

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Betrachte lineare Hypothesen

$$H_0 : \mathbf{C}\theta = \mathbf{d} \quad \text{vs.} \quad H_1 : \mathbf{C}\theta \neq \mathbf{d},$$

wobei \mathbf{C} vollen Zeilenrang $s \leq p = \dim(\theta)$ besitze.

Wichtiger Spezialfall:

$$H_0 : \theta_s = 0 \quad \text{vs.} \quad H_1 : \theta_s \neq 0,$$

wobei θ_s einen beliebigen s -dimensionalen Subvektor von θ bezeichnet, zum Beispiel in einem GLM, wo $\beta_s = \mathbf{0}$ bedeutet, dass die zugehörigen Kovariablen nicht signifikant sind.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Likelihood-Quotienten-Statistik

Die Likelihood-Quotienten-Statistik

$$\lambda = 2 \left(\ell(\hat{\theta}) - \ell(\tilde{\theta}) \right) = 2 \log \left[\frac{L(\hat{\theta})}{L(\tilde{\theta})} \right]$$

vergleicht das unrestringierte Maximum der Log-Likelihood $\ell(\hat{\theta})$ (über Θ) mit dem Maximum der Log-Likelihood unter der H_0 -Restriktion, d.h. $\tilde{\theta}$ maximiert $\ell(\theta)$ unter der Nebenbedingung $\mathbf{C}\theta = \mathbf{d}$. Die Struktur eines zugehörigen Tests lautet:

λ zu groß $\Rightarrow H_0$ ablehnen.

Nachteil: Es ist eine numerische Maximierung von $\ell(\theta)$ unter linearer Nebenbedingung notwendig, um $\tilde{\theta}$ zu erhalten.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Wald-Statistik

Die Wald-Statistik

$$w = (\mathbf{C}\hat{\boldsymbol{\theta}} - \mathbf{d})^\top (\mathbf{C}\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}})\mathbf{C}^\top)^{-1} (\mathbf{C}\hat{\boldsymbol{\theta}} - \mathbf{d})$$

misst die (gewichtete) Distanz zwischen der unrestringierten Schätzung $\mathbf{C}\hat{\boldsymbol{\theta}}$ von $\mathbf{C}\boldsymbol{\theta}$ und dem hypothetischen Wert \mathbf{d} unter H_0 . Ein Test wird so konstruiert, dass

w zu groß $\Rightarrow H_0$ ablehnen.

Vorteil gegenüber λ : Keine Berechnung von $\tilde{\boldsymbol{\theta}}$ nötig.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Score- (oder Rao-) Statistik

Die Score-Statistik lautet

$$u = s(\tilde{\theta})^\top \mathbf{I}^{-1}(\tilde{\theta}) s(\tilde{\theta}).$$

Idee: Für $\hat{\theta}$ gilt $s(\hat{\theta}) = 0$. Falls H_1 richtig ist, wird $s(\tilde{\theta})$ deutlich von $0 = s(\hat{\theta})$ verschieden sein, d.h.

u wird groß $\Rightarrow H_0$ ablehnen.

Die Statistik berechnet also den Abstand $s(\tilde{\theta})$ vom Ursprung, gewichtet mit $\mathbf{I}^{-1}(\tilde{\theta})$.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Beispiel 3.8 (Test für einen Subvektor)

Betrachte

- ▶ $H_1 : \eta = \mathbf{x}^\top \boldsymbol{\beta}$ Prädiktor in vollem GLM,
- ▶ $H_0 : \eta_s = \mathbf{x}_s^\top \boldsymbol{\beta}_s$ Prädiktor in reduziertem GLM (nach Weglassen von Kovariablen).

Die Log-Likelihood $\ell(\boldsymbol{\beta}_s)$ im reduzierten Submodell werde durch $\hat{\boldsymbol{\beta}}_s$ maximiert. Mit $\hat{\boldsymbol{\beta}}_s$ und $\hat{\boldsymbol{\beta}}$ lässt sich die Likelihood-Quotienten-Statistik bestimmen.

Für die Wald-Statistik ergibt sich

$$\mathbf{w} = (\hat{\boldsymbol{\beta}})_s^\top \hat{\mathbf{A}}_s^{-1} (\hat{\boldsymbol{\beta}})_s,$$

dabei bezeichne $(\hat{\boldsymbol{\beta}})_s$ die Elemente des Subvektors $\boldsymbol{\beta}_s$ in $\hat{\boldsymbol{\beta}}$ und $\hat{\mathbf{A}}_s$ sei die Teilmatrix von $\hat{\mathbf{A}} = \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$, die diesen Elementen entspricht.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Satz 3.5

Unter H_0 und den gleichen Regularitätsannahmen wie in Satz 3.4 gilt:

$$\lambda, w, u \stackrel{a}{\sim} \chi^2(s).$$

D.h. man lehnt H_0 ab, falls $\lambda, w, u > \chi_{1-\alpha}^2(s)$ ist. Für finite Stichproben besitzen λ, w, u aber unterschiedliche Werte; im Zweifelsfall sollte man λ bevorzugen.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Beweis

