Dichte von $\mathbb{P}_\lambda^{X,Y}$

geg: • Mit $\pi_{X,Y}(x,y)$ bzw. $\pi^\lambda_{X,Y}(x,y)$ bezeichne Dichte von $\mathbb{P}^{X,Y}$ bzw. $\mathbb{P}_\lambda^{X,Y}$ bzgl. $\lambda \otimes \mu$, wobei λ Lebesgue-Maß, μ Zählmaß

ges: $\frac{d\mathbb{P}_\lambda^{X,Y}}{d\mathbb{P}^{X,Y}} = \frac{a_\lambda(x,y)}{\bar{a}(\lambda)}$

Beweis:

$$\pi^\lambda_{X,Y}(x,y) = \pi_{X,Y|Z=1}(x,y)$$

Bayes ↓

$$= \frac{\mathbb{P}(Z=1 \mid X=x, Y=y)}{\mathbb{P}(Z=1)} \cdot \pi_{X,Y}(x,y)$$

$\mathbb{P}(Z=1) = \int a_\lambda(x,y)\, d\mathbb{P}^{X,Y}(x,y) = \bar{a}(\lambda)$

$\mathbb{P}(Z=1 \mid X=x, Y=y) = a_\lambda(x,y)$

$$= \frac{a_\lambda(x,y)}{\bar{a}(\lambda)} \cdot \pi_{X,Y}(x,y)$$

$$\Rightarrow \frac{d\mathbb{P}_\lambda^{X,Y}}{d\mathbb{P}^{X,Y}} = \frac{a_\lambda(x,y)}{\bar{a}(\lambda)}$$

□

Offset bei LCC:

geg: · $g(x)$ bedingte log-odds Funktion von P_λ
· $f(x)$ bedingte log-odds Funktion von P
} gegeben $X = x$

z: $g(x) = f(x) - \lambda^T x$

Beweis:

$$g(x) = \log \frac{P_\lambda(Y=1 \mid X=x)}{P_\lambda(Y=0 \mid X=x)}$$

$$= \log \frac{\frac{\pi^\lambda_{x,Y}(1,x)}{\pi^\lambda_x(x)}}{\frac{\pi^\lambda_{x,Y}(0,x)}{\pi^\lambda_x(x)}}$$

$$= \log \frac{\pi_{x,Y}(1,x) \cdot \frac{a_\lambda(1,x)}{\bar{a}(x)}}{\pi_{x,Y}(0,x) \cdot \frac{a_\lambda(0,x)}{\bar{a}(x)}}$$

$$= \log \frac{\pi_{x,Y}(1,x)}{\pi_{x,Y}(0,x)} + \log \frac{a_\lambda(1,x)}{a_\lambda(0,x)}$$

$\frac{\pi_{x,Y}(1,x)}{\pi_{x,Y}(0,x)} = \frac{P(Y=1 \mid X=x)}{P(Y=0 \mid X=x)} = \frac{p(x)}{1-p(x)}$

$\frac{a_\lambda(1,x)}{a_\lambda(0,x)} = \frac{1-\hat{p}_\lambda(x)}{\hat{p}_\lambda(x)}$

$$= f(x) - \lambda^T x$$

□

Proposition 2

geg: • $\mathbb{E}\|X\| < \infty$
• Daten nicht separabel
• $\theta^* = \bar{\theta}(0) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} R(\theta)$

z: $\theta^* = \bar{\theta}(\theta^*) = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} Q_{\theta^*}(\theta)$ ← gilt per Def.

Beweis:

Da $Q_{\theta^*}(\theta)$ strikt konvex ist, gilt

$$\theta = \bar{\theta}(\lambda) \iff \nabla_\theta Q_\lambda(\theta) = 0 .$$

Betrachte also ← Komponentenweises Integral

$$\nabla_\theta Q_\lambda(\theta) = \int \nabla_\theta \rho(\theta - \lambda; x, y)\, d\mathbb{P}_\lambda^{x,y}(x,y)$$

$$= \int yx - x \cdot \frac{\exp((\theta-\lambda)^T x)}{1+\exp((\theta-\lambda)^T x)}\, d\mathbb{P}_\lambda^{x,y}(x,y)$$

