

### VL9: 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

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

und wurde maximiert von den Maximum-Likelihood-Schätzern

$$\hat{\beta}_{\text{ML}} = (X^T X)^{-1} X^T \vec{y} , \quad (1)$$

$$\hat{\sigma}_{\text{ML}}^2 = \frac{1}{n} [P_{X^\perp} \vec{y}]^2 . \quad (2)$$

Dann hatten wir uns noch klargemacht, was es heisst, dass diese Schätzer erwartungstreu sind. Dazu müssen wir die folgenden Gleichungen überprüfen:

$$\mathbf{E}[\hat{\beta}_{\text{ML}}] \stackrel{?}{=} \vec{\beta} \quad (3)$$

$$\mathbf{E}[\hat{\sigma}_{\text{ML}}^2] \stackrel{?}{=} \sigma^2 \quad (4)$$

und wir hatten uns überlegt, was das  $\mathbf{E}[\cdot]$  hier genau bedeutet:

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

$$\begin{aligned} \hat{\beta}_{\text{ML}} &= \hat{\beta}_{\text{ML}}(X, \vec{y}) : \mathbb{R}^{n \times (p+1)} \times \mathbb{R}^n \rightarrow \mathbb{R}^{p+1} \\ \hat{\sigma}_{\text{ML}}^2 &= \hat{\sigma}_{\text{ML}}^2(X, \vec{y}) : \mathbb{R}^{n \times (p+1)} \times \mathbb{R}^n \rightarrow \mathbb{R} \end{aligned}$$

Für das

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

setzen wir auf jeden Fall unsere konkret gegebenen Daten ein. Für die  $\vec{y} = (y_1, \dots, y_n)$  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  $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$  normalverteilte, unabhängige Zufallszahlen mit Mittelwert 0 und Standardabweichung  $\sigma$  sind,

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

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

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

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

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

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

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

Wir wollen jetzt den folgenden Satz beweisen:

**Theorem 9.1:** a) Der Maximum-Likelihood-Schätzer  $\hat{\beta}_{\text{ML}}$  für die Regressionskoeffizienten ist erwartungstreu, es gilt

$$\mathbb{E}[\hat{\beta}_{\text{ML}}] = \vec{\beta}\tag{11}$$

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

$$\mathbb{E}[\hat{\sigma}_{\text{ML}}^2] = \frac{n-(p+1)}{n} \sigma^2\tag{12}$$

Damit ist der modifizierte Schätzer

$$\hat{s}^2 := \frac{n}{n-(p+1)} \hat{\sigma}_{\text{ML}}^2 = \frac{1}{n-(p+1)} [P_{X^\perp} \vec{y}]^2\tag{13}$$

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

**Beweis:** a) Mit

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

bekommen wir

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

und damit

$$\begin{aligned}\mathbb{E}[\hat{\beta}_{\text{ML}}] &= \mathbb{E}[\vec{\beta} + (X^T X)^{-1} X^T \vec{\varepsilon}] \\ &= \vec{\beta} + (X^T X)^{-1} X^T \mathbb{E}[\vec{\varepsilon}] = \vec{\beta}\end{aligned}$$

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

**Lemma 9.2:** Es seien  $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$  normalverteilte, unabhängige Zufallszahlen mit Mittelwert 0 und Standardabweichung  $\sigma$  wie in unserer Modellannahme (5,6). Weiterhin sei  $A \in \mathbb{R}^{n \times n}$  eine beliebige  $n \times n$ -Matrix. Dann gilt

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

Insbesondere,

$$\mathbb{E}[(A\vec{\varepsilon})^2] = \sum_{j=1}^n \sum_{\ell=1}^n a_{j\ell} a_{j\ell} \sigma^2 = \text{Tr}(AA^T) \sigma^2. \quad (15)$$

**Beweis:** Wir haben

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

Wegen

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

bekommen wir

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

und das Lemma ist bewiesen. ■

**Beweis Teil b Theorem 9.1:** Es war

$$\widehat{\sigma}_{\text{ML}}^2 = \frac{1}{n} [P_{X^\perp} \vec{y}]^2$$

mit

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

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

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

haben wir also

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

Den Erwartungswert können wir jetzt mit Hilfe des Lemma 9.2 berechnen:

$$\begin{aligned} \mathbb{E}[\widehat{\sigma}_{\text{ML}}^2] &= \mathbb{E}\left[\frac{1}{n} \{P_{X^\perp} \vec{y}\}^2\right] \\ &= \frac{1}{n} \mathbb{E}[\{P_{X^\perp} \vec{\varepsilon}\}^2] \\ &\stackrel{\text{Lemma 9.2}}{=} \frac{1}{n} \text{Tr}[P_{X^\perp}^T P_{X^\perp}] \sigma^2 \end{aligned}$$

Nun ist

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

so dass

$$\begin{aligned} \text{Tr}[P_{X^\perp}^T P_{X^\perp}] &= \text{Tr}[Id - P_X] \\ &= n - \text{Tr}[P_X] \end{aligned}$$

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

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

Also insgesamt

$$\begin{aligned}\mathbb{E}[\widehat{\sigma}_{\text{ML}}^2] &= \frac{1}{n} \text{Tr}[P_{X^\perp}^T P_{X^\perp}] \sigma^2 \\ &= \frac{n-(p+1)}{n} \sigma^2\end{aligned}$$

und das Theorem 9.1 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.