Modelos Matemáticos e Aplicações
Exercises - Linear Model - 2020-21

1 Linear Regression

Warning: The datasets necessary in many exercises may: (i) be available in the base `R` distribution; (ii) be available om the course webpage, in:

Materiais de Apoio $\longrightarrow$ *Módulo II* $\longrightarrow$ *Modelo Linear* $\longrightarrow$ *Dados.*

Some exercises give detailed instructions on how to access the data. In other exercises, the datasets become available when the file `exerML.RData` is loaded into an `R` session[1]: download the file `exerML.RData` into your working directory[2]; then use (if available) the option `Load Workspace` in the `Files` menu[3].

1. Based on data from Portugal's National Statistics Board (Instituto Nacional de Estatística, INE) a file in the CSV (*Comma separated values*) format was created, called `Cereais.csv`, containing the evolution of the land surface that is anually used in Portugal for the production of grain cereals (variable `area`, in km^2), in the period from 1986 to 2011 (variable `ano`). The file `Cereais.csv` is on the course webpage. Download the file `Cereais.csv` to your working directory. Save the data in a *data frame* called `Cereais`, through the following command:

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

   (a) Draw and discuss the scatterplot of agricutural land area *vs.* year.

   (b) Based on the above scatterplot, suggest a value for the coefficient of linear correlation between land area and year. Use `R` commands to calculate that coefficient of correlation and discuss its value.

   (c) Fit a regression line of agricultural area over years. Discuss the meaning of the parameters in the fitted line, in the context of the problem being considered.

   (d) Comment the quality of the fitted line. Calculate its coefficient of determination and interpret its value.

   (e) Draw the fitted regression line over the scatterplot. Comment the result.

   (f) Calculate the Total Sum of Squares (SQT ou, with English initials, SST), based on the value of y's sample variance.

   (g) Calculate the Regression Sum of Squares (SQR, or SSR in English).

   (h) Calculate the Residual Sum of Squares (SQRE or SSE), directly from the residuals and check numerically the fundamental relation of a linear regression: SQT=SQR+SQRE.

   (i) Change the units of measurment of variable `area`, from km^2 to hectares (`area` $\rightarrow$ `area`$\times$100). Fit the regression again after thos transformation. What happened to the fitted parameters and to the coefficient of determination R^2? Comment.

   (j) Using the original data once again, transform variable `ano` in a counter of the years under study (`ano` $\rightarrow$ `ano`$-$1985). Again fit the regression, after this transformation. What happened to the fitted parameters and to the coefficient of determination R^2? Comment.

[1] The extension `.RData` indicates that this file was created within an `R` session, using the `save` command.
[2] The working directory of an `R` session may be identified with the command `getwd()`.
[3] Alternatively, you may give, within an `R` session in that directory, the command `load("exerML.RData")`.

2. The file `Azeite.xls` is available on the course webpage. It is a spreadsheet, of the type that can be opened with office applications such as LibreOffice. The spreadsheet has data relative to the production of olive oil in Portugal in the period 1995-2010, from the National Statistics Board (Instituto Nacional de Estatística, www.ine.pt). The columns `Azeitona` and `Azeite` give the production of olives (for olive oil), em t, and of olive oil, em hl, respectively.

   The contents of the file `Azeite.xls` should be read into an `R` session, using the add-on package `xslx`. In what follows, it is assumed that this module has already been installed[4].

   (a) Download the file `Azeite.xls` into your working directory. Afterwards, and from within an `R` session, save its content into a *data frame* called `azeite`, using the following commands:

   ```
   > library(xlsx)
   > azeite <- read.xlsx("Azeite.xls", sheetIndex=1, header=TRUE)
   ```

   Note: The first command loads package `xlsx` into memory. This must be done at the beginning of each `R` session. The second command reads the *first sheet* in file `Azeite.xls` (as indicated by the argument `sheetIndex=1`), and will interret the first row on that sheet as giving the column names (as specified by the argument `header=TRUE`). The command works because the sheet in file `Azeite.xls` does not contain other things (such as graphics). Additional arguments of the command allower for a finer control (see `help(read.xlsx)`)[5].

   (b) Creat the scatterplot relating the production of olive oil (`Azeite`, vertical axis, variable y) with that of olives (`Azeitona`, horizontal axis, variable x).

   (c) Based on the scatterplot, suggest a value for the linear correlation coefficient between both variables. Check on your guess by calculating the value of r_{xy}. Discuss the value you obtained.

