Bias Correction in Maximum Likelihood Logistic Regression (Schaefer, R. 1983)

Jens Rahnfeld

SS 2021

Maximum Likelihood auf kleinen Datensätzen

- Maximum Likelihood Schätzer ist nur asymptotisch unverzerrt.

Maximum Likelihood auf kleinen Datensätzen

- Maximum Likelihood Schätzer ist nur asymptotisch unverzerrt.
- Bei kleineren Datensätzen kann es zu einem hohen Bias kommen.

Maximum Likelihood auf kleinen Datensätzen

- Maximum Likelihood Schätzer ist nur asymptotisch unverzerrt.
- Bei kleineren Datensätzen kann es zu einem hohen Bias kommen.
- Idee: Korrigiere den Maximum Likelihood Schätzer um dessen Bias

Logistisches Regressionsmodell

- Datensatz $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ aus n i.i.d. Datenpunkten (x_i, y_i), mit zugrundeliegender Verteilung:

$$\mathbb{P}(y = 1 \mid x') = \frac{\exp(\beta_t^\top x)}{1 + \exp(\beta_t^\top x)}$$

wobei $x = \begin{pmatrix} 1 \\ x' \end{pmatrix} \in \mathbb{R}^p$, $y \in \{0, 1\}$, $\beta_t \in \mathbb{R}^p$.

Logistisches Regressionsmodell

- Datensatz $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ aus n i.i.d. Datenpunkten (x_i, y_i), mit zugrundeliegender Verteilung:

$$\mathbb{P}(y = 1 \mid x') = \frac{\exp(\beta_t^\top x)}{1 + \exp(\beta_t^\top x)}$$

  wobei $x = \begin{pmatrix} 1 \\ x' \end{pmatrix} \in \mathbb{R}^p$, $y \in \{0, 1\}$, $\beta_t \in \mathbb{R}^p$.

- β_t bezeichnet wahren Parameter und ist konstant.

Logistisches Regressionsmodell

- Datensatz $D_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ aus n i.i.d. Datenpunkten (x_i, y_i), mit zugrundeliegender Verteilung:

$$\mathbb{P}(y = 1 \mid x') = \frac{\exp(\beta_t^\top x)}{1 + \exp(\beta_t^\top x)}$$

  wobei $x = \begin{pmatrix} 1 \\ x' \end{pmatrix} \in \mathbb{R}^p$, $y \in \{0, 1\}$, $\beta_t \in \mathbb{R}^p$.

- β_t bezeichnet wahren Parameter und ist konstant.
- Wir schreiben x_i für $\begin{pmatrix} 1 \\ x_i \end{pmatrix} = \begin{pmatrix} 1 & x_i^{(2)} & \ldots & x_i^{(p)} \end{pmatrix}^\top$

Notation

- Datensatz:

$$\boldsymbol{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(2)} & \dots & x_1^{(p)} \\ \vdots & & & \\ 1 & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix} \in \mathbb{R}^{n \times p}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

Notation

- Datensatz:

$$\boldsymbol{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(2)} & \dots & x_1^{(p)} \\ \vdots & & & \\ 1 & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix} \in \mathbb{R}^{n \times p}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

- Wahrscheinlichkeiten:

$$p_i(\beta) = \mathbb{P}_\beta(y_i = 1 \mid x_i) = \frac{\exp(\beta^\top x_i)}{1 + \exp(\beta^\top x_i)}, \quad \boldsymbol{p}(\beta) = \begin{pmatrix} p_1(\beta) \\ \vdots \\ p_n(\beta) \end{pmatrix}$$

Notation

- Datensatz:

$$\boldsymbol{X} = \begin{pmatrix} x_1^\top \\ \vdots \\ x_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_1^{(2)} & \dots & x_1^{(p)} \\ \vdots & & & \\ 1 & x_n^{(2)} & \dots & x_n^{(p)} \end{pmatrix} \in \mathbb{R}^{n \times p}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$$

- Wahrscheinlichkeiten:

$$p_i(\beta) = \mathbb{P}_\beta(y_i = 1 \mid x_i) = \frac{\exp(\beta^\top x_i)}{1 + \exp(\beta^\top x_i)}, \quad \boldsymbol{p}(\beta) = \begin{pmatrix} p_1(\beta) \\ \vdots \\ p_n(\beta) \end{pmatrix}$$

- wahre Varianz:

$$\boldsymbol{V} = \operatorname{diag}(v_1, \dots, v_n) = \begin{pmatrix} p_1(\beta_t)(1 - p_1(\beta_t) & & \\ & \ddots & \\ & & p_n(\beta_t)(1 - p_n(\beta_t)) \end{pmatrix}$$

Annahmen

We assume that the independent variables are bounded so that $\lim (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})/n$ is finite and positive definite as $n \to \infty$. Then, under certain regularity conditions, it is well known[3] that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a p-dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$. Thus if the sample size is large, $\hat{\boldsymbol{\beta}}$ is approximately unbiased and has approximate covariance matrix given by $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$. If the sample size is small, however, $\hat{\boldsymbol{\beta}}$ might be biased.

Figure: Schaefer R., Bias Correction in Maximum Likelihood Logistic Regression, 1983

- $\boldsymbol{X}$ wird im Paper als deterministische Matrix betrachtet.

Annahmen

We assume that the independent variables are bounded so that $\lim(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})/n$ is finite and positive definite as $n \to \infty$. Then, under certain regularity conditions, it is well known[3] that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a p-dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$. Thus if the sample size is large, $\hat{\boldsymbol{\beta}}$ is approximately unbiased and has approximate covariance matrix given by $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$. If the sample size is small, however, $\hat{\boldsymbol{\beta}}$ might be biased.

Figure: Schaefer R., Bias Correction in Maximum Likelihood Logistic Regression, 1983

- $\boldsymbol{X}$ wird im Paper als deterministische Matrix betrachtet.
- Beachte:

$$\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X} = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} v_1 \cdot x_1^\top \\ \vdots \\ v_n \cdot x_n^\top \end{pmatrix} = \sum_{i=1}^{n} v_i \cdot x_i x_i^\top$$

Annahmen

We assume that the independent variables are bounded so that $\lim (\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})/n$ is finite and positive definite as $n \to \infty$. Then, under certain regularity conditions, it is well known[3] that $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to a p-dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$. Thus if the sample size is large, $\hat{\boldsymbol{\beta}}$ is approximately unbiased and has approximate covariance matrix given by $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$. If the sample size is small, however, $\hat{\boldsymbol{\beta}}$ might be biased.

