Der LE M-Schätzer der logistischen Regression

Vortrag: Nils Kornacker

28. Juli 2020

Albert-Ludwigs-Universität Freiburg

Inhaltsverzeichnis

Albert-Ludwigs-Universität Freiburg

Problemstellung & Ziel

Mit welcher Wahrscheinlichkeit folgt den Inputdaten $X_{n+1} = x_{n+1} \in \mathbb{R}^p$, Klassifizierung $Y_{n+1} = y_{n+1} \in \{0, 1, 2, 3, \ldots, K\}$ (i.e. Zustand/Ereignis)?

Dichotomes ($K = 1$) und polytomes ($K > 1$) logistisches Modell.

(Auch binär/multinomial).

> *„Logistic regression is the most common method used to model binary response data.“ (Hilbe)*

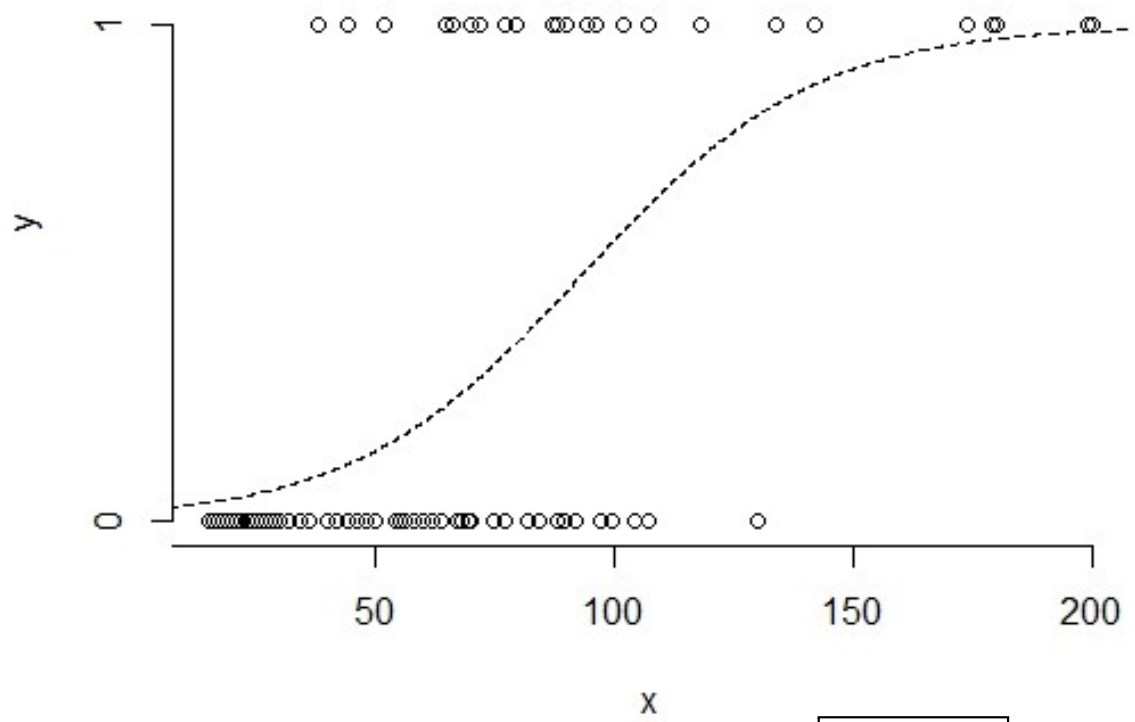
> *„[...] [L]ogistic regression is by far the most popular modeling procedure used to analyze epidemiologic data when the illness measure is dichotomous.“ (Kleinbaum, Klein)*

Aktualität Überwältigend

$(X_v)_1 = 1 \ \forall v.$

Einführung - Beispiel

81 Realisierungen einer nach P verteilten Zufallsvariable $(Y, X) \in \{0, 1\} \times \mathbb{R}^2$ (eigenständig erdacht).



Millionär - **Einkommen**. Fit nach LRM. ↷ UQA's. Univariate Qualitative Antwortmodelle

Tod in den nächsten 10 Jahren - Blutwert(μg/L),

(Antwort)Reaktion - Expositionsdauer/ -konzentration.

Univariate Qualitative Antwortmodelle (UQA's) im dichotomen Fall

Gegeben eine Funktion $F : \mathbb{R} \to \mathbb{R}$ *und Daten* $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ *ist ein univariates qualitatives Antwortmodell durch*

$$\mathbf{P}(Y_v = 1 | X_v = x_v) = F(x_v^T \theta_0) \qquad v = 1, \ldots, n$$

definiert. Y_v *sind unabhängige,* $\{0,1\}$*-wertige Zufallsvariablen,* $\theta_0 \in \mathbb{R}^p$ *ein (unbekannter) Parametervektor und* $X_v \in \mathbb{R}^p$*.* $\forall v \in \{1, \ldots, n\}$*.*

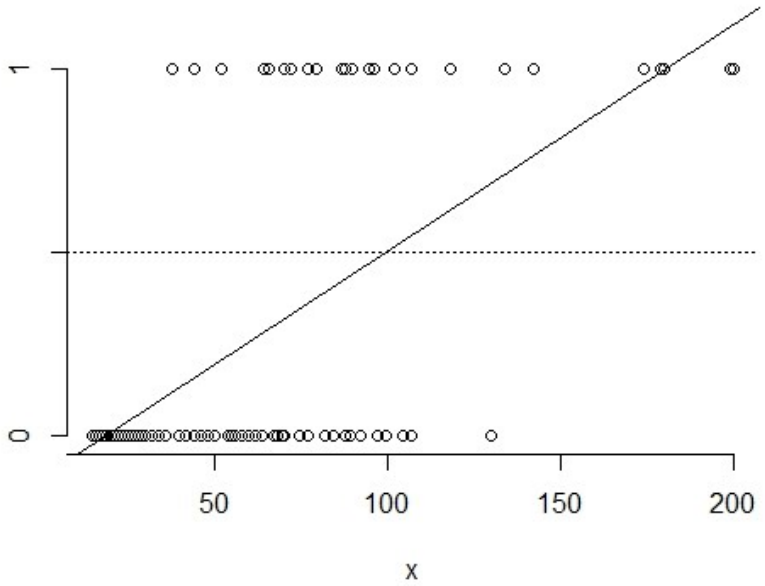
Die Einschränkung $(X_v)_1 = 1$ erlaubt eine größere Klasse von Modellen und insbesondere solche, die nichttrivial unabhängig von den Daten $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ sind.

Eine abhängige Variable $\Leftrightarrow$ Univariat.

[(Amemiya)]

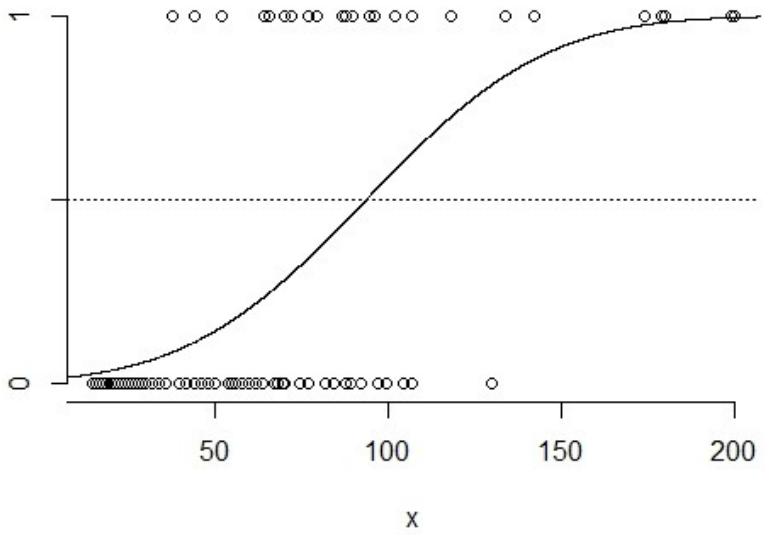
Wahl der Funktion - UQA's

$$F(x) = x$$



offensichtlicher Nachteil - „defect“:
Wahrscheinlichkeiten in $\mathbb{R}\backslash[0,1]$.

