week8: Erwartungswert und Varianz der Maximum-Likelihood-Schätzer, Teil1

In der letzten Veranstaltung hatten wir das lineare Regressionsproblem als ein statistisches Problem formuliert und die Maximum-Likelihood-Schätzer hergeleitet. Die Likelihood-Funktion war gegeben durch

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

mit

$$\mu_i \;=\; \beta_0 \;+\; \beta_1\, x_{i,1} \;+\; \cdots \;+\; \beta_p\, x_{i,p}$$

und wurde maximiert von den Maximum-Likelihood-Schätzern

$$\hat{\beta}_{\mathrm{ML}} \;=\; (X^T X)^{-1} X^T \vec{y}\ , \tag{1}$$

$$\widehat{\sigma^2}_{\mathrm{ML}} \;=\; \tfrac{1}{n}\left[\, P_{X^\perp}\vec{y}\,\right]^2\ . \tag{2}$$

Machen wir uns klar, was es heisst, dass diese Schätzer erwartungstreu sind. Dazu müssen wir die folgenden Gleichungen überprüfen:

$$\mathsf{E}\big[\,\hat{\beta}_{\mathrm{ML}}\,\big] \;\stackrel{?}{=}\; \vec{\beta} \tag{3}$$

$$\mathsf{E}\big[\,\widehat{\sigma^2}_{\mathrm{ML}}\,\big] \;\stackrel{?}{=}\; \sigma^2 \tag{4}$$

Was bedeutet das $\mathsf{E}[\,\cdot\,]$ hier jetzt genau?

In die Schätzer können wir die $\vec{x}_0,\cdots,\vec{x}_p$ und das $\vec{y}$ einsetzen und bekommen dann Zahlen heraus. Wir können etwa schreiben

$$\hat{\beta}_{\mathrm{ML}} \;=\; \hat{\beta}_{\mathrm{ML}}(X,\vec{y}) \;:\; \mathbb{R}^{n\times(p+1)}\times\mathbb{R}^n \;\to\; \mathbb{R}^{p+1}$$

$$\widehat{\sigma^2}_{\mathrm{ML}} \;=\; \widehat{\sigma^2}_{\mathrm{ML}}(X,\vec{y}) \;:\; \mathbb{R}^{n\times(p+1)}\times\mathbb{R}^n \;\to\; \mathbb{R}$$

Für das

$$X \;=\; \begin{pmatrix} | & | & & | \\ \vec{x}_0 & \vec{x}_1 & \cdots & \vec{x}_p \\ | & | & & | \end{pmatrix}$$

setzen wir auf jeden Fall unsere konkret gegebenen Daten ein. Für die $\vec{y}=(y_1,\cdots,y_n)$ können wir entweder die ebenfalls konkret gegebenen Daten einsetzen, oder wir setzen Daten ein, die

von unserer Modell-Annahme Lineare Regression als Statistisches Problem herkommen. Da hatten wir ja gesagt, dass die Ypsilons 'in Wirklichkeit' von einem stochastischen Modell

$$\begin{aligned} \vec{y} &= \beta_0 \vec{x}_0 + \beta_1 \vec{x}_1 + \cdots + \beta_p \vec{x}_p + \vec{\varepsilon} \\ &= X\vec{\beta} + \vec{\varepsilon} \end{aligned} \tag{5}$$

herkommen, wobei die $\vec{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)$ normalverteilte, unabhängige Zufallszahlen mit Mittelwert 0 und Standardabweichung σ sind,

$$\varepsilon_i \in N(0,\sigma) \quad \text{unabhängig} \tag{6}$$

Die Ypsilons sind dann also ebenfalls Zufallszahlen und die setzen wir dann in die Maximum-Likelihood-Schätzer ein:

$$\hat{\beta}_{\mathrm{ML}} = \hat{\beta}_{\mathrm{ML}}(X,\vec{y}) = \hat{\beta}_{\mathrm{ML}}\big(X, X\vec{\beta}+\vec{\varepsilon}\big) \tag{7}$$

$$\widehat{\sigma^2}_{\mathrm{ML}} = \widehat{\sigma^2}_{\mathrm{ML}}(X,\vec{y}) = \widehat{\sigma^2}_{\mathrm{ML}}\big(X, X\vec{\beta}+\vec{\varepsilon}\big) \tag{8}$$