Für $\theta = \lambda = \theta^*$ gilt:

$$\nabla_\theta Q_{\theta^*}(\theta^*) = \int yx - x \cdot \underbrace{\frac{\exp(\overbrace{(\theta^*-\theta^*)^T x}^{=0})}{1+\exp(\underbrace{(\theta^*-\theta^*)^T x}_{=0})}}_{=\frac{1}{2}}\, d\mathbb{P}_{\theta^*}^{x,y}(x,y)$$

$$= \int (y-\tfrac{1}{2})x \frac{a_{\theta^*}(x,y)}{\bar{a}(\theta^*)}\, d\mathbb{P}^{x,y}(x,y)$$

$$= \frac{1}{\bar{a}(\theta^*)} \mathbb{E}\left[(y-\tfrac{1}{2})x \left(y(1-\rho^*(x)) + (1-y)\rho^*(x)\right)\right]$$

$$= \frac{1}{\bar{a}(\theta^*)} \mathbb{E}\Big[x\Big(\underbrace{y^2}_{=y}(1-\rho^*(x)) + \underbrace{y(1-y)}_{=0}\rho^*(x)\Big)$$

$$\underbrace{- \tfrac{1}{2} y(1-\rho^*(x)) - \tfrac{1}{2}(1-y)\rho^*(x)\Big]}_{= -\frac{1}{2} a_{\theta^*}(x,y)}$$

$$= \frac{1}{\bar{a}(\theta^*)} \mathbb{E}\left[\tfrac{1}{2} xy(1-\rho^*(x)) - \tfrac{1}{2} x(1-y)\rho^*(x)\right]$$

$$= \frac{1}{2} \cdot \frac{1}{\bar{a}(\theta^*)} \mathbb{E}\Big[x\Big(\underbrace{\rho(x)(1-\rho^*(x)) - (1-\rho(x))\rho^*(x)}_{=\rho(x)-\rho^*(x)}\Big)\Big]$$

$$= \frac{1}{2} \frac{1}{\bar{a}(\theta^*)} \underbrace{\mathbb{E}\left[x(y-\rho^*(x))\right]}_{=\nabla_\theta R(\theta^*)=0} = 0$$

□

Theorem 5

geg: • $\mathbb{E}\|x\| < \infty$
• Daten nicht separabel
• λ_n ist konsistent, d.h. $\lambda_n \xrightarrow{\mathbb{P}} \theta^*$

z: $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta^*$

Beweis:
Sei $\Theta \subset \mathbb{R}^p$ kompakt mit $\theta^* \in \Theta^\circ$. ($\Theta^\circ$: Inneres von Θ)

Sei $\varepsilon := \inf_{\theta \in \partial\Theta} Q_{\theta^*}(\theta) - Q_{\theta^*}(\theta^*) > 0$. (I)

$\theta^* = \operatorname{argmin}_{\theta \in \Theta} Q_{\theta^*}(\theta)$, $\theta^* \notin \partial\Theta$, Minimum eindeutig,
(wegen $\partial\Theta$ kompakt wird Infimum angenommen)

Prop. 4 $\Rightarrow$ $\mathbb{P}\left(\sup_{\theta \in \Theta} |\hat{Q}_{\lambda_n}(\theta) - Q_{\theta^*}(\theta)| \geq \frac{\varepsilon}{4}\right) \xrightarrow{n \to \infty} 0$ (II)

Sei weiter $\nu_n \in \partial\Theta$ mit

$$\hat{Q}_{\lambda_n}(\nu_n) - \inf_{\theta \in \partial\Theta} \hat{Q}_{\lambda_n}(\theta) < \frac{\varepsilon}{4} \quad \forall n \in \mathbb{N} \qquad \text{(III)}$$

Dann gilt mit W'heit, die gegen 1 konvergiert:

$$\inf_{\theta \in \partial\Theta} \hat{Q}_{\lambda_n}(\theta) - \hat{Q}_{\lambda_n}(\theta^*) \qquad (= 0)$$

$$= \underbrace{\inf_{\theta \in \partial\Theta} \hat{Q}_{\lambda_n}(\theta) - \hat{Q}_{\lambda_n}(\nu_n)}_{> -\frac{\varepsilon}{4} \text{ (III)}} + \underbrace{\hat{Q}_{\lambda_n}(\nu_n) - Q_{\theta^*}(\nu_n)}_{> -\frac{\varepsilon}{4} \text{ (II)}} + \underbrace{Q_{\theta^*}(\nu_n) - \inf_{\theta \in \partial\Theta} Q_{\theta^*}(\theta)}_{\geq 0}$$

$$+ \underbrace{\inf_{\theta \in \partial\Theta} Q_{\theta^*}(\theta) - Q_{\theta^*}(\theta^*)}_{= \varepsilon \text{ (I)}} + \underbrace{Q_{\theta^*}(\theta^*) - \hat{Q}_{\lambda_n}(\theta^*)}_{> -\frac{\varepsilon}{4} \text{ (II)}}$$

$$> \tfrac{1}{4}\varepsilon > 0$$

$\Rightarrow \inf_{\theta \in \partial\Theta} \hat{Q}_{\lambda_n}(\theta) > \hat{Q}_{\lambda_n}(\theta^*)$