- *Beweis für w* : Es gilt

$$\hat{\theta} \stackrel{a}{\sim} N(\theta, \mathbf{I}^{-1}(\hat{\theta}))$$

und damit

$$\mathbf{C}\hat{\theta} \stackrel{a}{\sim} N(\mathbf{C}\theta, \mathbf{C}\mathbf{I}^{-1}(\hat{\theta})\mathbf{C}^T).$$

Unter H_0 folgt

$$\mathbf{C}\hat{\theta} - \underbrace{\mathbf{C}\theta}_{\mathbf{d}} \stackrel{a}{\sim} N(\mathbf{0}, \underbrace{\mathbf{C}\mathbf{I}^{-1}(\hat{\theta})\mathbf{C}^T}_{\mathbf{A}}),$$

also

$$\mathbf{A}^{-1/2}(\mathbf{C}\hat{\theta} - \mathbf{d}) \stackrel{a}{\sim} N(\mathbf{0}, \mathbf{I})$$

und somit

$$\mathbf{w} = (\mathbf{C}\hat{\theta} - \mathbf{d})^T \mathbf{A}^{-1}(\mathbf{C}\hat{\theta} - \mathbf{d}) \stackrel{a}{\sim} \chi^2(s).$$

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

- *Beweis für λ* : Durch Taylorentwicklung kann gezeigt werden, dass $w \stackrel{a}{\sim} \lambda$ und somit $\lambda \stackrel{a}{\sim} \chi^2(s)$. Die Beweisskizze wird hier lediglich für den Spezialfall

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

geführt (das entspricht $\mathbf{C} = \mathbf{I}$, $\mathbf{d} = \theta_0$, $\text{rang}(\mathbf{C}) = p = \dim(\theta)$).

Eine Taylorentwicklung 2. Ordnung von $\ell(\theta_0)$ um den unrestringierten Maximum-Likelihood-Schätzer $\hat{\theta}$ liefert

$$\ell(\theta_0) \approx \ell(\hat{\theta}) + s(\hat{\theta})^\top (\theta_0 - \hat{\theta}) - \frac{1}{2} (\theta_0 - \hat{\theta})^\top \mathbf{J}(\hat{\theta}) (\theta_0 - \hat{\theta}),$$

also wegen $s(\hat{\theta}) = 0$

$$\begin{aligned} \lambda = 2 \left(\ell(\hat{\theta}) - \ell(\theta_0) \right) &\approx (\hat{\theta} - \theta_0)^\top \mathbf{J}(\hat{\theta}) (\hat{\theta} - \theta_0) \\ &\approx (\hat{\theta} - \theta_0)^\top \mathbf{I}(\hat{\theta}) (\hat{\theta} - \theta_0) = w \stackrel{a}{\sim} \chi^2(p). \end{aligned}$$

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

- *Beweis für u* : Wir nehmen denselben Spezialfall wie im Beweis für λ an, also $\tilde{\theta} = \theta_0$. Es ist

$$s(\theta_0) \stackrel{a}{\sim} N(\mathbf{0}, \mathbf{I}(\theta_0))$$

bzw.

$$\mathbf{I}^{-1/2}(\theta_0)s(\theta_0) \stackrel{a}{\sim} N(\mathbf{0}, I),$$

also

$$s(\theta_0)^\top \underbrace{\mathbf{I}^{-\top/2}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)}_{\mathbf{I}(\theta_0)^{-1}} s(\theta_0) \stackrel{a}{\sim} \chi^2(p).$$



3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.2 Konfidenzintervalle

- ▶ Gemeinsamer Konfidenzbereich:

$$(\hat{\theta} - \theta)^{\top} \mathbf{I}(\hat{\theta})(\hat{\theta} - \theta) \stackrel{a}{\sim} \chi^2(p)$$

$$\Rightarrow \mathbb{P}_{\theta} \left((\hat{\theta} - \theta)^{\top} \mathbf{I}(\hat{\theta})(\hat{\theta} - \theta) \leq \chi_{1-\alpha}^2(p) \right) \stackrel{a}{\approx} 1 - \alpha.$$

Daraus lässt sich ein $(1 - \alpha)$ -Konfidenz-Ellipsoid konstruieren.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.2 Konfidenzintervalle

- ▶ Komponentenweise Konfidenzintervalle für θ_j , $j = 1, \dots, p$:

$$\frac{\hat{\theta}_j - \theta_j}{\hat{\sigma}_j} \stackrel{a}{\sim} N(0, 1),$$

wobei $\hat{\sigma}_j^2$ das j -te Diagonalelement von $\widehat{\text{Cov}}(\hat{\theta}) = \mathbf{I}^{-1}(\hat{\theta})$ ist.

