

Statistik II

- GLM (Regression und Varianzanalyse)
- Logistische Regression
- Loglineare Modelle
- Verallgemeinerte Lineare Modelle (GLIM)
"Generalised Linear Interactive Models"

Regularitätsbedingungen

Die Loglikelihood für einen Parameter heißt regulär, wenn sie in einer offenen Nachbarschaft des wahren Werts durch eine konkave quadratische Funktion approximiert werden kann.

Sei der Parameterraum Θ offen, ist die Exponentialverteilungsfamilie regulär. (Eine schwächere Bedingung für den Rand heißt "steepness", Steilheit.)

Wir brauchen Regularität um:

- die Operationen Ableiten und Integrieren zu vertauschen
- lokale Approximationen mit Taylorentwicklungen zu machen
- sicherzustellen, dass verschiedene Parameterwerte zu verschiedenen Verteilungen führen.

Beispiele:

- $B(n, \pi)$ ist nur regulär für $0 < \pi < 1$

- Die Invers Gauß Verteilung

$$f(y; \mu, \alpha) = \sqrt{\frac{\alpha}{2\pi y^3}} e^{-\frac{\alpha(y-\mu)^2}{2\mu^2 y}}$$

$$y, \mu, \alpha > 0$$

ist nicht regulär, weil es für $\mu = \infty$ eine gültige Dichte bleibt (aber sie ist steil).

- Die Gleichverteilung

$$f(y; \theta) = \frac{1}{\theta} \quad 0 < y \leq \theta < \infty$$

ist nicht regulär, weil der Träger vom Parameter abhängt.

Eigenschaften der Score Statistik

$$\begin{aligned}U &= \frac{\partial \log f(y; \theta)}{\partial \theta} \\ &= \frac{1}{f} \frac{\partial f(y; \theta)}{\partial \theta}\end{aligned}$$

Erwartungswert der Score Statistik

$$\begin{aligned}E[U] &= \int U f dy \\ &= \int \frac{\partial f(y; \theta)}{\partial \theta} dy \\ &= \frac{\partial}{\partial \theta} \left(\int f(y; \theta) dy \right) \\ &= 0\end{aligned}$$

Varianz der Score Statistik

Die Fisher Information ist der Erwartungswert der beobachteten Information, $i(\theta) = E[J]$.

$$\begin{aligned} J &= -\frac{\partial^2 \log f}{\partial \theta \partial \theta'} \\ &= -\frac{\partial U}{\partial \theta} \\ &= -\frac{\partial}{\partial \theta} \left(\frac{1}{f} \frac{\partial f}{\partial \theta} \right) \\ &= \frac{1}{f^2} \frac{\partial f}{\partial \theta} \frac{\partial f}{\partial \theta'} - \frac{1}{f} \frac{\partial^2 f}{\partial \theta \partial \theta'} \end{aligned}$$

Dann gilt

$$\frac{\partial^2 f}{\partial \theta \partial \theta'} = U' U f - J f$$

$$\text{Aber } \int \frac{\partial^2 f}{\partial \theta \partial \theta'} dy = 0 \Rightarrow E[U' U] = E[J]$$

$$\text{Und } E[U] = 0 \Rightarrow E[U' U] = \text{Var}[U] \Rightarrow \text{Var}[U] = i(\theta)$$

Asymptotische Normalverteilung der ML Schätzer

Betrachten wir eine Taylorentwicklung der Score Statistik um den wahren Wert θ :

$$U(\hat{\theta}) = U(\theta) + \frac{\partial U}{\partial \theta} \frac{\hat{\theta} - \theta}{1!} + \dots$$

Da $U(\hat{\theta}) = 0$, haben wir zu einer ersten Approximation:

$$U(\theta) \doteq J(\theta)(\hat{\theta} - \theta)$$

$J(\theta)$ ist eine Summe von unabhängigen Komponenten und nach dem Gesetz der großen Zahlen gilt fast sicher, dass

$$\lim_{n \rightarrow \infty} J(\theta) = E[J(\theta)] = i(\theta)$$

Deshalb gilt auch

$$(\hat{\theta} - \theta) \doteq i^{-1}(\theta)U(\theta)$$

Erwartungswert und Varianz der rechten Seite sind 0 bzw. $i^{-1}(\theta)$. $U[\theta]$ ist eine Summe unabhängiger Komponente und nach dem Zentralen Grenzwertsatz gilt

$$\hat{\theta} \sim N(\theta, i^{-1}(\theta))$$

Güte einer Modellanpassung

Nullmodell: nur 1 Parameter, μ_0 , jede Beobachtung wird auf den gleichen Wert angepaßt

