week7: Maximum Likelihood Schätzung der Regressionskoeffizienten

Wir wollen von jetzt an das lineare Regressionsproblem nicht mehr als ein deterministisches Minimierungsproblem betrachten, sondern als ein statistisches Problem. Dazu betrachten wir das folgende Setup: Gegeben seinen wieder die Datenvektoren

$$\vec{x}_0 \;=\; \begin{pmatrix} x_{1,0} \\ x_{2,0} \\ \vdots \\ x_{n,0} \end{pmatrix}, \quad \vec{x}_1 \;=\; \begin{pmatrix} x_{1,1} \\ x_{2,1} \\ \vdots \\ x_{n,1} \end{pmatrix}, \quad \cdots \quad , \; \vec{x}_p \;=\; \begin{pmatrix} x_{1,p} \\ x_{2,p} \\ \vdots \\ x_{n,p} \end{pmatrix} \tag{1}$$

und

$$\vec{y} \;=\; \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \tag{2}$$

und $\vec{x}_0$ sei etwa wieder der konstante Vektor mit lauter Einsen, $\vec{x}_0 = (1, ..., 1)$ oder $x_{i,0} = 1$ für $1 \leq i \leq n$. Jetzt machen wir die folgende

Annahme (Regression als statistisches Problem): Die Ypsilon's $\vec{y} = (y_1, \cdots, y_n)$ sind nicht einfach gegebene Daten, sondern "in Wirklichkeit" kommen sie von einem stochastischen Modell

$$\vec{y} \;=\; \beta_0 \, \vec{x}_0 \;+\; \beta_1 \, \vec{x}_1 \;+\; \cdots \;+\; \beta_p \, \vec{x}_p \;+\; \vec{\varepsilon} \tag{3}$$

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

$$\varepsilon_i \;\in\; N(0, \sigma) \quad \text{unabhängig} \tag{4}$$

Die $\vec{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ sind gegebene Zahlen, die wir aber nicht kennen und bestimmen wollen.

Problem: Welche Auswahl von

$$\vec{\beta} \;=\; (\beta_0, \beta_1, \cdots, \beta_p)$$

passt 'am besten' zum statistischen Modell (3,4), wenn die $\vec{x}_j$ und $\vec{y}$ konkret durch die Daten (1,2) gegeben sind?

Lösung: Wir bestimmen die $\vec{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)$ mit Hilfe der Maximum-Likelihood-Methode, dadurch ist dann auch die genaue mathematische Bedeutung von 'am besten' definiert: Die beste Auswahl der $(\beta_0, \beta_1, \cdots, \beta_p)$ soll die Likelihood-Funktion maximieren. Wir müssen also zunächst die Likelihood-Funktion bestimmen:

Jedes einzelne y_i ist gegeben durch

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i \\ &= \mu_i + \varepsilon_i \end{aligned} \tag{5}$$

wobei wir abgekürzt haben

$$\mu_i := \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} \tag{6}$$

Insbesondere ist jedes y_i also selber eine normalverteilte Zufallszahl mit Mittelwert μ_i und Standardabweichung σ. Das bedeutet, dass die Wahrscheinlichkeit, dass sich ein solches y_i in einem Intervall $[\tilde{y}_i, \tilde{y}_i + d\tilde{y}_i)$ realisiert, gegeben ist durch

$$\mathsf{Prob}\big[\, y_i \in [\tilde{y}_i, \tilde{y}_i + d\tilde{y}_i) \,\big] = \tfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\tilde{y}_i - \mu_i)^2}{2\sigma^2}}\, d\tilde{y}_i \tag{7}$$