Figure: Schaefer R., Bias Correction in Maximum Likelihood Logistic Regression, 1983

- $\boldsymbol{X}$ wird im Paper als deterministische Matrix betrachtet.
- Beachte:

$$\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X} = \begin{pmatrix} x_1 & \dots & x_n \end{pmatrix} \begin{pmatrix} v_1 \cdot x_1^\top \\ \vdots \\ v_n \cdot x_n^\top \end{pmatrix} = \sum_{i=1}^{n} v_i \cdot x_i x_i^\top$$

- Als Zufallsvariable betrachtet, fordern wir, dass $E[v \cdot xx^\top]$ endlich, positiv definit und $X^\top VX$ fast sicher invertierbar definit ist. Weiter fordern wir $\|x\| \leq C$ fast sicher für ein $C < \infty$.

Maximum Likelihood Estimation

- Maximiere Log-Likelihood:

$$l(\beta) = \sum_{i=1}^{n} y_i \cdot \beta^\top x_i - \log(1 + \exp(\beta^\top x_i))$$

Maximum Likelihood Estimation

- Maximiere Log-Likelihood:

$$l(\beta) = \sum_{i=1}^{n} y_i \cdot \beta^\top x_i - \log(1 + \exp(\beta^\top x_i))$$

- Maximierer $\hat{\beta}$ ist Lösung von

$$\dot{l}(\beta) = \mathbf{0}$$

wobei

$$\dot{l}(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta)) \cdot x_i = \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{p}(\beta))$$

Taylorentwicklung der Score-Funktion

$$\dot{l}(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta)) \cdot x_i$$

- Entwickle p_i im Punkt β_t:

$$\begin{aligned}
\dot{p}_i(\beta) &= \frac{d}{d\beta} \frac{\exp(\beta^\top x_i)}{1 + \exp(\beta^\top x_i)} \\
&= \frac{\exp(\beta^\top x_i)(1 + \exp(\beta^\top x_i)) - \exp(\beta^\top x_i)^2}{(1 + \exp(\beta^\top x_i))^2} \cdot x_i^\top \\
&= \frac{\exp(\beta^\top x_i)}{(1 + \exp(\beta^\top x_i))^2} \cdot x_i^\top \\
&= p_i(\beta) \cdot (1 - p_i(\beta)) \cdot x_i^\top
\end{aligned}$$

$$\dot{p}_i(\beta) = \underbrace{p_i(\beta) \cdot (1 - p_i(\beta))}_{=:v_i(\beta)} \cdot x_i^\top$$

- Entwickle p_i im Punkt β_t:

$$\begin{aligned}\frac{d}{d\beta} p_i(\beta)(1 - p_i(\beta)) &= v_i(\beta)(1 - p_i(\beta)) \cdot x_i^\top - p_i(\beta) v_i(\beta) \cdot x_i^\top \\ &= v_i(\beta)(1 - 2p_i(\beta)) \cdot x_i^\top\end{aligned}$$

$$\dot{p}_i(\beta) = \underbrace{p_i(\beta) \cdot (1 - p_i(\beta))}_{=:v_i(\beta)} \cdot x_i^\top$$

- Entwickle p_i im Punkt β_t:

$$\begin{aligned}
\frac{d}{d\beta} p_i(\beta)(1 - p_i(\beta)) &= v_i(\beta)(1 - p_i(\beta)) \cdot x_i^\top - p_i(\beta) v_i(\beta) \cdot x_i^\top \\
&= v_i(\beta)(1 - 2p_i(\beta)) \cdot x_i^\top
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \ddot{p}_i(\beta) &= \frac{d}{d\beta} p_i(\beta)(1 - p_i(\beta)) \cdot x_i \\
&= v_i(\beta)(1 - 2p_i(\beta)) \cdot x_i x_i^\top
\end{aligned}$$

$$\dot{l}(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta)) \cdot x_i$$

- Entwickle p_i im Punkt β_t:

$$\dot{p}_i(\beta) = v_i(\beta) \cdot x_i^\top$$
$$\ddot{p}_i(\beta) = v_i(\beta)(1 - 2p_i(\beta)) \cdot x_i x_i^\top$$

$$\dot{l}(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta_t) - \dot{p}_i(\beta_t)(\beta - \beta_t) - \frac{1}{2} \cdot (\beta - \beta_t)^\top \ddot{p}_i(\beta_t)(\beta - \beta_t)) \cdot x_i$$
$$- o(\|\beta - \beta_t\|^2) x_i$$

$$\dot{l}(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta)) \cdot x_i$$

- Entwickle p_i im Punkt β_t:

$$\dot{p}_i(\beta) = v_i(\beta) \cdot x_i^\top$$
$$\ddot{p}_i(\beta) = v_i(\beta)(1 - 2p_i(\beta)) \cdot x_i x_i^\top$$

$$\dot{l}(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta_t) - \dot{p}_i(\beta_t)(\beta - \beta_t) - \frac{1}{2} \cdot (\beta - \beta_t)^\top \ddot{p}_i(\beta_t)(\beta - \beta_t)) \cdot x_i$$
$$- o(\|\beta - \beta_t\|^2) x_i$$

- Wir vernachlässigen den $o(\|\beta - \beta_t\|^2)$ Term vorerst.

Matrixnotation

- Terme 0. Ordnung:

$$\sum_{i=1}^{n}(y_i - p_i(\beta_t)) \cdot x_i = \boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{p}(\beta_t))$$

- Terme 1. Ordnung:

$$\sum_{i=1}^{n} \dot{p}_i(\beta_t)(\beta - \beta_t)x_i = \sum_{i=1}^{n} v_i \cdot x_i^\top(\beta - \beta_t)x_i = \boldsymbol{X}^\top \boldsymbol{V}\boldsymbol{X}(\beta - \beta_t)$$

- Terme 2. Ordnung:

$$
\begin{aligned}
&-\frac{1}{2}\sum_{i=1}^{n}(\beta-\beta_t)^\top \ddot{p}_i(\beta_t)(\beta-\beta_t)\cdot x_i \\
&= \boldsymbol{X}^\top \cdot \left(-\frac{1}{2}\right) \begin{pmatrix} (\beta-\beta_t)^\top \ddot{\boldsymbol{p}}_1(\beta-\beta_t) \\ \vdots \\ (\beta-\beta_t)^\top \ddot{\boldsymbol{p}}_n(\beta-\beta_t) \end{pmatrix}
\end{aligned}
$$

wobei $\ddot{\boldsymbol{p}}_i = \ddot{p}_i(\beta_t) = v_i(1-2p_i)\cdot x_i x_i^\top \in \mathbb{R}^{p\times p}$