Dadurch hängen die Schätzer dann also von den $N(0,\sigma)$-normalverteilten Zufallszahlen $\vec{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)$ ab. Der Erwartungswert in den Gleichungen (3) und (4) meint dann einfach Integration gegen die entsprechenden n Gauss'schen Glockenkurven:

$$\mathsf{E}\big[\,\hat{\beta}_{\mathrm{ML}}\,\big] := \int_{\mathbb{R}^n} \hat{\beta}_{\mathrm{ML}}\big(X, X\vec{\beta}+\vec{\varepsilon}\big) \prod_{k=1}^{n} \Big\{ e^{-\frac{\varepsilon_k^2}{2\sigma^2}} \frac{d\varepsilon_k}{\sqrt{2\pi\sigma^2}} \Big\} \tag{9}$$

$$\mathsf{E}\big[\,\widehat{\sigma^2}_{\mathrm{ML}}\,\big] := \int_{\mathbb{R}^n} \widehat{\sigma^2}_{\mathrm{ML}}\big(X, X\vec{\beta}+\vec{\varepsilon}\big) \prod_{k=1}^{n} \Big\{ e^{-\frac{\varepsilon_k^2}{2\sigma^2}} \frac{d\varepsilon_k}{\sqrt{2\pi\sigma^2}} \Big\} \tag{10}$$

Wir wollen jetzt den folgenden Satz beweisen:

Theorem 8.1: a) Der Maximum-Likelihood-Schätzer $\hat{\beta}_{\mathrm{ML}}$ für die Regressionskoeffizienten ist erwartungstreu, es gilt

$$\mathsf{E}\big[\,\hat{\beta}_{\mathrm{ML}}\,\big] = \vec{\beta} \tag{11}$$

b) Der Maximum-Likelihood-Schätzer $\widehat{\sigma^2}_{\mathrm{ML}}$ für die Varianz der Fehler ist nur asymptotisch, im Limes $n \to \infty$ erwartungstreu. Genauer gilt

$$\mathsf{E}\big[\,\widehat{\sigma^2}_{\mathrm{ML}}\,\big] = \tfrac{n-(p+1)}{n}\,\sigma^2 \tag{12}$$

Damit ist der modifizierte Schätzer

$$\widehat{s^2} := \tfrac{n}{n-(p+1)}\,\widehat{\sigma^2}_{\mathrm{ML}} = \tfrac{1}{n-(p+1)}\,\big[\,P_{X^\perp}\vec{y}\,\big]^2 \tag{13}$$

erwartungstreu, es gilt $\mathsf{E}[\widehat{s^2}] = \sigma^2$.

Beweis: a) Mit

$$\vec{y} = X\vec{\beta} + \vec{\varepsilon}$$

bekommen wir

$$\begin{aligned}\hat{\beta}_{\mathrm{ML}} &= (X^T X)^{-1} X^T \vec{y} \\ &= (X^T X)^{-1} X^T \big(X\vec{\beta} \,+\, \vec{\varepsilon} \big) \\ &= \vec{\beta} \,+\, (X^T X)^{-1} X^T \vec{\varepsilon}\end{aligned}$$

und damit

$$\begin{aligned}\mathsf{E}\big[\, \hat{\beta}_{\mathrm{ML}} \,\big] &= \mathsf{E}\big[\, \vec{\beta} \,+\, (X^T X)^{-1} X^T \vec{\varepsilon} \,\big] \\ &= \vec{\beta} \,+\, (X^T X)^{-1} X^T \, \mathsf{E}[\, \vec{\varepsilon}\,] \;=\; \vec{\beta}\end{aligned}$$

da $\mathsf{E}[\vec{\varepsilon}] = 0$. Zum Beweis von Teil (b) beweisen wir zunächst das folgende

Lemma 8.2: Es seien $\vec{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)$ normalverteilte, unabhängige Zufallszahlen mit Mittelwert 0 und Standardabweichung σ wie in unserer Modellannahme (5,6). Weiterhin sei $A \in \mathbb{R}^{n\times n}$ eine beliebige $n \times n$-Matrix. Dann gilt

$$\mathsf{E}\big[\, (A\vec{\varepsilon})_j (A\vec{\varepsilon})_k \,\big] \;=\; \textstyle\sum_{\ell=1}^n a_{j\ell}\, a_{k\ell}\, \sigma^2 \;=\; \big(AA^T \big)_{j,k}\, \sigma^2 \tag{14}$$

Insbesondere,

$$\mathsf{E}\big[\, (A\vec{\varepsilon})^2 \,\big] \;=\; \textstyle\sum_{j=1}^n \sum_{\ell=1}^n a_{j\ell}\, a_{j\ell}\, \sigma^2 \;=\; \mathrm{Tr}\big(AA^T \big)\, \sigma^2 \;\;. \tag{15}$$

Beweis: Wir haben

$$(A\vec{\varepsilon})_j (A\vec{\varepsilon})_k \;=\; \textstyle\sum_{\ell,m=1}^n a_{j\ell}\, \varepsilon_\ell \; a_{km}\, \varepsilon_m$$