$$F(x) = \Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-(x^2/2)}$$

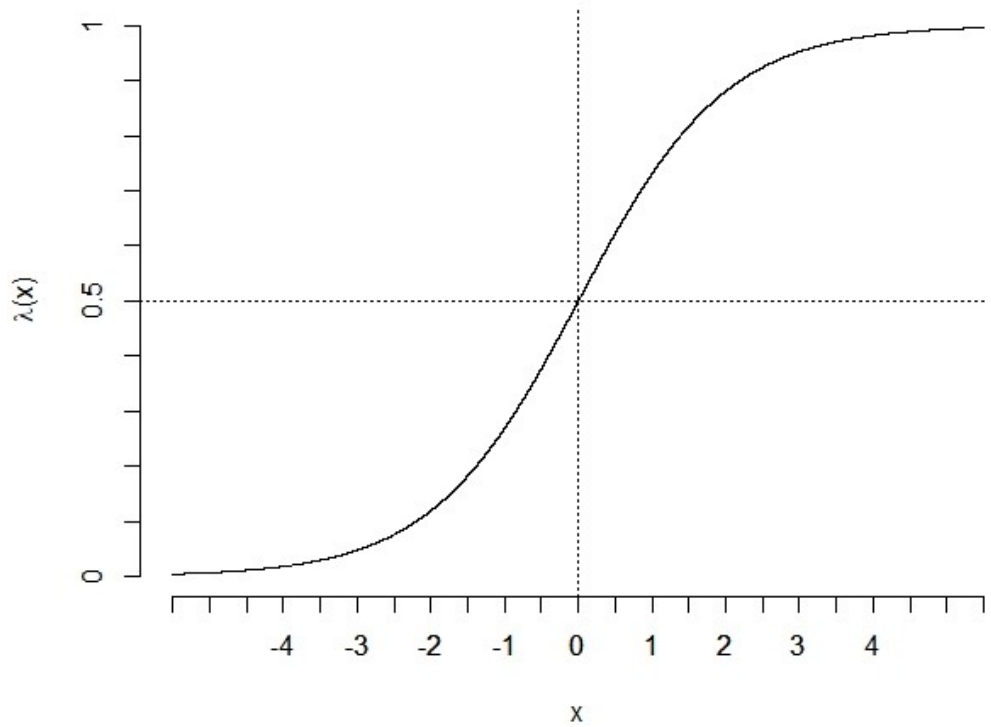


Oft genutzt, Grundlage des „Probit“-Modells. Aber: Konkrete Rechnungen aufwendiger.
↳ Toleranzverteilung

Wahl der Funktion - UQA's

Die logistische Funktion $\lambda : \begin{array}{l} \mathbb{R} \to (0,1) \\ x \mapsto \frac{e^x}{1+e^x} \end{array}$.

Vergleiche wie ähnlich $\lambda(x\pi/\sqrt{3}), \Phi(x)$.



Bem. nochmals:

$0 < \lambda(x) < 1$

sowie

$$\lambda'(x) = \lambda(x)(1 - \lambda(x)).$$

„[...] [T]he choice of F is not critical as long as it is a distribution function." (Amemiya)

„[...] [E]xtremely flexible and easy used function." (Hosmer, Lemeshow und Sturdivant).

Modell

Das dichotome logistische Regressionsmodell ist definiert durch

$$Y_v|X_v = x_v \;\sim\; \text{Ber}(\pi(x_v)) \;\; v = 1, \ldots, n,$$

wobei $(Y_v, X_v) \sim_{iid} P$ *unbekannt,* $Y_v \in \{0, 1\}$ *und* $X_v \in \mathbb{R}^p$.

$$\pi(x) := \lambda(x^T\theta_0)$$

Äquivalent können verwendet werden:

a) $\mathbf{E}[Y_v|X_v = x_v] = \pi(x_v)$;

Oder

b) $\frac{\mathbf{P}(Y_v^{=1}|x_v)}{1-\mathbf{P}(Y_v^{=1}|x_v)} = e^{x_v^T\theta_0}$;

c) $\ln\left(\frac{\mathbf{P}(Y_v^{=1}|x_v)}{1-\mathbf{P}(Y_v^{=1}|x_v)}\right) = x_v^T\theta_0$. mit $\{Y_v^{=1}|x_v\} := \{Y_v = 1|X_v = x_v\}$.

Exkurs (Odds und Daten)

Für Ereignisse E, E_i, $i = 1, 2$ definieren wir Odds und Odds-Ratio:

$$\textbf{odds}(\mathsf{E}) := \frac{\mathbf{P}(\mathsf{E})}{1-\mathbf{P}(\mathsf{E})},$$
$$\textbf{oddsratio}(\mathsf{E}_1, \mathsf{E}_2) := \frac{\mathbf{P}(\mathsf{E}_1)/(1-\mathbf{P}(\mathsf{E}_1))}{\mathbf{P}(\mathsf{E}_2)/(1-\mathbf{P}(\mathsf{E}_2))}.$$

Im logistischen Modell gilt unter dem wahren Parameter θ_0 und mit den Ereignissen $\mathsf{E} = \{Y_v = 1 | X_v = x_v\}$ & $\mathsf{E}_1 = \{Y_v = 1 | X_v = x_v + a \cdot e_k\}$; $a \in \mathbb{R}$:

$$\textbf{odds}(\mathsf{E}) = \frac{\mathbf{P}(\mathsf{E})}{1-\mathbf{P}(\mathsf{E})} = \frac{\pi(x_v)}{1-\pi(x_v)} = e^{x_v^\mathsf{T}\theta_0},$$
$$\textbf{oddsratio}(\mathsf{E}_1, \mathsf{E}) = \frac{\textbf{odds}(\mathsf{E}_1)}{\textbf{odds}(\mathsf{E})} = \frac{e^{(x_v + a\cdot e_k)^\mathsf{T}\theta_0}}{e^{x_v^\mathsf{T}\theta_0}} = e^{a\cdot(\theta_0)_k}.$$

$\Rightarrow$ Ein um $a \in \mathbb{R}$ höherer *Wert* der k-ten unabhängigen Variable erhöht die Chancen um den Faktor $e^{a\cdot(\theta_0)_k}$.

$\rightsquigarrow$ Interpretation der Parameter. Fit!
„Damit erleichtert sich die Interpretation.“ (Backhaus et al.).
$\hookrightarrow$ . . . der Parameter.

Problem?

Exkurs (Odds und Daten)

Stärke: „[...] ability to handle many variables.“ (Hosmer, L., S.).

- Metrisch- oder intervallskalierte Daten kein 'besonderer' Umgang.
- Nominal- (Geschlecht, Ethnie, Ernährung) und ordinalskalierte Daten (Bundesligatabelle, Platzierung 3000m) mittels *Dummy-Variablen*.

Beispiel.: Betrachte nominalskalierte Variable Ernährungsform; $V \in \{1, 2, 3, 4\}$ und
1 = omnivor, 2 = vegetarisch, 3 = vegan, 4 = sonstige.
↳ stattdessen *DV's* einbinden: $V_i \in \{0, 1\}_{i \in \{1,2,3\}}$.

	V_1	V_2	V_3
vegetarisch	1	0	0
vegan	0	1	0
sonstige	0	0	1
omnivor	0	0	0

V_1 bis V_3, *DV's*. Eine weniger, da eine Referenzkategorie.

Eine Referenzkategorie. Wahl relevant.

Labelierungsunabhängig →→ Odds-Ratio interpretierbar.

Der Parametervektor$_{\theta_0}$

Wird aus Stichprobe $^{\text{iid}}[(Y_1, X_1), \ldots, (Y_n, X_n)] \in [\{0,1\} \times \mathbb{R}^p]^n$ geschätzt.
Methode: Maximum Likelihood.

$$\begin{aligned}
\mathrm{Li}(\theta) &= \prod_{v=1}^{n} \mathbf{P}(Y_v = y_v | X_v = x_v, \theta) = \prod_{v=1}^{n} (\lambda(x_v^\mathsf{T}\theta))^{y_v} \cdot (1 - \lambda(x_v^\mathsf{T}\theta))^{1-y_v} \\
&= \prod_{v=1}^{n} \frac{e^{y_v x_v^\mathsf{T}\theta}}{1 + e^{x_v^\mathsf{T}\theta}}, \\
\mathrm{l}(\theta) &:= \ln(\mathrm{Li}(\theta)) = \sum_{v=1}^{n} y_v x_v^\mathsf{T}\theta - \ln(1 + e^{x_v^\mathsf{T}\theta}), \\
\frac{\partial \mathrm{l}(\theta)}{\partial \theta} &= \sum_{v=1}^{n} x_v (y_v - \lambda(x_v^\mathsf{T}\theta)) \stackrel{!}{=} 0 \text{ (Score Gleichung)}, \\
\frac{\partial^2 \mathrm{l}(\theta)}{\partial \theta \partial \theta^\mathsf{T}} &= \sum_{v=1}^{n} x_v x_v^\mathsf{T} (-\lambda(x_v^\mathsf{T}\theta)(1 - \lambda(x_v^\mathsf{T}\theta))).
\end{aligned}$$

~~**Sehr wichtig:** Die Log-Likelihood ist strikt konkav.~~

Eine kleine Geschichte: Es war einmal:

The consistency and asymptotic normality of $\hat{\theta}_n$ can be proved, for instance, by combining Theorems 5.7 and 5.39. (Alternatively, we may follow the classical approach given in section 5.6. The latter is particularly attractive for the logit model, for which the log likelihood is strictly concave in θ, so that the point of maximum is unique.) For identifiability of θ we must assume that the distribution of the X_i is not concentrated on a $(k-1)$-dimensional

Van der Vaart.

