Statistics II: Exercise Session 7

1 Regression analysis

1.1 General regression model

$$Y_i = f(x_{1i}, x_{2i}, \ldots, x_{mi}) + U_i = \mathrm{E}[Y_i] + U_i \qquad \text{with } \mathrm{E}[U_i] = 0$$

1.2 Simple linear regression function

True regression line	$\mathrm{E}[Y_i] = \beta_0 + \beta_1 \cdot x_i$
Regression model	$Y_i = \mathrm{E}[Y_i] + U_i = \beta_0 + \beta_1 \cdot x_i + U_i$
Error term	$U_i = Y_i - \mathrm{E}[Y_i]$
	with $\mathrm{E}[U_i] = 0$, $\mathrm{Var}(U_i) = \sigma_u^2$, $\mathrm{Cov}(U_i U_j) = 0$ for $i \neq j$
	and $U_i \sim N(0; \sigma_u)$
Fitted regression line	$\widehat{y}_i = b_0 + b_1 \cdot x_i$
Sample regression line	$y_i = \widehat{y}_i + \widehat{u}_i = b_0 + b_1 \cdot x_i + \widehat{u}_i$
Residuals	$\widehat{u}_i = y_i - \widehat{y}_i$

Least squares estimator for β_0, β_1, σ_u^2

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \left(\sum_{i=1}^{n} x_i\right) \cdot \left(\sum_{i=1}^{n} y_i\right)}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{s_{xy}}{s_x^2} = r_{xy} \cdot \frac{s_y}{s_x}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i \cdot \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} x_i \cdot y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \overline{y} - b_1 \cdot \overline{x}$$

$$s_{\widehat{u}}^2 = \frac{\sum_{i=1}^{n} \widehat{u}_i^2}{n-2}$$

Coefficient of determination

$$R_{yx}^2 = R_{xy}^2 = \frac{s_{yx}^2}{s_y^2 \cdot s_x^2} = r_{yx}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Exercises

9-3: *Cross-sectional analysis of 11 companies*

In a cross-sectional analysis of 11 companies in each sector we examine the dependence of sales Y (in mil. EUR) on investments X_1 (in 1 000 EUR), expenses for research and development X_2 (in 1 000 EUR) and expenses for advertisement X_3 (in 1 000 EUR) for a given period of time. Values of the explanatory variables X_1, X_2 and X_3 and the dependent variable Y are stated in the following table:

i	y_i	x_{i1}	x_{i2}	x_{i3}
1	12.6	117.0	84.5	3.1
2	13.1	126.3	89.7	3.6
3	15.1	134.4	96.2	2.3
4	15.1	137.5	99.1	2.3
5	14.9	141.7	103.2	0.9
6	16.1	149.4	107.5	2.1
7	17.9	158.4	114.1	1.5
8	21.0	166.5	120.4	3.8
9	22.3	177.1	126.8	3.6
10	21.9	179.8	127.2	4.1
11	21.0	183.8	128.7	1.9

a) Determine the simple linear regression functions of sales with respect to investments, expenditures for research and development and advertising expenses respectively, as well as the associated coefficient of determination.

b) Calculate all simple linear correlation coefficients between the given characteristics.

9-5: *Consumption expenditure*

In March 1992 the total available income of 8 households was 30 880 EUR. In the same month the 8 households generated total consumption expenditure amounting to 26 800 EUR. Per EUR of additional income of these households 0.813 EUR on average was spent on consumption.

a) State the (economically meaningful) linear regression function.

b) What consumer spending can be expected on average for a level of disposable income in amount of 2 800 EUR?

9-7: *Additional statistical unit*

Random variables X and Y were observed in 9 statistical units. However, instead of the individual pairs of values (x_i, y_i), observations are the sums

$$\sum_{i=1}^{9} x_i = 34, \quad \sum_{i=1}^{9} y_i = 60, \quad \sum_{i=1}^{9} x_i^2 = 144, \quad \sum_{i=1}^{9} y_i^2 = 422, \quad \sum_{i=1}^{9} x_i y_i = 244$$

Subsequently it turns out that the pair of values $(x_{10}, y_{10}) = (6, 10)$ needs to be taken into account, too. Which of the following regression lines $y = b_0 + b_1 x$ is correct by the method of least squares for the 10 pairs of values?

a) $y = 0.2 + 1.7x$ b) $y = 0.6 + 1.6x$ c) $y = 1.2 + 1.5x$ d) $y = 1.4 + 1.4x$
e) $y = 1.8 + 1.3x$ f) $y = 2.0 + 1.8x$ g) $y = 2.2 + 1.2x$ h) $y = 2.8 + 1.0x$