Statistische Modellierung III

-Generalisierte lineare Modelle-

Dr. Martin Scharpenberg

MSc Medical Biometry/Biostatistics

WiSe 2019/2020

Annahmen im GLM

Generalisierte lineare Modelle

- Wollen nun lineare und binäre Regression auf Exponentialfamilien verallgemeinern
- Können so gleichzeitig eine Reihe von praktisch relevanten Regressionsmodellen mit metrischen, binären oder Kategoriellen Zielvariablen behandeln
- $(Y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, wieder Daten von stochastisch unabhängigen Beobachtungseinheiten mit zufälliger Zielvariable Y_i und fixen (nicht-zufälligen) Kovariablen
- Zielvariablen Y_i sollen einen gemeinsamen Träger $\mathbb{T}$ haben
- Machen dazu eine Verteilungs- und eine Strukturannahme

Verteilungsannahme (Verteilungsmodell)

- Annahme: Zielvariablen Y_i haben Dichten bzgl. des Lebesgue- oder Zählmaßes haben, die einer Exponentialfamilie entstammen, d.h. für alle Werte y_i im gemeinsamen Träger $\mathbb{T}$ sei die die Dichte von Y_i

$$f(y_i|\theta_i, \phi, \omega_i) = f(y_i|\theta_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i - c(y_i, \phi, \omega_i)\right\}.$$

- Kanonischer Parameter θ_i und das (bekannte) Gewicht ω_i können also von der Beobachtungseinheit i abhängen, ϕ ist immer unabhängig von i

- Bei Einzelbeobachtungen ist meist auch $\omega = 1$ von i unabhängig, bei gruppierten Daten allerdings nicht

- Wissen aus dem vorigen Kapitel, dass

$$E(Y_i) = \mu_i = b'(\theta_i) \quad \text{und} \quad Var(Y_i) = \sigma_i^2 = \phi\, b''(\theta_i)/\omega_i$$

Strukturannahme (Strukturmodell)

- Wollen die Verteilung von Y_i mit dem linearen Prädiktor verbinden
- Strukturmodell verknüpft den Erwartungswert μ_i mit dem linearen Prädiktor $\eta_i = \mathbf{x}_i\beta$ durch eine (uns bekannte) Reponsefunktion $h : \mathbb{R} \longrightarrow \mathbb{R}$ bzw. ihrer Umkehrfunktion $g = h^{-1}$, der Linkfunktion
- Es wird angenommen, dass

$$\mu_i = E(Y_i) = b'(\theta_i) = h(\eta_i) \quad \text{für } \eta_i = \mathbf{x}_i\beta,$$

bzw. $\mathbf{x}_i\beta = \eta_i = g(\mu_i) = g[b'(\theta_i)]$ für $g = h^{-1}$

Kanonische Link- bzw. Responsefunktion

- Falls $\eta_i = \theta_i$, also linearer Prädiktor und natürlicher Parameter zusammenfallen, dann ist

$$h(\eta_i) = b'(\theta_i) \qquad \text{und damit} \qquad g(\mu_i) = (b')^{-1}(\mu_i)$$

- Man nennt $g = (b')^{-1}$ die "kanonische" oder "natürliche" Linkfunktion (Wir nennen $h = b'$ die "kanonische" Responsefunktion)

Beispiele

Binomialverteilung - logistische Regression

- Wissen dass für die Binomialverteilung gilt:

$$b(\theta) = \log(1 + e^{\theta}) \quad \text{und damit} \quad b'(\theta) = \frac{e^{\theta}}{1 + e^{\theta}}$$

- Der kanonische Link ist also die Log-Odds:

$$g(\pi) = \log \frac{\pi}{1 - \pi} \qquad (= (b')^{-1}(\pi)\,)$$

- Probit und Log-Log-Modelle sind ebenfalls generalisierte lineare Modelle, da $Y \sim B(n, p)$ eine Exponentialfamilie, allerdings mit nicht-kanonischen Link-Funktionen:
    - Probit-Modell: $g(\pi) = \Phi^{-1}(\pi)$,
    - Log-Log-Modell: $g(\pi) = -\log(-\log(\pi))$.