Aber:

if $\boldsymbol{u} \neq \mathbf{0}$. The last inequality is a result of the assumption (3.2): If $n\, a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0)$ converged to 0 as n tends to infinity, we would have

$$\sqrt{\frac{p}{n}} \frac{1}{a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0)} = \frac{\sqrt{p\,n}}{n\, a_{pn}^{\delta}(\boldsymbol{X}, \boldsymbol{\beta}^0)} \xrightarrow[n \to \infty]{} \infty$$

which is a contradiction. The Hessian matrix $\mathbf{H}_{E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}}(\boldsymbol{\beta})$ is thus strictly positive definite and $E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}$ is strictly convex in $\bar{B}(\boldsymbol{\beta}^0, \delta)$. Therefore, under the hypotheses **M1** and **M2**, the absolute minimum $\hat{\boldsymbol{\beta}}$ of $E_{\boldsymbol{y},\sigma_{\mathrm{LR}}}$ (the absolute maximum of $\ln L$) is not only unique but converges in probability to

Beer.

We shall assume that the Hessian matrix of $\log L_T$ is invertible.

Gourieroux, Monfort.

Grundannahmen in soweit ich das überblicken kann allen Beweisen zu Existenz, Konsistenz, asymptotischer Normalität des MLE zielen auf Invertierbarkeit der Hesse-Matrix ab.

Van der Vaart irrt.

Gegenbsp. siehe Ausarbeitung.

Der Parametervektor$_{\theta_0}$

Mit

$$\mathbf{X} := \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \qquad \mathbf{y} := (y_1, \ldots, y_n)^\top$$

$$\begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} =: \mathbf{W}$$

$$\tilde{\lambda} := (\lambda(x_1^\top\theta), \lambda(x_2^\top\theta), \ldots, \lambda(x_n^\top\theta))^\top \quad \text{und} \quad \lambda_v := \tilde{\lambda}_v(1-\tilde{\lambda}_v) = \mathbf{W}_v$$

gilt:

Score Gleichung:	$\mathbf{X}^\top(\mathbf{y} - \tilde{\lambda}) = 0$
Hesse Matrix:	$\mathbf{H} = -\mathbf{X}^\top\mathbf{W}\mathbf{X}$.

H ist negativ semidefinit:

$$x^\top\mathbf{H}x = -(\mathbf{X}x)^\top\mathbf{W}\mathbf{X}x = -a^\top\mathbf{W}a = -\sum_{v=1}^{p} a_v^2\lambda_v \leq 0 \qquad \forall x \in \mathbb{R}^p\backslash\{0\}.$$

Ist **X** mit vollem Rang und $n \geq p$, so folgt Definitheit mit entsprechender Invertierbarkeit der Matrix **H**.

Der Parametervektor$_{\theta_0}$ (Newton-Raphson)

Ausgehend nun von Realisierungen $(Y_v, X_v) = (y_v, x_v),\ v = 1, \ldots, n$.
Lösungsverfahren (der Score Gleichung): i.A. Newton-Raphson

$$\theta^{k+1} = \theta^k - \left(\frac{\partial^2 \mathrm{l}(\theta)}{\partial\theta\partial\theta^\mathsf{T}}\right)^{-1} \frac{\partial \mathrm{l}(\theta)}{\partial\theta},$$

$$\theta^{k+1} = \theta^k + (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}(\mathbf{y} - \tilde{\lambda})$$

$$= (\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{W}\mathbf{z} \ \star$$

Mit $\mathbf{z} := \mathbf{X}\theta^k + \mathbf{W}^{-1}(\mathbf{y} - \tilde{\lambda})$.

$\star$ löst g.k.Q. Problem: $\min_{\theta\in\mathbb{R}^p} \sum_{v=1}^n \mathbf{W}_v(\mathbf{z}_v - x_v^\mathsf{T}\theta)^2$.

(Im Falle $\mathbf{X}$ nicht vollen Rang (bzw. $n < p \rightsquigarrow$ Gene): andere Methoden $\rightsquigarrow$ Levenberg-Marquardt.)

Inhaltsverzeichnis

M-Schätzer

Sei $X_1, \ldots, X_n$ Stichprobe einer Verteilung P und $\theta_0 \in \Theta$ ein unbekannter Parameter (bzw. Funktional) (zugehörig) der Verteilung (im log. Modell bedingten Verteilung $Y_v|X_v$).

Definition

Jeder Schätzer $\hat{\theta}_n \in \Theta$, der eine sogenannte Kriteriumsfunktion

$$\theta \mapsto M_n(\theta) := \frac{1}{n} \sum_{v=1}^{n} m_\theta(X_v)$$

maximiert heißt ein *M-Schätzer*. Die Summanden sind bekannte Funktionen $m_\theta : \mathcal{X} \to \overline{\mathbb{R}}$ ausgewertet an der Stichprobe $X_1, \ldots, X_n$.

Im Falle der Differenzierbarkeit von m_θ:

$$\sum_{v=1}^{n} \frac{\partial}{\partial \theta} m_\theta \Big|_{\theta = \theta_{\arg\max M_n(\theta)}} (X_v) = 0,$$

sofern natürlich kein Randmaximum vorliegt, bzw. Θ offen ist.

Z-Schätzer

Sei $X_1, \ldots, X_n$ Stichprobe einer Verteilung P und $\theta_0 \in \Theta$ ein unbekannter Parameter (zugehörig) der Verteilung*.

Definition

Jeder Schätzer $\hat{\theta}_n \in \Theta$, der eine sogenannte Schätzgleichung

$$\Psi_n(\theta) = \frac{1}{n} \sum_{v=1}^{n} \psi_\theta(X_v) = 0$$

erfüllt heißt ein *Z-Schätzer*. Die Summanden sind bekannte (vektorwertige) Funktionen ψ_θ ausgewertet an der Stichprobe $X_1, \ldots, X_n$.

Oftmals Koordinatenanzahl von θ und ψ_θ übereinstimmend, z.B. im Falle des Gradienten $\psi_\theta = \nabla m_\theta =: \frac{\partial}{\partial \theta} m_\theta$.

Äquivalent: $\sum_{v=1}^{n} \psi_{\theta,k}(X_v) = 0, \;\; k = 1, 2, \ldots, p.$

Im Allg. sind Z-Schätzer keine M-Schätzer, müssen also insbesondere auch kein Maximierungsproblem lösen.

Trotzdem Verwendung „M-Schätzer“!

Beispiel 1, Lageparameterschätzer

Lage: Erwartungswert, Median, Symmetriezentrum (f.e.) etc.
Schätzer: Stichprobenmittel, Stichprobenmedian (z.B.).

$$\sum_{v=1}^{n}(X_v - \theta) = 0 \quad \text{und} \quad \sum_{v=1}^{n}\operatorname{sign}(X_v - \theta) = 0$$

$\Rightarrow$ Z-Schätzer.

Z-Schätzer $\hat{\theta}_n$ welche $\sum_{v=1}^{n}\psi(X_v - \theta) = 0$ für ψ eine reellwertige Funktion erfüllen, heißen *Lageparameterschätzer*.
Location Estimator

Engl.: *Location Equivariance*:

$$\hat{\theta}_n(X_1 + s, \ldots, X_n + s) = \hat{\theta}_n(X_1, \ldots, X_n) + s, \quad s \in \mathbb{R}$$

Nicht skaleninvariant!:
$\hat{\theta}_n(aX_1, \ldots, aX_n) \neq a\hat{\theta}_n(X_1, \ldots, X_n), a \in \mathbb{R}.$ (i.A.)

. . . weiterführend: Redescending Estimators und Robust Statistics

Hier interessant: Huber- (, Hampel-, Tukey-, etc.) Schätzer ~ 'Robust Statistics'
Vgl. Staudte S. 118, Huber S.100-103, Van der Vaart S. 43!

