week11: Effizienz der Maximum-Likelihood-Schätzer, Teil2:
Mit Cramer-Rao Abschätzung (nicht klausurrelevant)

In einer Stochastik II Vorlesung wird gelegentlich die sogenannte Cramer-Rao Abschätzung bewiesen, das ist die folgende Sache: Wir haben Zufallszahlen oder zufällige Grössen $x_1, x_2, \cdots, x_n$, die von einer Wahrscheinlichkeitsverteilung

$$p_\theta(x) \;=\; p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n) \tag{1}$$

generiert worden sind. Es gelte also

$$\int_{\mathbb{R}^n} p_\theta(x)\; d^n x \;=\; \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n)\; dx_1 dx_2 \cdots dx_n \;=\; 1 \tag{2}$$

Die theta's $\theta_1, \cdots, \theta_m$ sind die Modellparameter. Wenn wir das $p_\theta(x)$ nur als Funktion von θ auffassen, weil wir für die $x_1, \cdots, x_n$ die uns gegebenen realisierten Daten einsetzen, dann ist das genau die Likelihood-Funktion,

$$L(\theta) \;=\; L(\theta_1,\cdots,\theta_m) \;=\; p_{\theta_1,\cdots,\theta_m}(x_1,\cdots,x_n) \tag{3}$$

Schreiben wir das $p_\theta(x)$ für unsere Situation, Lineare Regression als statistisches Problem, eben hin: Die Modellparameter sind die Regressionskoeffizienten $\beta_0, \cdots, \beta_p$ und die Standardabweichung σ der normalverteilten Fehler $\varepsilon_1, \cdots, \varepsilon_n$, wenn wir für unsere gegebenen Daten $\vec{y}$ und $\vec{x}_0, \cdots, \vec{x}_p$ den Regressionsansatz

$$\vec{y} \;=\; \beta_0\, \vec{x}_0 \;+\; \cdots \;+\; \beta_p\, \vec{x}_p \;+\; \vec{\varepsilon} \tag{4}$$

oder in Koordinaten

$$y_i \;=\; \beta_0\, x_{i,0} \;+\; \cdots \;+\; \beta_p\, x_{i,p} \;+\; \varepsilon_i \;=:\; \mu_i \;+\; \varepsilon_i \tag{5}$$

machen. Die zufälligen Grössen, die wir oben mit $x_1, \cdots, x_n$ bezeichnet haben, sind hier jetzt die $y_1, \cdots, y_n$, wir haben hier also eine W'keitsdichte oder Likelihood-Funktion

$$p_{\beta,\sigma}(y) \;=\; p_{\beta_0,\cdots,\beta_p,\sigma}(y_1,\cdots,y_n) \;=\; L(\beta_0,\cdots,\beta_p,\sigma) \;=\; L(\beta,\sigma) \tag{6}$$

Die hatten wir in dem week7.pdf hergeleitet, das war da die Gleichung (9):

$$L(\beta,\sigma) \;=\; \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}}\, \exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma^2} \right] \right\} \tag{7}$$

Schreiben wir jetzt die Cramer-Rao Abschätzung hin, das war etwa das Theorem 5.3.2 aus dem week10a.pdf aus der Stochastik II Vorlesung aus dem SS2021. Für die Formulierung des Theorems benutzen wir zunächst wieder die Notation von ganz oben, die zufälligen Grössen seien also ganz allgemeine $x_1, \cdots, x_n$ mit W'keitsdichte $p_{\theta_1,\cdots,\theta_m}(x_1, \cdots, x_n)$.

Theorem (Cramer-Rao Abschätzung, Stochastik II): Gegeben sei eine Wahrscheinlichkeitskeitsverteilung $p_\theta(x) = p_{\theta_1,\cdots,\theta_m}(x_1, \cdots, x_n)$ mit

$$\int_{\mathbb{R}^n} p_\theta(x)\, d^n x \;=\; \int_{\mathbb{R}^n} p_{\theta_1,\cdots,\theta_m}(x_1, \cdots, x_n)\, dx_1 dx_2 \cdots dx_n \;=\; 1 \tag{8}$$

Für $k = 1, \cdots, m$ seien

$$\hat\theta \;=\; \begin{pmatrix} \hat\theta_1 \\ \vdots \\ \hat\theta_m \end{pmatrix} \;:\; \mathbb{R}^n \;\to\; \mathbb{R}^m \tag{9}$$

erwartungstreue Schätzer, d.h. es gilt

$$\mathsf{E}[\,\hat\theta_k\,] \;:=\; \int_{\mathbb{R}^n} \hat\theta_k(x)\, p_\theta(x)\, d^n x \;=\; \theta_k \tag{10}$$

