Review - Linear regression

Estimating the covariance and correlation

The **covariance** of two random variables is defined as

$$\operatorname{cov}(X, Y) = \operatorname{E}((X - \operatorname{E} X)(Y - \operatorname{E} Y)) = \operatorname{E}(XY) - \operatorname{E} X \operatorname{E} Y.$$

and can be estimated based on a random sample of paired observations $(X_1, Y_1), \dots, (X_n, Y_n)$ using the **sample covariance** as

$$s_{X,Y} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i Y_i - n \bar{X}_n \bar{Y}_n \right).$$

The **correlation coefficient** of two random variables gives a measure of their mutual linear dependence is defined as

$$\rho_{X,Y} = \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var} X} \sqrt{\operatorname{var} Y}}.$$

$\rho_{X,Y}$ is always in $[-1, 1]$ and can be estimated using the **sample correlation coefficient** as

$$r_{X,Y} = \frac{s_{X,Y}}{s_X \cdot s_Y}.$$

Linear regression

If we want to model the dependence of Y on x taken as fixed, we can use **linear regression**. We assume that there is a linear dependence of the form

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

where ε_i are independent zero-mean random errors and α and β are parameters which we want to estimate.

Based on independent pairs of observations $(x_1, Y_1), \dots, (x_n, Y_n)$, we estimate the parameters by estimators a and b using the **least squares method**. If we view the explanatory variables $(x_1, \dots, x_n)$ as a realization of a random sample $(X_1, \dots, X_n)$, we obtain

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n \bar{X}_n \bar{Y}_n}{\sum_{i=1}^{n} X_i^2 - n \bar{X}_n^2} = \frac{s_{X,Y}}{s_X^2} = r_{X,Y} \frac{s_Y}{s_X},$$

$$a = \bar{Y}_n - b \cdot \bar{X}_n.$$

If we want to **predict** the value of Y for a certain value of x, we can find the prediction as

$$\hat{Y} = a + b \cdot x.$$

If we want to find x fitting for a certain $Y = y$, we can use the **reverse prediction**

$$\hat{x} = \frac{y - a}{b}.$$

Exercises 12 - Linear regression

1. We study the connection between the bodily weight and height. We have sampled five individuals and measured their heights in centimeters $\mathbf{X} = (158, 161, 168, 175, 182)$ and their weights in kilograms $\mathbf{Y} = (55, 63, 75, 71, 83)$.

a) Estimate the correlation between the weight and height.

b) Suppose there is a linear dependence of weight on height. Estimate the parameters of the regression line.

c) What is the expected weight of a person who is 165 cm tall?

2. We study the distortion of a plastic sheet depending on used pressure. We measured:

x_i	2	4	6	8	1 0	MPa
Y_i	14	35	48	61	80	mm

a) Assume linear dependence. Find the estimates of the parameters of the regression line.

b) What pressure do we need to produce a distortion of 70 mm?

3. In a computer classroom there are 25 computers. We study the total electricity consumption Y_i depending on number of running computers x_i. For measured data $(x_1, Y_1), \dots, (x_{25}, Y_{25})$, we have following statistics available:

$$\bar{X}_n = 12, \qquad \bar{Y}_n = 3800, \qquad s_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = 5000,$$

$$s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2} = 4, \qquad s_Y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2} = 1500.$$

Consider the linear model $Y_i = \alpha + \beta x_i + \varepsilon_i$, for $i = 1\dots, n$ with normally distributed independent errors ε_i.

a) Find the estimates of parameters α and β.

b) Estimate the electricity consumption of 40 running computers.

c) Find the estimate of the correlation coefficient $\rho_{X,Y}$.

4. We study the linear dependence of monthly incomes Y_i (in thousands of CZK) on the length of studies x_i (in years). From 20 records we have computed the following characteristics

$$\sum_{i=1}^{20} X_i = 300, \qquad \sum_{i=1}^{20} Y_i = 480, \qquad \sum_{i=1}^{20} X_i^2 = 5000, \qquad \sum_{i=1}^{20} Y_i^2 = 13520,$$
$$\sum_{i=1}^{20} X_i Y_i = 8000.$$

a) Find the estimates of the coefficients α and β.

b) Estimate the income of a person who has studied for 13 years.

c) Estimate the correlation between the length of studies and monthly incomes.

5. Suppose we observe the following data:

x_i	5.7	13.8	9	0.1	9.5
Y_i	16.2	31.2	24	8	22.3

a) Estimate the regression coefficients of the linear model $Y_i = \alpha + \beta x_i + \varepsilon_i$.

b) Estimate the regression coefficients of the quadratic model $Y_i = \gamma + \delta x_i^2 + \varepsilon_i$.

6. Consider the linear model $Y_i = \alpha + \beta x_i + \varepsilon_i$, for $i = 1, \ldots, n$ with normally distributed errors ε_i. For $n = 100$ observed pairs (x_i, Y_i) we have computed:

$$\bar{X}_n = 10, \qquad s_X^2 = 1.2, \qquad \bar{Y}_n = 158.3, \qquad s_Y^2 = 19.6, \qquad s_{X,Y} = 19.4.$$

Find the estimates of parameters α and β.

7. For the following data

x_i	6	14	9	1	9
Y_i	16	32	24	8	22

consider the linear model: $Y_i = \alpha + \beta x_i + \varepsilon_i$
and the quadratic model: $Y_i = \gamma + \delta x_i^2 + \eta_i$.

a) Find the estimates of the regression parameters for both models.

b) Which of the models fits the data better? (Find the coefficient of determination R^2 for both models.)