$$l_0 = l(\hat{\mu}_0 \cdot \mathbf{1}_n, \varphi; Y)$$

saturiertes Modell: für jede Beobachtung einen Parameter

$$l_s = l(Y, \varphi; Y)$$

aktuelles Modell:

$$l_m = l(\hat{\mu}, \varphi; Y)$$

Devianz für aktuelles Modell:

$$\begin{aligned} D_m &= -2(l_m - l_s)\varphi \\ &= 2 \sum_{i=1}^n w_i \{y_i(\theta_i^s - \theta_i^m) - b(\theta_i^s) + b(\theta_i^m)\} \end{aligned}$$

Verteilung der Devianz

Unter der Annahme, dass die grundlegende Verteilung normal ist, kann man die Devianz aus der Verteilung von $\hat{\theta}$ durch

$$D_{N_1}(\theta) \doteq (\hat{\theta} - \theta)' i(\theta) (\hat{\theta} - \theta)$$

approximieren. Wegen $\hat{\theta} \sim N(\theta, i^{-1}(\theta))$ sollte asymptotisch gelten dass

$$(\hat{\theta} - \theta)' i(\theta) (\hat{\theta} - \theta) \sim \chi_p^2$$

wo p die Dimension von θ ist.

Eine zweite Approximation gewinnen wir aus der Taylorentwicklung der Loglikelihoodfunktion:

$$l(\theta) = l(\hat{\theta}) + (\hat{\theta} - \theta)' \frac{\partial l(\hat{\theta})}{\partial \theta} + \frac{1}{2} (\hat{\theta} - \theta)' \frac{\partial^2 l(\hat{\theta})}{\partial \theta \partial \theta'} (\hat{\theta} - \theta) + \dots$$

Da $\frac{\partial l(\hat{\theta})}{\partial \theta} = U(\hat{\theta}) = 0$ haben wir

$$D_{N_2}(\theta) \doteq (\hat{\theta} - \theta)' J(\hat{\theta} - \theta)$$

Ein erweitertes asymptotisches Argument gilt dann hier auch und wir nehmen an, dass

$$(\hat{\theta} - \theta)' J(\hat{\theta} - \theta) \sim \chi_p^2$$

Beispiele

Im folgenden bezeichne $\theta^s = \theta(Y)$ den Schätzer für θ aus dem saturierten Modell und $\theta^m = \theta(\hat{\mu})$ den Schätzer für θ aus dem aktuellen Modell.

Normalverteilung:

$$Y_i \sim N(\mu_i, \sigma^2)$$

Es gilt dann:

$$\begin{aligned}\theta_i^s &= y_i \\ b(\theta_i^s) &= \frac{(\theta_i^s)^2}{2} = \frac{y_i^2}{2} \\ \theta_i^m &= \hat{\mu}_i \\ b(\theta_i^m) &= \frac{(\theta_i^m)^2}{2} = \frac{\hat{\mu}_i^2}{2}\end{aligned}$$

$$\begin{aligned}
D(Y; \hat{\mu}) &= D_m \\
&= -2(\ell_m - \ell_s)\varphi \\
&= 2 \sum_{i=1}^n w_i \{y_i(\theta_i^s - \theta_i^m) - b(\theta_i^s) + b(\theta_i^m)\} \\
&= 2 \sum_{i=1}^n \left\{ y_i(y_i - \hat{\mu}_i) - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right\} \\
&= 2 \sum_{i=1}^n \left(y_i^2 - y_i \hat{\mu}_i - \frac{y_i^2}{2} + \frac{\hat{\mu}_i^2}{2} \right) \\
&= \sum_{i=1}^n (y_i^2 - 2y_i \hat{\mu}_i + \hat{\mu}_i^2) \\
&= \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 \\
&= QS_{res}.
\end{aligned}$$

Poisson:

$$Y_i \sim Po(\lambda_i)$$

$$D(Y; \hat{\lambda}) = D_m$$

$$= 2 \sum_{i=1}^n \left\{ y_i (\ln y_i - \ln \hat{\lambda}_i) - (e^{\ln y_i} - e^{\ln \hat{\lambda}_i}) \right\}$$

$$= 2 \sum_{i=1}^n \left\{ y_i \ln \frac{y_i}{\hat{\lambda}_i} - (y_i - \hat{\lambda}_i) \right\}$$

$$= 2 \sum_{i=1}^n \left\{ y_i \ln \frac{y_i}{\hat{\lambda}_i} \right\} - 2 \sum_{i=1}^n (y_i - \hat{\lambda}_i)$$

