

VL13: Vertrauensintervalle für die Regressionskoeffizienten, Teil2

Wir betrachten dasselbe Setting wie im `week12.pdf` und erinnern an die Schätzer

$$\hat{\beta}_{\text{ML}} \equiv \hat{\beta} = (X^T X)^{-1} X^T \vec{y}$$
$$\hat{s}^2 = \frac{1}{n-(p+1)} [P_{X^\perp} \vec{y}]^2$$

mit der Matrix X der Regressoren

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

Die Aussage des Theorems 12.1 war, dass die Testgrösse

$$T_j := \frac{\hat{\beta}_j - \mathbf{E}[\hat{\beta}_j]}{\sqrt{\hat{\mathbf{V}}[\hat{\beta}_j]}} = \frac{\hat{\beta}_j - \beta_j}{\hat{s} \sqrt{[(X^T X)^{-1}]_{j,j}}}$$

mit

$$\hat{s} := \sqrt{\hat{s}^2}$$

für jedes $j \in \{0, 1, \dots, p\}$ t-verteilt ist, genauer, $t_{n-(p+1)}$ -verteilt ist. Das bedeutet:

$$\text{Prob}[T_j \in (a, b)] = \int_a^b p_{t_{n-(p+1)}}(x) dx$$

mit der Dichte der t-Verteilung

$$p_{t_m}(x) \stackrel{\text{Satz 2.2}}{=} c_m \frac{1}{\left(1 + \frac{x^2}{m}\right)^{\frac{m+1}{2}}}$$

und der Normierungskonstanten

$$c_m = \frac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})\sqrt{\pi m}} .$$

In der R-Software bekommt man das $p_{t_m}(x)$ folgendermassen:

$$p_{t_m}(x) = \text{dt}(\mathbf{x}, \text{df} = \mathbf{m})$$

Dabei steht das erste 'd' an dem dt für 'density', 't' ist der Verteilungsname, die t-Verteilung eben, und das 'df' steht für 'degrees of freedom'.

Geben wir jetzt ein Konfidenzlevel α vor, etwa

$$\alpha = 90\% .$$

Nehmen wir an, wir haben für das Regressionsproblem

$$\begin{aligned} \vec{y} &= \beta_0 \vec{x}_0 + \beta_1 \vec{x}_1 + \cdots + \beta_p \vec{x}_p + \vec{\varepsilon} \\ &= X\vec{\beta} + \vec{\varepsilon} \end{aligned} \quad (1)$$

den Koeffizienten β_j geschätzt mit

$$\hat{\beta}_j = \{ (X^T X)^{-1} X^T \vec{y} \}_j$$

Wir möchten jetzt ein Intervall

$$I_\alpha = [\hat{\beta}_j - \delta\beta_j, \hat{\beta}_j + \delta\beta_j]$$

bestimmen, so dass wir sagen können: In $\alpha = 90\%$ aller Fälle (..welcher Fälle? Antwort: Die \vec{y} 's sollen durch das stochastische Modell (1) mit normalverteilten ε_i 's generiert werden. Aber typischerweise ist das ja gar nicht der Fall, sondern das sind vorgegebene Daten, oder? Richtig...) liegt das tatsächliche β_j in I_α ,

$$\beta_j \in I_\alpha \quad \text{mit W'keit } \alpha = 90\% .$$

Das bekommen wir jetzt folgendermassen: Zunächst mal ist

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{s} \sqrt{[(X^T X)^{-1}]_{j,j}}}$$

äquivalent zu

$$\beta_j = \hat{\beta}_j - T_j \cdot \hat{s} \cdot \sqrt{[(X^T X)^{-1}]_{j,j}}$$

In welchem Bereich befinden sich die T_j 's in $\alpha = 90\%$ aller Fälle? Wir müssen die folgende Gleichung nach x_α auflösen:

$$\int_{-x_\alpha}^{+x_\alpha} p_{n-(p+1)}(x) dx \stackrel{!}{=} \alpha = 90\% \quad (2)$$