Beispiel 2, Maximum-Likelihood-Schätzer

Sei $X_1, \ldots, X_n$ unabh. Stichprobe einer Verteilung P mit Dichte (bzw. Zähldichte) p_θ. Dann löst der MLE $\hat{\theta}_n^{MLE}$ das Problem

$$\max_{\theta \in \Theta} \sum_{v=1}^{n} \ln(p_\theta)(X_v).$$

$\Rightarrow$ Der Maximum-Likelihood Schätzer ist ein M-Schätzer.
Im Falle der partiellen Diff'barkeit (für jedes x) auch Z-Schätzer:

$$\Psi_n(\theta) = \sum_{v=1}^{n} \frac{\partial}{\partial \theta} \ln(p_\theta)(X_v) = \sum_{v=1}^{n} \left(\frac{\frac{\partial}{\partial \theta} p_\theta}{p_\theta} \right)(X_v) = 0,$$

ausgewertet bei $\hat{\theta}_n^{MLE} = \theta$.

Nun: Beweise von Theorem 1 und Theorem 2. Zuvor einige vorbereitende Lemmata und Definitionen.

Asymptotisch Normalverteilt.

Eine Folge von Schätzern $(\hat{\theta}_n)_{n\in\mathbb{N}}$ für $\theta_0 \in \Theta$ heißt *asymptotisch normalverteilt*, falls eine Folge $(\mu_n(\theta_0), \sigma_n(\theta_0))_{n\in\mathbb{N}}$ existiert, so, dass

$$\sigma_n(\theta_0)^{-1}(\hat{\theta}_n - \mu_n(\theta_0)) \xrightarrow{D} \mathcal{N}(0, C)$$

für eine Kovarianzmatrix C.

Konsistenz.

$(\hat{\theta}_n)_{n\in\mathbb{N}}$ ist ein konsistenter Schätzer für $\theta_0 \in \Theta$, g.d.w.

$$\mathbf{P}_{\theta_0}\left(\|\hat{\theta}_n - \theta_0\| > \varepsilon\right) \to 0$$

$\forall \varepsilon > 0$ und eine jeweilige Norm $\|\cdot\|$.

Lemma 1.

Seien X_n und X vektorwertige Zufallsvariablen. Dann gelten

$$X_n \xrightarrow{P} X \quad \Rightarrow \quad X_n \xrightarrow{D} X$$

$$X_n \xrightarrow{P} c \quad \Leftrightarrow \quad X_n \xrightarrow{D} c, \qquad c \in \mathbb{R}.$$

Lemma 2. *Lemma von Slutsky*

Seien X_n, Y_n und X [vektorwertige] Zufallsvariablen mit $X_n \xrightarrow{D} X$ und $Y_n \xrightarrow{D} c$ für eine Konstante c. Dann ist

i) $X_n + Y_n \xrightarrow{D} X + c$

ii) $Y_n X_n \xrightarrow{D} cX$

iii) $Y_n^{-1} X_n \xrightarrow{D} c^{-1} X, \qquad$ wobei $c \neq 0.$

Die letzteren beiden Produktkonvergenzen sind auch im Falle einer matrixwertigen Variable Y_n korrekt.
Fordere für *iii*): c invertierbar.

O-Notation

Für eine Folge von Zufallsvariablen $(X_n)_{n\in\mathbb{N}}$ gilt:

$$X_n = o_P(1) \;\Leftrightarrow\; X_n \xrightarrow{P} 0 \;\Leftrightarrow\; \mathbf{P}(\|X_n\| > \varepsilon) \to 0 \;\forall \varepsilon > 0.$$

$(X_n)_{n\in\mathbb{N}}$ heißt *stochastisch beschränkt* genau dann, wenn es für jedes $\varepsilon > 0$ ein $M > 0$ gibt, so, dass für $n \in \mathbb{N}$ beliebig

$$\mathbf{P}(\|X_n\| > M) < \varepsilon.$$

Dies wird $X_n = O_P(1)$ notiert.

Lemma 3.

Es gelten

$$o_P(1) + o_P(1) = o_P(1) \;\&$$
$$O_P(1)o_P(1) = o_P(1).$$

Theorem 1

$\Psi_n(\theta) = \frac{1}{n} \sum_{v=1}^{n} \psi_\theta(X_v) \quad / = \mathbb{P}_n(\theta)/$
$\Psi(\theta) = P\psi_\theta$

Notation: $Pf := \int f dP$

einfacher Oberpunkt, z.B. $\dot{\Psi}_n \rightarrow$ Ableitung
doppelter Oberpunkt, z.B. $\ddot{\Psi}_n \rightarrow$ zweite Ableitung

$X_1, \ldots, X_n$ iid. Stichprobe der Verteilung P.

Theorem 1

Für jedes $\theta \in \Theta$ ($\Theta \subset \mathbb{R}^p$, offen) sei die Abbildung $\theta \mapsto \psi_\theta(x)$ zweimal stetig differenzierbar für alle x. Weiter sei $P\psi_{\theta_0} = 0$, $P\|\psi_{\theta_0}\|^2 < \infty$ und die Matrix $P\dot{\psi}_{\theta_0}$ existiert und ist nichtsingulär. Für die partiellen zweiten Ableitungen gelte $\left|\frac{\partial^2 (\psi_{\theta_0})_k(x)}{\partial\theta_i \partial\theta_j}\right| \leq \ddot{\psi}(x)$, mit einer integrierbaren, messbaren Funktion $\ddot{\psi}$ für alle θ in einer Umgebung $\mathcal{B}$ von θ_0. $x \mapsto \psi_\theta(x)$, $x \mapsto (\dot{\psi}_\theta)_{jk}(x)$ und $x \mapsto (\ddot{\psi}_\theta)_{ijk}(x)$ seien messbar für alle $\theta \in \mathcal{B}$.

Dann erfüllt jede konsistente Schätzfolge $\hat{\theta}_n$ mit $\Psi_n(\hat{\theta}_n) = 0 \ \forall n \in \mathbb{N}$ die Eigenschaft

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -(P\dot{\psi}_{\theta_0})^{-1} \frac{1}{\sqrt{n}} \sum_{v=1}^{n} \psi_{\theta_0}(X_v) + o_P(1)$$

und ist damit asymptotisch normalverteilt:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}(0, (P\dot{\psi}_{\theta_0})^{-1} P\psi_{\theta_0}\psi_{\theta_0}^{\mathsf{T}} (P\dot{\psi}_{\theta_0})^{-1}).$$

Beweis von Theorem 1

Taylor:

$$\Psi_n(\hat{\theta}_n) = 0 = \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_0)(\hat{\theta}_n - \theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)^\mathsf{T}\ddot{\Psi}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

für ein $\tilde{\theta}_n = (1-s)\theta_0 + s\hat{\theta}_n,\ s \in [0,1]$.

Merke $\{\hat{\theta}_n \in \mathcal{B}\}$.

Zeige:

$$\text{Zentraler Grenzwertsatz}: \quad \sqrt{n}\Psi_n(\theta_0) \xrightarrow{D} Z \sim \mathcal{N}(0, P\psi_{\theta_0}\psi_{\theta_0}^\mathsf{T})$$

$$\text{SGGZ}: \quad \dot{\Psi}_n(\theta_0) \xrightarrow{P} V = P\dot{\psi}_{\theta_0}$$

$$\text{SGGZ}: \quad \ddot{\Psi}(\tilde{\theta}_n) = O_P(1).$$

Daraus folgt: $-\Psi_n(\theta_0) = (V + o_P(1))(\hat{\theta}_n - \theta_0)$.

Ziel des Theorems: $-V^{-1}\sqrt{n}\Psi_n(\theta_0) + o_P(1) = \sqrt{n}(\hat{\theta}_n - \theta_0)$.

Schritt 1, 2

$$\sqrt{n}\Psi_n(\theta_0) \xrightarrow{D} Z, \text{ mit } Z \sim \mathcal{N}(0, P\psi_{\theta_0}\psi_{\theta_0}^T)$$

Beweis. $\Psi_n(\theta_0) = \frac{1}{n}\sum_{v=1}^{n} \psi_{\theta_0}(X_v)$. $\mathbf{E}[\psi_{\theta_0}(X_1)] = 0$ (nV).
$(X_v)_{v=1,\dots,n}$ iid. $\Rightarrow$ $(\psi_{\theta_0}(X_v))_{v=1,\dots,n}$ iid., wg. ψ_{θ_0} messbar.
Zentraler Grenzwertsatz. □

$$\dot{\Psi}_n(\theta_0) \xrightarrow{P} V = P\dot{\psi}_{\theta_0}$$

Beweis. $\dot{\Psi}_n(\theta_0) = \frac{1}{n}\sum_{v=1}^{n} \dot{\psi}_{\theta_0}(X_v)$.
$(X_v)_{v=1,\dots,n}$ iid. $\Rightarrow$ $((\dot{\psi}_{\theta_0}(X_v))_{ij})_{v=1,\dots,n}$ iid., da $(\dot{\psi}_{\theta_0})_{ij}$ messbar.
Schwaches Gesetz der großen Zahlen, Lemma 3 und Frobeniusnorm. □

Schritt 3.