Das zugehörige approximative $(1 - \alpha)$ -Konfidenzintervall lautet:

$$\hat{\theta}_j \pm z_{1-\alpha/2} \hat{\sigma}_j.$$

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.3 Modellwahl

Zum Vergleich verschiedener Modelle existieren Modellwahlkriterien, die die Güte der Anpassung, gemessen durch $\ell(\hat{\theta})$, und die Modellkomplexität $p = \dim(\theta)$ bewerten, indem sie die beiden Größen durch eine Straffunktion $\text{pen}(p)$ in einem Kompromiss zu

$$-\ell(\hat{\theta}) + \text{pen}(p)$$

zusammenführen. Dabei wird $-\ell(\hat{\theta})$ klein bei guter Anpassung, $\text{pen}(p)$ groß bei stark bzw. überparametrisierten Modellen.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.3 Modellwahl

Am bekanntesten ist *Akaike's Informationskriterium*

$$\text{AIC} = -2\ell(\hat{\theta}) + 2p$$

mit $\text{pen}(p) = 2p$.

Motivation: $\{f_{\theta}(x) = f(x|\theta), \theta \in \Theta\}$ parametrisiere die betrachteten Modelle und $g(x)$ sei die wahre Dichte für X .

Ziel: Minimiere die Kullback-Leibler-Distanz

$$D(g, f_{\theta}) = \mathbb{E}_X \left(\log \frac{g(X)}{f(X|\theta)} \right) \geq 0,$$

bzw. $\mathbb{E}_Z[\mathcal{K}(f_{\hat{\theta}(Z)}, g)] = \mathbb{E}_Z \mathbb{E}_X[\log g(X) - \log f_{\hat{\theta}(Z)}(X)]$, wenn θ aus gegebenen Daten $Z = z$ geschätzt wird, $X, Z \stackrel{i.i.d.}{\sim} g$.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.3 Modellwahl

Die Akaike Information (ohne Konstanten) $\mathbb{E}_Z \mathbb{E}_X[-\log f_{\hat{\theta}(Z)}(X)]$ ist ein prädiktives Maß für zwei unabhängige Realisationen x und z aus g .

Zur Schätzung liegt die maximierte Loglikelihood $-\log f_{\hat{\theta}(Z)}(Z)$ vor, die jedoch nicht erwartungstreu ist, sondern durch die doppelte Verwendung von Z „überoptimistisch“ bzgl. der Anpassung des Modells. Unter den Regularitätsbedingungen von Satz 3.4 lässt sich zeigen, dass der Bias genau durch $2p$ ausgeglichen wird.

3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.3 Modellwahl

Eine Alternative ist zum Beispiel das *Schwartz- (Bayes-) Informationskriterium*

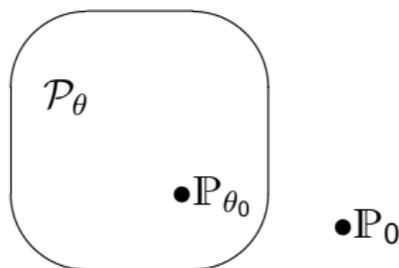
$$\text{BIC} = -2\ell(\hat{\theta}) + p \log n$$

wobei n die Größe des Datensatzes ist. Für $n \geq 8$ „bestraft“ das BIC die Modellkomplexität stärker als das AIC.

Es lässt sich zeigen, dass die Modellwahl basierend auf dem BIC asymptotisch äquivalent ist zur Modellwahl basierend auf sogenannten Bayes-Faktoren, siehe Held, Kapitel 7.2, für eine Herleitung. Die Bayes-Faktoren vergleichen die Posteriori-Modellwahrscheinlichkeiten mit den Priori-Modellwahrscheinlichkeiten.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Bisher haben wir volle (*genuine*) Likelihood-Inferenz betrieben: Gegeben war ein parametrisches statistisches Modell, das heißt eine Familie von Verteilungen oder Dichten mit Parameter $\theta \in \Theta$.
Bisherige Grundannahme: Es existiert ein „wahres“ $\theta_0 \in \Theta$ derart, dass \mathbb{P}_{θ_0} die Verteilung des datengenerierenden Prozesses \mathbb{P}_0 ist, das heißt $\mathbb{P}_{\theta_0} = \mathbb{P}_0$ gilt.



3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Fragen:

- ▶ Was passiert, wenn wir Likelihood-Inferenz innerhalb von \mathcal{P}_θ betreiben, aber der datengenerierende Prozess $\mathbb{P}_0 \notin \mathcal{P}_\theta$ ist (*Fehlspezifikation*)?
- ▶ Was passiert, wenn zwar der Verteilungstyp fehlspezifiziert, jedoch der Erwartungswert korrekt spezifiziert ist (*Quasi-Likelihood*)?
- ▶ Kann man auf die Likelihood verzichten und direkt von den Quasi-ML-Schätzgleichungen

$$\mathbb{E} s(\theta) \stackrel{!}{=} 0$$

starten?

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Beispiel 3.9 (Lineares Modell, vgl. Beispiel 3.12)

Wir betrachten wieder die Standard-Annahme

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \stackrel{i.i.d.}{\sim} N(0, \sigma^2)$$

bzw.