- Insgesamt erhalten wir:

$$\dot{l}(\beta) = \sum_{i=1}^{n} (y_i - p_i(\beta_t) - \dot{p}_i(\beta_t)(\beta - \beta_t) - \frac{1}{2} \cdot (\beta - \beta_t)^\top \ddot{p}_i(\beta_t)(\beta - \beta_t)) \cdot x_i$$

$$= \boldsymbol{X}^\top \left(\boldsymbol{y} - \boldsymbol{p}(\beta_t) - \boldsymbol{V}\boldsymbol{X}(\beta - \beta_t) - \frac{1}{2} \cdot \begin{pmatrix} (\beta - \beta_t)^\top \ddot{\boldsymbol{p}}_1 (\beta - \beta_t) \\ \vdots \\ (\beta - \beta_t)^\top \ddot{\boldsymbol{p}}_n (\beta - \beta_t) \end{pmatrix} \right)$$

Logistische Regression als Latent Variable Modell

(Alternatively, one could use iterative weighted least squares on the model

$$\mathbf{y} = \boldsymbol{\pi} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\pi}$ is the $(n \times 1)$ vector of the π_i and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of independent Bernoulli random errors, ε_i, with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\operatorname{Var}(\boldsymbol{\varepsilon}) = \mathbf{V} = \operatorname{diag}\{v_i\} = \operatorname{diag}\{\pi_i(1-\pi_i)\}$ to arrive at the same estimate.[5])

Figure: Schaefer R., Bias Correction in Maximum Likelihood Logistic Regression, 1983

- Relevant ist, dass $\boldsymbol{y} - \boldsymbol{p}(\beta_t)$ durch $\boldsymbol{\epsilon}$ ersetzt werden kann.

Logistische Regression als Latent Variable Modell

(Alternatively, one could use iterative weighted least squares on the model

$$\mathbf{y} = \boldsymbol{\pi} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\pi}$ is the $(n \times 1)$ vector of the π_i and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of independent Bernoulli random errors, ε_i, with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\operatorname{Var}(\boldsymbol{\varepsilon}) = \mathbf{V} = \operatorname{diag}\{v_i\} = \operatorname{diag}\{\pi_i(1-\pi_i)\}$ to arrive at the same estimate.[5])

Figure: Schaefer R., Bias Correction in Maximum Likelihood Logistic Regression, 1983

- Relevant ist, dass $\boldsymbol{y} - \boldsymbol{p}(\beta_t)$ durch $\boldsymbol{\epsilon}$ ersetzt werden kann.
- Bemerkung:

$$\begin{aligned}
&E[y_i] = E[E[1_{\{y_i=1\}} \mid x_i]] = E[\mathbb{P}(y_i = 1 \mid x_i)] = E[p_i(\beta_t)] \\
&E[(y_i - p_i(\beta_t))^2] = E[(1 - p_i(\beta_t))^2 \cdot 1_{\{y_i=1\}} + p_i(\beta_t)^2 \cdot 1_{\{y_i=0\}}] \\
&= E[(1 - p_i(\beta_t))^2 E[1_{\{y_i=1\}} \mid x_i] + p_i(\beta_t)^2 E[1_{\{y_i=0\}} \mid x_i]] \\
&= E[(1 - p_i(\beta))^2 p_i(\beta_t) + p_i(\beta_t)^2 (1 - p_i(\beta_t))] \\
&= E[p_i(\beta_t)(1 - p_i(\beta_t))]
\end{aligned}$$

Logistische Regression als Latent Variable Modell

(Alternatively, one could use iterative weighted least squares on the model

$$\mathbf{y} = \boldsymbol{\pi} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\pi}$ is the $(n \times 1)$ vector of the π_i and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ vector of independent Bernoulli random errors, ε_i, with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\operatorname{Var}(\boldsymbol{\varepsilon}) = \mathbf{V} = \operatorname{diag}\{v_i\} = \operatorname{diag}\{\pi_i(1-\pi_i)\}$ to arrive at the same estimate.[5])

Figure: Schaefer R., Bias Correction in Maximum Likelihood Logistic Regression, 1983

- Relevant ist, dass $\boldsymbol{y} - \boldsymbol{p}(\beta_t)$ durch $\boldsymbol{\epsilon}$ ersetzt werden kann.
- Bemerkung:

$$\begin{aligned}
&E[y_i] = E[E[1_{\{y_i=1\}} \mid x_i]] = E[\mathbb{P}(y_i = 1 \mid x_i)] = E[p_i(\beta_t)] \\
&E[(y_i - p_i(\beta_t))^2] = E[(1 - p_i(\beta_t))^2 \cdot 1_{\{y_i=1\}} + p_i(\beta_t)^2 \cdot 1_{\{y_i=0\}}] \\
&= E[(1 - p_i(\beta_t))^2 E[1_{\{y_i=1\}} \mid x_i] + p_i(\beta_t)^2 E[1_{\{y_i=0\}} \mid x_i]] \\
&= E[(1 - p_i(\beta))^2 p_i(\beta_t) + p_i(\beta_t)^2 (1 - p_i(\beta_t))] \\
&= E[p_i(\beta_t)(1 - p_i(\beta_t))]
\end{aligned}$$

- Achtung: Hier ist $\boldsymbol{X}$ wieder deterministisch!

- Äquivalente Modellierung:

$$\boldsymbol{y} = \boldsymbol{p}(\beta_t) + \boldsymbol{\epsilon}$$

wobei $\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \in \mathbb{R}^n$ für, bedingt auf $\boldsymbol{X}$ unabhängige Bernoulli-Fehler, ϵ_i mit

$$E[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = \boldsymbol{0}, Var[\boldsymbol{\epsilon} \mid \boldsymbol{X}] = \boldsymbol{V} = \begin{pmatrix} v_1 & & \\ & \ddots & \\ & & v_n \end{pmatrix}.$$

$$\boldsymbol{y} = \boldsymbol{p}(\beta_t) + \boldsymbol{\epsilon}$$

- Ersetze $\boldsymbol{y} - \boldsymbol{p}(\beta_t)$ durch $\boldsymbol{\epsilon}$:

$$\dot{l}(\beta) = \boldsymbol{X}^\top \left(\boldsymbol{\epsilon} - \boldsymbol{V}\boldsymbol{X}(\beta - \beta_t) - \frac{1}{2} \cdot \begin{pmatrix} (\beta - \beta_t)^\top \ddot{\boldsymbol{p}}_1 (\beta - \beta_t) \\ \vdots \\ (\beta - \beta_t)^\top \ddot{\boldsymbol{p}}_n (\beta - \beta_t) \end{pmatrix} \right)$$

Bias des MLE

- Für den MLE $\hat{\beta}$, eingesetzt ist die Score-Funktion $\dot{l}(\cdot)$ gilt:

$$\mathbf{0} = \boldsymbol{X}^\top \left(\boldsymbol{\epsilon} - \boldsymbol{V}\boldsymbol{X}(\hat{\beta} - \beta_t) - \frac{1}{2} \cdot \begin{pmatrix} (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_1 (\hat{\beta} - \beta_t) \\ \vdots \\ (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_n (\hat{\beta} - \beta_t) \end{pmatrix} \right)$$