   (d) Calculate the least squares estimates of the parameters in the regression line, and comment their significance.

   (e) Calculate the precision of the fitted regression line of y over x and discuss the value you calculated.

3. Show that, for any set of n values, $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, with means $\overline{x}$ and $\overline{y}$, respectively:

   (a) $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$.

   (b) $(n-1)\text{cov}_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n} x_i\, y_i - n\overline{xy}$.

4. Show that, in a simple linear regression, based on n pairs of observations $\{(x_i, y_i)\}_{i=1}^n$:

   (a) the mean of the observed values of y is equal to the mean of the fitted values of y.

   (b) the mean of the residuals ($e_i = y_i - \hat{y}_i$) is zero.

   (c) the slope of the regression line of y over x can be written in termos of the standard deviations of each variable and their correlation coefficient, and is given by: $b_1 = r_{xy} \cdot \frac{s_y}{s_x}$.

   (d) the coefficient of determination R^2 is equal to the squared correlation coefficient between the predictor variable x and the response variable y.

[4] Add-on packages must be previously installed on your computer. This module can be installed using the command `install.packages("xlsx")`. This operation need only be done once on each plataform. Do not confuse the *installation* of a package with *loading into memory*. The latter operation is carried out with the command `library(xlsx)` and must be repeated at the beginning of each new `R` session.

[5] Alternatively, you can open the file `Azeite.xls` and save the spreadsheet in a common text file (with *Save as* using option *Ficheiro de Texto - Text File*), with the name `Azeite.txt`. Place that file in your `R` working directory. From within an `R` session, read the contents of the file `Azeite.txt` into a *data frame* called `azeite`, with the command:
`azeite <- read.table("Azeite.txt", header=TRUE)`.

(e) the squared correlation coefficient between the n observed values y_i and the n corresponding fitted values, $\hat{y}_i$, is also equal to the coefficient of determination: $(r_{y\hat{y}})^2 = R^2$.

5. Consider the equation of a straight line without an additive constant (intercept zero) (the so-called "line forced to the origin"): $y = b\,x$.

   (a) Determine the least squares estimator for the parameter b, based on n pairs of observations $\{(x_i, y_i)\}_{i=1}^n$. Relate your result to the expression for the slope of the standard regression line.

   (b) Consider the data in the *data frame* `iris` (available on `R`), with morphometric measurements on 150 iris flowers. Consider as a response variable (`Petal.Width`) and as a predictor (`Petal.Length`). Fit, with the help of the `R` software, the line forced to the origin. **Aviso:** Use the command `lm`, with the formula corresponding to this model:

   `Petal.Width ~ -1 + Petal.Length`.

   Show that several of the properties of the usual (free) regression line are no longer valid, namely:

   i. the sum of residuals $e_i = y_i - \hat{y}_i = y_i - b\,x_i$ is not zero;

   ii. the sum of squared residuals is not equal to $(n-1)\,s_e^2$, where s_e^2 denotes the variance of the residuals;

   iii. with the standard definitions of the three sums of squares, it is now the case that $SQT \neq SQR + SQRE$.

   Comment. Explain the reasons for these differences between the standard least squares regression and the regression line forced to the origin.

6. There is a large number of additional packages for `R`, among which package `MASS`. It can be loaded[6] into a working session with the command `library(MASS)`.

   Consider the dataset `Animals`, available in package `MASS`, which gives mean brain weights of brains (in g) and bodies (in kg) of 28 animal species. We seek to study the relation between brain weight (response variable y) and body weight (predictor x).

   (a) Draw the scatterplot of body weight (horizontal axis) and brain weight (vertical axis). Calculate the corresponding correlation coefficient and comment.

   (b) Draw the scatterplot of the (natural) *logarithms* of body and brain weights. Calculate the coefficients of linear correlation and of determination for the relation between $\ln(x)$ and $\ln(y)$. Interpret and comment the values obtained.

   (c) Consider a linear relation of $\ln(y)$ over $\ln(x)$. Deduce the underlying trend between the original (non log-transformed) variables. Comment.

   In what follows, always consider the *log-transformed data*.

   (d) Fit the regression line of brain log-weight over body log-weight, using all the observations. Draw that line on the scatterplot and comment.

   (e) Consider the estimate of the line's slope, $b_1 = 0.49599$. What is the biological meaning of this value, both in termos of the log-transformed variables and in termos of the original (non log-transformed) variables?

