3.3 Testen linearer Hypothesen und Konfidenzintervalle

3.3.1 Testen von Hypothesen

Betrachte lineare Hypothesen

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

wobei $\boldsymbol{C}$ 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 $\boldsymbol{\beta}_s = \mathbf{0}$ bedeutet, dass die zugehörigen Kovariablen nicht signifikant sind.

Likelihood-Quotienten-Statistik

Die Likelihood-Quotienten-Statistik

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

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

$$\lambda \text{ zu groß} \Rightarrow H_0 \text{ ablehnen.}$$

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

Wald-Statistik

Die Wald-Statistik

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

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

$$w \text{ zu groß} \Rightarrow H_0 \text{ ablehnen.}$$

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

Score- (oder Rao-) Statistik

Die Score-Statistik lautet

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

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

$$u \text{ wird groß} \Rightarrow H_0 \text{ ablehnen.}$$

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

Beispiel 3.8 (Test für einen Subvektor)**.** *Betrachte*

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

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

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

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

Satz 3.5. *Unter H_0 und den gleichen Regularitätsannahmen wie in Satz 3.4 gilt:*

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

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

Beweis.

- *Beweis für w:* Es gilt

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

  und damit

  $$\boldsymbol{C}\widehat{\theta} \ \overset{a}{\sim} N(\boldsymbol{C}\theta, \boldsymbol{C}\mathbf{I}^{-1}(\widehat{\theta})\boldsymbol{C}^\top).$$

  Unter H_0 folgt

  $$\boldsymbol{C}\widehat{\theta} - \underbrace{\boldsymbol{C}\theta}_{\boldsymbol{d}} \ \overset{a}{\sim} \ N(\boldsymbol{0}, \underbrace{\boldsymbol{C}\mathbf{I}^{-1}(\widehat{\theta})\boldsymbol{C}^\top}_{\boldsymbol{A}}),$$

  also

  $$\boldsymbol{A}^{-1/2}(\boldsymbol{C}\widehat{\theta} - \boldsymbol{d}) \ \overset{a}{\sim} \ N(0, I)$$

  und somit

  $$\boldsymbol{w} = (\boldsymbol{C}\widehat{\theta} - \boldsymbol{d})^T \boldsymbol{A}^{-1} (\boldsymbol{C}\widehat{\theta} - \boldsymbol{d}) \ \overset{a}{\sim} \ \chi^2(s).$$

- *Beweis für λ:* Durch Taylorentwicklung kann gezeigt werden, dass $w \overset{a}{\sim} \lambda$ und somit $\lambda \overset{a}{\sim} \chi^2(s)$. 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 $\boldsymbol{C} = I, \boldsymbol{d} = \theta_0$, $\text{rang}(\boldsymbol{C}) = p = \dim(\theta)$). Eine Taylorentwicklung 2. Ordnung von $\ell(\theta_0)$ um den unrestringierten Maximum-Likelihood-Schätzer $\widehat{\theta}$ liefert

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

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

$$\lambda = 2\left(\ell(\widehat{\theta}) - \ell(\theta_0)\right) \approx (\widehat{\theta} - \theta_0)^\top \mathbf{J}(\widehat{\theta})(\widehat{\theta} - \theta_0) \approx (\widehat{\theta} - \theta_0)^\top \mathbf{I}(\widehat{\theta})(\widehat{\theta} - \theta_0) \overset{a}{\sim} \chi^2(p).$$

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

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

bzw.

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

also

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

□

3.3.2 Konfidenzintervalle

- Gemeinsamer Konfidenzbereich:

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

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

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

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

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

wobei $\widehat{\sigma}_j^2$ das j-te Diagonalelement von $\widehat{\text{Cov}}(\widehat{\theta}) = \mathbf{I}^{-1}(\widehat{\theta})$ ist. Das zugehörige approximative $(1-\alpha)$-Konfidenzintervall lautet:

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

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. Am bekanntesten ist *Akaikes 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. $\mathbb{E}_Z[K(f_{\hat{\theta}(Z)}, g)] = \mathbb{E}_Z \mathbb{E}_X[\log g(X) - \log f_{\hat{\theta}(Z)}(X)]$, wenn θ aus gegebenen Daten $Z = z$ geschätzt wird, $X, Z \overset{i.i.d}{\sim} g$. 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.