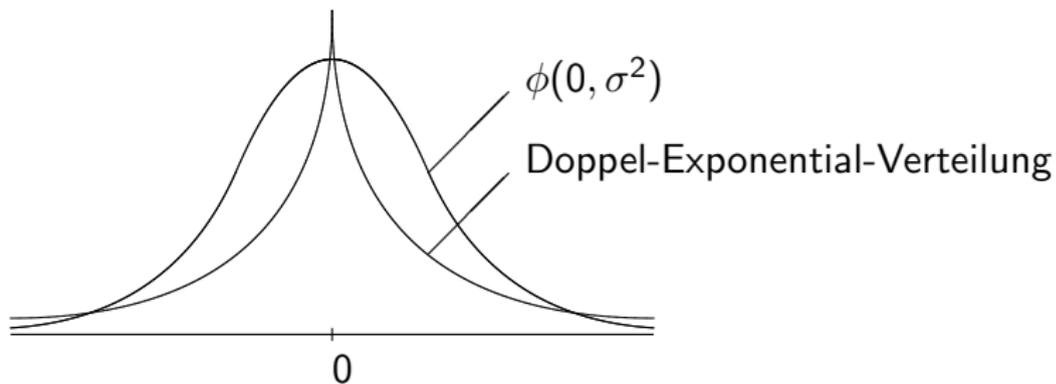
$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \equiv \mathcal{P}_\theta, \quad \theta = (\boldsymbol{\beta}, \sigma^2).$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Mögliche Fehlspezifikationen:

- (a) Die $N(0, \sigma^2)$ -Annahme für die ε_i ist falsch, zum Beispiel könnte die *wahre* Verteilung die Doppel-Exponential-Verteilung (Laplace-Verteilung) sein:

$$f(\varepsilon_i) \propto \exp(-|\varepsilon_i/\sigma|).$$



3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Die Doppel-Exponential-Verteilung (oder auch die Cauchy-/ $t(1)$ -Verteilung) ist spitzer im Zentrum und hat breitere Enden (heavy-tails).

⇒ Sie ist ausreißerunempfindlicher als die Normalverteilung.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

- (b) Die Kovarianzstruktur ist falsch, d.h. $\text{Cov}(\mathbf{y}) \neq \sigma^2 \mathbf{I}$. Wahre Kovarianzstruktur: $\text{Cov}(\mathbf{y}) = \sigma^2 \mathbf{W}$, zum Beispiel
- ▶ $\mathbf{W} = \text{diag}(W_1, \dots, W_n)$ (heteroskedastische Fehler) oder
 - ▶ \mathbf{W} nichtdiagonal (korrelierte Fehler).
- (c) Die Erwartungswertstruktur ist falsch: $\mathbb{E} \mathbf{y} \neq \mathbf{X}\boldsymbol{\beta}$, zum Beispiel wegen
- ▶ Fehlspezifikation nichtlinearer Effekte, zum Beispiel $x\beta_1 + x^2\beta_2$ oder $\beta \log x$,
 - ▶ fehlender Regressoren.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Wir beschränken uns auf den i.i.d. Fall: Seien X_1, \dots, X_n i.i.d. wie $X \sim g(x)$ und $g(x)$ die wahre Dichte. Als statistisches Modell betrachten wir die Familie von Dichten

$$\mathcal{P}_\theta = \left\{ f(x|\theta), \theta \in \Theta \right\}.$$

Falls ein $\theta_0 \in \Theta$ existiert mit $g(x) \equiv f(x|\theta_0)$, so ist das Modell korrekt spezifiziert. Falls kein $\theta_0 \in \Theta$ existiert mit $g(x) \equiv f(x|\theta_0)$, ist das Modell fehlspezifiziert.

$$\left(\begin{array}{c} f(x|\theta) \\ \theta \in \Theta \end{array} \right) \bullet g(x) \sim \mathbb{P}_0$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Definition 3.6 (Kullback-Leibler-Distanz)

Die Kullback-Leibler-Distanz von g und f_θ ist definiert durch

$$D(g, f_\theta) = \mathbb{E}_g \left(\log \frac{g(X)}{f(X|\theta)} \right),$$

d.h.

$$D(g, f_\theta) = \int \log \frac{g(x)}{f(x|\theta)} g(x) dx$$

für X stetig. Dabei wird der Erwartungswert bzgl. der „wahren“ Dichte bzw. Wahrscheinlichkeitsfunktion $g(x)$ gebildet.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Definition 3.6 (fortgeführt)

Es gilt:

$$D(g, f_{\theta}) \geq 0$$

mit

$$D(g, f_{\theta_0}) = 0 \quad \Leftrightarrow \quad g \equiv f_{\theta_0}.$$

Also:

$$D(g, f_{\theta_0}) = 0 \quad \text{für ein } \theta_0 \quad \Leftrightarrow \quad \text{Modell korrekt spezifiziert.}$$

Der Beweis erfolgt mit Ungleichung von Jensen (bewiesen im Beweis von Satz 3.4).

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Bemerkung.

Der (negative) Erwartungswert

$$-\mathbb{E}_g \log g(X) = - \int g(x) \log(g(x)) dx$$

heißt *Entropie* von g .

