2.1 Klassische Schätztheorie

2.1.4 Erwartungstreue Schätzer

- „Schöne" Resultate für finites n, aber für vergleichsweise einfache statistische Modelle.
- Problem: Für komplexere Modelle existieren keine „vernünftigen" erwartungstreuen Schätzer.
- Aber: Etliche Resultate besitzen allgemeine Eigenschaften für $n \to \infty$.

Informationsungleichungen

I. $\theta \in \mathbb{R}$ (skalar). Neben θ werden auch transformierte Parameter $\tau(\theta)$ betrachtet. Wenn Ableitungen benötigt werden, nehmen wir stillschweigend an, dass sie existieren.

Satz 2.18

Sei $f(x|\theta)$ Fisher-regulär.

1. *Ist $\widehat{\theta}$ erwartungstreu für θ, so gilt:*

$$\mathrm{Var}_\theta(\widehat{\theta}) \geq \frac{1}{I(\theta)} \qquad \textit{(Cramer-Rao-Ungleichung)}.$$

2. *Ist $T = T(X)$ erwartungstreu für $\tau(\theta)$, so gilt:*

$$\mathrm{Var}_\theta(T) \geq \frac{(\tau'(\theta))^2}{I(\theta)}.$$

   $\frac{(\tau'(\theta))^2}{I(\theta)}$ *heißt Cramer-Rao-Schranke.*

3. *Besitzt $\widehat{\theta}$ den Bias $B(\theta) = \mathbb{E}_\theta[\widehat{\theta}] - \theta$, so gilt*

$$\mathrm{MSE}_\theta(\widehat{\theta}) \geq B^2(\theta) + \frac{(1 + B'(\theta))^2}{I(\theta)}.$$

Beweis.
Gezeigt wird 2. Daraus folgt 1. für $\tau(\theta) = \theta$ und 3. für $\tau(\theta) = \theta + B(\theta)$.
Differentiation von

$$\tau(\theta) = \mathbb{E}_\theta[T] = \int T(x) f(x|\theta)\, dx$$

bezüglich θ, und Verwendung der Fisher-Regularität liefert:

$$\begin{aligned} \tau'(\theta) &= \int T(x) \frac{d}{d\theta} f(x|\theta)\, dx \\ &= \int T(x) s(\theta; x) f(x|\theta)\, dx \\ &= \mathrm{Cov}_\theta(T(X), s(\theta; X)). \end{aligned}$$

Beweis fortgeführt.
Unter Verwendung der Cauchy-Schwarz-Ungleichung

$$|\mathrm{Cov}(U, V)| \leq \sqrt{\mathrm{Var}(U)}\sqrt{\mathrm{Var}(V)}$$

folgt

$$\begin{aligned} (\tau'(\theta))^2 &\leq \mathrm{Var}_\theta(T(X))\mathrm{Var}_\theta(s(\theta; X)) \\ &= \mathrm{Var}_\theta(T(X))I(\theta). \end{aligned}$$

Also:

$$\mathrm{Var}_\theta(T(X)) \geq \frac{(\tau'(\theta))^2}{I(\theta)}.$$

□

Bemerkung.
Die Gleichheit wird genau dann angenommen, wenn eine einparametrische Exponentialfamilie $f(x|\theta) = h(x)\exp(b(\theta) + \gamma(\theta)T(x))$ vorliegt. In diesem Fall ist $T(x)$ ein effizienter Schätzer für seinen Erwartungswert $\tau(\theta) = -b'(\theta)/\gamma'(\theta)$. Also: eher eine kleine Modellklasse.

II. **Informationsungleichungen** für θ bzw. $\boldsymbol{\tau}(\theta)$ mehrdimensional.

Satz 2.19

Sei $f(x|\theta)$ Fisher-regulär.

1. *Ist $\widehat{\theta}$ erwartungstreu für $\theta = (\theta_1, \ldots, \theta_p)$, so gilt:*

$$\mathrm{Cov}_\theta(\widehat{\theta}) \geq \boldsymbol{I}^{-1}(\theta),$$

wobei sich „$\geq$" auf die Löwner-Ordnung bezieht (vgl. F. 121). Daraus folgt insbesondere $\mathrm{Var}_\theta(\widehat{\theta}_j) \geq v_{jj}$, $j = 1, \ldots, p$, wobei v_{jj} das j-te Diagonalelement von $\boldsymbol{I}^{-1}(\theta)$ bezeichnet.

2. *Ist $\boldsymbol{T}$ erwartungstreu für $\boldsymbol{\tau}(\theta)$, so gilt*

$$\mathrm{Cov}_\theta(\boldsymbol{T}) \geq \boldsymbol{H}(\theta)\boldsymbol{I}^{-1}(\theta)\boldsymbol{H}(\theta)^\top$$

mit der Funktionalmatrix $(\boldsymbol{H}(\theta))_{ij} = \frac{\partial}{\partial \theta_j}\tau_i(\theta)$. Die Matrix $\boldsymbol{H}(\theta)\boldsymbol{I}^{-1}(\theta)\boldsymbol{H}(\theta)^\top$ ist die Cramer-Rao-Schranke.

