Hochschule RheinMain
Prof. Dr. D. Lehmann

Verifikation der Formeln für die Vertrauensintervalle durch Simulation

Folgende Schritte: Wir betrachten das Regressionsproblem

$$\begin{aligned} y_i &= 3 - 0.5\,x_i + \varepsilon_i \qquad 1 \le i \le n \\ &=: \beta_0 + \beta_1\,x_i + \varepsilon_i \end{aligned}$$

wobei die ε_i normalverteilte Zufallszahlen sind mit Mittelwert 0 und, sagen wir, mit Standardabweichung $\sigma = 2$, und die x_i gegeben sind durch

$$(x_1, \cdots, x_n) = (-5, -4, -3, -2, -1, 0, +1, +2, +3, +4, +5)$$

so dass also $n = 11$. In Vektor-Notation haben wir wie üblich

$$\vec{y} = \beta_0\,\vec{x}_0 + \beta_1\,\vec{x}_1 + \vec{\varepsilon}$$

mit der Matrix X der Regressoren

$$X = \begin{pmatrix} | & | \\ \vec{x}_0 & \vec{x}_1 \\ | & | \end{pmatrix} = \begin{pmatrix} 1 & -5 \\ 1 & -4 \\ \vdots & \vdots \\ 1 & +5 \end{pmatrix} .$$

In der Vorlesung haben wir gezeigt: Geben wir uns ein Konfidenz-Level α, etwa $\alpha = 90\%$, vor, und schätzen wir die β's wie üblich durch $\hat{\beta}_j$ wobei

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

dann liegt in $\alpha = 90\%$ aller Fälle das tatsächliche β_j in dem Intervall (das ist dann das Vertrauensintervall zum Konfidenz-Level α)

$$\left[\, \hat{\beta}_j - \delta\beta_j \,,\, \hat{\beta}_j + \delta\beta_j \,\right] \tag{1}$$

wobei

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

mit (hier ist $n = 11$ und $p + 1 = 2$ oder $p = 1$)

$$\hat{s} \;=\; \sqrt{\tfrac{1}{n-(p+1)}\,(\vec{y}-\hat{y})^2}\,,$$

mit

$$\hat{y} \;=\; X\hat{\beta} \;=\; \hat{\beta}_0\vec{x}_0 \;+\; \hat{\beta}_1\vec{x}_1$$

der Regression-Fit, und das x_α ist gegeben durch die Gleichung

$$\int_{-\infty}^{-x_\alpha} p_{t_{n-(p+1)}}(x)\,dx \;\stackrel{!}{=}\; \frac{1-\alpha}{2} \;\stackrel{\text{hier}}{=}\; 0.05 \tag{2}$$

wobei (wir schreiben t_m anstatt $t_{n-(p+1)}$)

$$p_{t_m}(x) \;=\; \frac{c_m}{\left(1+\frac{y^2}{m}\right)^{\frac{m+1}{2}}} \qquad\qquad \text{mit} \quad c_m = \tfrac{\Gamma(\frac{m+1}{2})}{\Gamma(\frac{m}{2})\sqrt{\pi m}}$$

die Dichte der t_m-Verteilung ist. In der R-Software sind jetzt die folgenden Funktionen vorimplementiert (schauen Sie sich dazu noch einmal das file `W'keitsverteilungen-in-R.pdf` von der Vorlesungshomepage an):

$$\begin{aligned}
\texttt{dt}(\texttt{x},\texttt{df}=\texttt{m}) \;&:=\; p_{t_m}(x)\\
\texttt{pt}(\texttt{x},\texttt{df}=\texttt{m}) \;&:=\; \textstyle\int_{-\infty}^{x} p_{t_m}(y)\,dy \;=:\; F_{t_m}(x)\\
\texttt{qt}(\texttt{w},\texttt{df}=\texttt{m}) \;&:=\; F_{t_m}^{-1}(w)
\end{aligned}$$

