Inferenz nach Modellwahl
In der Praxis ist es häufig so, dass man in einem ersten Schritt ein Modell auswählt und in einem zweiten Schritt in diesem Modell Inferenz durchführt (Schätzen, Testen, Konfidenzintervalle).

Werden Modellselektion und Inferenz auf getrennten Datensätzen durchgeführt (z.B. Datensatzsplitting), führt dies zu valider Inferenz im zweiten Schritt.

Werden Modellselektion und Inferenz auf dem gleichen Datensatz durchgeführt, haben Tests oder Konfidenzintervalle im gewählten Modell nicht mehr die nominalen Eigenschaften.

Beispiel

Sei $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n.$

Für $\beta_1 = \beta_2 = 0$:

1. Teste im vollen Modell $H_0 : \beta_1 = 0$. Dann hat der p-Wert eine Gleichverteilung auf $[0, 1]$ und einen Fehler 1. Art von $\alpha = 0.05$.
2. Die Hypothese ist von der Selektion abhängig:
   Selektiere zunächst das Modell
   $M2 : y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$ oder
   $M3 : y_i = \beta_0 + \beta_2 x_2 + \varepsilon_i$
   mit dem kleineren AIC und teste dann für einen Effekt der selektierten Variable. Der p-Wert ist nicht mehr gleichverteilt und führt zu einen Fehler 1. Art von ca. 10%.

Beispiel fortgeführt

3. Testdurchführung ist von der Selektion ins Modell abhängig:
   Selektiere zunächst das Modell
   $M1 : y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$ oder
   $M2 : y_i = \beta_0 + \beta_1 x_1 + \varepsilon_i$ oder
   $M3 : y_i = \beta_0 + \beta_2 x_2 + \varepsilon_i$ oder
   $M4 : y_i = \beta_0 + \varepsilon_i$
   mit dem kleinsten AIC. Teste für $H_0 : \beta_1 = 0$, falls x_1 ins Modell aufgenommen wurde. Der p-Wert ist nicht mehr gleichverteilt.
   (i) Zählt man die Nicht-Aufnahme ins Modell als Nichtablehnung von H_0, so ist der Fehler 1. Art (hier) ca. 5%.
   (ii) Unter den durchgeführten Tests - also bedingt auf die Selektion von x_1 ins Modell - ist der Fehler 1. Art ca. 30%.

Von der Idee her muss die Antwort, die valide Inferenz gibt, berücksichtigen, *dass die Frage gestellt wurde*, also z.B. dass die Variable im Modell ist (siehe Fithian, Sun, Taylor, 2015, Optimal Inference After Model Selection, ArXiv-preprint 1410.2597v2).

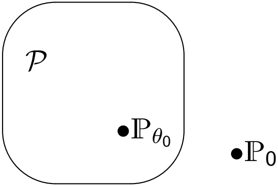
3.(ii) legt nahe, dass man bei der Inferenz *auf die Selektion des Modells bedingen muss*. Intuitiv: verwende die Information in den Daten, die bereits zur Modellselektion verwendet wurde, nicht mehr für die Inferenz, sondern bedinge auf sie. (Analog zum Datensplitting, aber effizienter.)

Die aktuelle Forschung beschäftigt sich mit valider Inferenz nach Modellselektion, siehe z.B. obiges Paper sowie weitere Artikel u.a. unter dem Stichwort *Post-Selection Inference*.

Das sich noch in der Entwicklung befindende R-Paket `selectiveInference` ermöglicht valide Inferenz nach Modellwahl mit Lasso, dem LARS-Algorithmus oder Forward Stepwise Regression. Beispielcodes zu den Funktionen finden sich in der Hilfe `?selectiveInference`.

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 $\boldsymbol{\theta} \in \Theta$.
Bisherige Grundannahme: Es existiert ein „wahres" $\boldsymbol{\theta}_0 \in \Theta$ 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.



3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Fragen:

3.4.1 Was passiert, wenn wir Likelihood-Inferenz innerhalb von $\mathcal{P}$ betreiben, aber der datengenerierende Prozess $\mathbb{P}_0 \notin \mathcal{P}$ ist *(Fehlspezifikation)*?

3.4.2 Was passiert, wenn zwar der Verteilungstyp fehlspezifiziert, jedoch der Erwartungswert korrekt spezifiziert ist *(Quasi-Likelihood)*?

3.4.2 Kann man auf die Likelihood verzichten und direkt von den Quasi-ML-Schätzgleichungen

$$\mathbf{qs}(\widehat{\theta}|\mathbf{X}) \stackrel{!}{=} \mathbf{0}$$

starten?

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

Beispiel 3.8 (Lineares Modell, vgl. Beispiel 3.11)
Wir betrachten wieder die Standard-Annahme

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

bzw.