Poissonverteilung

- Bei der Poissonverteilung gilt

$$\mu_i = b'(\theta_i) = e^{\theta_i} \qquad \Longleftrightarrow \qquad (b')^{-1}(\mu_i) = \log(\mu_i)$$

- Damit ist der kanonische Link $g(\mu_i) = \log(\mu_i)$ für $\mu_i = \lambda_i > 0$

Weitere Beispiele

- Normalverteilung: Hier sind Response- und Linkfunktion die Identität, d.h. der kanonische Link führt zur herkömmlichen linearen Regression
- Gammaverteilung (kanonischer Link $g(\mu_i) = -\frac{1}{\mu_i} = \eta_i \in \mathbb{R}_-$)
- Neagiv-Binomialverteilung (kanonischer Link $g(\pi_i) = \log \frac{\pi_i}{1+\pi_i} = \eta_i \in \mathbb{R}_-$)
- Für Gamma- und Negativ-Binomialverteilung ist der kanonische Parameter beschränkt, wohingegen $\eta_i = \mathbf{x}_i \beta$ i.A. unbeschränkt ist
- Hier ist es i.A. sinnvoller nicht-kanonische Links zu benutzen

Log-Likelihood-Kern eines GLMs

Log-Likelihood-Kern

- Die Log-Likelihood einer Exponentialfamilie ist gegeben durch

$$\frac{Y_i\theta_i - b(\theta_i)}{\phi}\omega_i - c(Y_i, \phi, \omega_i)$$

- Da $c(Y_i, \phi, \omega_i)$ nicht von β abhängt ist der wesentliche Teil der Log-Likelihood (der *Log-Likelihood-Kern*) gegeben durch

$$l_i(\beta) = \frac{Y_i\theta_i - b(\theta_i)}{\phi}\omega_i$$

- Bei Einzelbeobachtungen ist typischerweise $\omega_i = 1$

Log-Likelihood-Kern für gruppierte Beobachtungen

- Haben wir $l = 1, \ldots, m$ Gruppen mit $\omega_i = 1$ und $\mathbf{x}_i = \mathbf{z}_l$ für alle Einzelbeobachtungen $i \in G_l$ und sind diese stochastisch unabhängig, dann ist der Log-Likelihood-Kern für die Gruppe G_l

$$\tilde{l}_l(\beta) = \sum_{i \in G_l} l_i(\beta) = \sum_{i \in G_l} \frac{Y_i \theta_l - b(\theta_l)}{\phi} = \frac{(\sum_{i \in G_l} Y_i)\theta_l - n_l b(\theta_l)}{\phi} = \frac{\bar{Y}_l\, \theta_l - b(\theta_l)}{\phi} \omega_l,$$

wobei $\bar{Y}_l = \sum_{i \in G_l} Y_i / n_l$ und $\omega_l = n_l$

- Durch das Berücksichtigen von Gewichten, hat der Log-Likelihood-Kern von Einzelbeobachtungen Y_i und Mittelwerten $\bar{Y}_l$ von Gruppen mit Konstanten $\mathbf{x}_i = \mathbf{z}_l$ die gleich Form

- Gruppen können also über den Mittelwert wie Einzelbeobachtungen behandelt werden

Bestimmung des MLE

Verknüpfung von θ_i and β

- Für die Bestimmung der MLE werden wir θ_i als Funktion von β (bzw. $\mathbf{x}_i\beta$) auffassen und nach β ableiten

- Bei Verwendung des kanonischen Links gilt

$$\theta_i = \eta_i = \mathbf{x}_i\beta \quad \Longrightarrow \quad \frac{\partial\theta_i}{\partial\beta} = \mathbf{x}_i^T$$

- Für nicht-kanonische Links lässt sich zeigen, dass

$$\frac{\partial\theta_i}{\partial\beta} = \mathbf{x}_i^T h'(\eta_i)/v(\mu_i)$$

- Beim kanonischen Link gilt $\eta = \theta$ und $h(\eta) = b'(\theta)$ und damit $h'(\eta) = b''(\theta) = v(\mu_i)$