Bias des MLE

- Für den MLE $\hat{\beta}$, eingesetzt ist die Score-Funktion $\dot{l}(\cdot)$ gilt:

$$\mathbf{0} = \boldsymbol{X}^\top \left(\boldsymbol{\epsilon} - \boldsymbol{V}\boldsymbol{X}(\hat{\beta} - \beta_t) - \frac{1}{2} \cdot \begin{pmatrix} (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_1 (\hat{\beta} - \beta_t) \\ \vdots \\ (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_n (\hat{\beta} - \beta_t) \end{pmatrix} \right)$$

- Löse nach $(\hat{\beta} - \beta_t)$ auf:

$$\boldsymbol{X}^\top \boldsymbol{V}\boldsymbol{X}(\hat{\beta} - \beta_t) = \boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_1 (\hat{\beta} - \beta_t) \\ \vdots \\ (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_n (\hat{\beta} - \beta_t) \end{pmatrix}$$

$$\Leftrightarrow \hat{\beta} - \beta_t = (\boldsymbol{X}^\top \boldsymbol{V}\boldsymbol{X})^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_1 (\hat{\beta} - \beta_t) \\ \vdots \\ (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_n (\hat{\beta} - \beta_t) \end{pmatrix} \right)$$

$$\hat{\beta} - \beta_t = (\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X})^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_1 (\hat{\beta} - \beta_t) \\ \vdots \\ (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_n (\hat{\beta} - \beta_t) \end{pmatrix} \right)$$

- Für $\hat{\beta} \to \beta_t$ ist $\hat{\beta} - \beta_t = (\underbrace{\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X}}_{=:\boldsymbol{M}})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}$.

$$\hat{\beta} - \beta_t = (\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X})^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_1 (\hat{\beta} - \beta_t) \\ \vdots \\ (\hat{\beta} - \beta_t)^\top \ddot{\boldsymbol{p}}_n (\hat{\beta} - \beta_t) \end{pmatrix} \right)$$

- Für $\hat{\beta} \to \beta_t$ ist $\hat{\beta} - \beta_t = (\underbrace{\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X}}_{=:\boldsymbol{M}})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}$.
- Wähle $(\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}$ als initiale Lösung und substituiere $\hat{\beta} - \beta_t$ auf der rechten Seite:

$$\hat{\beta} - \beta_t \approx \boldsymbol{M}^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} (\boldsymbol{M}^{-1})^\top \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} (\boldsymbol{M}^{-1})^\top \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \right)$$

$$\hat{\beta} - \beta_t \approx \boldsymbol{M}^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} (\boldsymbol{M}^{-1})^\top \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} (\boldsymbol{M}^{-1})^\top \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \right)$$

$$= \boldsymbol{M}^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \right)$$

- Bemerkung: $\boldsymbol{M}^{-1}$ ist symmetrisch, da $\boldsymbol{M}$ symmetrisch ist.

$$\boldsymbol{M}(\boldsymbol{M}^{-1})^\top = \boldsymbol{M}^\top (\boldsymbol{M}^{-1})^\top = \boldsymbol{I}_p$$

$$\Rightarrow \boldsymbol{M}^{-1} = (\boldsymbol{M}^{-1})^\top.$$

$$\hat{\beta} - \beta_t \approx \boldsymbol{M}^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \right)$$

- Linker Summand:

$$\begin{aligned} E[(\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon}] &= E[E[(\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \mid \boldsymbol{X}]] \\ &= E[(\boldsymbol{X}^\top \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \underbrace{E[\boldsymbol{\epsilon} \mid \boldsymbol{X}]}_{=\boldsymbol{0}}] \\ &= \boldsymbol{0} \end{aligned}$$

$$\hat{\beta} - \beta_t \approx \boldsymbol{M}^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \right)$$

- Rechter Summand: Für $\boldsymbol{X}$-messbares $\boldsymbol{A}$ gilt:

$$\begin{aligned} E[\boldsymbol{\epsilon}^\top \boldsymbol{A} \boldsymbol{\epsilon} \mid \boldsymbol{X}] &= E\left[\sum_{i,j=1}^{n} \boldsymbol{A}_{ij} \epsilon_i \epsilon_j \mid \boldsymbol{X} \right] \\ &= \sum_{i,j=1}^{n} \boldsymbol{A}_{ij} \cdot \underbrace{E[\epsilon_i \epsilon_j \mid \boldsymbol{X}]}_{= Cov(\epsilon_i, \epsilon_j \mid \boldsymbol{X})} \\ &= \sum_{i=1}^{n} \boldsymbol{A}_{ii} \cdot Var[\epsilon_i \mid \boldsymbol{X}] \\ &= \mathrm{tr}(\boldsymbol{A} Var[\boldsymbol{\epsilon} \mid \boldsymbol{X}]) \\ &= \mathrm{tr}(\boldsymbol{A} \boldsymbol{V}) \end{aligned}$$

$$\hat{\beta} - \beta_t \approx \boldsymbol{M}^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \right)$$

- Rechter Summand:

$$E \left[\begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \mid \boldsymbol{X} \right] = \begin{pmatrix} \operatorname{tr}(\boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{V}) \\ \vdots \\ \operatorname{tr}(\boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{V}) \end{pmatrix}$$

$$\ddot{\boldsymbol{p}}_k = v_k(1-2p_k)\cdot x_i x_i^\top$$

- Betrachte zunächst $\operatorname{tr}(\boldsymbol{X}\boldsymbol{M}^{-1}\ddot{\boldsymbol{p}}_k\boldsymbol{M}^{-1}\boldsymbol{X}^\top\boldsymbol{V})$ ohne $v_k(1-2p_k)$:

$$\begin{aligned} x_k^\top \boldsymbol{M}^{-1}\boldsymbol{X}^\top &= x_k^\top \begin{pmatrix} \boldsymbol{M}^{-1}x_1 & \dots & \boldsymbol{M}^{-1}x_n \end{pmatrix} \\ &= \begin{pmatrix} x_k^\top\boldsymbol{M}^{-1}x_1 & \dots & x_k^\top\boldsymbol{M}^{-1}x_n \end{pmatrix} \in \mathbb{R}^{1\times n} \end{aligned}$$