Wenn die ε_i unabhängig sind, sind auch die y_i unabhängig und die W'keit, dass sich alle y_i in gegebenen Intervallen $[\tilde{y}_i, \tilde{y}_i + d\tilde{y}_i)$ realisieren tun, ist dann also gegeben durch das Produkt

$$\begin{aligned} &\mathsf{Prob}\Big[\, y_1 \in [\tilde{y}_1, \tilde{y}_1 + d\tilde{y}_1),\ y_2 \in [\tilde{y}_2, \tilde{y}_2 + d\tilde{y}_2),\ \cdots,\ y_n \in [\tilde{y}_n, \tilde{y}_n + d\tilde{y}_n) \,\Big] = \\ &= \mathsf{Prob}\big[\, y_1 \in [\tilde{y}_1, \tilde{y}_1 + d\tilde{y}_1) \,\big]\ \mathsf{Prob}\big[\, y_2 \in [\tilde{y}_2, \tilde{y}_2 + d\tilde{y}_2) \,\big]\ \cdots\ \mathsf{Prob}\big[\, y_n \in [\tilde{y}_n, \tilde{y}_n + d\tilde{y}_n) \,\big] \\ &\overset{(7)}{=} \prod_{i=1}^{n} \tfrac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\tilde{y}_i - \mu_i)^2}{2\sigma^2}}\, d\tilde{y}_i \\ &= (2\pi\sigma^2)^{-\frac{n}{2}}\ \exp\Big\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\tilde{y}_i - \mu_i)^2 \Big\}\, d\tilde{y}_1 \cdots d\tilde{y}_n \end{aligned} \tag{8}$$

Nun haben wir ja konkret realisierte Werte für die $y_1, \cdots, y_n$, die wir dann also auf der rechten Seite von (8) für die $\tilde{y}_i$ einsetzen können. Wenn wir das machen, bekommen wir eine Funktion $L = L(\vec{\beta}, \sigma)$, die nur noch von den Modellparametern $\beta_0, \cdots, \beta_p$, die sind in den μ_i's drin, und σ (und den Intervallbreiten $d\tilde{y}_i \equiv dy_i$, die aber wieder aus der eigentlichen Rechnung rausfallen werden) abhängen tut, das ist dann die Likelihood Funktion:

$$L(\vec{\beta}, \sigma) := \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}}\, \exp\left[-\frac{(y_i - \mu_i)^2}{2\sigma^2} \right] dy_i \right\} \tag{9}$$

Wir betrachten wieder den Logarithmus (mit den μ_i gegeben durch die Gleichung (6) von oben) und bekommen

$$\log L(\vec{\beta}, \sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)^2 + \text{const} \tag{10}$$

wobei die Konstante

$$\text{const} := -\frac{n}{2} \log(2\pi) + \sum_{i=1}^{n} \log(dy_i)$$

nur Terme enthält, die nicht von den Modellparametern $\vec{\beta}$ und σ abhängen. Das Maximieren von L ist äquivalent zum Maximieren von $\log L$, wir maximieren $\log L$. Notwendige Bedingung ist

$$\left(\frac{\partial \log L}{\partial \beta_0}, \frac{\partial \log L}{\partial \beta_1}, \cdots, \frac{\partial \log L}{\partial \beta_p}, \frac{\partial \log L}{\partial \sigma} \right) = (0, 0, \cdots, 0, 0) \tag{11}$$

Wir haben

$$\begin{aligned} \frac{\partial \log L}{\partial \beta_j} &= + \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(y_i - \mu_i) \frac{\partial \mu_i}{\partial \beta_j} \\ &= \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu_i)\, x_{i,j} \\ &= \frac{1}{\sigma^2} \left\{ \vec{y} \cdot \vec{x}_j - \left[\beta_0 \vec{x}_0 + \cdots + \beta_p \vec{x}_p \right] \cdot \vec{x}_j \right\} \stackrel{!}{=} 0 \end{aligned}$$

für alle $0 \le j \le p$. Damit bekommen wir das Gleichungssystem

$$\begin{aligned} \beta_0 \vec{x}_0\vec{x}_0 + \beta_1 \vec{x}_0\vec{x}_1 + \cdots + \beta_p \vec{x}_0\vec{x}_p &= \vec{x}_0\vec{y} \\ \beta_0 \vec{x}_1\vec{x}_0 + \beta_1 \vec{x}_1\vec{x}_1 + \cdots + \beta_p \vec{x}_1\vec{x}_p &= \vec{x}_1\vec{y} \\ &\vdots \\ \beta_0 \vec{x}_p\vec{x}_0 + \beta_1 \vec{x}_p\vec{x}_1 + \cdots + \beta_p \vec{x}_p\vec{x}_p &= \vec{x}_p\vec{y} \end{aligned}$$