Sei θ_0 „der“ Minimierer der Kullback-Leibler-Distanz:

$$\theta_0 = \operatorname{argmin}_{\theta \in \Theta} \left[\mathbb{E}_g \left\{ \log g(X) \right\} - \mathbb{E}_g \left\{ \log f(X|\theta) \right\} \right].$$

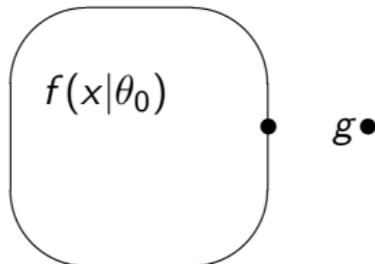
Da $\mathbb{E}_g \left\{ \log g(X) \right\}$ nicht von θ abhängt, gilt auch

$$\theta_0 = \operatorname{argmax}_{\theta \in \Theta} \mathbb{E}_g \left\{ \log f(X|\theta) \right\}.$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Die Dichte $f(x|\theta_0)$ liegt dann im Sinne der Kullback-Leibler-Distanz am „nächsten“ bei g .



3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Der ML-Schätzer ist

$$\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \log f(x_i | \theta).$$

Da $\frac{1}{n} \sum_{i=1}^n \log f(x_i | \theta) \xrightarrow{\mathbb{P}} \mathbb{E}_g \log f(X | \theta)$ (Gesetz der großen Zahlen), gilt vermutlich

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0,$$

das heißt der (Quasi-) ML-Schätzer konvergiert gegen jenes θ_0 , dessen Dichte $f(x | \theta_0)$ am nächsten bei g (bezüglich der Kullback-Leibler-Distanz) liegt.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Satz 3.7 (Asymptotische Eigenschaften des ML-Schätzers bei Misspezifikation)

1. *Konsistenz: Sei θ_0 ein (lokaler) Maximierer von*

$$\mathbb{E}_g \log f(X|\theta)$$

(bzw. ein Minimierer von $D(g, f_\theta)$). Unter Regularitätsannahmen (ähnlich wie bei Satz 3.4) existiert eine Folge $\hat{\theta}_n$ von („Quasi-“) ML-Schätzern, das heißt lokalen Maximierern von

$$\frac{1}{n} \sum_{i=1}^n \log f(x_i|\theta)$$

mit

$$\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta_0.$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Satz 3.7 (fortgeführt)

2. *Asymptotische Normalität: Es gilt*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N \left(\mathbf{0}, \mathbf{J}_1^{-1}(\theta_0) \mathbf{I}_1(\theta_0) \mathbf{J}_1^{-1}(\theta_0) \right)$$

mit

$$\mathbf{I}_1(\theta) \equiv \mathbb{E}_g \left(\underbrace{\left(\frac{\partial \log f(X|\theta)}{\partial \theta} \right)}_{\mathbf{s}_1(\theta)} \underbrace{\left(\frac{\partial \log f(X|\theta)}{\partial \theta} \right)^\top}_{\mathbf{s}_1(\theta)^\top} \right)$$

und der (Quasi-) Fisher-Information

$$\mathbf{J}_1(\theta) = \mathbb{E}_g \left(- \frac{\partial^2 \log f(X|\theta)}{\partial \theta \partial \theta^\top} \right).$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Bemerkung.

- ▶ Beweis: Pawitan S. 372 ff (sowie genaue Regularitätsbedingungen)
- ▶ Falls $g(x) \equiv f(x|\theta_0)$, also das Modell korrekt spezifiziert ist, gilt

$$\mathbf{I}_1(\theta) = \mathbf{J}_1(\theta)$$

(vergleiche Satz 2.16), und man erhält die übliche asymptotische Normalverteilung des ML-Schätzers bei korrekter Modellspezifikation.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

- Informell gilt

$$\hat{\theta}_n \stackrel{a}{\sim} N \left(\theta_0, \underbrace{\frac{1}{n} \mathbf{J}_1^{-1}(\theta_0) \mathbf{I}_1(\theta_0) \mathbf{J}_1^{-1}(\theta_0)}_{\mathbf{V}(\theta_0)} \right),$$

und $\mathbf{V}(\theta_0)$ wird geschätzt durch

$$\hat{\mathbf{V}}(\hat{\theta}_n) = \mathbf{J}^{-1}(\hat{\theta}_n) \mathbf{I}(\hat{\theta}_n) \mathbf{J}^{-1}(\hat{\theta}_n) \quad (\text{„Sandwich“-Matrix})$$

mit

$$\mathbf{I}(\hat{\theta}_n) = \sum_{i=1}^n \mathbf{s}_i(\hat{\theta}_n) \mathbf{s}_i^\top(\hat{\theta}_n) \quad \text{empirische Fisher-Matrix der Stichprobe,}$$

$$\mathbf{J}(\hat{\theta}_n) = - \sum_{i=1}^n \underbrace{\frac{\partial^2 \log f(x_i|\theta)}{\partial \theta \partial \theta^\top}}_{\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^\top}} \Bigg|_{\theta=\hat{\theta}_n} \quad \text{emp. beob. Informations-Matrix.}$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Bemerkung