$$\begin{aligned} \operatorname{tr}(\boldsymbol{X}\boldsymbol{M}^{-1}x_k x_k^\top \boldsymbol{M}^{-1}\boldsymbol{X}^\top\boldsymbol{V}) &= \operatorname{tr}((x_k^\top\boldsymbol{M}^{-1}\boldsymbol{X}^\top)^\top(x_k^\top\boldsymbol{M}^{-1}\boldsymbol{X}^\top)\boldsymbol{V}) \\ &= \sum_{i=1}^n (x_k^\top\boldsymbol{M}^{-1}x_i)^2\cdot v_i \\ &= \sum_{i=1}^n (x_k^\top\boldsymbol{M}^{-1}x_i)(x_i^\top\boldsymbol{M}^{-1}x_k)\cdot v_i \\ &= x_k^\top\boldsymbol{M}^{-1}\underbrace{\left(\sum_{i=1}^n v_i\cdot x_i x_i^\top\right)}_{=\boldsymbol{X}^\top\boldsymbol{V}\boldsymbol{X}=\boldsymbol{M}}\boldsymbol{M}^{-1}x_k \\ &= x_k^\top\boldsymbol{M}^{-1}x_k \end{aligned}$$

$$\operatorname{tr}(\boldsymbol{X}\boldsymbol{M}^{-1}x_k x_k^\top \boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{V}) = x_k^\top \boldsymbol{M}^{-1} x_k$$

- Konstanten lassen sich einfach aus der Spur herausziehen:

$$\begin{pmatrix} \operatorname{tr}(\boldsymbol{X}\boldsymbol{M}^{-1}\ddot{\boldsymbol{p}}_1\boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{V}) \\ \vdots \\ \operatorname{tr}(\boldsymbol{X}\boldsymbol{M}^{-1}\ddot{\boldsymbol{p}}_n\boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{V}) \end{pmatrix} = \begin{pmatrix} v_1(1-2p_1)\operatorname{tr}(\boldsymbol{X}\boldsymbol{M}^{-1}x_1 x_1{}^\top \boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{V}) \\ \vdots \\ v_n(1-2p_n)\operatorname{tr}(\boldsymbol{X}\boldsymbol{M}^{-1}x_n x_n^\top \boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{V}) \end{pmatrix}$$

$$= \boldsymbol{V} \begin{pmatrix} (1-2p_1)x_1^\top \boldsymbol{M}^{-1}x_1 \\ \vdots \\ (1-2p_n)x_n^\top \boldsymbol{M}^{-1}x_n \end{pmatrix}$$

$$\hat{\beta} - \beta_t \approx \boldsymbol{M}^{-1} \left(\boldsymbol{X}^\top \boldsymbol{\epsilon} - \frac{1}{2} \cdot \boldsymbol{X}^\top \begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \right)$$

- Wir fassen zusammen:

$$E[\hat{\beta} - \beta_t \mid \boldsymbol{X}]$$

$$\approx \underbrace{E[\boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \mid \boldsymbol{X}]}_{=0} - \frac{1}{2} \cdot \boldsymbol{M}^{-1} \boldsymbol{X}^\top E \left[\begin{pmatrix} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_1 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} \ddot{\boldsymbol{p}}_n \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix} \mid \boldsymbol{X} \right]$$

$$= -\frac{1}{2} \cdot \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{V} \begin{pmatrix} (1 - 2p_1) x_1^\top \boldsymbol{M}^{-1} x_1 \\ \vdots \\ (1 - 2p_n) x_n^\top \boldsymbol{M}^{-1} x_n \end{pmatrix}$$

- Für den Bias von $\hat{\beta}$ integrieren wir schließlich $E[\hat{\beta} - \beta_t \mid \boldsymbol{X}]$ aus.

Bias Correction

- In der Praxis: Um den hergeleiteten approximativen Bias korrigieren nicht möglich, da β_t unbekannt ist.

Bias Correction

- In der Praxis: Um den hergeleiteten approximativen Bias korrigieren nicht möglich, da β_t unbekannt ist.
- Stattdessen: Bestimme $\hat{\beta}$ und schätze damit $\boldsymbol{V}$ und $\boldsymbol{p}$.

$$\widehat{\text{bias}(\hat{\beta})} = -\frac{1}{2} \cdot \boldsymbol{M}^{-1}\boldsymbol{X}^\top \hat{\boldsymbol{V}} \begin{pmatrix} (1-2\hat{p}_1)x_1^\top \boldsymbol{M}^{-1} x_1 \\ \vdots \\ (1-2\hat{p}_n)x_n^\top \boldsymbol{M}^{-1} x_n \end{pmatrix}$$

Bias Correction

- In der Praxis: Um den hergeleiteten approximativen Bias korrigieren nicht möglich, da β_t unbekannt ist.
- Stattdessen: Bestimme $\hat{\beta}$ und schätze damit $\boldsymbol{V}$ und $\boldsymbol{p}$.

$$\widehat{\text{bias}(\hat{\beta})} = -\frac{1}{2} \cdot \boldsymbol{M}^{-1}\boldsymbol{X}^\top \hat{\boldsymbol{V}} \begin{pmatrix} (1-2\hat{p}_1)x_1^\top \boldsymbol{M}^{-1} x_1 \\ \vdots \\ (1-2\hat{p}_n)x_n^\top \boldsymbol{M}^{-1} x_n \end{pmatrix}$$

- Bias Corrected MLE:

$$\hat{\beta}_{\text{corrected}} = \hat{\beta} - \widehat{\text{bias}(\hat{\beta})}$$

Asymptotik der Bias Correction

- Wie verhält sich unser (idealer) Bias Correction Term asymptotisch?

The bias is

$$\text{bias}(\hat{\beta}) \approx (-1/2)(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\{\text{trace}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}E^{k}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\,\text{Var}(\boldsymbol{\varepsilon}))\}$$
$$(-1/2)(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}\{(1-2\pi_k)\mathbf{x}^{\mathrm{T}}{}_k(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_k\} \tag{6}$$

where $\{a_k\}$ represents a vector whose kth element is a_k.

By writing a typical element of (6) as a sum and noting that, since we assumed $\lim(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})/n$ finite, the elements of $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$ are $O(1/n)^6$, one can show that the elements in (6) are $O(1/n)$.

Asymptotik der Bias Correction

- Wie verhält sich unser (idealer) Bias Correction Term asymptotisch?

The bias is

$$\text{bias}(\hat{\boldsymbol{\beta}}) \approx (-1/2)(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\{\text{trace}(\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}E^{k}(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\,\text{Var}(\boldsymbol{\varepsilon}))\}$$
$$(-1/2)(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}\{(1-2\pi_k)\mathbf{x}^{\mathrm{T}}{}_k(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_k\} \tag{6}$$

where $\{a_k\}$ represents a vector whose kth element is a_k.

By writing a typical element of (6) as a sum and noting that, since we assumed $\lim(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})/n$ finite, the elements of $(\mathbf{X}^{\mathrm{T}}\mathbf{V}\mathbf{X})^{-1}$ are $O(1/n)$[6], one can show that the elements in (6) are $O(1/n)$.

