\pi_i(1-\pi_i)\phi_i \quad \text{mit} \quad \phi_i = 1 + (n_i - 1)\delta \end{aligned}$$

(b) Im Falle von $D_i \sim Be(a_i, b_i)$ spricht man vom sog. Beta-Binomial-Modell. Wie lautet die entsprechende Parametrisierung über π_i und δ? Überlegen Sie sich hierzu ein adäquates Beispiel.

Aufgabe 20

Es sollen die Samen der Sommerwurz (Fachbegriff orobanche; auch Unkraut genannt) untersucht werden. Betrachten Sie hierzu den Datensatz *orob2* aus dem Paket `aod`, der ursprünglich von Crowder (1978) analysiert wurde. Berechnen und vergleichen Sie die in den entsprechenden Teilaufgaben genannten Modelle mit folgenden Kovariablenstrukturen:

(i) `seed`

(ii) `seed + root`

(iii) `seed * root`

Die Zielgröße soll dabei jeweils *Samen gekeimt ja/nein* sein. Des Weiteren soll stets der Logit-Link verwendet werden.

(a) Binomialmodelle mittels der Funktion `glm()`!

(b) Beta-Binomial-Modelle mittels der Funktion `betabin()`!

(c) Quasi-Binomial-Modelle mittels der Funktion `glm()`!

(d) Vergleichen Sie die einzelnen Modell-Outputs! Was fällt Ihnen auf? Diskutieren Sie Möglichkeiten des Modellvergleichs!