   (f) Consider the scatterplot for the log-transformed data. Identify the three points that stand out on the right-hand side of the scatterplot. (**Warning:** explore R's `identify` command). Comment.

   In the following questions, consider only the (log-transformed) data of species that *are not dinosaurs*.

[6] Package `MASS` is usually installed when a standard distribution of `R` is installed.

(g) Fit the regression line of brain log-weight over body log-weight. Draw that line on the scatterplot and comment.

(h) Consider the estimate for the slope of the new regression line fitted after excluding the three dinosaur species, $b_1 = 0.75226$. What is the biological meaning of this values, both in terms of the relation between the log-transformed variables, and in terms of the relation between the original (non-transformed) variables?

7. A study on pollution in a big city collected measurements, on 116 days, of the levels of ozone in the air (in parts per thousand millions) at 14h00 and of the maximum temperature (in °C) on the same day. The observations are in the file `ozono.csv` (in `csv` format), available on the course webpage. After downloading the file to your R session's working directory, read the file contents into your session command `read.csv`:

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

(a) Draw the scatterplot of ozone levels (vertical axis) *vs.* maximum temperature (horizontal axis).

(b) Taking into account the curvature observed in the plot, it was suggested that an exponential model, with equation $y = a\,e^{b\,x}$, should be fitted.

i. Draw the scatterplot with suitable transformations of the variables in order to check whether an exponential model is an appropriate choice.

ii. Fit the *linearized* model, using the `R`'s `lm` command. Calculate the coefficient of determination and comment.

iii. Interpret the parameters of the fitted line, in terms of the original exponential equation.

iv. Indicate, justifying, what is the mean ozone level (in parts per thousand millions) estimated by the fitted model, for a day in which the maximum temperature is 25°C.

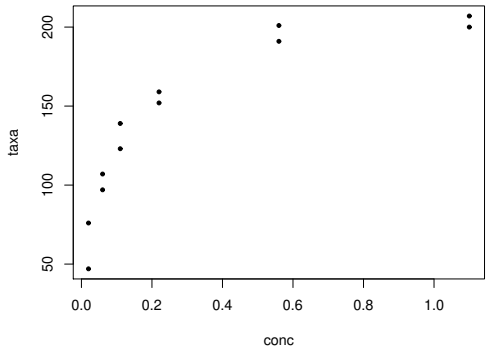
(c) Considere once again the original scatterplot. Draw the exponentail curve that results from the model fitted in the previous question.

8. In a study of enzyme kinematics, the goal was to analyze the rate of reaction in cells treated with puromycin. For different substrate concentrations (variable *conc*), measured in parts per million (ppm), the number of radioactive emissions per minute was registered, and based on those values an initial rate, or "speed" of the reaction was calculated, in counts/minute/minute (variable *taxa*). The results obtained are given in the following table and can be found *in the first two columns* and *first twelve rows* of the *data frame* `Puromycin`, available in the standard distributions of `R`, with column names `conc` and `rate`, respectively:

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

The relation between the rate of reaction and the substrate concentrations is shown in the plot to the right. Assume that a Michaelis-Menten suitably describes the relation, using the following parametrization of this model, where y represents the rate (*taxa*) and x the substrate concentration (*conc*),

$$y = \frac{ax}{b+x} \qquad (a > 0, b > 0 \quad \text{e} \quad x > 0)\ .$$



(a) Show that the above model can be linearized, indicating variable transformations that are appropriate and the linearized relation that results.

(b) Fit the linearized model that you chose above, using the `R` command `lm`.

(c) Estimate the parameters a and b in the original Michaelis-Menten model. How do you interpret the estimated value of parameter a? Draw the resulting Michaelis-Menten curve on the scatterplot in the original scales. Comment.