- Wir zeigen:

Proposition

$n \cdot E[\hat{\beta} - \beta_t \mid \boldsymbol{X}] \xrightarrow{\mathbb{P}} -\frac{1}{2}E[v \cdot xx^\top]^{-1}E\left[(1-2p)v \cdot xx^\top E[v \cdot xx^\top]^{-1}x\right] \in \mathbb{R}^p.$
Insbesondere ist $E[\hat{\beta} - \beta_t \mid \boldsymbol{X}] \in O_p(1/n)$.

Stetigkeit: Matrix invertieren

Lemma

Die Abbildung $(A \mapsto A^{-1}) : \mathbb{GL}_n(\mathbb{R}) \to \mathbb{GL}_n(\mathbb{R})$ *ist stetig bezüglich des von* $(\mathbb{R}^{n\times n}, \|\cdot\|)$ *induzierten metrischen Raumes.*

Beweis: Invertiere $A \in \mathbb{GL}_n(\mathbb{R})$ mithilfe der Cramerschen Regel

$$A^{-1} = \frac{1}{\det A}((-1)^{i+j} \det A_{ji})_{ij}$$

wobei A_{ji} die Streichmatrix bezeichnet, welche durch Weglassen der i-ten Zeile und j-ten Spalte entsteht. Dass det stetig ist, sieht man z.B. anhand der Leibnizformel:

$$\det A = \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)}.$$

Asymptotik der Bias Correction

Proposition

$n \cdot E[\hat{\beta} - \beta_t \mid \boldsymbol{X}] \xrightarrow{\mathbb{P}} -\frac{1}{2} E[v \cdot xx^\top]^{-1} E\left[v(1-2p) \cdot xx^\top E[v \cdot xx^\top]^{-1} x\right] \in \mathbb{R}^p.$
Insbesondere ist $E[\hat{\beta} - \beta_t \mid \boldsymbol{X}] \in O_p(1/n)$.

Beweis: Da $E[v \cdot xx^\top]$ endlich ist, folgt mit dem starken Gesetz der Großen Zahlen

$$\frac{1}{n} M = \frac{1}{n} \sum_{i=1}^{n} v_i \cdot x_i x_i^\top \xrightarrow{f.s.} E[v \cdot xx^\top].$$

Insbesondere existiert $E[v \cdot xx^\top]^{-1}$ wegen der positiven Definitheit.
Continous Mapping Theorem:

$$nM^{-1} = \left(\frac{1}{n} M\right)^{-1} \xrightarrow{f.s.} E[v \cdot xx^\top]^{-1}.$$

$$n \cdot E[\hat{\beta} - \beta_t \mid \boldsymbol{X}] = -\frac{1}{2} n \cdot M^{-1} X^\top V \begin{pmatrix} (1-2p_1) x_1^\top M^{-1} x_1 \\ \vdots \\ (1-2p_n) x_n^\top M^{-1} x_n \end{pmatrix}$$

$$= -\frac{1}{2} n M^{-1} \sum_{i=1}^{n} v_i (1-2p_i) x_i x_i^\top M^{-1} x_i$$

$$= -\frac{1}{2} \underbrace{nM^{-1}}_{\xrightarrow{f.s.} E[v \cdot xx^\top]} \cdot \frac{1}{n} \sum_{i=1}^{n} v_i (1-2p_i) x_i x_i^\top n M^{-1} x_i$$

Wegen Slutsky's Lemma reicht es also zu zeigen, dass

$$\frac{1}{n} \sum_{i=1}^{n} v_i (1-2p_i) x_i x_i^\top n M^{-1} x_i \xrightarrow{\mathbb{P}} E\left[v(1-2p) \cdot xx^\top E[v \cdot xx^\top]^{-1} x \right].$$

$$\frac{1}{n}\sum_{i=1}^{n} v_i(1-2p_i)x_i x_i^\top nM^{-1}x_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} v_i(1-2p_i)x_i x_i^\top \left((nM^{-1} - E[v\cdot xx^\top]^{-1}) + E[v\cdot xx^\top]^{-1}\right) x_i$$

Da $\|x\| \leq C < \infty$ fast sicher, existieren alle Momente $E[\|x\|^n]$ für $n \in \mathbb{N}$. Wegen der Äquivalenz der Normen, gilt für geeignetes $d > 0$:

$$E\left[\left\|v(1-2p)xx^\top E[v\cdot xx^\top]^{-1}x\right\|\right] \leq E\left[d\|x\|^3 \left\|E[v\cdot xx^\top]^{-1}\right\|\right] < \infty.$$

Nach dem schwachen Gesetz der großen Zahlen:

$$\frac{1}{n}\sum_{i=1}^{n} v_i(1-2p_i)x_i x_i^\top E[v\cdot xx^\top]^{-1}x_i \xrightarrow{\mathbb{P}} E\left[v(1-2p)\cdot xx^\top E[v\cdot xx^\top]^{-1}x\right].$$

Weiter gilt fast sicher:

$$\begin{aligned}
&\left\| \frac{1}{n} \sum_{i=1}^{n} v_i (1 - 2p_i) x_i x_i^\top (nM^{-1} - E[v \cdot xx^\top]^{-1}) x_i \right\| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} dC^3 \left\| nM^{-1} - E[v \cdot xx^\top]^{-1} \right\| \\
&= dC \left\| nM^{-1} - E[v \cdot xx^\top]^{-1} \right\| \xrightarrow{n \to \infty} 0
\end{aligned}$$

Insgesamt also:

$$\frac{1}{n} \sum_{i=1}^{n} v_i (1 - 2p_i) x_i x_i^\top nM^{-1} x_i \xrightarrow{\mathbb{P}} E\left[v(1 - 2p) \cdot xx^\top E[v \cdot xx^\top]^{-1} x \right].$$

□

- Bemerkung: Falls wir wissen, dass $\sup_{n\in\mathbb{N}} \left\| nM^{-1} \right\| < \infty$, können wir mit dominierter Konvergenz auch zeigen, dass

$$
\begin{aligned}
& E[n \cdot E[\hat{\beta} - \beta_t \mid \boldsymbol{X}]] \\
& \xrightarrow{n\to\infty} -\frac{1}{2} E[v \cdot xx^\top]^{-1} E\left[v(1-2p) \cdot xx^\top E[v \cdot xx^\top]^{-1} x \right]
\end{aligned}
$$

- Bemerkung: Falls wir wissen, dass $\sup_{n\in\mathbb{N}} \left\| nM^{-1} \right\| < \infty$, können wir mit dominierter Konvergenz auch zeigen, dass

$$
\begin{aligned}
& E[n \cdot E[\hat{\beta} - \beta_t \mid \boldsymbol{X}]] \\
& \xrightarrow{n\to\infty} -\frac{1}{2} E[v \cdot xx^\top]^{-1} E\left[v(1-2p) \cdot xx^\top E[v \cdot xx^\top]^{-1} x \right]
\end{aligned}
$$

- Hierfür zeigen wir, dass für jede Teilfolge der Erwartungswerte eine konvergente Teilteilfolge existiert.