oder in Matrix-Notation

$$A\,\vec{\beta} = \vec{b} \tag{12}$$

mit

$$A = \begin{pmatrix} \vec{x}_0\vec{x}_0 & \vec{x}_0\vec{x}_1 & \cdots & \vec{x}_0\vec{x}_p \\ \vec{x}_1\vec{x}_0 & \vec{x}_1\vec{x}_1 & \cdots & \vec{x}_1\vec{x}_p \\ \vdots & \vdots & & \vdots \\ \vec{x}_p\vec{x}_0 & \vec{x}_p\vec{x}_1 & \cdots & \vec{x}_p\vec{x}_p \end{pmatrix}, \qquad \vec{b} = \begin{pmatrix} \vec{x}_0\vec{y} \\ \vec{x}_1\vec{y} \\ \vdots \\ \vec{x}_p\vec{y} \end{pmatrix} \tag{13}$$

Das ist exakt dasselbe System, was wir auch schon in `week3.pdf` in den Gleichungen (10) und (11) erhalten hatten. Also bekommen wir auch exakt dieselbe Formel für die Regressionskoeffizienten: Mit der Matrix der Regressoren

$$X := \begin{pmatrix} | & | & & | \\ \vec{x}_0 & \vec{x}_1 & \cdots & \vec{x}_p \\ | & | & & | \end{pmatrix} \in \mathbb{R}^{n \times (p+1)} \tag{14}$$

sind die beta's gegeben durch

$$\vec{\beta} = A^{-1}\vec{b} = (X^T X)^{-1} X^T \vec{y} =: \hat{\beta}_{\mathrm{ML}} \tag{15}$$

$\hat{\beta}_{\mathrm{ML}}$ ist also der Maximum-Likelihood-Schätzer für die Regressionskoeffizienten. Das σ, die Standardabweichung der Fehler, können wir auch schätzen: Wir haben

$$\begin{aligned} \frac{\partial \log L}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{2}{2\sigma^3} \sum_{i=1}^{n} (y_i - \mu_i)^2 \stackrel{!}{=} 0 \\ \Leftrightarrow \qquad \sum_{i=1}^{n} (y_i - \mu_i)^2 &= n\,\sigma^2 \end{aligned}$$

oder

$$\sigma^2 \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i-\mu_i)^2 \;=:\; \widehat{\sigma^2}_{\rm ML} \tag{16}$$

mit

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

oder

$$\vec\mu \;=\; X\,\vec\beta \;=\; X\,(X^TX)^{-1}X^T\,\vec y \;=\; P_X\,\vec y \tag{17}$$

Damit können wir auch schreiben

$$\begin{aligned}
\widehat{\sigma^2}_{\rm ML} &= \tfrac{1}{n}\,(\vec y \,-\, \vec\mu)^2 \\
&= \tfrac{1}{n}\,\big(\vec y \,-\, P_X\vec y\big)^2 \\
&= \tfrac{1}{n}\,\big[\,(Id - P_X)\vec y\big]^2 \\
&= \tfrac{1}{n}\,\big[\,P_{X^\perp}\vec y\,\big]^2
\end{aligned} \tag{18}$$

Fassen wir unsere Ergebnisse in einem Theorem zusammen:

Theorem 7.1: Wir betrachten die Formulierung des linearen Regressionsproblems als statistisches Problem gegeben durch die Gleichungen (3) und (4). Dann ist die Likelihood-Funktion $L = L(\vec\beta,\sigma)$ 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 wird maximiert von den Maximum-Likelihood-Schätzern

$$\hat\beta_{\rm ML} \;=\; (X^TX)^{-1}X^T\vec y \;, \tag{19}$$

$$\widehat{\sigma^2}_{\rm ML} \;=\; \tfrac{1}{n}\,\big[\,P_{X^\perp}\vec y\,\big]^2 \;. \tag{20}$$