9. A study of raspberries carried out by the Horticulture Section of ISA analysed the fruits of 14 plants, in terms of 6 different variables: (i) the fruits' diameter (`Diametro`, cm); (ii) their height (`Altura`, cm); (iii) weight (`Peso`, g); (iv) content in soluble solids (`Brix`, degrees Brix); (v) `pH`; (vi) sugar content, excepting sucrose (`Acucar`, $g/100ml$). The data are in the *data frame* `brix`, available from file `exerML.RData`. The mean values of each variable, for the raspberries from each plant, are:

	Diametro	Altura	Peso	Brix	pH	Acucar
1	2.0	2.1	3.71	8.4	2.78	5.12
2	2.1	2.0	3.79	8.4	2.84	5.40
3	2.0	1.7	3.65	8.7	2.89	5.38
4	2.0	1.8	3.83	8.6	2.91	5.23
5	1.8	1.8	3.95	8.0	2.84	3.44
6	2.0	1.9	4.18	8.2	3.00	3.42
7	2.1	2.2	4.37	8.1	3.00	3.48
8	1.8	1.9	3.97	8.0	2.96	3.34
9	1.8	1.8	3.43	8.2	2.75	2.02
10	1.9	1.9	3.78	8.0	2.75	2.14
11	1.9	1.9	3.42	8.0	2.73	2.06
12	2.0	1.9	3.60	8.1	2.71	2.02
13	1.9	1.7	2.87	8.4	2.94	3.86
14	2.1	1.9	3.74	8.8	3.20	3.89

(a) Draw the scatterplots for each pair of variables with the command `plot(brix)`. Calculate the coefficients of linear correlation for each plot. Comment.

(b) We seek to model *Brix* contents based on the remaining variables. Write the multiple linear regression model equation, with *Brix* as the response variable and all other variable as predictors. How many parameters does this model have?

(c) Determine the estimated values of the model parameters, using the command `lm`.

(d) Use the `R` command `model.matrix` to obtain the model matrix $\mathbf{X}$. Using this matrix, compute the vector $\vec{\mathbf{b}}$ of the fitted parameter values, using the 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.

10. Consider a simple linear regression of a variable y on a variable x, based on n pairs of observations $\{(x_i, y_i)\}_{i=1}^n$. Consider the notation used in class (in which $\mathbf{X}$ is now a two-column matrix: a column of n ones and a column with the n values x_i of the predictor variable X; and $\vec{\mathbf{y}}$ denotes a vector with the n values of variable y). Show that:

(a) $$\mathbf{X}^t\vec{\mathbf{y}} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = \begin{bmatrix} n\overline{y} \\ \\ \sum_{i=1}^n x_i y_i \end{bmatrix} = \begin{bmatrix} n\overline{y} \\ \\ (n-1)\, cov_{xy} + n\overline{x}\overline{y} \end{bmatrix} .$$

(b) $$\mathbf{X}^t\mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix} = \begin{bmatrix} n & n\,\overline{x} \\ \\ n\,\overline{x} & (n-1)\, s_x^2 + n\,\overline{x}^2 \end{bmatrix} .$$

(c) $(\mathbf{X}^t\mathbf{X})^{-1} = \frac{1}{n(n-1)\,s_x^2}\begin{bmatrix} (n-1)\,s_x^2 + n\overline{x}^2 & -n\,\overline{x} \\ -n\,\overline{x} & n \end{bmatrix}$.

(d) Deduce from $\vec{\mathbf{b}} = (\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\vec{\mathbf{y}})$, the formulas for b_0 and b_1 in a Simple Linear Regression.

NOTE: Take into consideration that:

$$\begin{aligned} {}_{(n-1)}\,cov_{xy} &= \sum_i (x_i - \overline{x})(y_i - \overline{y}) = \sum_i x_i y_i - n\overline{x}\,\overline{y}\ ; \\ {}_{(n-1)}\,s_x^2 &= \sum_i (x_i - \overline{x})^2 = \sum_i x_i^2 - n\overline{x}^2\ . \end{aligned}$$

11. (a) Show, from its definition, that the matrix of orthogonal projections $\mathbf{H}$ in a multiple linear regression is idempotent ($\mathbf{H}\,\mathbf{H} = \mathbf{H}$) and symmetric ($\mathbf{H}^t = \mathbf{H}$).

    (b) Knowing that an vector that belongs to the column-space of matrix $\mathbf{X}$, that is, the subspace $\mathcal{C}(\mathbf{X})$ in a multiple linear regression model, can be written as a product $\mathbf{X}\vec{\mathbf{a}}$, for some vector of coefficients $\vec{\mathbf{a}}$, show that the vectors that belong to $\mathcal{C}(\mathbf{X})$ remain invariant when projected onto that same subspace, in other words, show that $\mathbf{H}\mathbf{X}\vec{\mathbf{a}} = \mathbf{X}\vec{\mathbf{a}}$.