Wegen

$$\begin{aligned}\mathsf{E}[\varepsilon_\ell\, \varepsilon_m] &= \begin{cases} \sigma^2 & \text{falls } \ell = m \\ 0 & \text{falls } \ell \neq m \end{cases} \\ &= \sigma^2\, \delta_{\ell,m}\end{aligned}$$

bekommen wir

$$\begin{aligned}\mathsf{E}\big[\, (A\vec{\varepsilon})_j (A\vec{\varepsilon})_k \,\big] &= \textstyle\sum_{\ell,m=1}^n a_{j\ell}\, a_{km}\, \mathsf{E}[\varepsilon_\ell\, \varepsilon_m] \\ &= \textstyle\sum_{\ell,m=1}^n a_{j\ell}\, a_{km}\, \sigma^2\, \delta_{\ell,m} \\ &= \textstyle\sum_{\ell=1}^n a_{j\ell}\, a_{k\ell}\, \sigma^2 \\ &= \textstyle\sum_{\ell=1}^n a_{j\ell}\, a^T_{\ell k}\, \sigma^2 \;=\; \big(AA^T \big)_{j,k}\, \sigma^2\end{aligned}$$

und das Lemma ist bewiesen. ■

Beweis Teil b Theorem 8.1: Es war

$$\widehat{\sigma^2}_{\mathrm{ML}} \;=\; \tfrac{1}{n}\left[\, P_{X^\perp}\vec{y}\,\right]^2$$

mit

$$P_X \;=\; X(X^TX)^{-1}X^T$$

und

$$P_{X^\perp} \;=\; Id - P_X \;\;.$$

Mit

$$\vec{y} \;=\; X\vec{\beta} \,+\, \vec{\varepsilon}$$

bekommen wir dann

$$P_{X^\perp}\vec{y} \;=\; P_{X^\perp}X\vec{\beta} \,+\, P_{X^\perp}\vec{\varepsilon}$$

Wegen

$$\begin{aligned} P_{X^\perp}X \;&=\; [Id - P_X]X \\ &=\; X \;-\; X(X^TX)^{-1}X^TX \\ &=\; X \;-\; X \;=\; 0 \end{aligned}$$

haben wir also

$$P_{X^\perp}\vec{y} \;=\; P_{X^\perp}\vec{\varepsilon}\;\;.$$

Den Erwartungswert können wir jetzt mit Hilfe des Lemma 8.2 berechnen:

$$\begin{aligned} \mathsf{E}\big[\,\widehat{\sigma^2}_{\mathrm{ML}}\,\big] \;&=\; \mathsf{E}\Big[\,\tfrac{1}{n}\left\{\,P_{X^\perp}\vec{y}\,\right\}^2\Big] \\ &=\; \tfrac{1}{n}\,\mathsf{E}\big[\,\left\{\,P_{X^\perp}\vec{\varepsilon}\,\right\}^2\,\big] \\ &\overset{\text{Lemma 8.2}}{=}\; \tfrac{1}{n}\,\mathrm{Tr}\big[\,P_{X^\perp}^TP_{X^\perp}\,\big]\,\sigma^2 \end{aligned}$$

Nun ist

$$\begin{aligned} P_{X^\perp}^TP_{X^\perp} \;&=\; [Id - P_X]^T[Id - P_X] \\ &\overset{P_X^T=P_X}{=}\; [Id - P_X][Id - P_X] \\ &=\; Id - 2P_X + P_X^2 \\ &\overset{P_X^2=P_X}{=}\; Id - P_X \end{aligned}$$

so dass

$$\begin{aligned} \mathrm{Tr}\big[\,P_{X^\perp}^TP_{X^\perp}\,\big] \;&=\; \mathrm{Tr}\big[\,Id - P_X\,\big] \\ &=\; n \;-\; \mathrm{Tr}[P_X] \end{aligned}$$

Schliesslich haben wir mit $\mathrm{Tr}[AB] = \mathrm{Tr}[BA]$

$$\begin{aligned} \mathrm{Tr}[P_X] &= \mathrm{Tr}\big[\, X(X^TX)^{-1}X^T \,\big] \\ &= \mathrm{Tr}\big[\, (X^TX)^{-1}X^TX \,\big] \\ &= \mathrm{Tr}\big[\, Id_{(p+1)\times(p+1)} \,\big] \;=\; p+1 \end{aligned}$$

Also insgesamt

$$\begin{aligned} \mathsf{E}\big[\, \widehat{\sigma^2}_{\mathrm{ML}} \,\big] &= \tfrac{1}{n}\,\mathrm{Tr}\big[\, P_{X^\perp}^T P_{X^\perp} \,\big]\,\sigma^2 \\ &= \tfrac{n-(p+1)}{n}\,\sigma^2 \end{aligned}$$

und das Theorem 8.1 ist bewiesen. ■

Nächste Woche berechnen wir dann noch die Varianzen der Maximum-Likelihood-Schätzer. Die brauchen wir, wenn wir dann zeigen wollen, dass die Maximum-Likelihood-Schätzer effizient in einer noch zu definierenden Klasse von Schätzern sind. Das meint gerade, dass die Maximum-Likelihood-Schätzer minimale Varianz in dieser Klasse von Schätzern haben. Die Maximum-Likelihood-Schätzer sind also sehr schöne Schätzer.