- Bemerkung: Falls wir wissen, dass $\sup_{n\in\mathbb{N}} \left\| nM^{-1} \right\| < \infty$, können wir mit dominierter Konvergenz auch zeigen, dass

$$E[n \cdot E[\hat{\beta} - \beta_t \mid \boldsymbol{X}]]$$
$$\xrightarrow{n\to\infty} -\frac{1}{2} E[v \cdot xx^\top]^{-1} E\left[v(1-2p) \cdot xx^\top E[v \cdot xx^\top]^{-1} x \right]$$

- Hierfür zeigen wir, dass für jede Teilfolge der Erwartungswerte eine konvergente Teilteilfolge existiert.
- Verwende dazu die äquivalente Formulierung für stochastisch konvergente Zufallsvariablen:

Proposition

$X_n \xrightarrow{\mathbb{P}} X$ genau dann, wenn für jede Teilfolge $(X_{n_m})_{m\in\mathbb{N}}$ eine Teilteilfolge $(X_{n_{m_k}})_{k\in\mathbb{N}}$ existiert, sodass $X_{n_{m_k}} \xrightarrow{f.s.} X$.

Term 3. Ordnung

- Was passiert mit dem $o\left(\left\|\hat{\beta} - \beta_t\right\|^2\right)$ Term?

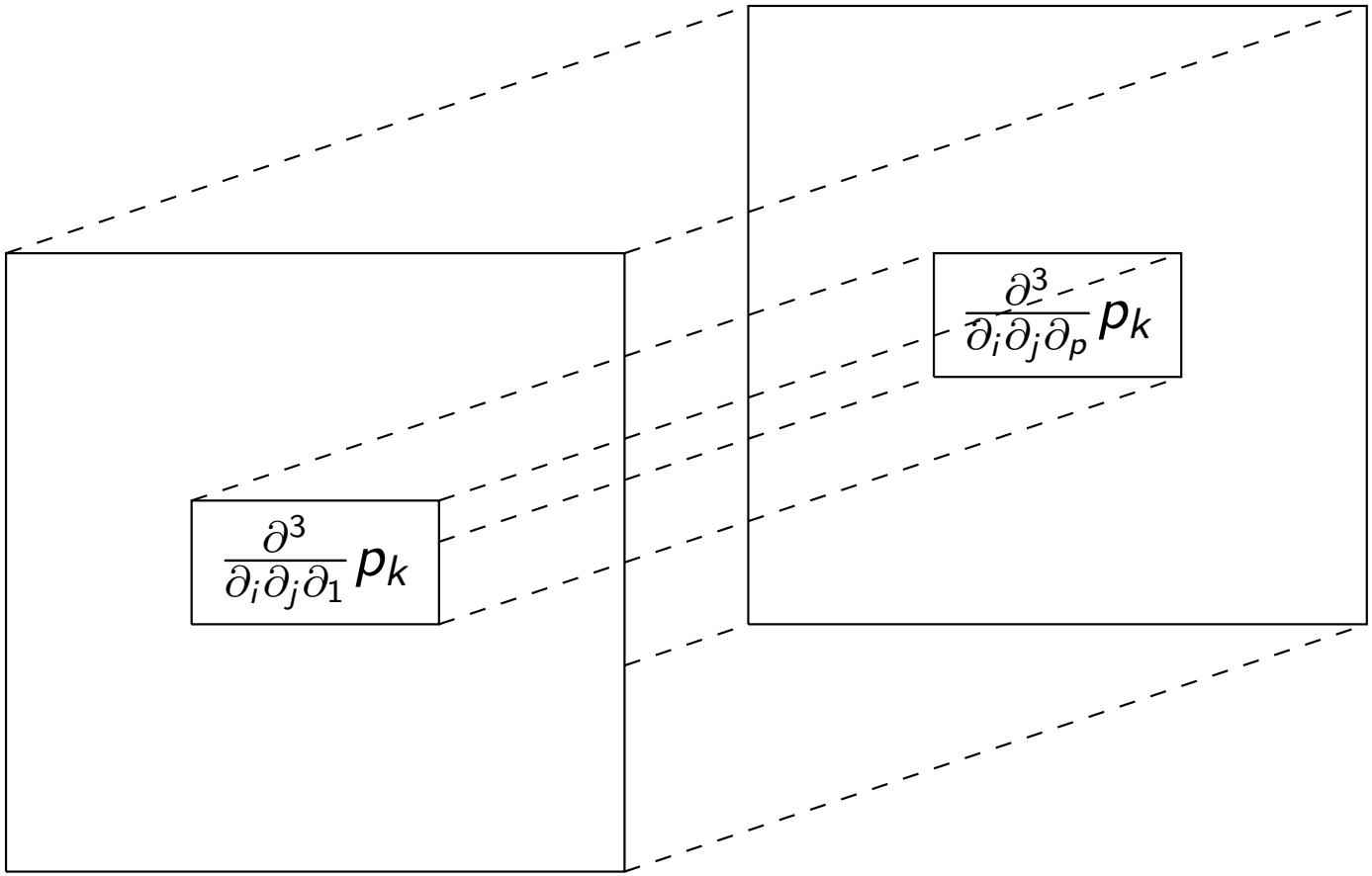
To ensure accounting for all terms of $O(1/n)$, we carried the Taylor series expansion in (3) to the third order. The resulting additional terms were all $O(1/n^2)$, hence (6) represents the bias of $\hat{\boldsymbol{\beta}}$ to order $O(1/n)$.

- Ausrechnen...