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

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

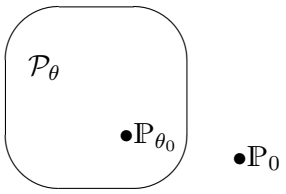
wobei n die Größe des Datensatzes ist. Für $n \geq 8$ „bestraft“ das BIC die Modellkomplexität stärker als das AIC.

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 $\mathbb{P}_{\theta_0}$ die Verteilung des datengenerierenden Prozesses $\mathbb{P}_0$ ist, das heißt $\mathbb{P}_{\theta_0} = \mathbb{P}_0$ gilt.



Fragen:

- Was passiert, wenn wir Likelihood-Inferenz innerhalb von $\mathcal{P}_\theta$ 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?

Beispiel 3.9 (Lineares Modell)**.** *Wir betrachten wieder die Standard-Annahme*

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

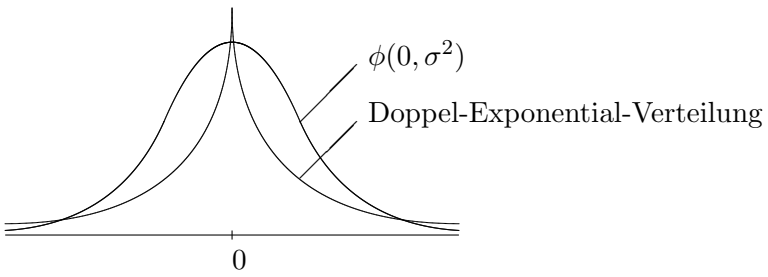
bzw.

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

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:

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



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

$\Rightarrow$ Sie ist ausreißerunempfindlicher.

(b) Die Kovarianzstruktur ist falsch, d.h. $\mathrm{Cov}(\boldsymbol{y}) \neq \sigma^2 \boldsymbol{I}$.

Wahre Kovarianzstruktur: $\mathrm{Cov}(\boldsymbol{y}) = \sigma^2 \boldsymbol{W}$, zum Beispiel

- $\boldsymbol{W} = \mathrm{diag}(W_1, \ldots, W_n)$ (heteroskedastische Fehler) oder
- $\boldsymbol{W}$ nichtdiagonal (korrelierte Fehler).

(c) Die Erwartungswertstruktur ist falsch: $\mathbb{E}\,\boldsymbol{y} \neq \boldsymbol{X}\boldsymbol{\beta}$, zum Beispiel wegen

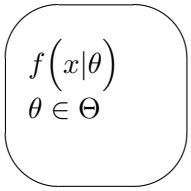
- Fehlspezifikation nichtlinearer Effekte, zum Beispiel $x\beta_1 + x^2\beta_2$ oder $\beta \log x$,
- fehlender Regressoren.

3.4.1 ML-Schätzung bei Fehlspezifikation

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

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

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.



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.

Es gilt:

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

mit

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

Also:

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

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

Bemerkung. *Der (negative) Erwartungswert*

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

heißt Entropie *von* g.

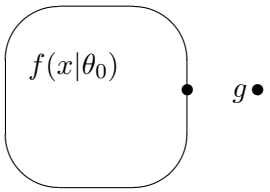
Sei θ_0 „der" Minimierer der Kullback-Leibler-Distanz:

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

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

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

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



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.

Genauer gilt:

Satz 3.7 (Asymptotische Eigenschaften des ML-Schätzers bei Missspezifikation)**.**

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

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

2. *Asymptotische Normalität: Es gilt*

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

mit

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

und der (Quasi-) Fisher-Information

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

Bemerkung.