    (c) Starting from the expression for the fitted values of Y, $\hat{\vec{\mathbf{y}}} = \mathbf{H}\vec{\mathbf{y}}$, show that the mean of the fitted values of Y, $\{\hat{y}_i\}_{i=1}^n$, is equal to the mean of the observed values, $\{y_i\}_{i=1}^n$.

    (d) Show that the sum of the residuals, in any linear regression, must be zero.

12. Consider the vector $\vec{\mathbf{1}}_n \in \mathbb{R}^n$, of n ones. Consider any other vector $\vec{\mathbf{x}} = (x_1, x_2, ..., x_n)^t$ in $\mathbb{R}^n$, which we assume is a vector of n observations of some variable X.

    (a) Build the matrix $\mathbf{P} = \vec{\mathbf{1}}_n(\vec{\mathbf{1}}_n^t\vec{\mathbf{1}}_n)^{-1}\vec{\mathbf{1}}_n^t$ of orthogonal projections onto the subspace $\mathcal{C}(\mathbf{1}_n) \subset \mathbb{R}^n$ which is spanned by the single vector $\vec{\mathbf{1}}_n$ (i.e., $\mathcal{C}(\vec{\mathbf{1}}_n)$ is the set of vectors that are scalar multiples of $\vec{\mathbf{1}}_n$).

    (b) Identify the elements of the vector $\mathbf{P}\vec{\mathbf{x}}$ which results from the orthogonal projection of vector $\vec{\mathbf{x}}$ onto the subspace $\mathcal{C}(\vec{\mathbf{1}}_n)$, and comment.

    (c) Show that the *centred* variable $\mathbf{x}^c$, whose generic element is $x_i - \overline{x}$, can be written as $\vec{\mathbf{x}} - \mathbf{P}\vec{\mathbf{x}} = (\mathbf{I} - \mathbf{P})\vec{\mathbf{x}}$, where $\mathbf{I}$ denotes the $n \times n$ identity matrix.

    (d) Show that the *standard deviation* of the n observations of variable X is proportional to the norm (length) of vector $\mathbf{x}^c$, defined in the previous question.

    (e) Graphically represent the situation described above. Show that a right triangle in $\mathbb{R}^n$ was defined. Apply the Pythagorean Theorem to this triangle and comment.

13. In a (simple or multiple) linear regression, we have:

$$\begin{aligned} SQT &= \|\vec{\mathbf{y}} - \mathbf{P}_{\vec{\mathbf{1}}_n}\vec{\mathbf{y}}\|^2 \\ SQR &= \|\mathbf{H}\vec{\mathbf{y}} - \mathbf{P}_{\vec{\mathbf{1}}_n}\vec{\mathbf{y}}\|^2 \\ SQRE &= \|\vec{\mathbf{y}} - \mathbf{H}\vec{\mathbf{y}}\|^2 \end{aligned}$$

    where $\vec{\mathbf{y}}$ denotes the vector of observations of the response variable, $\mathbf{H}$ is the 'hat' matrix of orthogonal projections onto the subespaço $\mathcal{C}(\mathbf{X})$ spanned by the columns of the model matrix $\mathbf{X}$ and $\mathbf{P}_{\vec{\mathbf{1}}_n}$ is the matrix of orthogonal projections onto the subspace $\mathcal{C}(\vec{\mathbf{1}}_n)$ spanned by the vector of n ones, $\vec{\mathbf{1}}_n$. Show, algebraically, that $SQT = SQR + SQRE$.

In the following Exercises, of an inferential nature, assume that the Linear Model is valid.

14. Assume that the iris flowers behind the morphometric dataset (*data frame* `iris`) are a random sample from a vaster population. Consider, in particular, the relation between petal width (`Petal.Width`, variable y) and petal length (`Petal.Length`, variable x), both in cm. Answer the following questions.

    (a) Obtain estimates of the variances and standard deviations of parameter estimators β_0 and β_1.

    (b) Calculate a 95% confidence interval for the slope β_1 of the population line.

    (c) Calculate a 95% confidence interval for the intercept β_0 of the population line.

    (d) Use a hypothesis test to validate the following statement: "for each additional centimeter in petal length, petal width grows, on average, $0.5cm$".

    (e) Use a hypothesis test to validate the following statement: "for each additional centimeter in petal length, petal width grows, on average, less than $0.5cm$".

    (f) Use a hypothesis test on the slope of the population regression line β_1 to validate the following statement: "there is no significant linear relation between petal length and width, in the iris flowers".

    (g) Validate the previous statement once again, but now using a model goodness-of-fit test (F test).