$\ddot{\Psi}(\tilde{\theta}_n) = O_P(1)$

$$\begin{pmatrix} \frac{\partial^2 \psi_1}{\partial\theta\partial\theta^\top} \\ \frac{\partial^2 \psi_2}{\partial\theta\partial\theta^\top} \\ \vdots \\ \frac{\partial^2 \psi_p}{\partial\theta\partial\theta^\top} \end{pmatrix}$$

$\ddot{\Psi}(\tilde{\theta}_n)$ ist Tensor dritter Stufe, wobei hier jeweils die Indizierung mit $\tilde{\theta}_n$ weggelassen wurde.

Beweis. Verwendet $\ddot{\psi}$ messbar, die Abschätzung $\left|\frac{\partial^2 (\psi_{\theta_0})_k(x)}{\partial\theta_i\partial\theta_j}\right| \leq \ddot{\psi}(x)$ auf $\{\hat{\theta}_n \in \mathcal{B}\}$ und das schwache Gesetz der großen Zahlen durch

$$\|A\|_{HS} := \left(\sum_{i,j,k=1}^{p} a_{ijk}^2\right)^{1/2} \quad \text{(Hilbert-Schmidt Norm).}$$

$\rightsquigarrow$ Außerdem Prohorov: $X_n \xrightarrow{D} X \Rightarrow X_n = O_P(1)$. $\square$

Damit ist insgesamt

$$\overset{\curvearrowleft}{-V^{-1}\sqrt{n}\Psi_n(\theta_0) + o_P(1)} = \sqrt{n}(\hat{\theta}_n - \theta_0).$$

$$-\Psi_n(\theta_0) = (V + o_P(1))(\hat{\theta}_n - \theta_0).$$

Störungslemma.

Sei $V \in \mathbb{R}^{p\times p}$ nichtsingulär. Ist $S \in \mathbb{R}^{p\times p}$ eine Störung von V mit $\|V^{-1}\|\|S\| < 1$, so ist $V + S$ nichtsingulär und

$$\|(V+S)^{-1}\| \leq \frac{\|V^{-1}\|}{1 - \|V^{-1}\|\|S\|}.$$

Beweis. Siehe Werner (1992).

Nun $A_n := \{\hat{\theta}_n \in \mathcal{B}\} \cap \{\|V^{-1}\| \cdot \|o_P(1)\| < 1\}$: $\mathbf{P}(A_n) \to 1$.
Mittels A_n, Slutsky, Lemma 1 und der Regularität von $V = P\dot{\psi}_{\theta_0}$:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} -V^{-1}Z \sim \mathcal{N}\left(0, V^{-1}P\psi_{\theta_0}\psi_{\theta_0}^{\mathsf{T}}\left(V^{-1}\right)^{\mathsf{T}}\right). \quad \square$$

Theorem 2

Voraussetzungen identisch zu Theorem 1.

Die Wahrscheinlichkeit $\mathbf{P}(\Psi_n(\theta) = 0 \text{ für mindestens ein } \theta \in \Theta)$ *konvergiert gegen 1.*
Es existiert eine Folge von Nullstellen von $\Psi_n(\theta)$*, nenne* $\hat{\theta}_n$*, so, dass* $\hat{\theta}_n \to \theta_0$ *in Wahrscheinlichkeit.*

Beweis von Theorem 2

Taylor:

$$\Psi(\theta) = P\psi_\theta = P\psi_{\theta_0} + P\dot{\psi}_{\theta_0}(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^\top P\ddot{\psi}_{\tilde{\theta}}(\theta - \theta_0)$$

$$\text{für ein } \tilde{\theta} = (1 - s_2)\theta_0 + s_2\theta,\ s_2 \in [0, 1].$$

Idee: Zeige mit Taylorentwicklung für $\Psi_n(\theta)$, dass

$$\mathbf{P}\left(\sup_{\theta \in A_{G_{\delta_n}}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq \delta_n\right) \to 1$$

für genauer zu konstruierende Folge(n) δ_n und davon abhängige $A_{G_{\delta_n}}$. Der FPS von Brouwer (essentiell für die Definition der entsprechenden Funktion ist der Satz über inverse Funktionen) liefert das Theorem.

Zeige $P\psi_\theta$ differenzierbar in $\theta_0 \in \Theta$; Finde Matrix $A \in \mathbb{R}^{p\times p}$ mit

$$\lim_{\theta\to\theta_0} \frac{P\psi_\theta - (P\psi_{\theta_0} + A(\theta-\theta_0))}{\|\theta-\theta_0\|} = 0.$$

Für $A = P\dot{\psi}_{\theta_0}$ erfüllt (verwende komponentenweise Beschränktheit der 2. Ableitung). Weiter ist (selbiges Argument) Ψ in ganz $\mathcal{B}$ differenzierbar. Bemerke $P\psi_\theta = \Psi(\theta)$. Existenz $P\dot{\psi}_{\theta_0} \Rightarrow$ Existenz $P\dot{\psi}_\theta$.

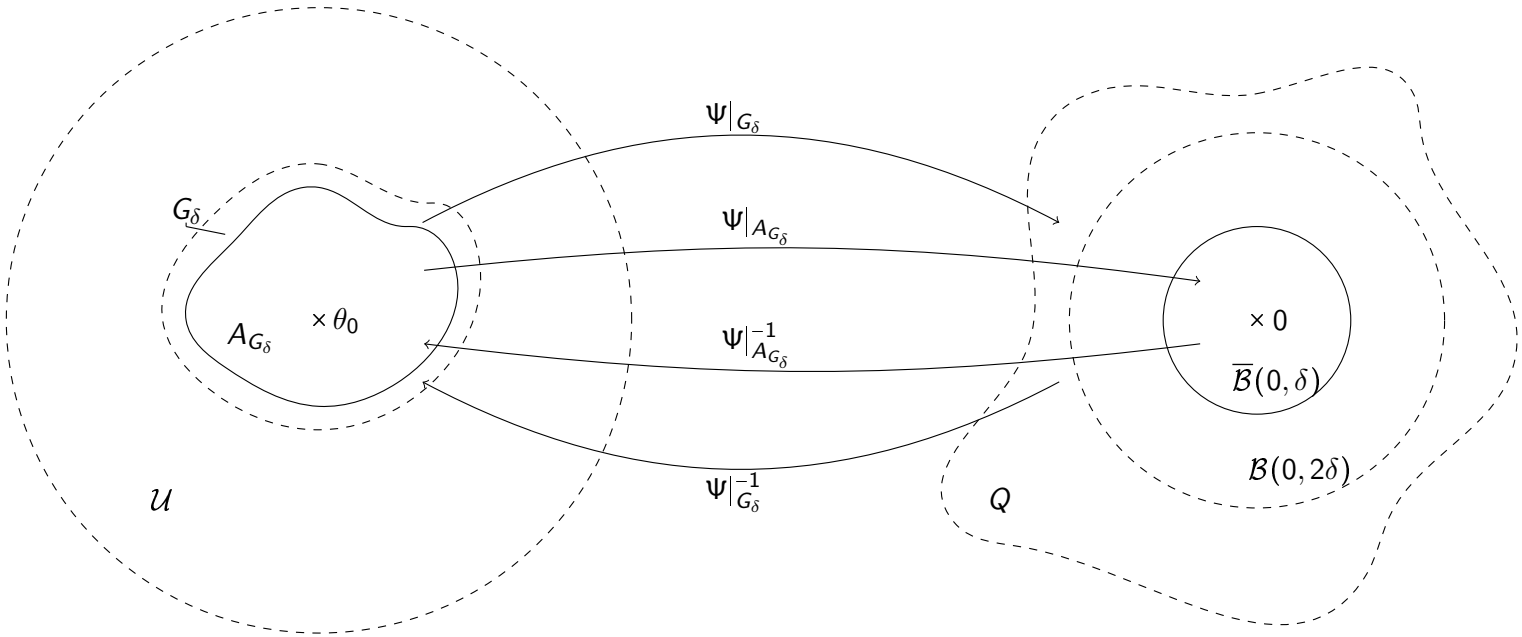
Ableitung stetig? - Ja, auf ganz $\mathcal{B}$. Wichtigste Beweiskomponente: Taylorentwicklung (zeilenweise!) von $\dot{\Psi}(\theta)$;

$P\dot{\psi}_\theta$ invertierbar? - Ja aber nicht unbedingt auf ganz $\mathcal{B}$, i.A. in (kleiner(er)) Umgebung$_{\text{nenne}\,\mathcal{U}}$ von $\theta_0 \to$ Beweis Störungslemma!