Weiter sei

$$\mathsf{Cov}(\hat\theta) \;:=\; \Big(\, \mathsf{Cov}[\hat\theta_k, \hat\theta_\ell] \,\Big)_{k,\ell=1,\ldots,m} \;\in\; \mathbb{R}^{m\times m} \tag{11}$$

die Covarianz-Matrix von $\hat\theta$, insbesondere ist also $\mathsf{V}[\,\hat\theta_k\,] = \mathsf{Cov}(\hat\theta)_{k,k}$ die Varianz des k-ten Schätzers. Wir definieren die sogenannte Fisher-Informationsmatrix $I(\theta)$ durch

$$I(\theta) \;:=\; \Big(\, -\mathsf{E}\Big[\, \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell} \,\Big] \,\Big)_{k,\ell=1,\ldots,m} \;\in\; \mathbb{R}^{m\times m} \tag{12}$$

mit

$$\mathsf{E}\Big[\, \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell} \,\Big] \;=\; \int_{\mathbb{R}^n} \tfrac{\partial^2 \log p_\theta}{\partial\theta_k \partial\theta_\ell}(x_1, \cdots, x_n)\, p_\theta(x_1, \cdots, x_n)\, d^n x \tag{13}$$

und $I^{-1}(\theta)$ sei das Inverse von $I(\theta)$. Dann gilt:

$$\big\langle\, v\,, \mathsf{Cov}(\hat\theta)\, v \,\big\rangle \;\geq\; \big\langle\, v\,, I^{-1}(\theta)\, v \,\big\rangle \qquad \forall v \in \mathbb{R}^m \tag{14}$$

Insbesondere gilt also, wenn wir für v den k-ten Standardbasisvektor nehmen,

$$\mathsf{V}[\,\hat\theta_k\,] \;\geq\; \big[\, I^{-1}(\theta) \,\big]_{k,k} \tag{15}$$

für jeden erwartungstreuen Schätzer $\hat\theta_k : \mathbb{R}^n \to \mathbb{R}$.

Wir wollen jetzt das $I(\theta)$ und das $I^{-1}(\theta)$ für unsere Situation, mit der W'keitsdichte gegeben durch (7),

$$p_\theta(x) \;\to\; p_{\beta,\sigma}(y) \;=\; \prod_{i=1}^{n} \Big\{\, \tfrac{1}{\sqrt{2\pi\sigma^2}}\, \exp\big[-\tfrac{(y_i-\mu_i)^2}{2\sigma^2}\big] \,\Big\} \tag{16}$$

berechnen. Damit bekommen wir dann eine untere Schranke für die Varianz von erwartungstreuen Schätzern für die Regressionskoeffizienten. Wenn diese untere Schranke dann identisch ist mit der Varianz der Maximum-Likelihood-Schätzer, und das ist der Fall, heisst das dann also, dass die Maximum-Likelihood-Schätzer minimale Varianz haben, also effizient sind. Im Gegensatz zu den Herleitungen von letzter Woche müssen wir diesmal keine spezielle funktionale Form der betrachteten Schätzer voraussetzen, letztes Mal hatten wir ja eine Menge $\mathcal{L}$ von linearen (in $\vec{y}$) Schätzern betrachtet. Diese Einschränkung ist jetzt nicht mehr nötig, die einzige Voraussetzung ist die Erwartungstreue, wir betrachten also Schätzer

$$\tilde{\beta}_k \;:\; \mathbb{R}^n \;\to\; \mathbb{R} \tag{17}$$

mit

$$\mathsf{E}[\,\tilde{\beta}_k\,] \;=\; \beta_k \tag{18}$$

Wir müssen die Grössen (13) berechnen. Wir haben mit $\nu := \sigma^2$

$$p_{\beta,\nu}(y) \;=\; \prod_{i=1}^{n} \Big\{ \tfrac{1}{\sqrt{2\pi\nu}} \exp\big[-\tfrac{(y_i-\mu_i)^2}{2\nu}\big] \Big\} \tag{19}$$

$$\log p_{\beta,\nu}(y) \;=\; -\tfrac{n}{2}\log(2\pi\nu) \;-\; \sum_{i=1}^{n} \tfrac{(y_i-\mu_i)^2}{2\nu} \tag{20}$$

wobei

$$\mu_i \;=\; \mu_i(\beta) \;=\; \beta_0\, x_{i,0} \;+\; \cdots \;+\; \beta_p\, x_{i,p} \tag{21}$$

Also bekommen wir

$$\tfrac{\partial \log p_{\beta,\nu}}{\partial \beta_j} \;=\; \sum_{i=1}^{n} \tfrac{\partial \log p_{\beta,\nu}}{\partial \mu_i}\, \tfrac{\partial \mu_i}{\partial \beta_j} \;=\; \sum_{i=1}^{n} \tfrac{y_i-\mu_i}{\nu}\, x_{i,j} \tag{22}$$

und

$$\tfrac{\partial \log p_{\beta,\nu}}{\partial \nu} \;=\; -\tfrac{n}{2\nu} \;+\; \sum_{i=1}^{n} \tfrac{(y_i-\mu_i)^2}{2\nu^2} \tag{23}$$