Bemerkung. Obige Bemerkung für skalares θ gilt analog für

$$f(x|\theta) = h(x)\exp(b(\theta) + \boldsymbol{\gamma}^\top(\theta)\boldsymbol{T}(x)),$$

d.h. für mehrparametrische Exponentialfamilien.

Beispiel 2.14 (Cramer-Rao-Schranke bei $X \sim N(\mu, \sigma^2)$)
$X_1, \ldots, X_n$ i.i.d. wie $X \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2)$. Dann gilt für die Informationsmatrix

$$I(\theta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} \quad \text{bzw.} \quad I^{-1}(\theta) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

Beste erwartungstreue Schätzer

Erwartungstreue Schätzer minimaler Varianz innerhalb einer vorgegebenen Klasse nennt man *effizient*. Die Informationsungleichungen motivieren:

Definition 2.20 (Gleichmäßig bester erwartungstreuer (UMVU) Schätzer)

1. θ *skalar:*
   Der Schätzer $\widehat{\theta}_{\mathit{eff}}$ *für* θ *heißt gleichmäßig bester erwartungstreuer oder UMVU („uniformly minimum variance unbiased")-Schätzer* $\overset{\mathit{def}}{\Leftrightarrow}$ $\widehat{\theta}_{\mathit{eff}}$ *ist erwartungstreu, und es gilt* $\mathrm{Var}_\theta(\widehat{\theta}_{\mathit{eff}}) \leq \mathrm{Var}_\theta(\widehat{\theta})$ *für alle* θ *und jeden erwartungstreuen Schätzer* $\widehat{\theta}$.
2. θ *mehrdimensional:*
   Ersetze $\mathrm{Var}_\theta(\widehat{\theta}_{\mathit{eff}}) \leq \mathrm{Var}_\theta(\widehat{\theta})$ *durch* $\mathrm{Cov}_\theta(\widehat{\theta}_{\mathit{eff}}) \leq \mathrm{Cov}_\theta(\widehat{\theta})$.

Satz 2.21 (Effizienz und Informationsungleichungen)

Sei $f(x|\theta)$ Fisher-regulär und $\widehat{\theta}$ erwartungstreu für θ. Falls $\mathrm{Cov}_\theta(\widehat{\theta}) = I^{-1}(\theta)$ für alle θ, so ist $\widehat{\theta}$ ein UMVU-Schätzer.

Beweis. Die Aussage folgt direkt aus der Informationsungleichung und obiger Definition. □

Beispiel 2.15 (Gauß-Experiment)
Seien $X_1, \ldots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ mit μ, σ^2 unbekannt.
Aus Beispiel 2.14 auf Folie 147 wissen wir, dass $I(\mu) = n/\sigma^2$ und somit $I^{-1}(\mu) = \sigma^2/n = \mathrm{Var}(\bar{X})$. Dann ist $\bar{X}$ UMVU für μ.
Aber

$$\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1} > \frac{2\sigma^4}{n} = \boldsymbol{I}^{-1}(\sigma^2).$$

Die Cramer-Rao-Schranke wird also nicht erreicht, somit kann nicht gefolgert werden, dass S^2 UMVU für σ^2 ist.

Beispiel 2.16 (Lineares Modell)

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim\ N(\mathbf{0}, \sigma^2 I) \quad \text{bzw.} \quad \boldsymbol{y}\ \sim\ N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I)$$

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\mathrm{KQ}} &= \widehat{\boldsymbol{\beta}}_{\mathrm{ML}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \text{ ist effizient für } \boldsymbol{\beta}, \\ \widehat{\sigma}^2 &= \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \text{ ist UMVU-Schätzer für } \sigma^2. \end{aligned}$$

Bemerkung. Zu unterscheiden sind folgende Situationen:

1. Es existiert ein UMVU-Schätzer, dessen Varianz gleich der Cramer-Rao-Schranke ist.
2. Es existiert ein UMVU-Schätzer, dessen Varianz größer als die Cramer-Rao-Schranke ist (findet man mit dem Satz von Lehmann-Scheffé, siehe Satz 2.23).
3. Der häufigste Fall: Es existiert (für finiten Stichprobenumfang) kein UMVU-Schätzer.

Fazit: Finite Theorie erwartungstreuer Schätzer ist von eingeschränkter Anwendungsrelevanz.

Aber: Es existiert eine analoge asymptotische Theorie mit breiter Anwendungsrelevanz, die sich an finiter Theorie orientiert (siehe Abschnitt 2.1.5).