Satz über die inverse Funktion.

Sei $\mathcal{U} \subset \mathbb{R}^p$ offen und $\Psi = P\psi \in C^1(\mathcal{U}, \mathbb{R}^p)$. Ist $\dot{\Psi}(\theta_0)$ invertierbar, so gibt es eine offene Umgebung G_δ von θ_0, so dass gilt:

- *$Q = \Psi(G_\delta)$ ist offene Umgebung von $0 = \Psi(\theta_0)$*
- *$\Psi|_{G_\delta}$ ist Diffeomorphismus der Klasse C^1.*



------ = offen

——— = abgeschlossen

$\Psi|_{A_{G_\delta}} : A_{G_\delta} \to \overline{\mathcal{B}}(0,\delta)$ ist Homöomorphismus.
bijektiv, stetig, Inverse stetig

Wir sind im Beweis von Theorem 2

Mittelwertsatz im Mehrdimensionalen (vgl. Forster, 2017).

$$\Rightarrow \text{diam } A_{G_\delta} \leq c \cdot 2\delta$$

Zweite Taylorentwicklung zeigt (vgl. Idee)

$$\sup_{\theta \in A_{G_\delta}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq \underbrace{\|\mathbb{P}_n \psi_{\theta_0}\|}_{o_P(1)} + c \cdot 2\delta \cdot \underbrace{\|\mathbb{P}_n \dot{\psi}_{\theta_0} - P\dot{\psi}_{\theta_0}\|}_{=o_P(1)} + c^2 \cdot 2\delta^2 \cdot p^{3/2} P\ddot{\psi}$$

$$+ \underbrace{\frac{1}{n} \sum_{v=1}^{n} \ddot{\psi}(X_v) \cdot c^2 \cdot 2\delta^2 \cdot p^{3/2}}_{O_P(1)}.$$

Mit $\boxed{a \in \mathbb{R},\ X_n = O_P(1) \Rightarrow a + X_n = O_P(1)}$ und

$\boxed{\delta \in \mathbb{R} \Rightarrow \delta o_P(1) = o_P(1)}$:

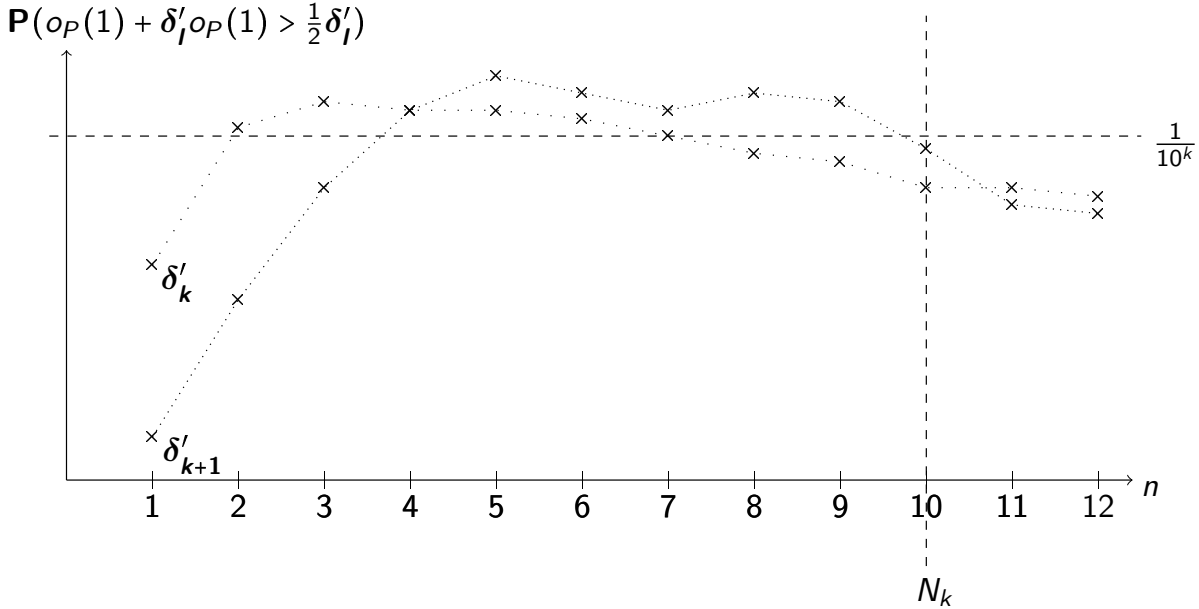
$$\sup_{\theta \in A_{G_\delta}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq o_P(1) + \delta o_P(1) + \delta^2 O_P(1).$$

Nullfolgen δ_n

$$\sup_{\theta \in A_{G_\delta}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq o_P(1) + \delta o_P(1) + \delta^2 O_P(1).$$

Es gibt $\delta_n \downarrow 0$ mit $\mathbf{P}\big(o_P(1) + \delta_n o_P(1) > \frac{1}{2}\delta_n\big) \to 0$.

Bsp: $\boldsymbol{\delta'_k} := \frac{\delta}{2^{k-1}}$, $\delta > 0$, $k \geq 1$, N_k minimal mit $\max\{\mathbf{P}(o_P(1) + \boldsymbol{\delta'_k} o_P(1) > \frac{1}{2}\boldsymbol{\delta'_k}), \mathbf{P}(o_P(1) + \boldsymbol{\delta'_{k+1}} o_P(1) > \frac{1}{2}\boldsymbol{\delta'_{k+1}})\} < \frac{1}{10^k} \quad \forall n \geq N_k$.



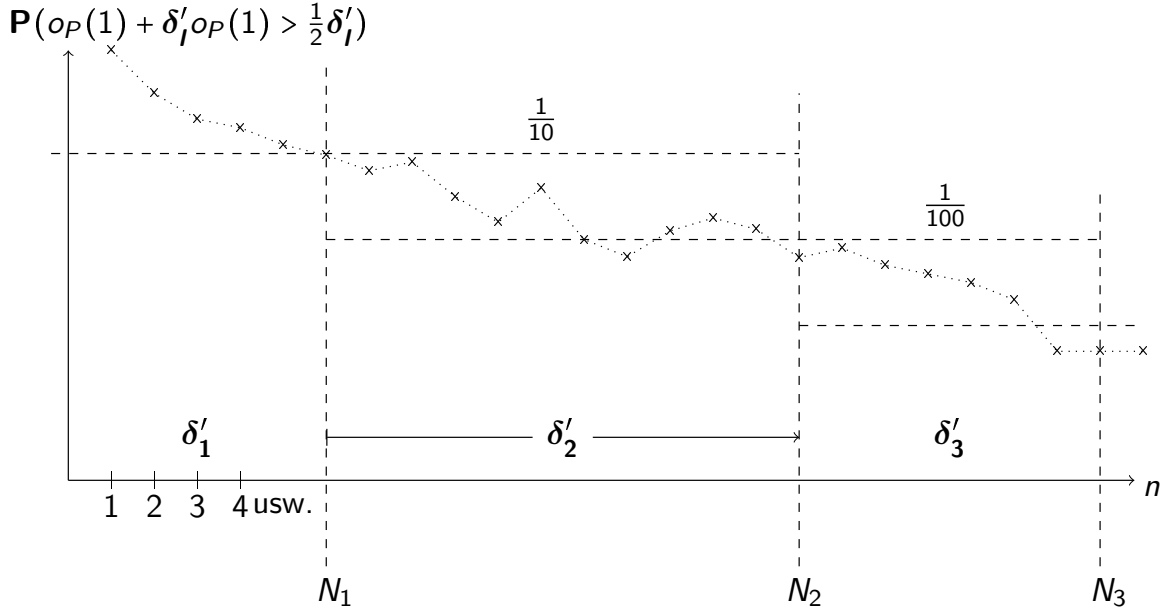
Dann $\delta_{N_{k-1}}, \ldots, \delta_{N_k - 1} = \boldsymbol{\delta'_k}$, $N_0 := 1$.

Nullfolgen δ_n

$$\sup_{\theta \in A_{G_\delta}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq o_P(1) + \delta o_P(1) + \delta^2 O_P(1).$$

Es gibt $\delta_n \downarrow 0$ mit $\mathbf{P}\big(o_P(1) + \delta_n o_P(1) > \frac{1}{2}\delta_n\big) \to 0$.

Bsp: $\boldsymbol{\delta'_k} := \frac{\delta}{2^{k-1}}$, $\delta > 0$, $k \geq 1$, N_k minimal mit $\max\{\mathbf{P}(o_P(1) + \boldsymbol{\delta'_k} o_P(1) > \frac{1}{2}\boldsymbol{\delta'_k}), \mathbf{P}(o_P(1) + \boldsymbol{\delta'_{k+1}} o_P(1) > \frac{1}{2}\boldsymbol{\delta'_{k+1}})\} < \frac{1}{10^k} \quad \forall n \geq N_k$.



