

week12: Vertrauensintervalle für die Regressionskoeffizienten, Teil1

Wir betrachten wieder das Regressionsproblem

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

mit der Matrix X der Regressoren,

$$X = \begin{pmatrix} | & | & \cdots & | \\ \vec{x}_0 & \vec{x}_1 & & \vec{x}_p \\ | & | & & | \end{pmatrix}\tag{2}$$

und $\vec{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)$ normalverteilte, unabhängige Zufallszahlen mit Mittelwert 0 und Standardabweichung σ ,

$$\varepsilon_i \in N(0, \sigma) \quad \text{unabhängig}\tag{3}$$

Wir hatten die folgenden Schätzer für die ‘wirklichen’ β_j ’s und das σ^2 hergeleitet,

$$\hat{\beta}_{\text{ML}} = (X^T X)^{-1} X^T \vec{y}\tag{4}$$

$$\hat{s}^2 = \frac{1}{n-(p+1)} [P_{X^\perp} \vec{y}]^2\tag{5}$$

und hatten gezeigt, dass diese Schätzer erwartungstreu und effizient sind, oder für das \hat{s}^2 , asymptotisch effizient sind. In dem Zusammenhang hatten wir auch die Varianz der $\hat{\beta}_{j,\text{ML}}$ berechnet, sie war gegeben durch

$$\text{V}[\hat{\beta}_{j,\text{ML}}] = \sigma^2 [(X^T X)^{-1}]_{j,j}\tag{6}$$

Den Ausdruck (6) kann man nur dann berechnen, wenn man das ‘wirkliche’ σ kennt, was ja typischerweise nicht der Fall ist. Deshalb schätzen wir das σ^2 in Formel (6) mit Hilfe des Schätzers \hat{s}^2 aus (5) und definieren

$$\hat{\text{V}}[\hat{\beta}_{j,\text{ML}}] := \hat{s}^2(\vec{y}) [(X^T X)^{-1}]_{j,j}\tag{7}$$

Die Grösse (7) lässt sich also vollständig aus den gegebenen Daten $\vec{x}_0, \dots, \vec{x}_p$ und \vec{y} berechnen. Es gilt nun das folgende Theorem, welches die Grundlage für die Berechnung der Vertrauensintervalle darstellt (wir lassen die subscripts ‘ML’ an den $\hat{\beta}$ der Einfachheit halber wieder weg):

Theorem 12.1: Die Testgrösse ($j = 0, 1, \dots, p$)

$$T_j := \frac{\hat{\beta}_j - \mathbf{E}[\hat{\beta}_j]}{\sqrt{\hat{\mathbf{V}}[\hat{\beta}_j]}} = \frac{\hat{\beta}_j - \beta_j}{\hat{s} \sqrt{[(X^T X)^{-1}]_{j,j}}} \quad (8)$$

mit

$$\hat{s} := \sqrt{\hat{s}^2} \quad (9)$$

ist $t_{n-(p+1)}$ -verteilt.

Wir wollen das Theorem mit Hilfe einer geeigneten R-Simulation verifizieren. Dazu betrachten wir dasselbe Setup wie in Aufgabe 2 vom Übungsblatt 9: Wir wählen etwa

$$\vec{x} = (-5, -4, -3, -2, -1, 0, +1, +2, +3, +4, +5) \quad (10)$$

und betrachten das Regressionsmodell

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \quad (11)$$

mit

$$\begin{aligned} \beta_0 &= -6 \\ \beta_1 &= -1 \\ \beta_2 &= 0.5 \end{aligned} \quad (12)$$

und die $(\varepsilon_1, \dots, \varepsilon_{n=11})$ normalverteilte, unabhängige Zufallszahlen mit Mittelwert 0 und Standardabweichung $\sigma = 2$. Ein Zufallsexperiment besteht aus dem Generieren der y_i 's gemäss Gleichung (11), dem Durchführen einer linearen Regression und dem Berechnen der Testgrössen T_0 , T_1 und T_2 . Wir machen dann wieder $N = 10000$ solche Zufallsexperimente und schauen uns dann jeweils die Histogramme für T_0 , T_1 und T_2 an. In diese Histogramme plotten wir dann, etwa wieder in rot, die theoretischen Verteilungen, also die entsprechenden t -Verteilungen, und sollten dann also im wesentlichen eine Übereinstimmung finden.

Anstatt des Vektors \vec{x} aus (10) mit $n = 11$ Komponenten nehmen wir dann vielleicht nochmal das folgende \vec{x} ,

$$\vec{x} = (-5, -3, -1, +1, +3, +5) \quad (13)$$

mit nur $n = 6$ Komponenten, so dass die Testgrössen T_j dann also $t_{6-3} = t_3$ -verteilt sind. Insbesondere für diesen Fall sollte man also eine deutliche Abweichung von der Normalverteilung sehen können.

→ Start R-Session