- *Falls $g(x) \equiv f(x|\theta_0)$, also das Modell korrekt spezifiziert ist, gilt*

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

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

- *Informell gilt*

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

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

$$\hat{\boldsymbol{V}}(\hat{\theta}_n) = \boldsymbol{J}^{-1}(\hat{\theta}_n)\, \boldsymbol{I}(\hat{\theta}_n)\, \boldsymbol{J}^{-1}(\hat{\theta}_n) \qquad (\text{„Sandwich"-Matrix})$$

mit

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

$$\boldsymbol{J}(\hat{\theta}_n) = -\sum_{i=1}^n \underbrace{\frac{\partial^2 log\, f(x_i|\theta)}{\partial \theta\, \partial \theta^\top}}_{\frac{\partial^2 l(\theta)}{\partial \theta\, \partial \theta^\top}}\Bigg|_{\theta=\hat{\theta}_n} \qquad \textit{empirische beobachtete Informations-Matrix.}$$

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* $\mathbb{E}_g\, \boldsymbol{s}(\theta)$*, das heißt* $\mathbb{E}_g(\boldsymbol{s}(\theta_0)) = \mathbf{0}$*. Außerdem*

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

   Dann gilt

   $$\hat{\theta}_n \overset{a}{\sim} N\Big(\theta_0, \hat{\boldsymbol{V}}(\hat{\theta}_n)\Big)$$

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

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

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

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

   $$\hat{\theta}_n \overset{a}{\sim} N\Big(\theta_0, \hat{\boldsymbol{V}}(\hat{\theta}_n)\Big) \quad \Rightarrow \quad \textit{Quasi-Likelihood.}$$

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 $\boldsymbol{\beta}$ im Regressionsfall noch konsistent und asymptotisch normalverteilt schätzen, wenn das statistische Modell nur teilweise fehlspezifiziert bzw. unvollständig spezifiziert ist?

Beispiel 3.10. *Seien* $Y_1, \ldots, Y_n$ *i.i.d. wie* $Y \sim f(Y|\mu, \sigma^2)$*,* f *symmetrisch um* μ*, aber nicht normal, etwa*

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

Trotzdem wählt man die (Log-) Likelihood

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

der Normalverteilung als Quasi-(Log-)Likelihood und maximiert diese. So kommt man auf die Quasi-Scorefunktion

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

Es gilt

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

also $\hat{\mu}_{QML} = \bar{y}$ 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).

Beispiel 3.11. *Seien $Y_1, \ldots, Y_n$ 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:

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

das heißt man ignoriert die Abhängigkeit der Varianz von i und berechnet

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

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

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

aber

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

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

Beispiel 3.12 (Lineares Modell)**.** *Standard–Annahme:*

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

bzw.

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

Mögliche Fehlspezifikationen:

(a) Normalverteilungsannahme falsch,

(b) Kovarianzstruktur $\operatorname{Cov} \boldsymbol{y} = \sigma^2 \boldsymbol{I}$ *falsch,*

(c) Erwartungswertstruktur $\mathbb{E}\, \boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ *falsch.*

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

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

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

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

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

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

mit

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

also

$$\hat{\boldsymbol{\beta}}_{QML} \overset{a}{\sim} N(\boldsymbol{\beta}_0, \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1})$$

wie unter NV-Annahme.