Zur Konstruktion von UMVU-Schätzern sind folgende zwei Aussagen nützlich:

Satz 2.22 (Rao-Blackwell)

Sei $T = T(X)$ suffizient für θ bzw. $\mathcal{P}_\theta$ und $\widehat{\theta}$ erwartungstreu für θ. Für den Schätzer

$$\widehat{\theta}_{RB} = \mathbb{E}_\theta[\widehat{\theta}|T] \quad \text{(„Rao-Blackwellization")}$$

gilt:

1. *$\widehat{\theta}_{RB}$ ist erwartungstreu für θ.*
2. $\mathrm{Var}_\theta(\widehat{\theta}_{RB}) \leq \mathrm{Var}_\theta(\widehat{\theta})$.
3. *In 2. gilt die Gleichheit, wenn $\widehat{\theta}$ nur von T abhängt, d.h. $\widehat{\theta}_{RB} = \widehat{\theta}$ mit Wahrscheinlichkeit 1.*

Satz 2.23 (Lehmann-Scheffé)

Ist $T = T(X)$ suffizient und vollständig (also minimalsuffizient) und $\widehat{\theta} = \widehat{\theta}(x)$ ein erwartungstreuer Schätzer, so ist

$$\widehat{\theta}^* = \mathbb{E}_\theta[\widehat{\theta}|T]$$

der mit Wahrscheinlichkeit 1 eindeutig bestimmte UMVU-Schätzer für θ.

Beispiel Satz von Lehmann-Scheffé

In Beispiel 2.16 (lineares Modell) ist $\hat{\sigma}^2$ erwartungstreu für σ^2.
Außerdem lässt sich zeigen, dass $\hat{\sigma}^2$ eine Funktion einer vollständigen suffizienten Statistik ist.
Damit ist $\hat{\sigma}^2$ UMVU-Schätzer für σ^2.

Wichtige Schätzer (Momentenschätzer, Shrinkage-Schätzer, ML- und Quasi-ML-Schätzer etc.) sind im Allgemeinen nicht erwartungstreu, besitzen aber günstige asymptotische ($n \to \infty$) Eigenschaften.

Im Folgenden sei

$$\hat{\theta}_n = \hat{\theta}(X_1, \dots, X_n)$$

Schätzer für θ.

Definition 2.24 (Asymptotische Erwartungstreue)

$\widehat{\theta}_n$ heißt asymptotisch erwartungstreu $\overset{def}{\Leftrightarrow}$

$$\lim_{n \to \infty} \mathbb{E}_\theta[\widehat{\theta}_n] = \theta \quad \textit{für alle } \theta.$$

Definition 2.25 (Konsistenz)

1. $\widehat{\theta}_n$ *ist (schwach) konsistent für* θ *(in Zeichen:* $\widehat{\theta}_n \overset{\mathbb{P}}{\to} \theta$ *(für alle* θ*))* $\overset{def}{\Leftrightarrow}$

$$\lim_{n \to \infty} \mathbb{P}_\theta(|\widehat{\theta}_n - \theta| \leq \varepsilon) = 1 \quad \textit{für alle } \varepsilon > 0 \textit{ und alle } \theta.$$

2. $\widehat{\theta}_n$ *heißt MSE-konsistent für* θ $\overset{def}{\Leftrightarrow}$

$$\lim_{n \to \infty} \mathrm{MSE}_\theta(\widehat{\theta}_n) = 0 \quad \textit{für alle } \theta.$$

3. $\widehat{\theta}_n$ *ist stark konsistent für* θ $\overset{def}{\Leftrightarrow}$

$$\mathbb{P}_\theta\left(\lim_{n \to \infty} \widehat{\theta}_n = \theta\right) = 1 \quad \textit{für alle } \theta.$$

Bemerkung.

1. Aus der (verallgemeinerten) Tschebyscheff-Ungleichung folgt

$$\widehat{\theta}_n \text{ MSE-konsistent } \Rightarrow \widehat{\theta}_n \text{ schwach konsistent.}$$

2. Wegen $\mathrm{MSE}_\theta(\widehat{\theta}_n) = \mathrm{Var}_\theta(\widehat{\theta}_n) + (\mathrm{Bias}_\theta(\widehat{\theta}_n))^2$ folgt:

$$\widehat{\theta}_n \text{ ist MSE-konsistent } \Leftrightarrow \mathrm{Var}_\theta(\widehat{\theta}_n) \to 0 \text{ und}$$

$$\mathrm{Bias}_\theta(\widehat{\theta}_n) \to 0 \text{ für alle } \theta.$$

Bemerkung fortgeführt.

3. Ist $\widehat{\theta}_n$ konsistent für θ und g eine stetige Abbildung, so ist auch $g(\widehat{\theta}_n)$ konsistent für $g(\theta)$ (Continuous Mapping Theorem/Stetigkeitssatz).
4. Konsistenznachweise bestehen in der Regel in der Anwendung (schwacher) Gesetze großer Zahlen (für i.i.d. Variablen; i.n.i.d. Variablen; abhängige Variablen, z.B. Martingale, Markov-Prozesse, ...).

