

Exercises - Statistics and Experimental Design - 2026-27

NOTICE:

Many of the exercises use data sets made available on the course *web* page, under *Support materials*, subsection *Data*.

1 Linear Regression - descriptive approach

1. Based on data from the Portuguese National Statistics Institute (INE), a CSV (*Comma separated values*) file named `cereais.csv` was created containing the annual evolution of the agricultural area used for the production of cereals for grain (variable `area`, in km^2) in Portugal, from 1986 to 2011 (variable `ano`). The file is available on the course *web* page and should be downloaded and saved in the directory where the R working session is located. Its contents can then be read into the R session using the command:

```
> Cereais <- read.csv("cereais.csv")
```

- (a) Construct a scatter plot of agricultural area *vs.* year and comment on it.
 - (b) Based on the graph obtained in the previous item, suggest a value for the correlation coefficient between agricultural area and year. Then use the R commands to calculate this correlation coefficient. Comment on its meaning.
 - (c) Fit a regression line of agricultural area used on year. Discuss the meaning of the parameters of the fitted line in the context of the problem under study.
 - (d) Comment on the quality of the line obtained by calculating its coefficient of determination and interpreting the value obtained.
 - (e) Plot the fitted regression line over the scatter plot and comment on it.
 - (f) Calculate the Total Sum of Squares (SQT), from the calculation of the sample variance of y .
 - (g) Calculate the value of the Regression Sum of Squares (SQR).
 - (h) Calculate the Residual Sum of Squares (SQRE), directly from the residuals, and numerically verify the fundamental relationship of Linear Regression: $\text{SQT} = \text{SQR} + \text{SQRE}$.
 - (i) Change the units of measurement of the variable `area`, from km^2 to hectares ($\text{area} \rightarrow \text{area} \times 100$). Fit the regression again after making this change. What happened to the estimated parameters and to the coefficient of determination R^2 ? Comment.
 - (j) Again using the original data, transform the variable `ano` into a counter for the years of the study ($\text{ano} \rightarrow \text{ano} - 1985$). Fit the regression again after making this change. What happened to the estimated parameters and to the coefficient of determination R^2 ? Comment.
2. The file `azeite.xls`, available on the course *web* page, is a spreadsheet file containing data on olive oil production in Portugal for the period 1995–2010, provided by the Portuguese National Statistics Institute (www.ine.pt). The columns “Azeitona” and “Azeite” correspond to the production of olives for oil (in t) and olive oil (in hl), respectively.
 - (a) Open the file `azeite.xls` with an *Office*-type programme and save the spreadsheet as a text file named `azeite.txt`, using *Save as* with the *Text File* option. Place this file in the R working directory.
 - (b) In an R session, store the data from the file `azeite.txt` (created in the previous item) in a *data frame* named `azeite`. You can use the command:

```
> azeite <- read.table("azeite.txt", header=TRUE)
```

- (c) Create a scatter plot relating olive oil production (vertical axis, variable y) and olive production (horizontal axis, variable x).
 - (d) Based on the scatter plot, suggest a value for the correlation coefficient between the two variables. Assess your suggestion by calculating the value of r_{xy} . Comment on the value obtained.
 - (e) Calculate the least-squares estimates for the parameters of the regression line, and comment on their meaning.
 - (f) Calculate the precision of the estimated regression line of y on x and comment on the value obtained.
3. The R programme has a large number of additional packages available. One of these additional packages is called MASS. It can be loaded using the command `library(MASS)`.

Consider the data set `Animals`, available in the MASS module, which lists average brain weights (in g) and body weights (in kg) for 28 species of terrestrial animals. The aim is to study a relationship between brain weights (response variable, y) and body weights (predictor variable, x).

- (a) Construct a scatter plot of body weights (horizontal axis) and brain weights (vertical axis). Calculate the corresponding correlation coefficient and comment.
- (b) Construct scatter plots with the following transformations of one or both variables:
 - i. $\ln(y)$ vs. x ;
 - ii. y vs. $\ln(x)$;
 - iii. $\ln(y)$ vs. $\ln(x)$.
- (c) Consider a linear relationship between $\ln(y)$ and $\ln(x)$. State the corresponding underlying relationship between the original variables (not logarithmised). Comment.

In the following items, always consider the *logarithmised data*.

