isep Instituto Superior de Engenharia do Porto

Licenciatura em Engenharia Informática – DEI/ISEP

**Análise de Dados em Informática 2022/2023**

**Theoretical-Practical Sheet 4**

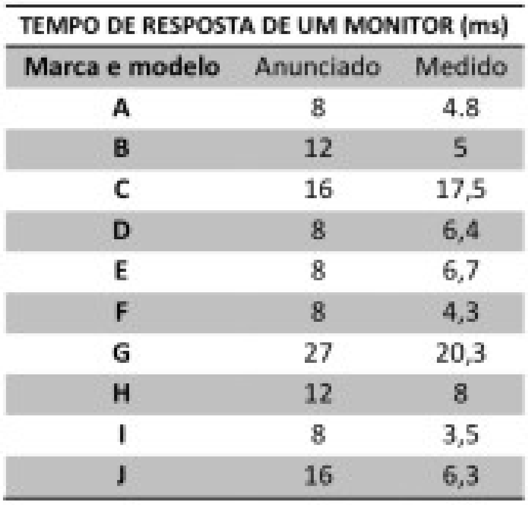
# Correlation and Linear Regression

**Objectives:**

- Becoming familiar with the R tool in addressing problems related to Correlation and Linear Regression;
- Analysis and interpretation of results.

## Practical Exercises

**1.** Consider the following table:



| TEMPO DE RESPOSTA DE UM MONITOR (ms) | | |
|---|---|---|
| Marca e modelo | Anunciado | Medido |
| A | 8 | 4.8 |
| B | 12 | 5 |
| C | 16 | 17,5 |
| D | 8 | 6,4 |
| E | 8 | 6,7 |
| F | 8 | 4,3 |
| G | 27 | 20,3 |
| H | 12 | 8 |
| I | 8 | 3,5 |
| J | 16 | 6,3 |

Use Kendall's rank correlation coefficient to determine the existence of any relationship between two variables (use α = 5%).

**2.** In a specific educational institution, Statistics (*Estatística*) marks are ranked by 6 levels: Excellent *(Excelente)*, Very Good *(Muito Bom)*, Good *(Bom)*, Sufficient *(Suficiente)*, Insufficient *(Insuficiente)* and Fail *(Mau)*. Regarding Calculus (*Cálculo*), marks can be A, B, C, D, E and F, from highest to lowest. In the file "`Notas.txt`", we can find the record of the marks of 25 students. Verify if a mark obtained in Calculus is positively related to a mark obtained in Statistics. Use a level of 5% and Spearman's rank correlation coefficient to analyse the problem.

**3.** A computer engineer, in charge of managing a group of servers, intends to analyse the existence of a correlation between connectivity errors and the daily number of server access requests. Therefore, he created a script to daily record the number of errors and the number of access requests. The values obtained can be found in the file "`Servidor.csv`".

**a)** Create a scatter plot and verify the existence of an association between two variables.

**b)** Calculate an appropriate correlation coefficient and determine, at a level of $5% $, if there is any positive correlation between two variables.

**4.** The file "`fang_data`" contains the daily market summary of technology companies: Facebook, Amazon, Netflix and Google, from 2013 to 2016. Create a correlation matrix considering the exchange rate of the 4 companies and comment on the results.

**5.** Consider the following values from variables X and Y:

| xi | yi |
|---|---|
| 21 | 185,79 |
| 24 | 214,47 |
| 32 | 288,03 |
| 47 | 424,84 |
| 50 | 454,58 |
| 59 | 539,03 |
| 68 | 621,55 |
| 74 | 675,06 |
| 62 | 562,03 |
| 50 | 452,93 |
| 41 | 369,95 |
| 30 | 273,98 |

**a)** Using a scatter plot, test whether a linear relationship exists between two variables.

**b)** Estimate the parameters of the regression line and *y* (40).

**c)** Calculate the Pearson's correlation coefficient and comment on the results.

**d)** Check if the assumptions related to the residuals can be verified.

**6.** The following table presents the life insurance amounts (*capital seguro*) and the annual income *(rendimento anual*), in thousands of currency units (*u.m., unidades monetárias*), of 12 households from a particular country.

| Rendimento anual (milhares de u.m.) | Capital seguro (milhares de u.m.) |
|---|---|
| 14 | 31 |
| 19 | 40 |
| 23 | 49 |
| 12 | 20 |
| 9 | 21 |
| 15 | 34 |
| 22 | 54 |
| 25 | 52 |
| 15 | 28 |
| 10 | 21 |
| 12 | 24 |
| 16 | 34 |

**a)** Create a scatter plot for this data and add a regression line. Comment on the results.
**b)** Estimate the life insurance amount for a household with an annual income of 20000 u.m.
**c)** Check the residuals for homoscedasticity and its independence.
**d)** Test the normality of the residuals.

**7.** A mechanical engineer wants to analyse the surface finish of metal parts made in a lathe and he suspects that it may be related to the speed of the lathe (in revolutions per minute) and to the type of cutting tool being used. The file "**`ExemploMontegomery-12-11.csv`**" contains data from the collected sample:

**a)** Present an appropriate linear regression model and interpret its coefficients.

**b)** Estimate the regression line parameters.

**c)** Calculate the adjusted coefficient of determination and comment on the results.

**d)** Determine whether the assumptions related to the residuals can be verified.

**e)** Verify the presence of multicollinearity.

## Consolidation Exercises

**1.** The following data shows the wheat production, in thousands of tons, from a particular region between 1986 and 1994.

| Ano | Volume de produção |
|---|---|
| 1986 | 285 |
| 1987 | 270 |
| 1988 | 294 |
| 1989 | 279 |
| 1990 | 260 |
| 1991 | 262 |
| 1992 | 258 |
| 1993 | 272 |
| 1994 | 255 |

**a)** Create a scatter plot and add the corresponding regression line. Comment on the results.

**b)** Calculate the coefficient of determination and the coefficient of correlation. Comment on the results.

**c)** Using the Durbin-Watson test, verify the independence of residuals.

**2.** The file "`regressao_exerc9.txt`" contains data from 9 variables related to 517 wildfires that happened in Montesinho Natural Park. Consider the dependent variable *area* while the other 8 variables are independent (*FFMC*, *DMC*, *DC*, *ISI*, *tem*, *RH*, *wind*, *rain*):

**a)** Estimate the multiple linear regression model.

**b)** Determine whether the assumptions related to the residuals can be verified.

**c)** Verify the presence of multicollinearity.

**d)** Find a simpler model with lower AIC (Akaike Information Criterion) value.

**3.** The file "`Covid19.csv`" contains data regarding the total number of covid19 infected individuals in Portugal from March 3rd to March 28th, 2020.

**a)** Create a scatter plot and test whether a linear relationship exists between two variables.

**b)** Execute the change of variables, *Z=log (nº de infetados)*. Create a scatter plot (between variables *Dia* and *Z*) and test whether a linear relationship exists between the two variables.

**c)** Find the regression line, *Z = m.Dia + b*, plot the data and the regression line. Check the normality of residuals (with zero mean) and whether they are independent and homoscedastic. Estimate the number of infected individuals on March 29th and 30th.