Beispiel 2.17

1. $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ ist wegen $\mathbb{E}(\bar{X}_n) = \mu$ und $\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n} \to 0$ für $n \to \infty$ konsistent.
2. $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ und $\widetilde{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ sind MSE-konsistent für σ^2.
3. Mit $g(x) = \sqrt{x}$ folgt, dass

$$S_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} \quad \text{und} \quad \widetilde{S}_n = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}$$

konsistent sind für σ.

Beispiel 2.17 fortgeführt

4. $S_n^2/\bar{X}_n$ ist konsistent für σ^2/μ für $\mu > 0$, da mit $\theta = (\mu, \sigma)$ und $g(\theta) = \sigma^2/\mu$ wieder der Stetigkeitssatz benutzt werden kann.

5. $\widehat{\pi}_n$ ist konsistent für π (im Bernoulli-Experiment).

6. $\widehat{\boldsymbol{\beta}}_{KQ}, \widehat{\boldsymbol{\beta}}_{Ridge}$ sind konsistent für $\boldsymbol{\beta}$ im linearen Modell unter gewissen schwachen Annahmen an $\boldsymbol{X}$, siehe Beispiel 2.19.

Asymptotische Normalität

Viele Schätzer (KQ-, Momenten-, ML-, Quasi-ML-, Bayes-Schätzer) sind unter Regularitätsannahmen asymptotisch normalverteilt. Informell ausgedrückt heißt das: Für große n ist $\widehat{\theta}_n$ nicht nur approximativ erwartungstreu, sondern zusätzlich approximativ normalverteilt, kurz

$$\widehat{\theta}_n \overset{a}{\sim} N(\theta, V(\theta))$$

mit (approximativer) Kovarianzmatrix

$$\mathrm{Cov}_\theta(\widehat{\theta}_n) \overset{a}{\sim} V(\theta),$$

die durch

$$\widehat{\mathrm{Cov}}_\theta(\widehat{\theta}_n) := V(\widehat{\theta}_n)$$

geschätzt wird.

In der Diagonalen von $V(\widehat{\theta}_n)$ stehen dann die (geschätzten) Varianzen

$$\widehat{\mathrm{Var}}(\widehat{\theta}_j) = v_{jj}(\widehat{\theta}_n)$$

der Komponenten $\theta_j, j = 1, \ldots, p$, von θ.

$\Rightarrow$ "Üblicher" Output statistischer Software ist

$$\underbrace{\widehat{\theta}_j}_{\text{Schätzer}} \qquad \underbrace{\widehat{\sigma}_{\widehat{\theta}_j} = \sqrt{v_{jj}(\widehat{\theta})}}_{\text{Standardfehler}} \qquad \underbrace{\text{t}}_{\text{t-Statistik}} \qquad \underbrace{\text{p}}_{\text{p-Wert}}$$

Beispiel 2.18.

Seien $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F(x|\theta)$ mit $\mathbb{E}(X_i) = \mu$ und $\mathrm{Var}(X_i) = \sigma^2$. Aber F sei *nicht* gleich Φ, sondern z.B. die Verteilungsfunktion von $B(\pi)$ oder $Po(\lambda)$. Für $\bar{X}_n$ gilt

$$\mathbb{E}(\bar{X}_n) = \mu \text{ und } \mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}.$$

Aufgrund des zentralen Grenzwertsatzes folgt

$$\bar{X}_n \overset{a}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right),$$

zum Beispiel

$$\bar{X}_n \overset{a}{\sim} N\left(\pi, \frac{\pi(1-\pi)}{n}\right) \quad \text{bei } B(\pi).$$

Beispiel 2.18 fortgeführt.
Genauere Formulierung:

$$\sqrt{n}(\bar{X}_n - \mu) \stackrel{d}{\to} N(0, \sigma^2) \quad \text{für } n \to \infty,$$

im Beispiel also

$$\sqrt{n}(\bar{X}_n - \pi) \stackrel{d}{\to} N(0, \pi(1-\pi)) \quad \text{für } n \to \infty$$

bzw.

$$\left.\begin{array}{rcl} \frac{\bar{X}_n - \mu}{\sigma}\sqrt{n} & \stackrel{d}{\to} & N(0,1), \\ \frac{\bar{X}_n - \pi}{\sqrt{\pi(1-\pi)}}\sqrt{n} & \stackrel{d}{\to} & N(0,1). \end{array}\right\} \text{zentraler Grenzwertsatz}$$

Die $\sqrt{n}$-Normierung ist vor allem bei i.i.d. Stichprobenvariablen geeignet.