Berechnung der Score-Funktion

- Für stochastisch unabhängige Beobachtungen bzw. Gruppenmittelwerte gilt für die Score-Funktion:

$$\begin{aligned}\mathbf{s}(\beta) &= \frac{\partial}{\partial\beta} l(\beta) = \sum_{i=1}^{n} \frac{\partial}{\partial\beta} l_i(\beta) = \sum_{i=1}^{n} \frac{\partial}{\partial\beta} \frac{Y_i\theta_i - b(\theta_i)}{\phi}\omega_i \\ &= \sum_{i=1}^{n} \left\{ Y_i \frac{\partial\,\theta_i}{\partial\beta} - \frac{\partial\,\theta_i}{\partial\beta}\, b'(\theta_i) \right\} \omega_i/\phi = \sum_{i=1}^{n} \frac{\partial}{\partial\beta}\theta_i \{Y_i - b'(\theta_i)\}\, \omega_i/\phi \\ &= \sum_{i=1}^{n} \mathbf{x}_i^T \frac{h'(\eta_i)}{v(\mu_i)} \{Y_i - \mu_i\}\, \omega_i/\phi = \sum_{i=1}^{n} \mathbf{x}_i^T h'(\eta_i) \{Y_i - \mu_i\}/\sigma_i^2,\end{aligned}$$

  wobei $\sigma_i^2 = \phi\, v(\mu_i)/\omega_i = Var(Y_i)$

- Der MLE $\hat{\beta}$ von β ist die Lösung des Gleichungssystems $\mathbf{s}(\beta) = \mathbf{0}$

Berechnung der Fisher-Matrix

- Die Fisher-Matrix berechnet sich zu

$$\mathbf{F}(\beta) = E\left(\mathbf{s}(\beta)\mathbf{s}(\beta)^T\right) = \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \frac{h'(\eta_i)^2}{\sigma_i^4} \underbrace{E\{(Y_i - \mu_i)^2\}}_{=\sigma_i^2} = \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \left(\frac{h'(\eta_i)}{\sigma_i}\right)^2$$

$$= \sum_{i=1}^{n} \mathbf{x}_i^T \mathbf{x}_i \, w_i^2.$$

wobei $w_i = h'(\eta_i)/\sigma_i$

- Das gerade definierte Gewicht w_i ist **nicht** das vorgegebene Gewicht ω_i in der Dichte von Y_i

Matrixschreibweise

- Wollen Score und Fisher-Matrix in Matrixschreibweise ausdrücken
- Definiere dazu $d_i = d_i(\beta) = h'(\mathbf{x}_i\beta)$ sowie
$$\mathbf{D} = \text{diag}(d_1, \ldots, d_n), \quad \Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2), \quad \mathbf{W} = \text{diag}(w_1, \ldots w_n)$$
mit $w_i = (d_i/\sigma_i)^2$
- Darüber hinaus betrachten wir die üblichen $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mu = (\mu_1, \ldots, \mu_n)^T$ und die Designmatrix $\mathbf{X}$
- Damit folgt
$$\mathbf{s}(\beta) = \mathbf{X}^T\mathbf{D}\Sigma^{-1}(\mathbf{Y} - \mu) \quad \text{und} \quad \mathbf{F} = \mathbf{X}^T\mathbf{W}\mathbf{X},$$
wobei $\mathbf{D}$, Σ und $\mathbf{W}$ von β abhängen

Fisher-Scoring-Algorithmus

- Der Maximum-Likelihood-Schätzer für β ist die Lösung des Gleichunsgssytems $\mathbf{s}(\beta) = \mathbf{0}$
- Diese kann wie bei der logistischen Regression mit dem Fisher-Scoring-Verfahren numerisch-iterativ bestimmt werden
- Im k-ten Iterationsschritt hat der MLE die Form

$$\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} + \mathbf{F}^{-1}(\hat{\beta}^{(k)})\,\mathbf{s}(\hat{\beta}^{(k)})$$

- Darstellung als gewichtete KQ-Schätzung:

$$\hat{\beta}^{(k+1)} = (\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(k)}\tilde{\mathbf{Y}}^{(k)}$$

mit $\mathbf{W}^{(k)} = \mathrm{diag}(d_1^{(k)}/\sigma_1^{(k)}, \ldots, d_n^{(k)}/\sigma_n^{(k)})$ und den "Arbeitsbeobachtungen"

$$\tilde{Y}_i^{(k)} = \mathbf{x}_i\hat{\beta}^{(k)} + d_i^{-1}(\hat{\beta}^{(k)})[Y_i - \hat{\mu}^{(k)}], \qquad \hat{\mu}_i^{(k)} = h(\mathbf{x}_i\hat{\beta}^{(k)})$$