1. Im i.n.i.d. Fall gilt (informell): Sei $\ell(\theta, x) = \log f(x|\theta)$ und

$$\theta_0 := \operatorname{argmax}_{\theta} \mathbb{E}_g \ell(\theta, X) = \operatorname{argmax}_{\theta} \mathbb{E}_g \left\{ \sum_{i=1}^n \ell_i(\theta, X_i) \right\},$$

bzw. sei θ_0 die Nullstelle von $\mathbb{E}_g \mathbf{s}(\theta)$, das heißt $\mathbb{E}_g(\mathbf{s}(\theta_0)) = \mathbf{0}$. Außerdem

$$\hat{\theta}_n = \operatorname{argmax}_{\theta} \ell(\theta, x) \quad \text{bzw.} \quad \mathbf{s}(\hat{\theta}_n) = \mathbf{0}.$$

Dann gilt

$$\hat{\theta}_n \stackrel{a}{\sim} N\left(\theta_0, \hat{\mathbf{V}}(\hat{\theta}_n)\right)$$

wie oben, nur mit $f_i(x_i|\theta)$ an Stelle von $f(x_i|\theta)$.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

2. Angenommen, der Modellparameter $\tilde{\theta} = (\theta, \alpha)^\top$ setze sich zusammen aus einem eigentlich interessierenden Parameter θ und einem Nuisance-Parameter α . Die Scorefunktion lautet dann

$$s(\theta, \alpha) = \begin{pmatrix} s_\theta(\theta, \alpha) \\ s_\alpha(\theta, \alpha) \end{pmatrix} = \begin{pmatrix} s_\theta(\tilde{\theta}) \\ s_\alpha(\tilde{\theta}) \end{pmatrix}.$$

Falls trotz fehlspezifizierter Likelihood der eigentlich interessierende Parameter die ML-Gleichung $\mathbb{E}_g(s_\theta(\tilde{\theta}_0)) = 0$ erfüllt, so gilt weiterhin

$$\hat{\theta}_n \stackrel{a}{\approx} N\left(\theta_0, \hat{\mathbf{V}}(\hat{\theta}_n)\right) \Rightarrow \text{Quasi-Likelihood.}$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Frage: Lassen sich Parameter von Interesse wie der Mittelwert μ im i.i.d. Fall oder der Kovariablenvektor β im Regressionsfall noch konsistent und asymptotisch normalverteilt schätzen, wenn das statistische Modell nur teilweise fehlspezifiziert bzw. unvollständig spezifiziert ist?

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Beispiel 3.10.

Seien Y_1, \dots, Y_n i.i.d. wie $Y \sim f(Y|\mu, \sigma^2)$, f symmetrisch um μ , aber nicht normal, etwa

$$\mathcal{P}_0 = \left\{ f(y|\mu_0) = \frac{1}{2\sigma} e^{-|y-\mu_0|/\sigma} \right\} \text{ (Laplace- oder Doppel-Exponential-Vtlg).}$$

Trotzdem wählt man die (Log-) Likelihood

$$\text{ql}(\mu) = -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 + \text{const}$$

der Normalverteilung als Quasi-(Log-)Likelihood und maximiert diese.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

So kommt man auf die Quasi-Scorefunktion

$$qs(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu).$$

Es gilt

$$\mathbb{E}_0 qs(\mu_0) = \frac{1}{\sigma^2} \sum_{i=1}^n \underbrace{(\mathbb{E}_0(Y_i) - \mu_0)}_{=\mu_0} = 0,$$

also ist der Minimierer der Kullback-Leibler-Distanz gleich dem wahren μ_0 , $\hat{\mu}_{QML} = \bar{y}$ wie üblich und wegen $\mathbb{E}_0 \bar{Y} = \mu_0$ erwartungstreu.

Allerdings ist \bar{y} kein (asymptotisch) effizienter Schätzer mehr (die Rao-Cramer-Schranke wird nicht erreicht).

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Beispiel 3.11.

Seien Y_1, \dots, Y_n unabhängig, $Y_i \sim N(\mu_0, \sigma_i^2)$ und

$$\mathcal{P}_0 = \left\{ \prod_{i=1}^n \phi(y_i | \mu_0, \sigma_i^2) = \frac{1}{(2\pi)^{n/2} \cdot \prod_{i=1}^n \sigma_i} \exp \left(- \sum_{i=1}^n \frac{1}{2} \frac{(y_i - \mu_0)^2}{\sigma_i^2} \right) \right\}.$$

Dann wählt man als Quasi-Log-Likelihood:

$$\text{ql}(\mu) = -\frac{1}{2} \sum_{i=1}^n \left(\frac{y_i - \mu}{\sigma} \right)^2,$$

das heißt man ignoriert die Abhängigkeit der Varianz von i .