Für nicht identisch verteilte Stichprobenvariablen wie zum Beispiel $y_1|x_1, \ldots, y_n|x_n$ in Regressionssituationen benötigt man bei $\sqrt{n}$-Normierung Voraussetzungen, die (teilweise) unnötig restriktiv sind.

Besser ist dann eine „Matrix-Normierung" mit Hilfe einer „Wurzel" $\boldsymbol{I}^{\frac{1}{2}}(\theta)$ der Informationsmatrix.

Einschub: Wurzel einer positiv definiten Matrix

- $\boldsymbol{A}$ ist positiv definit, wenn $\boldsymbol{A}$ symmetrisch ist und $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$ für alle $\boldsymbol{x} \neq \boldsymbol{0}$ gilt.
- Dann heißt eine Matrix $\boldsymbol{A}^{\frac{1}{2}}$ *(linke) Wurzel* von $\boldsymbol{A}$ $\overset{\text{def}}{\Leftrightarrow}$

$$\boldsymbol{A}^{\frac{1}{2}} \underbrace{(\boldsymbol{A}^{\frac{1}{2}})^\top}_{=\boldsymbol{A}^{\frac{\top}{2}},\ \text{rechte Wurzel}} = \boldsymbol{A}.$$

Allerdings ist $\boldsymbol{A}^{\frac{1}{2}}$ nicht eindeutig, da für eine beliebige orthogonale Matrix $\boldsymbol{Q}$ auch $\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{Q}$ eine linke Wurzel ist:

$$\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{Q}(\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{Q})^\top = \boldsymbol{A}^{\frac{1}{2}} \underbrace{\boldsymbol{Q}\boldsymbol{Q}^\top}_{=\boldsymbol{I}} \boldsymbol{A}^{\frac{\top}{2}} = \boldsymbol{A}.$$

- Zwei gebräuchliche Wurzeln sind:

  1. **Symmetrische Wurzel:** Betrachte die Spektralzerlegung von $\boldsymbol{A} \in \mathbb{R}^{p \times p}$. Mit der Matrix $\boldsymbol{P} \in \mathbb{R}^{p \times p}$ der orthonormalen Eigenvektoren als Spalten ist

$$\boldsymbol{P}^\top \boldsymbol{A} \boldsymbol{P} = \boldsymbol{\Lambda} = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{pmatrix},$$

wobei für alle i die $\lambda_i > 0$ die Eigenwerte von $\boldsymbol{A}$ sind. (Diese Zerlegung ist numerisch aufwändig!) Dann gilt auch

$$\boldsymbol{A} = \boldsymbol{P} \boldsymbol{\Lambda} \boldsymbol{P}^\top = \underbrace{\boldsymbol{P} \boldsymbol{\Lambda}^{\frac{1}{2}}}_{=\boldsymbol{A}^{\frac{1}{2}}} \underbrace{(\boldsymbol{\Lambda}^{\frac{1}{2}})^\top \boldsymbol{P}^\top}_{=\boldsymbol{A}^{\frac{\top}{2}}},$$

und $\boldsymbol{A}^{\frac{1}{2}}$ heißt *symmetrische Wurzel von* $\boldsymbol{A}$.

2. **Cholesky-Wurzel:** Sei $\boldsymbol{A}^{\frac{1}{2}} := \boldsymbol{C}$ untere Dreiecksmatrix mit positiven Diagonalelementen und $\boldsymbol{C}\boldsymbol{C}^\top = \boldsymbol{A}$. Dann ist $\boldsymbol{C}$ die *eindeutig* bestimmte *Cholesky-Wurzel* von $\boldsymbol{A}$.
(Diese ist numerisch vergleichsweise einfach zu erhalten!)

- **Anwendungen in der Statistik**
  1. Erzeugen von $N_p(\mathbf{0}, \boldsymbol{\Sigma})$-verteilten Zufallszahlen ($\boldsymbol{\Sigma}$ vorgegeben): Falls $\boldsymbol{Z} \sim N_p(\mathbf{0}, \mathbf{I})$, ist einfache Simulation möglich, indem p unabhängige $N(0,1)$-verteilte Zufallsvariablen $Z_1, \ldots, Z_p$ simuliert werden.
     Dann gilt auch
     $$\boldsymbol{\Sigma}^{1/2}\boldsymbol{Z} \;\sim\; N_p(\mathbf{0}, \underbrace{\boldsymbol{\Sigma}^{1/2}\boldsymbol{I}\boldsymbol{\Sigma}^{\top/2}}_{\boldsymbol{\Sigma}}).$$
     Also: Berechne Cholesky-Wurzel von $\boldsymbol{\Sigma}$, ziehe p $N(0,1)$-verteilte Zufallsvariablen $\boldsymbol{Z} = (z_1, \ldots, z_p)^\top$, berechne $\boldsymbol{Y} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$. Dann ist $\boldsymbol{Y} = (Y_1, \ldots, Y_p)^\top$ ein $N_p(\mathbf{0}, \boldsymbol{\Sigma})$-verteilter Zufallsvektor.