Im folgenden wollen wir die statistischen Eigenschaften der Maximum-Likelihood-Schätzer etwas genauer untersuchen. Dazu schauen wir uns zunächst die Erwartungstreue an. Wir wollen also überprüfen, ob die Gleichungen

$$\mathsf{E}\big[\,\hat\beta_{\rm ML}\,\big] \;\stackrel{?}{=}\; \vec\beta \tag{21}$$

$$\mathsf{E}\big[\,\widehat{\sigma^2}_{\rm ML}\,\big] \;\stackrel{?}{=}\; \sigma^2 \tag{22}$$

gültig sind. Dazu müssen wir uns zunächst klarmachen, was das $\mathsf{E}[\,\cdot\,]$ hier genau bedeutet, was ist da zu rechnen?

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

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

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 aus Gleichung (1) ein. Für die $\vec{y} = (y_1, \cdots, y_n)$ können wir entweder die ebenfalls konkret gegebenen Daten aus Gleichung (2) einsetzen, oder wir setzen Daten ein, die von dem stochastischen Modell (3) herkommen, die also durch Zufallszahlen generiert worden sind. Zur Überprüfung der Erwartungtreue machen wir jetzt letzteres, so dass wir etwa schreiben können:

$$\hat{\beta}_{\mathrm{ML}} = \hat{\beta}_{\mathrm{ML}}(X, \vec{y}) = \hat{\beta}_{\mathrm{ML}}\big(X, X\vec{\beta} + \vec{\varepsilon}\big) \tag{23}$$

$$\widehat{\sigma^2}_{\mathrm{ML}} = \widehat{\sigma^2}_{\mathrm{ML}}(X, \vec{y}) = \widehat{\sigma^2}_{\mathrm{ML}}\big(X, X\vec{\beta} + \vec{\varepsilon}\big) \tag{24}$$

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 (21) und (22) meint dann einfach Integration gegen die entsprechenden n Gauss'schen Glockenkurven:

$$\mathsf{E}\big[\,\hat{\beta}_{\mathrm{ML}}\,\big] := \int_{\mathbb{R}^n} \hat{\beta}_{\mathrm{ML}}\big(X, X\vec{\beta} + \vec{\varepsilon}\big) \prod_{k=1}^{n} \Big\{ e^{-\frac{\varepsilon_k^2}{2\sigma^2}} \, \frac{d\varepsilon_k}{\sqrt{2\pi\sigma^2}} \Big\} \tag{25}$$

$$\mathsf{E}\big[\,\widehat{\sigma^2}_{\mathrm{ML}}\,\big] := \int_{\mathbb{R}^n} \widehat{\sigma^2}_{\mathrm{ML}}\big(X, X\vec{\beta} + \vec{\varepsilon}\big) \prod_{k=1}^{n} \Big\{ e^{-\frac{\varepsilon_k^2}{2\sigma^2}} \, \frac{d\varepsilon_k}{\sqrt{2\pi\sigma^2}} \Big\} \tag{26}$$

Man bekommt dann die folgenden Resultate:

Theorem 7.2: a) Der Maximum-Likelihood-Schätzer $\hat{\beta}_{\mathrm{ML}}$ für die Regressionskoeffizienten ist erwartungstreu, es gilt

$$\mathsf{E}\big[\,\hat{\beta}_{\mathrm{ML}}\,\big] = \vec{\beta} \tag{27}$$

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

$$\mathsf{E}\big[\,\widehat{\sigma^2}_{\mathrm{ML}}\,\big] = \tfrac{n-(p+1)}{n}\,\sigma^2 \tag{28}$$

Damit ist der modifizierte Schätzer

$$\widehat{s^2} := \tfrac{n}{n-(p+1)}\,\widehat{\sigma^2}_{\mathrm{ML}} = \tfrac{1}{n-(p+1)}\,\big[\,P_{X^\perp}\vec{y}\,\big]^2 \tag{29}$$

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

Beweis: ..machen wir nächste Woche.