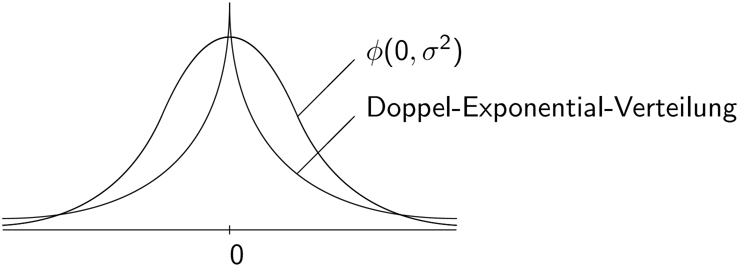
$$\mathbf{y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \equiv \mathcal{P}, \quad \boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2).$$

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:

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



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).

$\Rightarrow$ Sie ist ausreißerunempfindlicher als die Normalverteilung.

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

(b) Die Kovarianzstruktur ist falsch, d.h. $\mathrm{Cov}(\mathbf{y}) \neq \sigma^2 \mathbf{I}_n$. Wahre Kovarianzstruktur: $\mathrm{Cov}(\mathbf{y}) = \sigma^2 \mathbf{W}$, zum Beispiel

- $\mathbf{W} = \mathrm{diag}(W_1, \ldots, W_n)$ (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 $X_1, \ldots, X_n$ i.i.d. wie $X \sim g(x)$ und $g(x)$ die wahre Dichte. Als statistisches Modell betrachten wir die Familie von Dichten

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

Falls ein $\boldsymbol{\theta}_0 \in \Theta$ existiert mit $g(x) \equiv f(x|\boldsymbol{\theta}_0)$, so ist das Modell korrekt spezifiziert. Falls kein $\boldsymbol{\theta}_0 \in \Theta$ existiert mit $g(x) \equiv f(x|\boldsymbol{\theta}_0)$, ist das Modell fehlspezifiziert.

$$\boxed{\begin{array}{l} f\,(x|\boldsymbol{\theta}) \\ \boldsymbol{\theta} \in \Theta \end{array}} \qquad \bullet g(x) \sim \mathbb{P}_0$$

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

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

d.h.

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

für X stetig. Dabei wird der Erwartungswert $\mathbb{E}_g$ bzgl. der „wahren" Dichte bzw. Wahrscheinlichkeitsfunktion $g(x)$ gebildet.

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:

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

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

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Bemerkung.

Der (negative) Erwartungswert

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

heißt *Entropie* von g.

Sei $\boldsymbol{\theta}_0$ „der" Minimierer der Kullback-Leibler-Distanz:

$$\boldsymbol{\theta}_0 = \underset{\theta\in\Theta}{\operatorname{argmin}} \left(\, \mathbb{E}_g[\log\, g(X)] - \mathbb{E}_g[\log\, f(X|\boldsymbol{\theta})]\,\right).$$

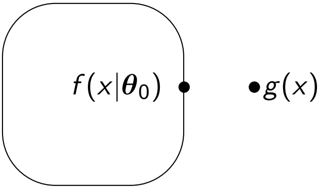
Da $\mathbb{E}_g[\log\, g(X)]$ nicht von $\boldsymbol{\theta}$ abhängt, gilt auch

$$\boldsymbol{\theta}_0 = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{E}_g\Big[\log\, f(X|\boldsymbol{\theta})\Big].$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Die Dichte $f(x|\boldsymbol{\theta}_0)$ 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

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

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

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0,$$

das heißt der (Quasi-) ML-Schätzer konvergiert gegen jenes $\boldsymbol{\theta}_0$, dessen Dichte $f(x|\boldsymbol{\theta}_0)$ am nächsten bei $g(x)$ (bezüglich der Kullback-Leibler-Distanz) liegt.

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

1. *Konsistenz: Sei* $\boldsymbol{\theta}_0$ *ein (lokaler) Maximierer von*

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

(bzw. ein Minimierer von $D(g, f_\theta)$*). Unter Regularitätsannahmen (ähnlich wie bei Satz 3.4) existiert für* $X_1, ..., X_n$ *iid. wie* X *eine Folge* $\hat{\boldsymbol{\theta}}_n$ *von („Quasi-") ML-Schätzern, das heißt lokalen Maximierern von*

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

mit

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{\mathbb{P}} \boldsymbol{\theta}_0.$$

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Satz 3.7 (fortgeführt)