2. Matrixnormierung bei asymptotischer Normalverteilung:

Beispiel 2.19 (Asymptotische Normalität des KQ-Schätzers im linearen Modell)
Seien $y_1|\boldsymbol{x}_1, \ldots, y_n|\boldsymbol{x}_n$ unabhängig. Dann gilt

$$\mathbb{E}[y_i|\boldsymbol{x}_i] = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \quad \mathrm{Var}(y_i|\boldsymbol{x}_i) = \sigma^2, \quad i = 1, \ldots, n,$$

$$\Leftrightarrow \; \boldsymbol{y}_n = \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n, \; \mathbb{E}[\boldsymbol{\varepsilon}_n] = \boldsymbol{0}, \; \mathrm{Cov}(\boldsymbol{\varepsilon}_n) = \sigma^2 \boldsymbol{I}_n.$$

Der KQ-Schätzer ist

$$\widehat{\boldsymbol{\beta}}_n = (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{X}_n^\top \boldsymbol{y}_n, \; \mathbb{E}[\widehat{\boldsymbol{\beta}}_n] = \boldsymbol{\beta}, \; \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_n) = \sigma^2 (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}.$$

Die Informationsmatrix unter der Normalverteilungsannahme ist

$$\boldsymbol{I}(\boldsymbol{\beta}) = \frac{\boldsymbol{X}_n^\top \boldsymbol{X}_n}{\sigma^2} = \mathrm{Cov}(\widehat{\boldsymbol{\beta}}_n)^{-1}.$$

Beispiel 2.19 fortgeführt

Zentrale Grenzwertsätze (für unabhängige, nicht identisch verteilte Zufallsvariablen, kurz: i.n.i.d.) liefern unter geeigneten Voraussetzungen (informell):

$$\widehat{\boldsymbol{\beta}}_n \overset{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}).$$

Genauere Formulierungen nehmen an, dass

$$\lim_{n \to \infty} \frac{1}{n} \boldsymbol{X}_n^\top \boldsymbol{X}_n =: \boldsymbol{A} > 0 \qquad (1)$$

existiert (also: $\boldsymbol{X}_n^\top \boldsymbol{X}_n \approx n\boldsymbol{A} \ \Leftrightarrow \ (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \approx \boldsymbol{A}^{-1}/n$ für große n).

Beispiel 2.19 fortgeführt

Anwendung des (multivariaten) zentralen Grenzwertsatzes liefert dann:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{d} N_p(0, \sigma^2 \boldsymbol{A}^{-1})$$

bzw.

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_n &\overset{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 \boldsymbol{A}^{-1}/n) \\ \widehat{\boldsymbol{\beta}}_n &\overset{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1}). \end{aligned}$$

Die Annahme (1) ist zum Beispiel erfüllt, wenn $\boldsymbol{x}_i$, $i = 1, \ldots, n$, i.i.d. Realisierungen stochastischer Kovariablen $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$ sind.

Beispiel 2.19 fortgeführt

Dann gilt nach dem Gesetz der großen Zahlen:

$$\frac{1}{n}\boldsymbol{X}_n^\top \boldsymbol{X}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \overset{n\to\infty}{\to} \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] =: \boldsymbol{A}.$$

Typischerweise ist die Annahme (1) *nicht* erfüllt bei deterministischen Regressoren mit Trend. Das einfachste Beispiel hierfür ist ein linearer Trend: $x_i = i$ für $i = 1, \dots, n$ und $y_i = \beta_1 i + \varepsilon_i$. Dann ist

$$\boldsymbol{X}_n^\top \boldsymbol{X}_n = \sum_{i=1}^{n} i^2$$

und daher

$$\frac{1}{n}\boldsymbol{X}_n^\top \boldsymbol{X}_n = \frac{\sum_{i=1}^{n} i^2}{n} \geq n \overset{n\to\infty}{\to} \infty.$$

Beispiel 2.19 fortgeführt

In diesem Fall ist eine andere Normierung nötig, zum Beispiel eine Matrixnormierung mit

$$\boldsymbol{C}_n = (\boldsymbol{X}_n^\top \boldsymbol{X}_n).$$

Dann lässt sich die asymptotische Normalität des KQ-Schätzers

$$\boldsymbol{C}_n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \overset{d}{\to} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

bzw.

$$\tilde{\boldsymbol{C}}_n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) := \frac{\boldsymbol{C}_n^{1/2}}{\sigma}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \overset{d}{\to} N_p(\boldsymbol{0}, \boldsymbol{I})$$

unter folgenden, sehr schwachen Bedingungen zeigen:

Beispiel 2.19 fortgeführt

(D) Divergenzbedingung: Für $n \to \infty$ gilt:

$$(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \to \mathbf{0}.$$

Eine äquivalente Forderung ist:

$$\lambda_{\min}(\boldsymbol{X}_n^\top \boldsymbol{X}_n) \to \infty,$$

wobei $\lambda_{\min}$ den kleinsten Eigenwert von $\boldsymbol{X}_n^\top \boldsymbol{X}_n$ bezeichnet. Die Divergenzbedingung sichert, dass die „Informationsmatrix"

$$\boldsymbol{X}_n^\top \boldsymbol{X}_n = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

für $n \to \infty$ gegen ∞ divergiert, die Information mit $n \to \infty$ also laufend wächst.
Es gilt: (D) ist hinreichend und notwendig für die (schwache und starke) Konsistenz des KQ-Schätzers $\widehat{\boldsymbol{\beta}}_n$.

Beispiel 2.19 fortgeführt

(N) Normalitätsbedingung:

$$\max_{i=1,\ldots,n} \boldsymbol{x}_i^\top (\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1} \boldsymbol{x}_i \;\to\; 0 \quad \text{für } n \to \infty$$

(N) sichert, dass die Information jeder Beobachtung i asymptotisch gegenüber der Gesamtinformation $\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top$ vernachlässigbar ist.

Beispiel 2.19 fortgeführt

Unter (D) und (N) gilt

$$(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \overset{d}{\to} N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

(Beweis mit Grenzwertsätzen für unabhängige, nicht identisch verteilte Zufallsvariablen), d.h. für praktische Zwecke:

$$\widehat{\boldsymbol{\beta}}_n \overset{a}{\sim} N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}_n^\top \boldsymbol{X}_n)^{-1})$$

für genügend großen Stichprobenumfang n. Dabei darf zusätzlich σ^2 durch einen konsistenten Schätzer $\widehat{\sigma}^2$ ersetzt werden.

Definition 2.26 (Asymptotische Normalität)

$\widehat{\theta}_n$ heißt asymptotisch normalverteilt für $\theta \overset{def}{\Leftrightarrow}$ es existiert eine Folge von Matrizen $\boldsymbol{A}_n$ mit $\lambda_{\min}(\boldsymbol{A}_n) \to \infty$, so dass

$$\boldsymbol{A}_n^{1/2}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta)) \quad \text{für } n \to \infty.$$

mit nicht-negativ definiter (in der Regel positiv definiter) Matrix $V(\theta)$.

Für den Spezialfall $\boldsymbol{A}_n = n\boldsymbol{I}_k$ entspricht dies der $\sqrt{n}$-Normierung

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta)) \quad \text{für } n \to \infty.$$

Bemerkung.

1. Praxisformulierung:

$$\widehat{\theta}_n \overset{a}{\sim} N(\theta, V(\theta)/n)$$

   bzw.

$$\widehat{\theta}_n \overset{a}{\sim} N(\theta, (\boldsymbol{A}_n^{1/2})^{-1} V(\theta) (\boldsymbol{A}_n^{1/2})^{-\top}).$$

   Dabei darf θ in $V(\theta)$ durch $\widehat{\theta}_n$ ersetzt werden.
2. Oft: $V(\theta) = \mathbf{I}_k$ möglich, wenn geeignet normiert wird, zum Beispiel bei ML-Schätzung.

Beispiel 2.20.

Seien $X_1, \ldots, X_n$ i.i.d. Zufallsvariablen mit (bekanntem) Erwartungswert μ und Varianz σ^2.

$$S_\mu^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$$

ist asymptotisch normal für σ^2 mit $V(\theta) = \mu_4 - \sigma^4$, $\mu_4 = \mathbb{E}[(X_i - \mu)^4] < \infty$. S_μ^2 ist erwartungstreu.

Beispiel 2.20 fortgeführt.

Für die Varianz erhält man:

$$\begin{aligned} \mathrm{Var}(S_\mu^2) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2\right) \\ &= \frac{1}{n^2}\cdot n\cdot \mathrm{Var}\left[(X_1-\mu)^2\right] \\ &= \frac{1}{n}\Big(\mathbb{E}[(X_1-\mu)^4]-\big(\mathbb{E}[(X_1-\mu)^2]\big)^2\Big) \\ &= \frac{1}{n}(\mu_4-\sigma^4). \end{aligned}$$

Es liegen die Voraussetzungen zur Anwendung des zentralen Grenzwertsatzes vor. Aus ihm folgt:

$$S_\mu^2 \overset{a}{\sim} N(\sigma^2,(\mu_4-\sigma^4)/n) \quad \text{bzw.} \quad \sqrt{n}(S_\mu^2-\sigma^2)\xrightarrow{d} N(0,\mu_4-\sigma^4).$$

Die Delta-Methode

$\widehat{\theta}_n$ sei asymptotisch normalverteilter Schätzer für θ.