- Wir berechnen für jedes $k = 1, \ldots, n$ die dritte Ableitung $\dddot{\boldsymbol{p}}_k$:

$$\dddot{\boldsymbol{p}}_k = \begin{bmatrix} \frac{\partial^3}{\partial_1 \partial_1 \partial_1} p_k & \cdots & \frac{\partial^3}{\partial_1 \partial_p \partial_1} p_k & \frac{\partial^3}{\partial_1 \partial_1 \partial_p} p_k & \cdots & \frac{\partial^3}{\partial_1 \partial_p \partial_p} p_k \\ \vdots & & \vdots & \vdots & & \vdots \\ \frac{\partial^3}{\partial_p \partial_1 \partial_1} p_k & \cdots & \frac{\partial^3}{\partial_p \partial_p \partial_1} p_k & \frac{\partial^3}{\partial_p \partial_1 \partial_p} p_k & \cdots & \frac{\partial^3}{\partial_p \partial_p \partial_p} p_k \end{bmatrix}$$

$$\dddot{p}_k^{(i,j)} =$$



$$= \frac{d}{d\beta} v_k (1 - 2p_k) x_k^{(i)} x_k^{(j)}$$

$$= \underbrace{\left(v_k (1 - 2p_k)^2 - 2v_k^2\right)}_{=:s_k} x_k^\top x_k^{(i)} x_k^{(j)}$$

$$\dddot{\boldsymbol{p}}_k^{(i,j)} = s_k x_k^\top x_k^{(i)} x_k^{(j)}$$

- Multiplizieren des Tensors $\dddot{\boldsymbol{p}}_k$ mit den Richtungen $\hat{\beta} - \beta_t$:

$$\frac{1}{6} s_k \left\{ x_k^{(d)} (\hat{\beta} - \beta_t)^\top x_k x_k^\top (\hat{\beta} - \beta_t) \right\}_d^\top (\hat{\beta} - \beta_t)$$

- Einsetzen der Initiallösung $\boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}$ liefert als Term 3. Ordnung:

$$\boldsymbol{M}^{-1} \sum_{i=1}^{n} o(\left\| \hat{\beta} - \beta_t \right\|^2) x_i$$

$$= \frac{1}{6} \cdot \boldsymbol{M}^{-1}\boldsymbol{X}^\top \begin{pmatrix} s_1 \left\{ x_1^{(d)} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} x_1 x_1^\top \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \right\}_d^\top \boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon} \\ \vdots \\ s_n \left\{ x_n^{(d)} \boldsymbol{\epsilon}^\top \boldsymbol{X} \boldsymbol{M}^{-1} x_n x_n^\top \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{\epsilon} \right\}_d^\top \boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon} \end{pmatrix}$$

- Das sieht ganz ähnlich aus, wie der Term 2. Ordnung mit zusätzlichem $O_p(1/n)$ Term $\boldsymbol{M}^{-1}\boldsymbol{X}^\top \boldsymbol{\epsilon}$.

Experimente - Setup

- Simulationen für Stichprobengrößen $n \in \{25, 200\}$ und Anzahl der Features $p \in \{2, 5\}$.

Experimente - Setup

- Simulationen für Stichprobengrößen $n \in \{25, 200\}$ und Anzahl der Features $p \in \{2, 5\}$.
- Auf $[0, 1]^p$ uniform verteilte $x_1, \ldots, x_n$.

Experimente - Setup

- Simulationen für Stichprobengrößen $n \in \{25, 200\}$ und Anzahl der Features $p \in \{2, 5\}$.
- Auf $[0, 1]^p$ uniform verteilte $x_1, \ldots, x_n$.
- Wahrer Parameter β_t gewählt als Eigenvektor des kleinsten bzw. größten Eigenwerts von $\boldsymbol{X}^\top \boldsymbol{X}$.

Experimente - Setup

- Simulationen für Stichprobengrößen $n \in \{25, 200\}$ und Anzahl der Features $p \in \{2, 5\}$.
- Auf $[0, 1]^p$ uniform verteilte $x_1, \ldots, x_n$.
- Wahrer Parameter β_t gewählt als Eigenvektor des kleinsten bzw. größten Eigenwerts von $\boldsymbol{X}^\top \boldsymbol{X}$.
- Bestimme Wahrscheinlichkeiten p_i und ziehe y_i wie folgt:

$$y_i = \begin{cases} 1 & p_i \leq u(0, 1) \\ 0 & \text{sonst} \end{cases}$$

Experimente - Setup

- Simulationen für Stichprobengrößen $n \in \{25, 200\}$ und Anzahl der Features $p \in \{2, 5\}$.
- Auf $[0, 1]^p$ uniform verteilte $x_1, \ldots, x_n$.
- Wahrer Parameter β_t gewählt als Eigenvektor des kleinsten bzw. größten Eigenwerts von $\boldsymbol{X}^\top \boldsymbol{X}$.
- Bestimme Wahrscheinlichkeiten p_i und ziehe y_i wie folgt:

$$y_i = \begin{cases} 1 & p_i \leq u(0, 1) \\ 0 & \text{sonst} \end{cases}$$

- Pro Konfiguration 500 Replikationen der obigen Prozedur.

Experimente - Ergebnisse

Eigenvalue	$p-1$	Sample Size	ML	Corrected ML
smallest	2	25	10.38 (12.22)	7.46 (7.94)
		200	4.79 (2.37)	4.68 (2.30)
smallest	5	25	39.75 (74.36)	13.03 (9.99)
		200	5.63 (2.63)	5.38 (2.48)
largest	2	25	16.73 (33.54)	8.91 (7.04)
		200	4.73 (1.05)	4.63 (1.01)
largest	5	25	50.22 (66.25)	12.48 (8.10)
		200	5.94 (1.53)	5.65 (1.41)

Table: Mean-Squared Error für den MLE und den korrigierten MLE. In Klammern jeweils dessen Standardabweichung.

Zusammenfassung

- Approximation des Bias durch Taylorentwicklung der Score-Funktion bis zur zweiten Ordnung.

Zusammenfassung

- Approximation des Bias durch Taylorentwicklung der Score-Funktion bis zur zweiten Ordnung.
- Korrektur des MLE um den approximierten Bias.

Zusammenfassung

- Approximation des Bias durch Taylorentwicklung der Score-Funktion bis zur zweiten Ordnung.
- Korrektur des MLE um den approximierten Bias.
- Approximierter Bias liegt in $O(1/n)$.

Zusammenfassung

- Approximation des Bias durch Taylorentwicklung der Score-Funktion bis zur zweiten Ordnung.
- Korrektur des MLE um den approximierten Bias.
- Approximierter Bias liegt in $O(1/n)$.
- Bei extrem kleinen Datensätzen können substantielle Verbesserung beobachtet werden.

Zusammenfassung

- Approximation des Bias durch Taylorentwicklung der Score-Funktion bis zur zweiten Ordnung.
- Korrektur des MLE um den approximierten Bias.
- Approximierter Bias liegt in $O(1/n)$.
- Bei extrem kleinen Datensätzen können substantielle Verbesserung beobachtet werden.
- Tradeoff: Laufzeit vs. Verbesserung.

References

- Schaefer, R. (1983). Bias Correction in Maximum Likelihood Logistic Regression. *Statistics in Medicine, Vol.2*
- Rohde, A. (2021). Wahrscheinlichkeitstheorie
- Huber-Klawitter, A. (2019). Lineare Algebra I.