2. *Asymptotische Normalität: Es gilt*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{qi}^{-1}(\boldsymbol{\theta}_0)\,\mathrm{Cov}(\mathbf{s}_1(\boldsymbol{\theta}_0|X))\,\mathbf{qi}^{-1}(\boldsymbol{\theta}_0)\right)$$

mit

$$\mathrm{Cov}(\mathbf{s}_1(\boldsymbol{\theta}_0|X)) = \mathbb{E}_g\Bigg[\underbrace{\left(\frac{\partial \log f(X|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)}_{\mathbf{s}_1(\boldsymbol{\theta}_0|X)}\underbrace{\left(\frac{\partial \log f(X|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)^{\top}}_{\mathbf{s}_1(\boldsymbol{\theta}_0|X)^{\top}}\Bigg]$$

und der (Quasi-) Fisher-Information

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

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Bemerkungen

- Beweis: Pawitan S. 372 ff (sowie genaue Regularitätsbedingungen)

- Falls $g(x) \equiv f(x|\boldsymbol{\theta}_0)$, also das Modell korrekt spezifiziert ist, gilt

$$\mathrm{Cov}(\mathbf{s}_1(\boldsymbol{\theta}_0|X)) = \mathbf{qi}(\boldsymbol{\theta}_0) = \mathbf{i}(\boldsymbol{\theta}_0)$$

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

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

- Informell gilt

$$\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N\left(\boldsymbol{\theta}_0, \underbrace{\frac{1}{n}\mathbf{qi}^{-1}(\boldsymbol{\theta}_0)\,\mathrm{Cov}(\mathbf{s}_1(\boldsymbol{\theta}_0|X))\,\mathbf{qi}^{-1}(\boldsymbol{\theta}_0)}_{\mathbf{V}(\boldsymbol{\theta}_0)}\right),$$

  und $\mathbf{V}(\boldsymbol{\theta}_0)$ wird geschätzt durch

$$\widehat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_n) = \mathbf{J}^{-1}(\hat{\boldsymbol{\theta}}_n|\mathbf{x})\,\mathbf{C}(\hat{\boldsymbol{\theta}}_n|\mathbf{x})\,\mathbf{J}^{-1}(\hat{\boldsymbol{\theta}}_n|\mathbf{x}) \qquad \text{(„Sandwich"-Matrix)}$$

  mit

$$\mathbf{C}(\hat{\boldsymbol{\theta}}_n|\mathbf{x}) = \sum_{i=1}^{n} \mathbf{s}_1(\hat{\boldsymbol{\theta}}_n|x_i)\,\mathbf{s}_1^\top(\hat{\boldsymbol{\theta}}_n|x_i)$$

  n-fache empirische Kovarianzmatrix von $\mathbf{s}_1(\boldsymbol{\theta}_0|X)$

$$\mathbf{J}(\hat{\boldsymbol{\theta}}_n|\mathbf{x}) = -\sum_{i=1}^{n} \underbrace{\frac{\partial^2 \log f(x_i|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\top}}_{\frac{\partial^2 \ell(\boldsymbol{\theta}|x_i)}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\top}}\Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_n}$$

  empirische beobachtete Informationsmatrix der Stichprobe

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.1 ML-Schätzung bei Fehlspezifikation

Bemerkungen

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

$$\boldsymbol{\theta}_0 := \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_g[\,\ell(\boldsymbol{\theta}|\mathbf{X})\,] = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_g\left[\sum_{i=1}^{n} \ell_i(\boldsymbol{\theta}|X_i)\right],$$

bzw. sei $\boldsymbol{\theta}_0$ die Nullstelle von $\mathbb{E}_g\left[\mathbf{s}(\boldsymbol{\theta}|\mathbf{X})\right]$, das heißt $\mathbb{E}_g\left[\mathbf{s}(\boldsymbol{\theta}_0|\mathbf{X})\right] = \mathbf{0}$. Außerdem

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

Dann gilt

$$\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N\Big(\boldsymbol{\theta}_0, \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_n)\Big)$$

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

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

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

Falls trotz fehlspezifizierter Likelihood der eigentlich interessierende Parameter $\mathbb{E}_g[\mathbf{s}_\theta(\widetilde{\boldsymbol{\theta}}_0|\mathbf{X})] = \mathbf{0}$ erfüllt, so gilt weiterhin

$$\hat{\boldsymbol{\theta}}_n \overset{a}{\sim} N\Big(\boldsymbol{\theta}_0, \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_n)\Big) \quad \Rightarrow \quad \text{Quasi-Likelihood.}$$

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

Beispiel 3.9

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

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

Trotzdem wählt man die (Log-) Likelihood

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

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

So kommt man auf die Quasi-Scorefunktion

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

Es gilt

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

also ist der Minimierer der Kullback-Leibler-Distanz gleich dem wahren μ_0, $\hat{\mu}_{\mathrm{QML}} = \bar{y}$ wie üblich, wegen $\mathbb{E}_0[\bar{Y}] = \mu_0$ erwartungstreu und wegen Satz 3.7 konsistent und asymptotisch normalverteilt.