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Man berechnet also

$$qs(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \mu),$$

$$\mathbb{E}_0 qs(\mu) = \frac{1}{\sigma^2} \sum_{i=1}^n (\mu_0 - \mu) = 0 \quad \Leftrightarrow \quad \mu_0 = \mu,$$

$$\hat{\mu}_{\text{QML}} = \bar{y}, \quad \mathbb{E}(\hat{\mu}_{\text{QML}}) = \mu_0 \quad \text{erwartungstreu,}$$

aber

$$\text{Var}_0(\hat{\mu}_{\text{QML}}) = \text{Var}_0(\bar{Y}) = \frac{1}{n^2} \sum_{i=1}^n \text{Var}_0(Y_i) = \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2,$$

das heißt $\hat{\mu}_{\text{QML}} = \bar{y}$ ist ineffizient, aber (falls zum Beispiel $\sigma_i^2 \leq c$) konsistent und normalverteilt.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Beispiel 3.12 (Lineares Modell, vgl. Beispiel 3.9)
Standard–Annahme:

$$y_i | \mathbf{x}_i \sim N(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma^2)$$

bzw.

$$\mathbf{y} | \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) .$$

Mögliche Fehlspezifikationen:

- (a) Normalverteilungsannahme falsch,
- (b) Kovarianzstruktur $\text{Cov } \mathbf{y} = \sigma^2 \mathbf{I}$ falsch,
- (c) Erwartungswertstruktur $\mathbb{E} \mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ falsch.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

zu (a): Dies ist der Fall, wenn \mathbf{y} nicht normalverteilt ist, aber die Kovarianzstruktur und das Erwartungswertmodell korrekt sind.

Es gilt: $\mathbb{E}_0 \mathbf{y} = \mathbf{X} \boldsymbol{\beta}_0$ ist das wahre Modell.

$$s(\boldsymbol{\beta}) = \frac{1}{\sigma^2} \mathbf{X}^\top (\mathbf{y} - \mathbf{X} \boldsymbol{\beta})$$

$$\mathbb{E}_0 s(\boldsymbol{\beta}_0) = 0$$

Dabei ist $\mathbb{E}_0 s(\boldsymbol{\beta}_0)$ der Erwartungswert im wahren Modell mit wahren Parameter.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

zu (a): *fortgeführt*

Es ergibt sich

$$\hat{\beta}_{\text{QML}} = \hat{\beta}_{\text{KQ}} = (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{y}$$

mit

$$\begin{aligned} \mathbb{E}_0(\hat{\beta}_{\text{QML}}) &= (\mathbf{X}^{\top} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbb{E} \mathbf{y} = \beta_0 \quad (\text{erwartungstreu}), \\ \text{Cov}_0(\hat{\beta}_{\text{QML}}) &= \sigma^2 (\mathbf{X}^{\top} \mathbf{X})^{-1}, \end{aligned}$$

also (siehe Beispiel 2.19)

$$\hat{\beta}_{\text{QML}} \stackrel{a}{\approx} N(\beta_0, \sigma^2 (\mathbf{X}^{\top} \mathbf{X})^{-1})$$

wie unter NV-Annahme.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

zu (b): Die wahre Kovarianzmatrix ist $\sigma^2\mathbf{W}$ statt $\sigma^2\mathbf{I}$:

$$\mathbb{P}_0 : \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}_0, \sigma^2\mathbf{W})$$

$$\mathbb{E}_0 s(\boldsymbol{\beta}_0) = 0$$

$$\hat{\boldsymbol{\beta}}_{\text{QML}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\mathbb{E}_0(\hat{\boldsymbol{\beta}}_{\text{QML}}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} \boldsymbol{\beta}_0 = \boldsymbol{\beta}_0$$

$$\begin{aligned} \text{Cov}_0(\hat{\boldsymbol{\beta}}_{\text{QML}}) &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \text{Cov}_0(\mathbf{Y}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ & \quad (\neq \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}) \end{aligned}$$

$\hat{\boldsymbol{\beta}}_{\text{QML}}$ ist konsistent, aber nicht effizient. (Ein effizienter Schätzer wäre der gewichtete KQ- bzw. Aitken-Schätzer $\hat{\boldsymbol{\beta}}_{\text{AITKEN}} = (\mathbf{X}^\top \mathbf{W}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}^{-1} \mathbf{y}$.)

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

zu (c): Der wahre Erwartungswert ist ungleich $\mathbf{X}\beta$:

$$\text{wahrer Erwartungswert: } \mathbb{E}_0 \mathbf{y} = \boldsymbol{\mu}_0 = \mathbf{X}_0 \boldsymbol{\beta}_0$$

$$\Rightarrow \text{wahres Modell: } \mathbf{y} \sim N(\mathbf{X}_0 \boldsymbol{\beta}_0, \sigma^2 \mathbf{I})$$

(falls NV -Annahme und $\sigma^2 \mathbf{I} = \text{Cov}_0(\mathbf{y})$ richtig). Dann ist

$$\hat{\boldsymbol{\beta}}_{QML} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\mathbb{E}_0(\hat{\boldsymbol{\beta}}_{QML}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}_0 \boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_0 .$$

Somit ist $\hat{\boldsymbol{\beta}}_{QML}$ verzerrter Schätzer, aber liefert das best-approximierende lineare Modell mit Designmatrix \mathbf{X} .

