Blatt 11

Für eine lineare Regression ($\hat{y} = aX + b$) ist der Pearsonsche Korrelationskoeffizient definiert als $\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$. Das Bestimmtheitsmaß ist definiert als $R^2 = \rho^2 = \frac{Cov(X,Y)^2}{Var(X)\,Var(Y)}$

Alternativ kann man mithilfe der Variablen

$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} ((Ax_i + B) - \bar{y})^2$ → erklärte Streuung

$SQR = SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (Ax_i + B))^2$ → residuale Streuung

je nach Literatur versch Bezeichnungen

$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2$ → totale Streuung

rechnen:

$R^2 = \frac{SQE}{SQT}$ ⇒ Das Bestimmtheitsmaß ist also der Anteil der durch die Regression beschriebenen Streuung an der Gesamtstreuung ⇒ $0 \leq R^2 \leq 1$; je größer R^2 desto besser erklärt die Regressionsgerade den Zusammenhang von X und Y

Um das Konfidenzintervall von ρ um einen Schätzer r zu berechnen, muss man die Fisher-Transformation $z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$ nutzen, da bekannt ist, dass $z \sim \mathcal{N}\left(\frac{1}{\sqrt{n-3}}, \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)\right)$.
Es müssen also folgende Schritte durchgeführt werden

a) Transformation: $z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right)$

b) Berechnung $(1-\alpha)$-Konfidenzintervall (in Abhängigkeit von z): $\left[\underbrace{z - \frac{z_{1-\alpha/2}}{\sqrt{n-3}}}_{z_1}; \underbrace{z + \frac{z_{1-\alpha/2}}{\sqrt{n-3}}}_{z_2}\right]$ z-Wert der Standardnormalvtlg

c) Retransformation: $r_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$, $r_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$

$\Rightarrow$ Das Konfidenzintervall von ρ zum Niveau $1-\alpha$ ist dann $[r_1; r_2]$

Achtung: das Konfidenzintervall von ρ liegt normalerweise nicht symmetrisch bzgl dem Schätzer r
[d.h. $r-r_1 \neq r_2-r$]

- <u>Test:</u> $H_0: \rho = 0$ $H_1: \rho \neq 0$

Teststatistik: $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

H_0 ablehnen, falls $|T| > t_{1-\alpha/2}(n-2)$

- <u>F-Test</u> Testet, ob das Regressionsmodell einen Erklärungswert für Y hat

Teststatistik: $T = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p}$

p = 'Anzahl Parameter'
[d.h. bei $Y = AX+B \Rightarrow 1$
bei $Y = A_1X_1 + A_2X_2 + B \Rightarrow 2$
...]

Es hat Erklärungswert, falls $T > F_{1-\alpha}(p, n-p-1)$

↑ F-Wert der Fisher-Verteilung

- <u>$(1-\alpha)$-Konfidenzintervall</u> für Y in x_0

Im Gegensatz zum Prognoseintervall, ist das Konfidenzintervall

$$\left[(Ax_0+B) \pm \left(\sqrt{S^2 \cdot \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{Var(x)\cdot n}\right)} \cdot t_{n-2}(1-\alpha/2)\right)\right]$$

In Sage werden die meisten Regressionsberechnungen über R getätigt, daher ist es am einfachsten mittels '%r' auf den R-Interpreter zu wechseln und dort

mittels der lm( ) - Funktion die notwendigen Werte berechnen zu lassen (z.B. die Schätzer A,B der Regression $Y = AX + B$ mit 'coef((lm(y~x))'