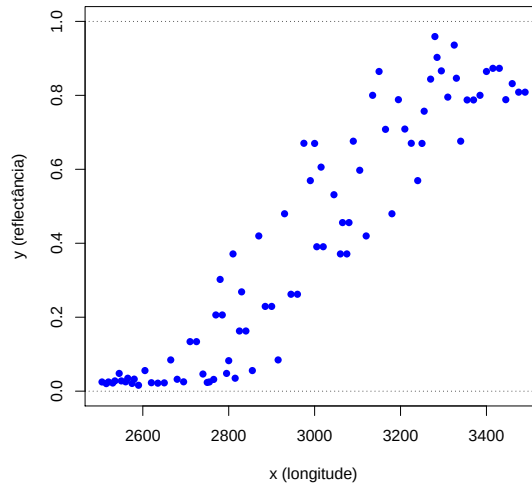
- (d) Calculate the correlation and determination coefficients associated with the relationship between $\ln(x)$ and $\ln(y)$. Interpret the values obtained. How can it be explained that the Coefficient of Determination is not particularly high, when it is clear from the scatter plot that there is a good linear relationship between log-body-weight and log-brain-weight for the majority of species?
 - (e) Fit the regression line of log-brain-weight on log-body-weight (using all observations). Plot this line over the scatter plot and comment.
 - (f) Consider now the estimate of the slope of the line, $b_1 = 0.49599$. What is the biological meaning of this value, both in the relationship between logarithmised variables and in the relationship between the original variables (not logarithmised)?
 - (g) Consider the scatter plot of the logarithmised variables. Identify the three points that stand out in the lower right-hand part of the scatter plot. (NOTE: explore the `identify` command in R). Comment.
 - (h) Fit the regression line of log-brain-weight on log-body-weight considering only the logarithmised data for species that *are not dinosaurs*. Plot this line over the scatter plot and comment. In particular, how can the considerable increase in the coefficient of determination be explained? (NOTE: Use the previous scatter plot, containing all species, to better understand the effect of excluding the three dinosaur species on the fitted line).
4. In a study of pollution in a large city, measurements were taken on 116 days of the amount of ozone in the air (in parts per billion) at 2:00 pm and of the maximum temperature (in $^{\circ}\text{C}$) on the respective day. These observations are contained in a CSV file named `ozono.csv`, available on the course *web* page and which can be downloaded to the working directory of the R session. The contents of this file can then be read into the R session and stored in an object named `ozono`, using the `read.csv` command:

```
> ozono <- read.csv("ozono.csv")
```

- (a) Construct a scatter plot of ozone (vertical axis) *vs.* maximum temperature (horizontal axis).
- (b) Taking into account the curvature observed in the graph, fitting an exponential model of the form $y = ae^{bx}$ has been suggested.
- Construct a scatter plot using the appropriate transformations to assess whether the exponential model is indeed a good option.
 - Fit the *linearised* model using the `lm` command in R. Determine its coefficient of determination and comment.
 - Interpret the parameters of the fitted line directly in terms of the exponential model.
 - Indicate, giving your reasoning, the mean ozone concentration (in parts per billion) estimated by the fitted model for a day on which the maximum temperature is 25°C.
- (c) Consider the original scatter plot again. Plot the exponential curve corresponding to the fit obtained in the previous item.
5. In a study of enzymatic reactions, the “rate” of the reaction in cells treated with puromycin is investigated. For different substrate concentrations (variable *conc*), measured in parts per million (ppm), the number of radioactive emissions per minute was recorded, and the initial rate or “velocity” of the reaction, in counts/minute/minute (variable *taxa*), was calculated from these measurements. The results obtained are given in the following table and are found in the *first two columns* and the *first twelve rows* of the *data frame* Puromycin in R, with the names *conc* and *rate*, respectively (these are the rows corresponding to the *treated* level of the *state* factor):

conc	0.02	0.02	0.06	0.06	0.11	0.11	0.22	0.22	0.56	0.56	1.10	1.10
taxa	76	47	97	107	123	139	159	152	191	201	207	200

- (a) Construct a scatter plot of reaction rates (vertical axis) *vs.* substrate concentrations (horizontal axis).
- (b) Taking into account the curvature observed in the graph, the Michaelis-Menten model is assumed to be suitable for describing the relationship in question, and it is decided to use this model with the following parameterisation (where y represents the *rate* and x the *conc*):
- $$y = \frac{ax}{b+x} \quad (a > 0, b > 0 \text{ and } x > 0).$$
- Show that the model can be linearised, indicating the linearised relationship and the transformations of variables required.
 - Fit the linearised model chosen in the previous item using the `lm` command in R.
 - Estimate the parameters a and b in the original Michaelis-Menten model. How should the estimated value of parameter a be interpreted? Plot the curve corresponding to the fit over the original scatter plot to better understand its meaning. Comment on the result.
6. In the study of images obtained by remote sensing, *reflectance*, which measures the proportion of incident radiation reflected by the surface under study, is important. In order to determine the location of a feature in the terrain that separates a low-reflectance area (values close to 0) from a region of very high reflectance (values close to 1), measurements of position (variable x corresponding to longitude, in metres) and reflectance (variable y , dimensionless) were made for a set of 85 pixels from a remotely sensed image. The observed values are shown in the graph.