Nun ist für jedes m

$$1 = \int_{-\infty}^{+\infty} p_{t_m}(x) dx = \int_{-\infty}^{-x_\alpha} p_{t_m}(x) dx + \int_{-x_\alpha}^{+x_\alpha} p_{t_m}(x) dx + \int_{+x_\alpha}^{+\infty} p_{t_m}(x) dx$$

und, da die Dichte $p_{t_m}(x)$ symmetrisch ist, $p_{t_m}(-x) = p_{t_m}(x)$, ist

$$\int_{+x_\alpha}^{+\infty} p_{t_m}(x) dx = \int_{-\infty}^{-x_\alpha} p_{t_m}(x) dx$$

Also gilt

$$\int_{-x_\alpha}^{+x_\alpha} p_m(x) dx + 2 \int_{-\infty}^{-x_\alpha} p_m(x) dx = 1$$

und Gleichung (2) ist äquivalent zu (jetzt mit $m = n - (p + 1)$)

$$\int_{-x_\alpha}^{+x_\alpha} p_m(x) dx = 1 - 2 \int_{-\infty}^{-x_\alpha} p_m(x) dx \stackrel{!}{=} \alpha = 90\%$$

oder

$$\int_{-\infty}^{-x_\alpha} p_m(x) dx = \frac{1-\alpha}{2} = 5\% \quad (3)$$

Für alle gängigen W'keitsdichten sind die kumulierten Verteilungsfunktionen

$$F(x) := \int_{-\infty}^x p_{t_m}(y) dy$$

in der R-Software vorimplementiert und können mit der Syntax

$$F(x) = \text{pVerteilungsname}(\mathbf{x}, \text{Parameter})$$

$$\stackrel{\text{hier}}{=} \text{pt}(\mathbf{x}, \text{df} = \mathbf{m})$$

aufgerufen werden. Allerdings wollen wir hier ja das x_α oder das $-x_\alpha$ bestimmen, also schreiben wir

$$\begin{aligned} \int_{-\infty}^{-x_\alpha} p_{t_m}(y) dy &= F(-x_\alpha) \stackrel{!}{=} \frac{1-\alpha}{2} = 5\% \\ \Rightarrow -x_\alpha &= F^{-1}\left(\frac{1-\alpha}{2}\right) = F^{-1}(0.05) , \end{aligned}$$

wir brauchen also die Umkehrfunktion F^{-1} von

$$F(x) = \int_{-\infty}^x p_{t_m}(y) dy$$

Für alle gängigen W'keitsverteilungen sind diese Umkehrfunktionen, die heissen auch 'Quantil-Funktionen', ebenfalls in R vorimplementiert und können mit der Syntax

$$\begin{aligned} -x_\alpha &= F^{-1}\left(\frac{1-\alpha}{2}\right) \\ &= \text{qt}\left(\frac{1-\alpha}{2}, \text{df} = \mathbf{m}\right) \end{aligned}$$

aufgerufen werden. Also: Die Bedingung: In $\alpha = 90\%$ aller Fälle ist

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\hat{s} \sqrt{[(X^T X)^{-1}]_{j,j}}} \in (-x_\alpha, +x_\alpha)$$

ist äquivalent zu: In $\alpha = 90\%$ aller Fälle ist

$$\begin{aligned}\beta_j &\in \left(\hat{\beta}_j - x_\alpha \cdot \hat{s} \cdot \sqrt{[(X^T X)^{-1}]_{j,j}} , \hat{\beta}_j + x_\alpha \cdot \hat{s} \cdot \sqrt{[(X^T X)^{-1}]_{j,j}} \right) \\ &=: (\hat{\beta}_j - \delta\beta_j , \hat{\beta}_j + \delta\beta_j)\end{aligned}\tag{4}$$

mit

$$\delta\beta_j = x_\alpha \cdot \hat{s} \cdot \sqrt{[(X^T X)^{-1}]_{j,j}}\tag{5}$$

und x_α gegeben durch

$$-x_\alpha = F^{-1}\left(\frac{1-\alpha}{2}\right) = \text{qt}\left(\frac{1-\alpha}{2}, \text{df} = \mathbf{m}\right) .\tag{6}$$

Das Intervall (4) ist dann das Vertrauensintervall für den Regressionskoeffizienten β_j zum Konfidenzlevel α .