K_{n,δ_n} - Konsistenzumgebung

Sei solche Nullfolge δ_n mit $\mathbf{P}\left(o_P(1) + \delta_n o_P(1) > \frac{1}{2}\delta_n\right) \to 0$ gegeben.
Mit der Abschätzung
$\sup_{\theta \in A_{G_\delta}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq o_P(1) + \delta o_P(1) + \delta^2 O_P(1)$ und

$$K_{n,\delta_n} := \{ \sup_{\theta \in A_{G_{\delta_n}}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq \delta_n \}$$

gilt:

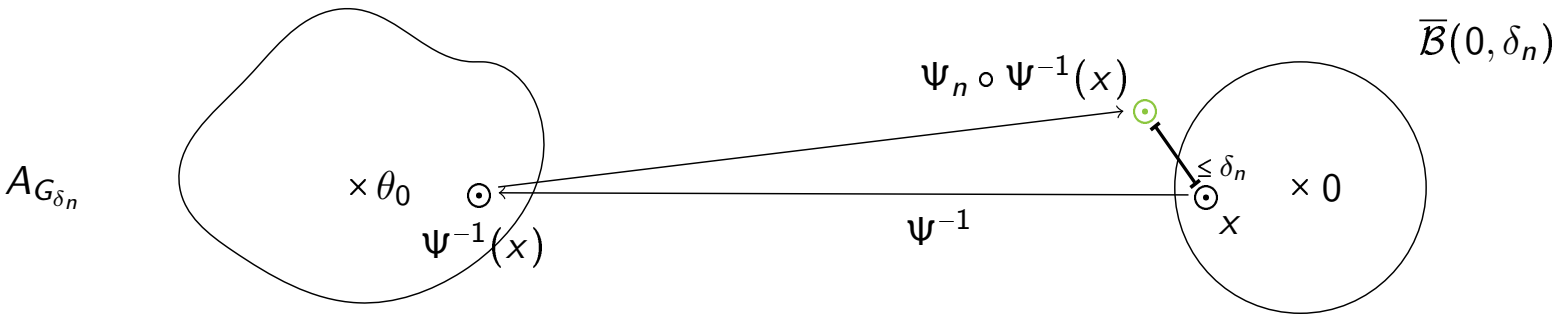
$$\begin{aligned}
\mathbf{P}(K_{n,\delta_n}) &:= \mathbf{P}\left(\sup_{\theta \in A_{G_{\delta_n}}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq \delta_n \right) \\
&\geq \mathbf{P}(o_P(1) + \delta_n o_P(1) + \delta_n^2 O_P(1) \leq \delta_n) \\
&\geq \mathbf{P}\left(\left\{ o_P(1) + \delta_n o_P(1) \leq \frac{\delta_n}{2} \right\} \cap \left\{ \delta_n^2 O_P(1) \leq \frac{\delta_n}{2} \right\} \right) \\
&\geq 1 - \mathbf{P}\left(\left\{ o_P(1) + \delta_n o_P(1) > \frac{\delta_n}{2} \right\} \right) - \mathbf{P}\left(\left\{ O_P(1) > \frac{1}{2\delta_n} \right\} \right) \to 1.
\end{aligned}$$

Nullstelle von Ψ_n in $A_{G_{\delta_n}}$

Damit ist auf $K_{n,\delta_n} := \{\sup_{\theta \in A_{G_{\delta_n}}} \|\Psi_n(\theta) - \Psi(\theta)\| \leq \delta_n\}$ die Funktion

$$x \mapsto x - \Psi_n \circ \Psi^{-1}(x) \tag{1}$$

Selbstabbildung der Kugel $\overline{\mathcal{B}}(0, \delta_n)$.



Da die Abbildung aus (1) stetig ist, hat sie mit dem Fixpunktsatz von Brouwer* einen Fixpunkt in $\overline{\mathcal{B}}(0, \delta_n)$. $\Rightarrow$ Ψ_n hat Nullstelle in $A_{G_{\delta_n}}$. $\square$

*Sei $f : B_d \to B_d$ stetig. Dann existiert ein $\xi \in B_d$ mit $f(\xi) = \xi$. *Beweis.* Siehe Werner (2009).

Q.e.d.?

Gezeigt wurden:

a) Es existiert mit gegen 1 konvergierender Wahrscheinlichkeit eine Nullstelle der Abbildung $\theta \mapsto \Psi_n(\theta) = \mathbb{P}_n \psi_\theta = \frac{1}{n} \sum_{v=1}^{n} \psi_\theta(X_v)$.

$$\begin{aligned}\mathbf{P}\big(\{\text{Es existiert NS in } \Theta\}\big) &\geq \mathbf{P}\Big(\{\text{Es existiert NS in } A_{G_{\delta_n}} \subset \Theta\}\Big) \\ &\geq \mathbf{P}(K_{n,\delta_n}) \to 1.\end{aligned}$$

b) Es existiert eine Folge von Nullstellen dieser Abbildung so, dass $\hat{\theta}_n \to \theta_0$ in Wahrscheinlichkeit.

$$\mathbf{P}(\|\hat{\theta}_n - \theta_0\| > \varepsilon) = \mathbf{P}(\{\|\hat{\theta}_n - \theta_0\| > \varepsilon\} \cap I_n) + \mathbf{P}(\{\|\hat{\theta}_n - \theta_0\| > \varepsilon\} \cap I_n^c) \to 0.$$

Mit $\hat{\theta}_n \in A_{G_{\delta_n}}$ Nullstelle $\big|$Existenz, sonst $\hat{\theta}_n$ beliebig, $I_n := \{\hat{\theta}_n \in A_{G_{\delta_n}}\}$.

□

Kommentar

Ab nun: $\text{rg}(\mathbf{X}) = p$ für n hinreichend groß.

Beweis der Voraussetzungen

Für Konsistenz, Existenz und asymptotische Normalität des MLE im logistischen Regressionsmodell ist also mit $(Y, X) \sim P$ zu zeigen:

1. $\theta \mapsto \nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)$ zweimal stetig diffbar für jedes (y, x).
2. $(y,x) \mapsto \nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)$, $(y,x) \mapsto \frac{\partial^2}{\partial\theta_j\partial\theta_k} \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)$ und $(y,x) \mapsto \frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)$ messbar für θ in einer Umgebung von θ_0.
3. $\mathbf{E}\left[\nabla\big|_{\theta=\theta_0} \ln\left(\frac{e^{YX^\top\theta}}{1+e^{X^\top\theta}}\right)\right] = 0.$
4. $\mathbf{E}\left[\left\|\nabla\big|_{\theta=\theta_0} \ln\left(\frac{e^{YX^\top\theta}}{1+e^{X^\top\theta}}\right)\right\|^2\right] < \infty.$
5. Erwartungswert der Jacobimatrix$\big|_{\theta=\theta_0}$ von $\nabla \ln\left(\frac{e^{YX^\top\theta}}{1+e^{X^\top\theta}}\right)$ existiert und die erwartete Matrix ist nichtsingulär.
6. Es existiert eine integrierbare und messbare Funktion $\ddot{\psi}(x)$ mit

$$\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(\nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)\right)_k\right| \leq \ddot{\psi}(x) \quad \forall i,j,k$$

in einer Umgebung von θ_0.

Hier nur „(x)“, siehe nächste Folie.