Damit erhalten wir die folgenden zweiten Ableitungen:

$$\tfrac{\partial^2 \log p_{\beta,\nu}}{\partial \beta_j \partial \beta_k} \;=\; \sum_{i=1}^{n} \tfrac{-x_{i,k}}{\nu}\, x_{i,j} \;=\; -\,\tfrac{\vec{x}_j \vec{x}_k}{\nu} \tag{24}$$

$$\tfrac{\partial^2 \log p_{\beta,\nu}}{\partial \beta_j \partial \nu} \;=\; -\sum_{i=1}^{n} \tfrac{y_i-\mu_i}{\nu^2}\, x_{i,j} \tag{25}$$

$$\tfrac{\partial^2 \log p_{\beta,\nu}}{\partial \nu^2} \;=\; \tfrac{n}{2\nu^2} \;-\; \sum_{i=1}^{n} \tfrac{(y_i-\mu_i)^2}{\nu^3} \tag{26}$$

Wegen

$$\mathsf{E}[\,y_i-\mu_i\,] \;=\; \mathsf{E}[\,\varepsilon_i\,] \;=\; 0 \tag{27}$$

$$\mathsf{E}[\,(y_i-\mu_i)^2\,] \;=\; \mathsf{E}[\,\varepsilon_i^2\,] \;=\; \sigma^2 \;=\; \nu \tag{28}$$

bekommen wir dann die folgenden Erwartungswerte:

$$\mathsf{E}\Big[\,\tfrac{\partial^2 \log p_{\beta,\nu}}{\partial\beta_j\partial\beta_k}\,\Big] \;=\; -\,\tfrac{\vec{x}_j\vec{x}_k}{\nu} \tag{29}$$

$$\mathsf{E}\Big[\,\tfrac{\partial^2 \log p_{\beta,\nu}}{\partial\beta_j\partial\nu}\,\Big] \;=\; 0 \tag{30}$$

$$\mathsf{E}\Big[\,\tfrac{\partial^2 \log p_{\beta,\nu}}{\partial\nu^2}\,\Big] \;=\; \tfrac{n}{2\nu^2} \;-\; \sum_{i=1}^{n} \tfrac{\nu}{\nu^3} \;=\; \tfrac{n}{2\nu^2} \;-\; \tfrac{n}{\nu^2} \;=\; -\,\tfrac{n}{2\nu^2} \tag{31}$$

Also,

$$I(\theta) \;=\; I(\beta,\nu) \;=\; \begin{pmatrix} \frac{1}{\nu}\big(\vec{x}_j\vec{x}_k\big)_{j,k=0}^{p} & 0 \\ 0 & \frac{n}{2\nu^2} \end{pmatrix} \;=\; \begin{pmatrix} \frac{1}{\nu}\,X^TX & 0 \\ 0 & \frac{n}{2\nu^2} \end{pmatrix} \tag{32}$$

und damit

$$I^{-1}(\beta,\nu) \;=\; \begin{pmatrix} \nu\,(X^TX)^{-1} & 0 \\ 0 & \frac{2\nu^2}{n} \end{pmatrix} \tag{33}$$

Die Cramer-Rao Abschätzung liefert also in diesem Fall (mit $j = 0, 1, \cdots, p$)

$$\mathsf{V}[\,\tilde{\beta}_j\,] \;\geq\; I^{-1}(\beta,\nu)_{j,j} \;=\; \nu\,[(X^TX)^{-1}]_{j,j} \;=\; \sigma^2\,[(X^TX)^{-1}]_{j,j} \tag{34}$$

$$\mathsf{V}[\,\tilde{\nu}\,] \;\geq\; I^{-1}(\beta,\nu)_{p+1,p+1} \;=\; \tfrac{2\nu^2}{n} \;=\; \tfrac{2\sigma^4}{n} \tag{35}$$

für beliebige erwartungstreue Schätzer $\tilde{\beta}_j$ und $\tilde{\nu} = \widetilde{\sigma^2}$. Der Ausdruck auf der rechten Seite von (34) ist aber gerade die Varianz des Maximum-Likelihood-Schätzers für den j-ten Regressionskoeffizienten,

$$\mathsf{V}\big[\,\hat{\beta}_{\mathrm{ML},j}\,\big] \;=\; \sigma^2\,[(X^TX)^{-1}]_{j,j} \tag{36}$$

das hatten wir ja in dem week9.pdf in dem Theorem 9.1, Teil (a), gezeigt. Also haben die Maximum-Likelihood-Schätzer für die Regressionskoeffizienten minimale Varianz in der Menge aller erwartungstreuen Schätzer, sie sind also effizient. Für die Varianz des $\widehat{s^2}$, die erwartungstreue Version des Maximum-Likelihood-Schätzers für das σ^2, hatten wir ebenfalls im Theorem 9.1, dann im Teil (b), den Ausdruck

$$\mathsf{V}\big[\,\widehat{s^2}\,\big] \;=\; \frac{2\sigma^4}{n-(p+1)} \tag{37}$$

hergeleitet. Für festes n ist das ein kleines bischen grösser als die rechte Seite von (35). Hier könnte man dann zunächst nur sagen, dass das $\widehat{s^2}$ asymptotisch, im Limes $n \to \infty$, effizient ist.