Die Kovarianzmatrix ist dann gegeben durch:

$$\text{Cov}_0(\hat{\boldsymbol{\beta}}_{QML}) = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \underbrace{\text{Cov}_0(\mathbf{y})}_{\sigma^2 \mathbf{I}} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} .$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Fazit aus den Beispielen:

- ▶ Falls die Likelihood oder die Varianzstruktur fehlspezifiziert sind, jedoch die Erwartungswertstruktur

$$\mathbb{E}y_i = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$

korrekt spezifiziert ist, erhält man konsistente Schätzer für $\boldsymbol{\mu}$ bzw. $\boldsymbol{\beta}$.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

- ▶ Es genügt sogar, die Nullstelle der Quasi-Scorefunktion

$$qs(\hat{\mu}) \stackrel{!}{=} 0 \quad \text{bzw.} \quad qs(\hat{\beta}) \stackrel{!}{=} 0$$

zu bestimmen. Falls für das „wahre“ μ_0 bzw. β_0

$$\mathbb{E}_0 qs(\mu_0) = 0 \quad \text{bzw.} \quad \mathbb{E}_0 qs(\beta_0) = 0$$

gilt, dann ist die Nullstelle $\hat{\mu}$ bzw. $\hat{\beta}$ konsistent und asymptotisch normalverteilt für μ bzw. β (vgl. Satz 3.7).

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

⇒ Idee der „Schätzgleichungen“ (*estimating equations*):
Definiere eine *Schätzfunktion* oder *Quasi-Scorefunktion*

$$qs(\theta) = \sum_{i=1}^n \psi_i(y_i, \theta)$$

so, dass für den „wahren“ Parameter θ_0

$$\mathbb{E}_0 qs(\theta_0) = \sum_{i=1}^n \mathbb{E}_0[\psi_i(y_i, \theta_0)] = 0$$

erfüllt ist. Dann ist der *Quasi-ML-Schätzer* oder „*M-Schätzer*“
definiert als Nullstelle

$$qs(\hat{\theta}_{QML}) \stackrel{!}{=} 0 \quad (\text{Schätzgleichung})$$

der *Schätzfunktion* $qs(\theta)$.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Beispiel 3.12 (Generalisierte Regression)

Sei

$$\begin{aligned} \mathbb{E}_0 y_i &= \mu_i(\beta) && \text{korrekt spezifiziert ,} \\ \text{Var}_0 y_i &= \phi v_i(\beta) && \text{(eventuell) fehlspezifiziert .} \end{aligned}$$

Es gilt: $\mathbb{E}_0 s(\beta) = 0$.

Es wird nur eine Annahme hinsichtlich der Schätzgleichung getroffen, jedoch nicht für die Verteilung:

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

$$\begin{aligned} s(\boldsymbol{\beta}) &= \frac{1}{\phi} \sum_{i=1}^n \left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) v_i(\boldsymbol{\beta})^{-1} \underbrace{(y_i - \mu_i(\boldsymbol{\beta}))}_{\mathbb{E}_0(y_i) - \mu_i(\boldsymbol{\beta}) = 0} \\ &\propto \sum_{i=1}^n \left(\frac{\partial \mu_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \right) v_i(\boldsymbol{\beta})^{-1} (y_i - \mu_i(\boldsymbol{\beta})) \end{aligned}$$

hat Erwartungswert 0 und

$$s(\hat{\boldsymbol{\beta}}) \stackrel{!}{=} 0 .$$

$\Rightarrow \hat{\boldsymbol{\beta}}$ ist konsistent und asymptotisch normalverteilt.

Speziell: „generalized estimating equation“ (wie in GLM: $\mu_i(\boldsymbol{\beta}) = h(\mathbf{x}_i^\top \boldsymbol{\beta})$).

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Beispiel 3.13 ((Binäre) Longitudinaldaten (repeated measures) oder Clusterdaten)

Die Datenpaare $(y_{ij}, \mathbf{x}_{ij})$, $i = 1, \dots, n$, $j = 1, \dots, n_i$, seien je n_i wiederholte Beobachtungen an Individuen oder in „Clustern“, wie zum Beispiel Familien oder Klassen $i = 1, \dots, n$.

n_i : Anzahl der (zeitlich) wiederholten Beobachtungen pro Individuum oder Cluster

y_{ij} : Zielvariable

\mathbf{x}_{ij} : Kovariablenvektor

$y_{ij}|\mathbf{x}_{ij}$ sei aus einer Exponentialfamilie (normal, binomial, Poisson, ...) mit Erwartungswert

$$\mathbb{E}(y_{ij}|\mathbf{x}_{ij}) = h(\mathbf{x}_{ij}^\top \boldsymbol{\beta}) = \mu_{ij} .$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

Die Schätzgleichungen bei Vernachlässigung von (zeitlichen) Korrelationen zwischen den Messwiederholungen lauten

$$qs(\beta) = \sum_{i=1}^n \sum_{j=1}^{n_i} \mathbf{x}_{ij} w_{ij}(\beta) (y_{ij} - h(\mathbf{x}_{ij}^T \beta)) \stackrel{!}{=} 0$$

mit

$$\mathbb{E}_{\beta_0} qs(\beta_0) = 0 ,$$

wobei die $w_{ij}(\beta)$ geeignete Gewichte sind. Somit ist $\hat{\beta}_{QML}$ konsistent und asymptotisch normal, jedoch unter Effizienzverlust.