    (h) Predict the expected value of petal width for iris flowers whose petal length is 4.5cm. Calculate a confidence interval for this expected value.

    (i) Calculate a prediction interval (95%) associated with the width of a petal whose length is 4.5cm. Compare this interval with the confidence interval obtained in the previous question and comment.

    (j) Study the residual plots to detect any possible problems with the model assumptions. Commente your conclusions.

    (k) Check the effects on the fitted model parameters and on the coefficient of determination for each of the following transformations of the data. Comment.

        i. petal lengths are given in millimeters ($x \to 10 \times x$), but widths (y) in centimeters.

        ii. petal widths are given in millimeters ($y \to 10 \times y$), but lengths (x) in centimeters.

        iii. both petal widths and lengths are given in millimeters ($x \to 10 \times x$ and $y \to 10 \times y$).

15. Let $\vec{\mathbf{Z}}_{k\times 1}$ be a random vector. Prove the following properties:

    (a) $E[\alpha\vec{\mathbf{Z}}] = \alpha E[\vec{\mathbf{Z}}]$, with α a (non-random) scalar.

    (b) $E[\vec{\mathbf{Z}} + \vec{\mathbf{a}}] = E[\vec{\mathbf{Z}}] + \vec{\mathbf{a}}$, with $\vec{\mathbf{a}}$ a non-random vector.

    (c) $V[\alpha\vec{\mathbf{Z}}] = \alpha^2 V[\vec{\mathbf{Z}}]$, with α a (non-random) scalar.

    (d) $V[\vec{\mathbf{Z}} + \vec{\mathbf{a}}] = V[\vec{\mathbf{Z}}]$, with $\vec{\mathbf{a}}$ a non-random vector.

    (e) Consider a second random vector $\vec{\mathbf{U}}_{k\times 1}$. Show that $E[\vec{\mathbf{Z}} + \vec{\mathbf{U}}] = E[\vec{\mathbf{Z}}] + E[\vec{\mathbf{U}}]$.

16. The F statistic for the goodness-of-fit test is $F = \frac{QMR}{QMRE}$. The Coefficient of Determination is $R^2 = \frac{SQR}{SQT}$. Taking into account the properties of the Sums of Squares,

    (a) Show that the F statistic can also be written as:

$$F = \frac{n-(p+1)}{p} \cdot \frac{R^2}{1-R^2}$$

    (b) Confirm, based on the expression above, that the F statistic is (for constant n) an *increasing function of the Coefficient of Determination*. Interpret this fact, in terms of the significance of R^2 and the nature of this goodness-of-fit test.

17. The H.J. Andrews Experimental Forest, in the US State of Oregon, makes available a number of forestry datasets (https://andrewsforest.oregonstate.edu/data). One of these datasets has measurements of nutrient concentrations of vegetation in small watersheds, and is called `TN025`. The data consists of 117 measurements of the concentration of several nutrients[7]. The file `TN025.csv` contains the dataset and is available on the course webpage.

    (a) Load the dataset into an `R` session. Inspect the nature of the data frame's X26 columns. The observed nutrient concentrations are in columns 12 to 25, the nutrients being identified by their chemical symbols (Note: the column for sodium concentration is called "`NA.`", with a final dot, so as to avoid confusion with the symbol `NA` that is used by `R` to denote missing values). These concentrations are all in $mg\ kg^{-1}$, except Nitrogen (`N`) and Carbon (`C`), which are given as percentages. Column 9, `TYPE`, indicates the type of material on which the observations were made.

    (b) Based on the correlation matrix between different concentrations of nutrients, choose the best linear predictor of the concentration of phosphorus (variable `P`).

        i. Fit the regression line of `P` over your chosen predictor. Discuss the quality of the fit, based on the output of the `lm` command.

        ii. Draw the 117-point scatterplot and draw the regression line on top. Comment your result. In particular, identify the observations associated with the column of points that appears on the left side of the plot. How many observations are there in this column? Comment.

        iii. Draw the plots of residuals and other diagnostics of the fitted regression. Comment, taking also into account your reply to the previous question. Identify the point with a very large Cook's distance and discuss it.

    (c) Again with the response variable phosphorus, consider now the predictor potassium (`K`). Draw the corresponding scatterplot and comment it. Fit the regression line and discuss it.

    (d) Consider a simple linear regression of the log-transformation of `P` over the log-transformation of `K`.

        i. Draw the corresponding scatterplot and comment.