$\hat{Q}_{\lambda_n}(\theta)$ strikt konvex $\Rightarrow$ $\hat{\theta}_n = \operatorname{argmin}_{\theta \in \mathbb{R}^p} \hat{Q}_{\lambda_n}(\theta) \in \Theta$

} Mit $\mathbb{P}(\ldots) \to 1$

Mit der speziellen Wahl $\Theta = \overline{B_\delta(\theta^*)}$ für $\delta > 0$ bel., folgt

$$P\left(|\hat{\theta}_n - \theta^*| > \delta\right) = P(\hat{\theta}_n \notin \Theta) \xrightarrow{n\to\infty} 0,$$

also $\hat{\theta}_n \xrightarrow{P} \theta^*$.

□

Theorem 8

geg: • Logistisches Regressionsmodell korrekt spezifiziert mit wahrem Parameter θ_0

• Σ_{full} : asymptotische Kovarianzmatrix des MLE auf dem gesamten Datensatz

• $\mathbb{E}\|X\|^2 < \infty$

• $\lambda_n \xrightarrow{\mathbb{P}} \theta_0$ unabh. von den Daten $(x_i, y_i)_{i \in \mathbb{N}}$

z: $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{D} \mathcal{N}(0, 2\Sigma_{full})$

Beweis:
Da das Modell für $\mathbb{P}$ korrekt spezifiziert ist, ist es das auch für $\mathbb{P}_\lambda$.

(Denn die Parameter unterscheiden sich unter um λ.)

Also gilt $\bar{\theta}(\lambda) = \underset{\theta \in \mathbb{R}^2}{\operatorname{argmin}} Q_\lambda(\theta) = \theta_0$. → Da $\nabla_\theta Q_\lambda(\theta) = 0$ für $\theta = \theta_0$

Da das Modell korrekt spezifiziert ist, gilt $J(\theta_0, \lambda) = H(\theta_0, \lambda)$.

Nach Theorem 6 ergibt sich also:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \longrightarrow \mathcal{N}(0, \underbrace{\bar{a}(\theta_0)^{-1} H(\theta_0, \theta_0)^{-1}}_{=: \Sigma_{LCC}})$$

Bemerke zunächst:

$$H(\theta_0, 0) = \underbrace{\bar{a}(0)^{-1}}_{=\frac{1}{2}} \int \frac{\exp((\theta_0 - 0)^T x)}{(1+\exp((\theta_0-0)^T x))^2} \cdot \underbrace{\frac{\overbrace{\exp(0^T x)}^{=1} + \exp(\theta_0^T x)}{(1+\exp(0^T x))(1+\exp(\theta_0^T x))}}_{= \frac{1}{2}} x x^T d\mathbb{P}^X(x)$$

$$= \int \frac{\exp(\theta_0^T x)}{(1+\exp(\theta^T x))^2} x x^T d\mathbb{P}^X(x)$$

Für $\theta = \theta_0$, $\lambda = \theta_0$, $f(x) = \theta_0^T x$ vereinfacht sich $H(\theta, \lambda)$ zu:

$$H(\theta_0, \theta_0) = \bar{a}(\theta_0)^{-1} \int \underbrace{\frac{\overbrace{\exp((\theta_0 - \theta_0)^T x)}^{=1}}{(1+\underbrace{\exp((\theta_0-\theta_0)^T x)}_{=1})^2}}_{=\frac{1}{4}} \cdot \underbrace{\frac{\exp(\theta_0^T x) + \exp(\theta_0^T x)}{(1+\exp(\theta_0^T x))(1+\exp(\theta_0^T x))}}_{= 2\frac{\exp(\theta_0^T x)}{(1+\exp(\theta_0^T x))^2}} x x^T d\mathbb{P}^X(x)$$

$$= \bar{a}(\theta_0)^{-1} \frac{1}{2} \int \frac{\exp(\theta_0^T x)}{(1+\exp(\theta_0^T x))^2} x x^T \, d\mathbb{P}^x(x)$$