The means and variances of the observations of each variable, together with the corresponding correlation coefficient, are:

$$\bar{x} = 2966.882 \quad s_x^2 = 84859.51 \quad \bar{y} = 0.4010485 \quad s_y^2 = 0.1077003 \quad r_{xy} = 0.9326$$

- (a) Fit the linear regression line of reflectance on longitude. What proportion of the total variability of reflectance is explained by this regression?
 - (b) Taking into account the scatter plot and the fact that reflectance values only lie in the interval $]0, 1[$, a logistic relationship between x and y is suggested, with equation $y = \frac{1}{1+e^{-(c+dx)}}$. Show that this relationship can be linearised using the transformation $y^* = \ln\left(\frac{y}{1-y}\right)$, indicating the relationship between the parameters of the resulting line and the parameters of the logistic equation.
 - (c) Now consider the regression line of the linearised model, that is, of y^* on x . The parameters of the fitted model are $b_0 = -20.50$ and $b_1 = 0.006629$. The Residual Mean Square obtained is $QMRE = 0.6081371$. Calculate the coefficient of determination of the linearised model.
7. In the data repository (<http://archive.ics.uci.edu/ml/>) of the University of California, Irvine, there is a data set called “Wine recognition data” (source: Forina, M. et al, *PARVUS - An Extendible Package for Data Exploration, Classification and Correlation*. Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy) containing the results of the chemical analysis of wines from three cultivars from a particular region of Italy. This file is now also available on the course web page in `txt` format under the name `vinhos.txt`. The 14 columns of the data table correspond, respectively, to the variables cultivar (factor V1 with 3 levels, which will be ignored in this Exercise), alcohol content (V2), malic acid content (V3), ash (V4), alkalinity of ash (V5), magnesium content (V6), total phenols index (V7), flavonoid content (V8), non-flavonoid phenols content (V9), proanthocyanidin content (V10), colour intensity (V11), hue (V12), ratio of optical densities at two frequencies, OD280/OD315, (V13), and proline content (V14). There is interest in modelling flavonoid content (variable V8), an antioxidant that is difficult and costly to measure.
- (a) In an R session, store the data from the file `vinhos.txt` in a *data frame* named `vinhos`. Then remove the first column from the data table (a factor indicating the cultivar), replacing the previous *data frame*, using the command `vinhos<-vinhos[, -1]`.

- (b) Run the command `plot(vinhos)` and obtain the correlation matrix between the variables under study. Comment on the results.
 - (c) Fit the regression line of flavonoid content (V8) on alcohol content (V2). Calculate the coefficient of determination and determine the values of the three Sums of Squares associated with this regression.
 - (d) From the correlation matrix between the variables under study, and without performing the corresponding fit, state which is the best simple regression line for predicting flavonoid content (variable V8). For the simple linear regression chosen, determine the coefficient of determination and perform the corresponding decomposition of the total sum of squares.
 - (e) The predictor variable used in the previous item is also not simple to measure, as is the case for variables V9 and V10. It was suggested that a multiple linear regression model be sought for the response variable flavonoid content (V8) that does not use these predictors. A model with five predictor variables was proposed: V4, V5, V11, V12 and V13. Fit this model, and comment on its coefficient of determination, comparing it with the R^2 of the model in the previous item. The R command for fitting this multiple linear regression is:

```
> lm(V8 ~ V4 + V5 + V11 + V12 + V13 , data=vinhos)
```
 - (f) Fit a multiple linear regression of flavonoid content (variable V8) on all the remaining variables with the command `summary(lm(V8 ~ . , data=vinhos))`.
 - i. Use the coefficient of determination obtained with this command to determine the decomposition of the total sum of squares. Comment on the results.
 - ii. Compare the estimated coefficients of the predictor variables with the corresponding coefficients of the predictor variables present in the previous models. Comment.
8. Consider again the study on raspberries described in Exercise 4 of the *Introductory Concepts in Statistics and the R Programme*. Fruits from 14 different plants were analysed with respect to 6 different variables: (i) fruit *diameter* (*Diametro* in *cm*); (ii) its *height* (*Altura* in *cm*); (iii) its *weight* (*Peso* in *g*); (iv) its soluble solids content, *brix* (*Brix* in degrees Brix); (v) its *pH*; (vi) its *sugar* content, excluding sucrose (*Acucar* in *g/100ml*). The data are in the file `brix.txt`.
- (a) The aim is to model *Brix* content from the remaining observed variables. Write the multiple linear regression equation with *Brix* as the response variable and the remaining variables as predictors. How many coefficients does this model have?
 - (b) Determine the values of the model coefficients indicated in the previous item,
 - i. using the R command for fitting this multiple linear regression;
 - ii. obtaining the vector $\vec{\mathbf{b}}$ of coefficient values from its formula, $\vec{\mathbf{b}} = (\mathbf{X}^t \mathbf{X})^{-1} (\mathbf{X}^t \vec{\mathbf{y}})$, where $\vec{\mathbf{y}}$ is the vector of observations of the response variable (use the `model.matrix` command in R to construct the model matrix $\mathbf{X}$).
 - (c) Discuss the biological meaning of the coefficient associated with the variable *Weight* (*Peso*). What are the units of measurement of this coefficient?
 - (d) Discuss the meaning of the intercept b_0 resulting from the fit. Comment.
 - (e) Perform the necessary procedures to calculate the coefficient of determination of the fitted multiple linear regression and compare this coefficient of determination with the coefficients of determination associated with the simple linear regressions (with the same response variable). Comment.
 - (f) It is suggested that the predictor *Weight* (*Peso*) be removed. Write the equation of this multiple linear regression submodel. Without making any calculations, state the possible range of values for the coefficient of determination of this submodel. What can be said about the values of b_0 and the coefficients associated with the predictors of this submodel?