Bemerkungen

- Wir können zur Schätzung von $\hat{\beta}$ also effiziente Methoden zum Berechnen des gewichteten KQ-Schätzers verwenden
- $\mathbf{W}^{(k)}$ hängt wegen $d_1^{(k)} = h'(\mathbf{x}_i\hat{\beta}^{(k)})$ und $\sigma^{(k)} = \phi\, v[h(\mathbf{x}_i\hat{\beta}^{(k)})]/\omega_i$ vom Schätzer des k-ten Schritts ab
- Das unbekannte ϕ in der Formel für $\hat{\beta}^{(k+1)}$ kürzt sich mit den Termen $(\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}$ und $\mathbf{X}^T\mathbf{W}^{(k)}$ heraus und geht daher nicht in die Berechnung von $\hat{\beta}^{(k+1)}$ ein

Weitere Bemerkungen

- Der Fisher-Scoring-Algorithmus konvergiert meist nach wenigen Schritten. Wenn nicht, dann wurde entweder der falsche Startwert gewählt oder der MLE existiert nicht
- Bei einer kanonischen Linkfunktion hat $\mathbf{s}(\beta) = 0$ eine eindeutige Lösung. Bei nichtkanonischem Link kann es mehr als eine Lösung geben. In diesem Fall sollten mehrere verschiedene Startwerte verwendet werden
- Invertierbarkeit von $\mathbf{X}^T\mathbf{W}\mathbf{X} = \sum_{i=1}^n \mathbf{x}_i^T \mathbf{x}_i w_i$ folgt aus $\text{Rang}(\mathbf{X}) = k$ und $w_i = d_i^2/\sigma_i^2 > 0$ für alle i

Eigenschaften des MLE

- Unter schwachen Regularitätsbedingungen gilt

$$\mathbf{F}^{1/2}(\beta)(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{I})$$

- Darüber hinaus lässt sich zeigen, dass

$$\mathbf{F}^{1/2}(\hat{\beta})(\hat{\beta} - \beta) \xrightarrow{d} N(0, \mathbf{I})$$

Schätzung des Dispersionsparameters

Schätzung des Dispersionsparameters

- Es ist per Definition $\sigma_i^2 = \phi v(\mu_i)/\omega_i$
- Damit haben wir $\phi = \omega_i \sigma_i^2 / v(\mu_i)$
- Der Dispersionsparameter lässt sich also via

$$\hat{\phi} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \omega_i$$

  schätzen, wobei $\hat{\mu}_i = h(\mathbf{x}_i \hat{\beta})$
- Für gruppierte Daten gilt mit $\hat{\mu}_l = h(\mathbf{z}_l \hat{\beta})$:

$$\hat{\phi} = \frac{1}{n-k} \sum_{l=1}^{m} \frac{(\bar{Y}_l - \hat{\mu}_l)^2}{v(\hat{\mu}_l)} \omega_l$$

Testen linearer Hypothesen

Testen linearer Hypothesen

- Sei $\mathbf{C}$ eine $(r \times k)$-Restriktionsmatrix für $r < k$ und $\mathbf{d} \in \mathbb{R}^r$
- Sind wie bisher daran interessiert Hypothesen der Form

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

  zu testen

- Zum Testen von H_0 können wir wieder verschiedene Teststatistiken verwenden (siehe nächste Folie)
- Natürlich muss vorab festgelegt werden, welche der drei Teststatistiken verwendet werden soll

Testen linearer Hypothesen - Teststatistiken

1. Likelihood-Quotienten-Statistik:

$$lq = -2\{l(\tilde{\beta}) - l(\hat{\beta})\},$$

   wobei $\tilde{\beta}$ der MLE von β unter der Restriktion $\mathbf{C}\tilde{\beta} = \mathbf{d}$ ist

2. Wald-Statistik

$$W = (\mathbf{C}\hat{\beta} - \mathbf{d})^T [\mathbf{C}\mathbf{F}^{-1}(\hat{\beta})\mathbf{C}^T]^{-1} (\mathbf{C}\hat{\beta} - \mathbf{d})$$

3. Score-Statistik

$$U = \mathbf{s}(\tilde{\beta})^T \mathbf{F}^{-1}(\tilde{\beta}) \mathbf{s}(\tilde{\beta})$$