Bei allen interessierenden Modellen gilt:

$$\sum_{i=1}^n (y_i - \hat{\lambda}_i) = 0$$

$$\Rightarrow D(Y; \hat{\lambda}) = 2 \sum_{i=1}^n \left\{ y_i \ln \frac{y_i}{\hat{\lambda}_i} \right\}$$

$$= G^2.$$

Binomial:

$\tilde{Y}_i \sim \text{Bi}_{n_i, \pi_i}$ ersetzt durch $Y_i = \frac{\tilde{Y}_i}{n_i}$

$$\ell_m = \ell(\hat{\pi} | Y)$$

$$\begin{aligned} &= \sum_{i=1}^n \left[n_i y_i \ln \left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) + n_i \ln(1 - \hat{\pi}_i) \right] \\ &= \sum_{i=1}^n [n_i y_i \ln(\hat{\pi}_i) + (n_i - n_i y_i) \ln(1 - \hat{\pi}_i)] \end{aligned}$$

$$\ell_s = \ell(Y | Y)$$

$$= \sum_{i=1}^n [n_i y_i \ln(\hat{y}_i) + (n_i - n_i y_i) \ln(1 - \hat{y}_i)]$$

Damit erhalten wir:

$$D(Y; \hat{p}) = D_m = -2(\ell_m - \ell_s)\varphi$$

$$= \sum_{i=1}^n n_i \left[y_i \ln \frac{y_i}{\hat{\pi}_i} + (1 - y_i) \ln \frac{1 - y_i}{1 - \hat{\pi}_i} \right]$$

Vergleich von Modellen Sei M_0 ein Untermodell von Modell M , wobei M_0 insgesamt q Parameter anpaßt und M p Parameter ($p > q$). Dann gilt:

- $\frac{D_{M_0} - D_M}{\varphi} \sim \chi_{p-q}^2$
- $\frac{D_{M_0} - D_M}{\hat{\varphi}(p-q)} \sim F_{p-q; n-p}$
- **AIC** = $D_M + 2p\hat{\varphi}$

Residuen

response $y_i - \hat{\mu}_i$

Pearson $\frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$

Working $(y_i - \hat{\mu}_i) \frac{\partial \eta_i}{\partial \mu_i}$

Devianz $\text{sgn}(y_i - \hat{\mu}_i) w_i \left[(y_i \theta(y_i) - b(y_i)) - (y_i \hat{\theta}_i - b(\hat{\theta}_i)) \right]$

Wieder Devianzen

Sei L_m das Likelihood des Modells m und l_m das Loglikelihood. Die Devianz des Modells m wird dann so definiert:

$$D_m = -2(l_m - l_s)$$

wobei s das saturierte Modell ist. (In einigen Fällen kann l_s gleich 0 sein, so daß wir $D_m = -2l_m$ schreiben dürfen.)

[Da φ für die Binomial und Poisson Modelle 1 beträgt, wird φ manchmal vernachlässigt. Vorsicht!]

Die Verteilung von D_m wird manchmal als asymptotisch χ_{n-p}^2 angegeben. Das gilt nur unter beschränkten Umständen, weil die Anzahl von Parametern in s direkt mit n steigen kann (weil wir einen Parameter für jeden Fall brauchen können), so daß die erforderliche Asymptotik nicht greift.

Was (angeblich) besser geht ist die Verteilung des Unterschieds zwischen zwei Devianzen für verschachtelte Modelle:

$$D_{m_2} - D_{m_1} = -2(l_{m_2} - l_{m_1}) \sim \chi_{\nu_2 - \nu_1}^2$$

Es ist mir nicht gelungen, einen vollständigen Beweis aufzutreiben. Einige Autoren weisen auf die Maximum Likelihood Theorie von Likelihood Ratio Tests, aber das kann nicht für unbekannte Varianzen gelten. Ein weiteres Problem besteht in unserem Verständnis der Asymptotik. Was passiert, wenn $n \rightarrow \infty$? Jede Beobachtung ist im Prinzip anders, wenn das Modell stetige erklärende Variablen enthält. Sogar bei einer festen Anzahl von kategorialen Kombinationen ist es nicht ganz klar, wie die Asymptotik verstanden werden sollte.

Trotzdem können (und werden) die Devianzunterschiede so eingesetzt, als ob sie (genau) χ^2 verteilt sind. Da man sowieso Residuen, Graphiken und Metainformationen bei der Wahl eines Modells in Betracht ziehen muß, sollen die Devianzen eher als richtungsweisend benutzt werden.