so dass die Gleichung (2) in R also durch

$$-x_\alpha \;=\; \texttt{qt(}\;\;(1-\texttt{alpha})/2\;\;,\;\;\texttt{df}=\texttt{n}-(\texttt{p}+1)\;\;\texttt{)}$$

gelöst werden kann (das `df` steht dabei für "degrees of freedom"). Wir wollen uns davon überzeugen, dass, wenn wir etwa $N = 10000$ Mal eine lineare Regression durchführen und die Vertrauensintervalle gemäss Gleichung (1) bestimmen, dann tatsächlich etwa 9000 Mal die richtigen $\beta_0 = 3$ und $\beta_1 = -0.5$ in diesen Vertrauensintervallen enthalten sind. Gehen Sie dazu folgendermassen vor:

a) Legen Sie die Vektoren $\vec{x}_0$, $\vec{x}_1$ und die Matrix der Regressoren X an.

b) Berechnen Sie die Grössen

$$\begin{aligned}
\texttt{sqrt00} \;&:=\; \sqrt{(X^TX)^{-1}_{0,0}}\\
\texttt{sqrt11} \;&:=\; \sqrt{(X^TX)^{-1}_{1,1}}
\end{aligned}$$

c) Berechnen Sie das x_α zum Konfidenz-Level $\alpha = 90\%$.

d) Legen Sie die Variable $N = 10000$ an und initialisieren Sie dann die Vektoren

$$\begin{aligned}
\texttt{hatbeta0} \;&=\; \texttt{rep}(0,\texttt{N})\\
\texttt{hatbeta1} \;&=\; \texttt{rep}(0,\texttt{N})\\
\texttt{deltabeta0} \;&=\; \texttt{rep}(0,\texttt{N})\\
\texttt{deltabeta1} \;&=\; \texttt{rep}(0,\texttt{N})
\end{aligned}$$

e) Programmieren Sie jetzt eine Schleife, die $N = 10000$ Mal durchlaufen wird. Bei jedem Durchlauf (der Schleifen-Index sei k) sollen

i) neue Zufallszahlen $\vec{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_{11})$ gezogen werden

ii) ein neuer Vektor $\vec{y} = \beta_0\vec{x}_0 + \beta_1\vec{x}_1 + \vec{\varepsilon}$ angelegt werden

iii) eine lineare Regression durchgeführt werden

iv) die so erhaltenen Werte für $\hat{\beta}_0$ und $\hat{\beta}_1$ in die Vektoren `hatbeta0[k]` und `hatbeta1[k]` geschrieben werden

v) die Grössen $\delta\beta_0$ und $\delta\beta_1$ berechnet und in die Vektoren `deltabeta0[k]` und `deltabeta1[k]` geschrieben werden.

f) Geben Sie jetzt etwa folgenden Befehl ein (ohne Zeilenumbruch, Sie müssten dazu vorher das `beta0` angelegt haben; 'res' etwa für result):

```
res_beta0 = ifelse( beta0 > hatbeta0 - deltabeta0 &
                    beta0 < hatbeta0 + deltabeta0 , 1 , 0 )
```

Mit dem Befehl

```
sum(res_beta0)
```

sollten Sie dann eine Zahl in der Nähe von 9000 erhalten. Berechnen Sie diese Zahl ebenfalls für das β_1.

Bemerkung: Der `lm()`-Befehl in der R-Software berechnet ebenfalls die Groesse

$$\sqrt{\widehat{s^2}\,(X^TX)^{-1}_{j,j}}\ .$$

Sie wird dort unter dem Variablennamen `stderr` angegeben, wenn man sich die Resultate einer linearen Regression mit dem `summary` Befehl anzeigen lässt. Diese Zahlen kann man dann also benutzen, wenn man ein Vertrauensintervall angeben soll, man muss diese Zahlen nicht selber von Hand berechnen.