(I) $= \bar{a}(\theta_0)^{-1} \frac{1}{2} \underbrace{H(\theta_0, 0)}_{= \Sigma_{full}^{-1}}$ denn $Q_0(\theta_0) = R_0(\theta_0) = R(\theta_0)$

Risikofkt für normales logistic regression

$$= (2 \bar{a}(\theta_0) \Sigma_{full})^{-1}$$

$$\Rightarrow \Sigma_{LCC} = \bar{a}(\theta_0)^{-1} \cdot H(\theta_0, \theta_0)^{-1} = \bar{a}(\theta_0)^{-1} \cdot 2 \bar{a}(\theta_0) \Sigma_{full} = 2 \Sigma_{full}$$

□

Z: $J(\theta_0,\lambda) = H(\theta_0,\lambda)$

Für eine bel. 2-fach diffbare Funktion $\varphi: \mathbb{R}^p \to \mathbb{R}$ gilt:

Kettenregel

$$\frac{\partial}{\partial\theta_i}\log\varphi(\theta) = \frac{1}{\varphi(\theta)}\cdot\frac{\partial}{\partial\theta_i}\varphi(\theta)$$

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log\varphi(\theta) = \frac{\partial}{\partial\theta_i}\left[\frac{1}{\varphi(\theta)}\cdot\frac{\partial}{\partial\theta_j}\varphi(\theta)\right]$$

$$= \frac{1}{\varphi(\theta)}\cdot\frac{\partial}{\partial\theta_i\partial\theta_j}\varphi(\theta) - \frac{1}{\varphi(\theta)^2}\cdot\frac{\partial}{\partial\theta_i}\varphi(\theta)\cdot\frac{\partial}{\partial\theta_j}\varphi(\theta)$$

Spaltenvektor

$$\Rightarrow \nabla_\theta^2\log\varphi(\theta) = \frac{1}{\varphi(\theta)}\cdot\nabla_\theta^2\varphi(\theta) - \frac{1}{\varphi(\theta)^2}\cdot\nabla_\theta\varphi(\theta)\cdot(\nabla_\theta\varphi(\theta))^T$$

Betrachte nun $\varphi_{x,y}(\theta) = y\,p_\theta(x) + (1-y)(1-p_\theta(x))$

(Bedingte Dichte von Y geg. $X=x$)

$\Rightarrow$ obiges gilt für festes x,y

$$\Rightarrow -\rho(\theta;x,y) = \log\varphi_{x,y}(\theta)$$ (log-likelihood Funktion)

Im Folgenden seien alle Erwartungswerte nur über X und Y gebildet, nicht über λ.

Wenn nun $\mathbb{E}_\lambda\left[\frac{1}{\varphi_{x,y}(\theta_0-\lambda)}\nabla_\theta^2\varphi_{x,y}(\theta_0-\lambda)\right] = 0$ ($\mathbb{E}[\dots]$ bzgl. $P_\lambda^{(x,y)}$), folgt:

$$H(\theta_0,\lambda) = -\nabla_\theta^2 Q_\lambda(\theta_0,\lambda)$$

$$= \nabla_\theta^2\,\mathbb{E}_\lambda[-\rho(\theta_0-\lambda;x,y)]$$

$$= \mathbb{E}_\lambda[\nabla_\theta^2\log\varphi_{x,y}(\theta_0-\lambda)]$$

$$= \mathbb{E}_\lambda\left[\frac{1}{\varphi_{x,y}(\theta_0-\lambda)}\cdot\nabla_\theta^2\varphi_{x,y}(\theta_0-\lambda)\right] + \mathbb{E}_\lambda\left[\frac{1}{\varphi_{x,y}(\theta_0-\lambda)^2}\nabla_\theta\varphi_{x,y}(\theta_0-\lambda)(\nabla_\theta\varphi_{x,y}(\theta_0-\lambda))^T\right]$$

$\stackrel{!}{=} 0$

Hier muss eigentlich ein − statt dem + stehen
↪ $H(\theta,\lambda)$ um Faktor -1 falsch definiert?