Frage: Wie ist für eine gegebene Abbildung

$$h: \ \mathbb{R}^p \ \to \ \mathbb{R}^k, k \leq p$$

der Schätzer $h(\widehat{\theta})$ für $h(\theta)$ verteilt?

Satz 2.27 (Delta-Methode)

Sei h wie oben.

1. *θ skalar: Für alle θ, für die h stetig differenzierbar ist mit $h'(\theta) \neq 0$, gilt:*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta))$$

$$\Rightarrow \sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, [h'(\theta)]^2 V(\theta))$$

Satz 2.28 (fortgeführt)

2. θ *vektoriell: Sei h gegeben durch*

$$\theta = (\theta_1, \dots, \theta_p)^\top \mapsto h(\theta) = (h_1(\theta), \dots, h_k(\theta))^\top$$

mit Funktionalmatrix

$$(H(\theta))_{ij} = \frac{\partial h_i(\theta)}{\partial \theta_j}$$

mit vollem Rang. Für alle θ, für die $h(\theta)$ komponentenweise stetig partiell differenzierbar ist und jede Zeile von $H(\theta)$ ungleich dem Nullvektor ist, gilt:

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta))$$

$$\Rightarrow \sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, H(\theta)V(\theta)H(\theta)^\top).$$

Beweisskizze für skalares θ

Taylorentwicklung von $h(\widehat{\theta}_n)$ um θ liefert:

$$h(\hat{\theta}_n) = h(\theta) + (\widehat{\theta}_n - \theta)h'(\theta) + o(\widehat{\theta}_n - \theta)^2.$$

Dabei ist für eine Folge von Zufallsvariablen X_n

$$X_n = o(a_n) \quad \text{falls } X_n/a_n \xrightarrow{P} 0 \text{ für } n \to \infty.$$

Also:

$$h(\widehat{\theta}_n) \approx h(\theta) + (\widehat{\theta}_n - \theta)h'(\theta)$$

bzw.

$$\sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \approx \sqrt{n}(\widehat{\theta}_n - \theta)h'(\theta)$$

Aus $\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, V(\theta))$ folgt dann, dass
$\sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \xrightarrow{d} N(0, h'(\theta)^2 V(\theta)).$ □

Asymptotische Cramer-Rao Schranke und asymptotische Effizienz

Seien $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} f(x|\theta)$ und

$$i(\theta) = -\mathbb{E}\left[\frac{\partial^2 \log f(x|\theta)}{\partial\theta\partial\theta^\top}\right]$$

die erwartete Fisher-Information einer Beobachtung X_i.
Die Information der gesamten Stichprobe $X_1, \ldots, X_n$ ist dann

$$\boldsymbol{I}(\theta) = n \cdot i(\theta).$$

Satz 2.28 (Asymptotische Cramer-Rao Ungleichung)

Unter Fisher-Regularität sowie leichten Zusatzannahmen gilt:

1. *Aus* $\sqrt{n}(\widehat{\theta}_n - \theta) \stackrel{d}{\to} N(0, V(\theta))$ *folgt* $V(\theta) \geq i^{-1}(\theta)$.
2. *Aus* $\sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \stackrel{d}{\to} N(0, D(\theta))$ *folgt*

$$D(\theta) \geq H(\theta) i^{-1}(\theta) H(\theta)^\top$$

mit "$\geq$" die Löwner-Ordnung (und den Bezeichnungen aus der Delta-Regel, Satz 2.27).

Definition 2.29 (Bester asymptotisch normaler (BAN)-Schätzer)

$\widehat{\theta}_n$ *heißt BAN-Schätzer, falls in 1. oben gilt:*

$$V(\theta) = i^{-1}(\theta).$$

Mit der Delta-Regel folgt unmittelbar:

Satz 2.30 (Transformation von BAN-Schätzern)

Ist $\widehat{\theta}_n$ *BAN-Schätzer für* θ*, so ist* $h(\widehat{\theta}_n)$ *BAN-Schätzer für* $h(\theta)$.

Bemerkung. Das Konzept der asymptotischen Effizienz lässt sich auf die Matrix-Normierung übertragen: $\widehat{\theta}$ ist BAN-Schätzer für θ genau dann, wenn

$$\boldsymbol{I}^{1/2}(\theta)(\widehat{\theta}_n - \theta) \overset{d}{\to} N(0, I)$$

bzw. $\widehat{\theta}_n \overset{a}{\sim} N(\theta, \boldsymbol{I}^{-1}(\widehat{\theta}_n))$, mit $\boldsymbol{I}^{1/2}(\theta)$ Wurzel der Fisher-Information $\boldsymbol{I}(\theta)$ der Stichprobe $X_1, \ldots, X_n$. Anstelle der erwarteten kann auch die beobachtete Fisher-Information $\boldsymbol{J}(\theta)$ verwendet werden.