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

$$\begin{aligned} &\mathbb{P}_0 : \boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta_0}, \sigma^2 \boldsymbol{W}) \\ &\mathbb{E}_0 s(\boldsymbol{\beta}_0) = 0 \\ &\hat{\boldsymbol{\beta}}_{QML} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \\ &\mathbb{E}_0(\hat{\boldsymbol{\beta}}_{QML}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta_0} = \boldsymbol{\beta_0} \end{aligned}$$

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

$\hat{\boldsymbol{\beta}}_{QML}$ *ist konsistent, aber nicht effizient.*

(Ein effizienter Schätzer wäre der gewichtete KQ– bzw. Aitken–Schätzer $\hat{\boldsymbol{\beta}}_{AITKEN} = (\boldsymbol{X}^\top \boldsymbol{W}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{-1} \boldsymbol{y}$*.)*

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

$$\begin{aligned} \textit{wahrer Erwartungswert:} \quad & \mathbb{E}_0 \boldsymbol{y} = \boldsymbol{\mu}_0 = \boldsymbol{X}_0 \boldsymbol{\beta}_0 \\ \Rightarrow \textit{wahres Modell:} \quad & \boldsymbol{y} \sim N(\boldsymbol{X}_0 \boldsymbol{\beta}_0, \sigma^2 \boldsymbol{I}) \end{aligned}$$

(falls N und $\sigma^2 \boldsymbol{I} = \mathrm{Cov}_0(\boldsymbol{y})$ richtig). Dann ist

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

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

Somit ist $\hat{\boldsymbol{\beta}}_{QML}$ verzerrter Schätzer, aber liefert das best–approximierende lineare Modell mit Designmatrix $\boldsymbol{X}$. Die Kovarianzmatrix ist dann gegeben durch:

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

Fazit aus den Beispielen:

- Falls die Likelihood oder die Varianzstruktur fehlspezifiziert sind, jedoch die Erwartungswertstruktur
$$\mathbb{E} y_i = \mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$$
korrekt spezifiziert ist, erhält man konsistente Schätzer für $\boldsymbol{\mu}$ bzw. $\boldsymbol{\beta}$.
- Es genügt sogar, die Nullstelle der Quasi–Scorefunktion
$$\mathrm{qs}(\hat{\boldsymbol{\mu}}) \stackrel{!}{=} 0 \quad \text{bzw.} \quad \mathrm{qs}(\hat{\boldsymbol{\beta}}) \stackrel{!}{=} 0$$
zu bestimmen. Falls für das „wahre“ $\boldsymbol{\mu}_0$ bzw. $\boldsymbol{\beta}_0$
$$\mathbb{E}_0 \, \mathrm{qs}(\boldsymbol{\mu}_0) = 0 \; , \;\; \mathbb{E}_0 \, \mathrm{qs}(\boldsymbol{\beta}_0) = 0$$
gilt, dann ist die Nullstelle $\hat{\boldsymbol{\mu}}$ bzw. $\hat{\boldsymbol{\beta}}$ konsistent und asymptotisch normalverteilt für $\boldsymbol{\mu}$ bzw. $\boldsymbol{\beta}$.

$\Rightarrow$ Idee der *„Schätzgleichungen“ (estimating equations)*:

Definiere eine *Schätzfunktion* oder *Quasi–Scorefunktion*

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

so, dass für den „wahren“ Parameter θ_0

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

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

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

der *Schätzfunktion* $\mathrm{qs}(\theta)$.

Beispiel 3.13 (Generalisierte Regression)**.** *Sei*

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

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

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

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

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:* $\mu_i(\boldsymbol{\beta}) = h(\boldsymbol{x}_i^\top \boldsymbol{\beta})$*).*

Beispiel 3.14 ((Binäre) Longitudinaldaten (repeated measures) oder Clusterdaten)**.** *Die Datenpaare* $(y_{ij}, \boldsymbol{x}_{ij})\,,\ i = 1, \ldots, n\,,\ j = 1, \ldots, n_i$*, seien je* n_i *wiederholte Beobachtungen an Individuen oder in „Clustern“, wie zum Beispiel Familien oder Klassen* $i = 1, \ldots, n$*.*

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

y_{ij}*: Zielvariable*

$\boldsymbol{x}_{ij}$*: Kovariablenvektor*

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

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

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

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

mit

$$\mathbb{E}_{\boldsymbol{\beta}_0}\, \mathrm{qs}(\boldsymbol{\beta}_0) = 0\,,$$

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