$$= \mathbb{E}_\lambda\left[\left(\nabla_\theta \log \varphi_{X,Y}(\theta_0-\lambda)\right)\left(\nabla_\theta \log \varphi_{X,Y}(\theta_0-\lambda)\right)^T\right]$$

$$= \mathrm{Var}_\lambda\left[\nabla_\theta \log \varphi_{X,Y}(\theta_0-\lambda)\right]$$

$$= \mathrm{Var}_\lambda\left[\nabla_\theta\left(-\rho(\theta_0-\lambda; X,Y)\right)\right]$$

$$= J(\theta, \lambda)$$

Also bleibt zu zeigen: $\mathbb{E}_\lambda\left[\frac{1}{\varphi_{X,Y}(\theta_0-\lambda)} \nabla_\theta^2 \varphi_{X,Y}(\theta_0-\lambda)\right] = 0$

Mit Turmeigenschaft gilt:

$$\mathbb{E}_\lambda\left[\frac{1}{\varphi_{X,Y}(\theta_0-\lambda)} \nabla_\theta^2 \varphi_{X,Y}(\theta_0-\lambda)\right] = \mathbb{E}\left[\underbrace{\mathbb{E}_\lambda\left[\frac{1}{\varphi_{X,Y}(\theta_0-\lambda)} \nabla_\theta^2 \varphi_{X,Y}(\theta_0-\lambda) \mid X\right]}_{(*)}\right]$$

Es genügt zu zeigen, dass der bedingte Erwartungswert $(*)$ gleich 0 ist.

$$\mathbb{E}_\lambda\left[\frac{1}{\varphi_{X,Y}(\theta_0-\lambda)} \nabla_\theta^2 \varphi_{X,Y}(\theta_0-\lambda) \mid X\right] = \sum_{y=0}^{\infty} \frac{1}{\varphi_{X,Y}(\theta_0-\lambda)} \nabla_\theta^2 \varphi_{X,Y}(\theta_0-\lambda) \cdot \varphi_{X,Y}(\theta_0-\lambda)$$

$\varphi_{X,Y}(\theta_0-\lambda)$ ist bedingte Zähldichte von Y gegeben X auf Subsample

$$= \nabla_\theta^2 \underbrace{\sum_{y=0}^{\infty} \varphi_{X,Y}(\theta_0-\lambda)}_{=1 \text{ (da Dichte)}}$$

$$= 0$$

ξ: $J(\theta_0, \lambda) = I_\lambda(\theta_0)$ (Fisher-Information)

Beweis:
Per Def. gilt:

$$J(\theta_0, \lambda) = \mathrm{Var}_\lambda\left(\nabla_\theta \rho(\theta_0 - \lambda; x, Y)\right)$$

$$= \mathbb{E}_\lambda\left[\left(\nabla_\theta \rho(\theta_0 - \lambda; x, Y)\right)\left(\nabla_\theta \rho(\theta_0 - \lambda; x, Y)\right)^T\right] - \underbrace{\left(\mathbb{E}_\lambda\left[\nabla_\theta \rho(\theta_0 - \lambda; x, Y)\right]\right)^2}_{\overset{!}{=} 0}$$

$$= \mathbb{E}_\lambda\left[\left(\nabla_\theta \log \varphi_{x,Y}(\theta_0 - \lambda)\right)\left(\nabla_\theta \log \varphi_{x,Y}(\theta_0 - \lambda)\right)^T\right] \quad \xi:$$

$$= I_\lambda(\theta_0)$$

(da $\varphi_{x,Y}(\theta_0 - \lambda)$ bedingte Zähldichte von Y geg. X ist und Multiplikation mit Dichte von X fällt weg, da $\log(a \cdot b) = \log(a) + \log(b)$ und Dichte von X nicht von θ abhängt.)

Also betrachte:

$$\mathbb{E}_\lambda\left[\nabla_\theta \rho(\theta_0 - \lambda; x, Y) \mid X\right] = \mathbb{E}_\lambda\left[\nabla_\theta \log \varphi_{x,Y}(\theta_0 - \lambda)\right]$$

$$= \mathbb{E}_\lambda\left[\frac{1}{\varphi_{x,Y}(\theta_0 - \lambda)} \cdot \nabla_\theta \varphi_{x,Y}(\theta_0 - \lambda)\right]$$

$$= \sum_{y=0}^{\infty} \frac{1}{\varphi_{x,Y}(\theta_0 - \lambda)} \cdot \nabla_\theta \varphi_{x,Y}(\theta_0 - \lambda) \cdot \varphi_{x,Y}(\theta_0 - \lambda)$$

$$= \nabla_\theta \underbrace{\sum_{y=0}^{\infty} \varphi_{x,Y}(\theta_0 - \lambda)}_{=1}$$

$$= 0$$