TASK LIST NO. 9: Correlation and Regression Analysis

Task 1

The dependence between two features X and Y was investigated based on a 10-element sample (e.g., X – file size in MB, Y – transmission time in sec). Measurement results:

- x_i: $3.5, 3.4, 2.1, 5.4, 1.1, 5.1, 6.9, 4.0, 4.5, 2.5$
- y_i: $1.6, 2.9, 1.5, 3.5, 0.6, 2.5, 7.1, 3.5, 2.1, 2.6$

Calculate the Pearson linear correlation coefficient r. Is the dependence strong?

Hint: Create a helper table with columns $x_i^2, y_i^2, x_i y_i$ and calculate the sums.

Task 2

In a certain network experiment, data regarding the number of errors (X) and system response time (Y) were collected. Since there was a lot of data, ready-made sums were calculated for $n = 25$ measurements:

$$\sum x_i = 375, \quad \sum y_i = 175$$

$$\sum x_i^2 = 6125, \quad \sum y_i^2 = 1245, \quad \sum x_i y_i = 2615$$

Calculate the correlation coefficient r and the means $\bar{x}$ and $\bar{y}$.

Task 3

For a population in which the investigated features (X, Y) have a two-dimensional normal distribution (e.g., processor temperature vs. its clock speed), a sample was taken: $(3,3), (5,3), (6,4), (5,8), (7,5), (8,6), (8,9), (5,4), (6,5)$... (full data in the set).

Determine the equation of the linear regression line $y = ax + b$ of the second kind (i.e., regression of Y with respect to X).

Hint: Use the least squares method.

Task 4

Using the regression line equation determined in Task 3, estimate the predicted value of feature Y (e.g., clock speed) if feature X (temperature) takes the value $x = 10$.

What risk does forecasting for x outside the data range of the sample (extrapolation) carry?

Task 5

For the data from Task 2, where the correlation coefficient was $r \approx -0.1$ (to be verified in calculations), verify at the significance level $\alpha = 0.05$ the hypothesis $H_0 : \rho = 0$ (no correlation in the population) against the hypothesis $H_1 : \rho \neq 0$.

Hint: Use the t-Student statistic:

$$t = \frac{r}{\sqrt{1 - r^2}} \sqrt{n - 2}$$

Task 6

For the regression line determined in Task 3, calculate the residual variance s_r^2 (or the standard deviation of residuals s_r). Interpretation: How much are the actual measurement points "scattered" around the determined regression line?

Formula: $s_r^2 = s_y^2(1 - r^2)$.

Task 7

Two experts evaluated the quality of the interface of 10 applications, ranking them from 1 to 10.

- Expert A: $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$
- Expert B: $2, 1, 4, 3, 6, 5, 8, 7, 10, 9$

Calculate Spearman's rank correlation coefficient. Do the experts agree in their evaluations?

Hint: Use the rank correlation formula based on the differences d_i between ranks.

Task 8

Two regression lines were determined: y with respect to x and x with respect to y.

$$y = -0.6x + 2$$

$$x = -1.2y + 1$$

Calculate the correlation coefficient r based on the slope coefficients of these lines.

Hint: The relationship $r^2 = a_{yx} \cdot a_{xy}$ holds. Remember the sign of the correlation coefficient!

Task 9

Suppose the dependence between algorithm execution time (Y) and data size (X) is exponential: $y = a \cdot e^{bx}$. How can these data be transformed so that known formulas for linear regression can be applied and parameters a and b determined?

Hint: Take the logarithm of both sides of the equation ($\ln y = \ln a + bx$). Then the new dependent variable is $Z = \ln Y$.

Task 10

Three parameters are monitored in a system: X_1 (CPU), X_2 (RAM), X_3 (Disk IO). Pairwise correlations were calculated: $r_{12} = 0.8$, $r_{13} = 0.1$, $r_{23} = 0.2$.

Which variables are strongly correlated with each other, and which are almost independent? What does this mean for the system administrator (e.g., will upgrading RAM affect CPU usage)?