Testen linearer Hypothesen - Teststatistiken

- Alle drei Statistiken sind unter H_0 approximativ χ^2_r verteilt
- Können H_0 also verwerfen, falls

$$lq, \omega \text{ oder } U \quad \geq \chi^2_r(1-\alpha)$$

- Natürlich muss vorab festgelegt werden, welche der drei Teststatistiken verwendet werden soll
- Die p-Werte sind:
    - Likelihood-Quotienten-Test: $p = 1 - \mathbf{F}_r^{\chi^2}(lq)$
    - Wald-Test: $p = 1 - \mathbf{F}_r^{\chi^2}(W)$
    - Score-Test: $p = 1 - \mathbf{F}_r^{\chi^2}(U)$

Einseitige Tests mit eindimensionaler Restriktion

- Betrachten den Fall $r = 1$, d.h. die Restriktionsmatrix ein Zeilenvektor $\mathbf{c} \in \mathbb{R}^k$, und $\mathbf{d}$ ist einfach eine Zahl $d \in \mathbb{R}$
- Können nun die folgenden einseitigen Hypothesen betrachten:
$$H_0 : \mathbf{c}\beta \leq d \quad \text{gegen} \quad H_1 : \mathbf{c}\beta > d.$$
- Es ist $\mathbf{c}\hat{\beta}/(\mathbf{c}\mathbf{F}^{-1}(\hat{\beta})\mathbf{c}^T) \xrightarrow{d} N(0, \mathbf{I})$
- Also erhalten wir einen Test zum asymptotischen Signifikanzniveau α, wenn wir H_0 verwerfen, falls
$$(\mathbf{c}\hat{\beta} - d)/[\mathbf{c}\mathbf{F}^{-1}(\hat{\beta})\mathbf{c}^T] \geq \Phi^{-1}(1-\alpha)$$
- Einseitige Tests an einen einzelnen Regressionskoeffizienten β_j erhalten wir mit $\mathbf{c} = e_j$ dem j-ten Einheitsvektor (als Zeilenvektor), denn dann ist $\mathbf{c}\beta = \beta_j$

Kriterien zur Beurteilung der Modellanpassung

Modellanpassung - AIC

- Wie bei der linearen Regression kann das Akaike-Informationskriterium

$$AIC = -2l(\hat{\beta}) + 2k.$$

  zur Modellwahl verwendet werden

- Falls das Modell einen Dispersionsparameter hat, wird dieser geschätzt und k wird durch $k + 1$ ersetzt (da ein Parameter mehr geschätzt wird)

Modellanpassung - Pearson-Statistik

- Ein anderes, uns bereits bekanntes Kriterium zur Beurteilung der Güte ist die Pearson-Statistik

$$\chi^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} \, \omega_i \, ,$$

  die wir bisher nur für gruppierte Daten mit binomialverteilten Zielvariablen eingeführt haben,

$$\chi^2 = \sum_{l=1}^{m} \frac{(\bar{Y}_l - \hat{\mu}_l)^2}{v(\hat{\mu}_l)} \, \omega_l$$

- Nur dann sinnvoll anwendbar, wenn $k < n$ und k nicht zu groß ist
- Bei gruppierten Daten sollte die Zahl der Beobachtungen pro Gruppe nicht zu klein sein

Modellanpassung - Devianz

- Weiteres (uns bekanntes) Kriterium zur Modellapassung ist die *Devianz*. Bei Einzelbeobachtungen hat sie die Form

$$D = -2\hat{\phi}\sum_{i=1}^{n}\{l_i(\hat{\mu}_i) - l_i(Y_i)\}$$

- Vergleicht das in Frage stehende generalisierte lineare Modell mit dem saturierten Modell (in dem μ_i durch Y_i geschätzt wird)
- Es ist immer $l_i(Y_i) \geq l_i(\hat{\mu}_i)$ und damit $D \geq 0$
- D wird als Abstand zwischen dem vorliegenden Modell und dem saturierten Modell interpretiert, hat allerdings nicht die Eigenschaften einer Metrik

Modellanpassung - Devianz

- Bei gruppierten Daten ist die Devianz

$$D = -2\hat{\phi} \sum_{l=1}^{m} \{ l_i(\hat{\mu}_l) - l_i(\bar{Y}_l) \}$$

- Wieder sollte $k < n$ nicht zu groß sein
- Bei gruppierten Daten hat man den Vorteil, dass bereits beim saturierten Modell der Erwartungswert μ_l mit mehr als einer Beobachtung geschätzt wird, nämlich durch den Mittelwert $\bar{Y}_l$