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

Beispiel 3.10
Seien $Y_1, \ldots, Y_n$ unabhängig, $Y_i \sim N(\mu_0, \sigma_i^2)$ und

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

Dann wählt man als Quasi-Log-Likelihood:

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

das heißt man ignoriert die Abhängigkeit der Varianz von i.

Man berechnet also

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

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

$$\hat{\mu}_{\mathrm{QML}} = \bar{y}, \quad \mathbb{E}[\hat{\mu}_{\mathrm{QML}}] = \mu_0 \quad \text{erwartungstreu},$$

aber

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

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

Beispiel 3.11 (Lineares Modell, vgl. Beispiel 3.8)
Standard–Annahme:

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

bzw.

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

Mögliche Fehlspezifikationen:

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

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 der Erwartungswert im wahren Modell.

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

$$\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta}_0)] = \mathbf{0}$$

Dabei ist $\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta}_0)]$ der Erwartungswert bzgl. der Verteilung von $\mathbf{y}$ im wahren Modell mit wahrem Parameter.

zu (a): *fortgeführt*

Es ergibt sich

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

mit

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

also (siehe Beispiel 2.19)

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

wie unter NV-Annahme.

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

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

$$\begin{aligned}
\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta}_0)] &= \mathbf{0} \\
\hat{\boldsymbol{\beta}}_{\mathsf{QML}} &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} \\
\mathbb{E}_0[\hat{\boldsymbol{\beta}}_{\mathsf{QML}}] &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0 \\
\mathrm{Cov}_0(\hat{\boldsymbol{\beta}}_{\mathsf{QML}}) &= (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathrm{Cov}_0(\mathbf{Y})\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} \\
(&\neq \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\ )
\end{aligned}$$

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

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

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

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

(falls NV-Annahme und $\mathrm{Cov}_0(\mathbf{y}) = \sigma^2\mathbf{I}_n$ 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:

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

Fazit aus den Beispielen:

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

$$\mathbb{E}_0[y_i] = \mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$$

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

- Es genügt sogar, die Nullstelle der Quasi–Scorefunktion

$$\mathrm{qs}(\hat{\mu}|\mathbf{y}) \stackrel{!}{=} 0 \quad \text{bzw.} \quad \mathbf{qs}(\hat{\boldsymbol{\beta}}) \stackrel{!}{=} \mathbf{0}$$

zu bestimmen. Falls für das „wahre“ μ_0 bzw. $\boldsymbol{\beta}_0$

$$\mathbb{E}_0[\mathrm{qs}(\mu_0|\mathbf{Y})] = 0 \quad \text{bzw.} \quad \mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta}_0)] = \mathbf{0}$$

gilt, dann ist die Nullstelle $\hat{\mu}$ bzw. $\hat{\boldsymbol{\beta}}$ konsistent und asymptotisch normalverteilt für μ_0 bzw. $\boldsymbol{\beta}_0$ (vgl. Satz 3.7).

3.4 Fehlspezifikation, Quasi-Likelihood und Schätzgleichungen

3.4.2 Quasi-Likelihood und Schätzgleichungen

$\Rightarrow$ Idee der *„Schätzgleichungen“ (estimating equations)*:
Definiere eine *Schätzfunktion* oder *Quasi–Scorefunktion*

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

so, dass für den „wahren“ Parameter $\boldsymbol{\theta}_0$

$$\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\theta}_0|\mathbf{y})] = \sum_{i=1}^{n} \mathbb{E}_0[\psi_i(y_i, \boldsymbol{\theta}_0)] = \mathbf{0}$$

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

$$\mathbf{qs}(\hat{\boldsymbol{\theta}}_{QML}|\mathbf{y}) \stackrel{!}{=} \mathbf{0} \quad (\textit{Schätzgleichung})$$

der *Schätzfunktion* $\mathbf{qs}(\boldsymbol{\theta}|\mathbf{y})$ konsistent und asymptotisch normalverteilt.

Beispiel 3.12 (Generalisierte Regression)

Sei

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

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

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

Es gilt: $\mathbb{E}_0[\mathbf{qs}(\boldsymbol{\beta})] = \mathbf{0}$ und

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

$\Rightarrow \hat{\boldsymbol{\beta}}$ ist konsistent und asymptotisch normalverteilt.
Speziell: *„generalized estimating equation“* (wie in GLM: $\mu_i(\boldsymbol{\beta}) = h(\mathbf{x}_i^\top\boldsymbol{\beta})$).

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

Die Datenpaare $(y_{ij}, \mathbf{x}_{ij})\,,\ i = 1, \ldots, n\,,\ j = 1, \ldots, n_i$, 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

$\mathbf{x}_{ij}$: Kovariablenvektor

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

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

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

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

mit

$$\mathbb{E}_0\Big[\mathbf{qs}(\boldsymbol{\beta}_0)\Big] = \mathbf{0}\ ,$$

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