Beweis der Voraussetzungen

1. $\theta \mapsto \nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)$ zweimal stetig differenzierbar für jedes (y,x).

$$\nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right) = x(y-\lambda(x^\top\theta)), \qquad \text{b.v.}$$

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right) = -\frac{\partial}{\partial\theta_j}x_k\frac{e^{x^\top\theta}}{1+e^{x^\top\theta}} = -x_jx_k\frac{e^{x^\top\theta}}{(1+e^{x^\top\theta})^2}, \qquad \text{b.v.}$$

$$\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k}\ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right) = -\frac{\partial}{\partial\theta_i}x_jx_k\frac{e^{x^\top\theta}}{(1+e^{x^\top\theta})^2} = -x_ix_jx_k\frac{(1-e^{x^\top\theta})e^{x^\top\theta}}{(1+e^{x^\top\theta})^3}. \qquad \square$$

2. Die Messbarkeit der jeweiligen Abbildungen folgt direkt der (komponentenweisen) Stetigkeit der Abbildungen

$$(y,x) \mapsto x\left(y - \frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right),$$

$$(y,x) \mapsto -x_jx_k\frac{e^{x^\top\theta}}{(1+e^{x^\top\theta})^2} \qquad \text{und}$$

$$(y,x) \mapsto -x_ix_jx_k\frac{(1-e^{x^\top\theta})e^{x^\top\theta}}{(1+e^{x^\top\theta})^3}. \qquad \square$$

Beweis der Voraussetzungen

1. $\theta \mapsto \nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)$ zweimal stetig differenzierbar für jedes (y, x).

$$\nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right) = x(y - \lambda(x^\top\theta)), \qquad \text{b.v.}$$

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k} \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right) = -\frac{\partial}{\partial\theta_j} x_k \frac{e^{x^\top\theta}}{1+e^{x^\top\theta}} = -x_j x_k \frac{e^{x^\top\theta}}{(1+e^{x^\top\theta})^2}, \qquad \text{b.v.}$$

$$\frac{\partial^3}{\partial\theta_i\partial\theta_j\partial\theta_k} \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right) = -\frac{\partial}{\partial\theta_i} x_j x_k \frac{e^{x^\top\theta}}{(1+e^{x^\top\theta})^2} = -x_i x_j x_k \frac{(1-e^{x^\top\theta})e^{x^\top\theta}}{(1+e^{x^\top\theta})^3}. \qquad \square$$

Setze weiter voraus:

$\exists M_0 \in \mathbb{R} : \ |(X_1)_k| \le M_0, \quad$ für alle $\quad k = 1, \ldots, p.$

Siehe z.B. Gourieroux, Amemiya, Nordberg...

6. $\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(\nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)\right)_k\right| \le \ddot{\psi}(x) \ \ \forall i, j, k.$

$$\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(\nabla \ln\left(\frac{e^{yx^\top\theta}}{1+e^{x^\top\theta}}\right)\right)_k\right| = \left| - x_i x_j x_k \frac{(1-e^{x^\top\theta})e^{x^\top\theta}}{(1+e^{x^\top\theta})^3}\right| \le M_0^3$$

$\rightsquigarrow \ddot{\psi}(x) := M_0^3$ messbar, integrierbar. $\square$

Beweis der Voraussetzungen

3. $\mathbf{E}\left[\nabla_{\theta_0} \ln\left(\frac{e^{YX^\top\theta}}{1+e^{X^\top\theta}}\right)\right] = 0$, mit $\mathbf{E}[Y|X] := \pi(X) = \frac{e^{X^\top\theta_0}}{1+e^{X^\top\theta_0}}$;

$$\begin{aligned}
\mathbf{E}\left[\nabla_{\theta_0} \ln\left(\frac{e^{YX^\top\theta}}{1+e^{X^\top\theta}}\right)\right] &= \mathbf{E}\left[X\left(Y - \frac{e^{X^\top\theta_0}}{1+e^{X^\top\theta_0}}\right)\right] \\
&\overset{\star}{=} \mathbf{E}\left[\mathbf{E}\left[X\left(Y - \frac{e^{X^\top\theta_0}}{1+e^{X^\top\theta_0}}\right)|X\right]\right] \\
&\overset{\text{mb., lin.}}{=} \mathbf{E}\left[X\left(\mathbf{E}[Y|X] - \mathbf{E}\left[\frac{e^{X^\top\theta_0}}{1+e^{X^\top\theta_0}}|X\right]\right)\right] \\
&\overset{\text{mb.}}{=} \mathbf{E}\left[X\left(\pi(X) - \pi(X)\mathbf{E}[1|X]\right)\right] \\
&= \mathbf{E}\left[X\left(\frac{e^{X^\top\theta_0}}{1+e^{X^\top\theta_0}} - \frac{e^{X^\top\theta_0}}{1+e^{X^\top\theta_0}}\right)\right] \\
&= \mathbf{E}[X \cdot 0] = 0.
\end{aligned}$$

$\mathbf{E}\left[\left|X_k\left(Y - \frac{e^{X^\top\theta_0}}{1+e^{X^\top\theta_0}}\right)\right|\right] < \infty$ dank Monotonie d.EW's/ Vor. 4.

□

Beweis der Voraussetzungen

4. $\mathbf{E}\left[\left\|\nabla_{\theta_0} \ln\left(\frac{e^{YX^\top\theta}}{1+e^{X^\top\theta}}\right)\right\|^2\right] < \infty;$ $\boxed{\exists M_0 \in \mathbb{R}: \ |(X_1)_k| \leq M_0, \ \forall k = 1, \ldots, p.}$

$\boxed{\lambda(x) := \frac{e^x}{1+e^x}}$

$$\begin{aligned}
\mathbf{E}\left[\left\|\nabla_{\theta_0} \ln\left(\frac{e^{YX^\top\theta}}{1+e^{X^\top\theta}}\right)\right\|^2\right] &= \mathbf{E}\left[\sum_{k=1}^{p}\left(X_k\left(Y - \lambda(X^\top\theta_0)\right)\right)^2\right] \\
&= \sum_{k=1}^{p}\mathbf{E}\left[X_k^2 Y^2\right] + \mathbf{E}\left[X_k^2\lambda(X^\top\theta_0)^2\right] - 2\mathbf{E}\left[X_k^2 Y\lambda(X^\top\theta_0)\right] \\
&\leq \sum_{k=1}^{p}\mathbf{E}\left[X_k^2\right] + \mathbf{E}\left[X_k^2\right] \\
&\leq 2\sum_{k=1}^{p}\mathbf{E}\left[M_0^2\right] \ = \ 2pM_0^2 \ < \infty,
\end{aligned}$$

wobei wir hier erneut die Monotonie des Erwartungswertes genutzt haben. □

5. Der Erwartungswert der Jacobimatrix$\big|_{\theta=\theta_0}$ von $\nabla \ln\left(\frac{e^{YX^\mathsf{T}\theta}}{1+e^{X^\mathsf{T}\theta}}\right)$ existiert und die erwartete Matrix ist nichtsingulär.

Die Komponenten der Jacobimatrix$\big|_{\theta=\theta_0}$ von $\nabla \ln\left(\frac{e^{YX^\mathsf{T}\theta}}{1+e^{X^\mathsf{T}\theta}}\right)$ sind:

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}\ln\left(\frac{e^{YX^\mathsf{T}\theta}}{1+e^{X^\mathsf{T}\theta}}\right)\Bigg|_{\theta=\theta_0} = -X_jX_k\frac{e^{X^\mathsf{T}\theta_0}}{(1+e^{X^\mathsf{T}\theta_0})^2}.$$

Damit ist

$$\left|\mathbf{E}\left[-X_jX_k\cdot\frac{e^{X^\mathsf{T}\theta_0}}{(1+e^{X^\mathsf{T}\theta_0})^2}\right]\right| \le \mathbf{E}\left[\left|-X_jX_k\cdot\frac{e^{X^\mathsf{T}\theta_0}}{(1+e^{X^\mathsf{T}\theta_0})^2}\right|\right] \le \mathbf{E}\left[M_0^2\cdot 1\right] = M_0^2\cdot 1$$

und der Erwartungswert existiert. Die Invertierbarkeit werden wir fordern.

Wir haben gesehen

Unter den Bedingungen

1. $(Y_v, X_v), v = 1, \ldots, n$ sind ii. nach P verteilt,
2. $\exists M_0 \in \mathbb{R}: \ |(X_1)_k| \leq M_0$ für alle $k = 1, \ldots, p$,
3. der Invertierbarkeit von $\mathbf{E}\left[-X_1 X_1^\mathsf{T} \frac{e^{X_1^\mathsf{T}\theta_0}}{\left(1+e^{X_1^\mathsf{T}\theta_0}\right)^2}\right]$ und
4. $\operatorname{rg}(\mathbf{X}) = p$ für ein $N \in \mathbb{N}$,

ist der Maximum Likelihood-Schätzer des logistischen Regressionsmodells existent, konsistent und asymptotisch normalverteilt!

Allgemeine Theorie: Theorem 1 und Theorem 2.

Literatur

Hauptquellen:
Trevor Hastie. *The elements of statistical learning. Data mining, inference, and prediction.* Hrsg. von Robert Tibshirani und Jerome Friedman. Literaturverzeichnis: Seite 699-727. New York, NY, 2009. URL: http://dx.doi.org/10.1007/978-0-387-84858-7

David W. Hosmer, Stanley Lemeshow und Rodney X. Sturdivant. *Applied Logistic Regression. Hosmer/Applied Logistic Regression*. 2013. DOI: 10.1002/9781118548387

A.W. van der Vaart. *Asymptotic statistics.* Bd. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998, S. xvi+443. ISBN: 0-521-49603-9; 0-521-78450-6. DOI: 10.1017/CBO9780511802256

Und für viele mehr siehe in der Ausarbeitung!

Vielen Dank für eure Aufmerksamkeit!

Fragen?