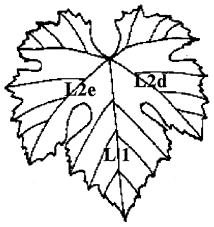
        ii. Fit the linear regression and comment the quality of fit, based on the results produced by the command `lm`. In particular, say whether the value of the coefficient of determination obtained is comparable with the value obtained in question 17c).

        iii. Inspect the plots of residuals and other diagnostics. Comment.

        iv. Deduce the curve that corresponds to the regression line fitted with the linearized model, when you revert back to the units of measurement of the original variables (`K` and `P`). Draw that curve on the scatterplot obtained in question 17c). Comment your result, and draw lessons of general interest.

18. The rigorous measurement of the surface area of leaves involves techniques that require that the leaves be plucked from the plants. We seek to estimate the surface area (variable **Area**) of vineleaves of different varieties, using predictors that can be measured without destroying the leaves. Specifically, we seek to predict surface area of vineleaves based on three length measurements on the leaves:

    - the length of the main vein **(NP)**;
    - the length of the left lateral vein **(NLesq)**; and
    - the length of the right lateral vein **(NLdir)**.



[7] More details regarding the dataset and its collection can be found on the website of the Experimental Forest. Follow the pointers *Data Catalogue* and then do a *Text Search* using the dataset's name.

Three different grape varieties (factor `Castas`) were observed: Fernão Pires, Vital and Água Santa, but with the purpose of obatining a single model that can be applied to any variety. ISA's Horticulture Section collected 200 leaves from each variety, and for each leaf, rigorous measurements of each predictor (in cm), as well as of the response variable (in cm^2), were collected. The resulting data are in the data frame `videiras` (in the `exerRL.RData` file on the webpage). The first 6 rows of the data frame are:

```
          Casta NLesq   NP NLdir Area
1  Fernao Pires  11.4 13.8  10.7  200
2  Fernao Pires   8.8  9.1   9.4  126
3  Fernao Pires  13.2 14.5  13.0  274
4  Fernao Pires  11.7 13.8  10.7  198
5  Fernao Pires   9.7 12.0  10.6  160
6  Fernao Pires  12.0 11.5  11.6  236
```

(a) Draw the scatterplots for each pair of observed variables. Discuss the result.

(b) Calculate the matrix of correlations bewteen all pairs of the 4 observed variables. Comment.

(c) Describe the Multiple Linear Regression Model for this problem.

(d) Fit the multiple regression described above and comment. In particular, test the model's goodness-of-fit.

(e) Assuming that the model is valid, test with a $\alpha = 0.01$ significance level, the hypothesis that for each additional centimeter in the main vein (and keeping the lateral veins constant) there is a mean increase in the leaves' surface area of 7 cm^2. Repeat the test, but now using a $\alpha = 0.05$ significance level. Comment.

(f) Is it admissible to state that coefficients of the two lateral vein lengths are equal? Provide a formal justification.

(g) The vein lengths of three new leaves were measured, on the vines. The results obtained were:

Leaf no.	NP	NLesq	NLdir
1	12.1	11.6	11.9
2	10.6	10.1	9.9
3	15.1	14.9	14.0

For each new leaf, calculate:

i. the estimated value of the leaf's surface area;

ii. a 95% confidence interval for the expected value of the leaf area associated with the set of values of the predictors;

iii. a 95% prediction interval for the surface area of each indvidual leaf.

(h) Study the residuals of the fitted model and comment.

(i) Fit a similar multiple linear regression, but previously log-transforming all four variables. Deduce the underlying trend relating the four *original* (non log-transformed) variables that results from this fitted model.

(j) Study the residuals and other diagnostics for the model fitted in the previous question. Compare the resulting plots with those from the model without log-transformations and comment.

19. CAED Report 17 (1963), from Iowa State University, gives the following meteorological and corn production data for the State of Iowa (USA), in the years 1930–1962.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	y
Year		Pre-season rainfall (in.)	Temp. May (^{o}F)	Rainfall June (in.)	Temp. June (^{o}F)	Rainfall July (in.)	Temp. July (^{o}F)	Rainfall August (in.)	Temp. August (^{o}F)	Corn prod. (bu/acre)
1930	1	17.75	60.2	5.83	69.0	1.49	77.9	2.42	74.4	34.0
1931	2	14.76	57.5	3.83	75.0	2.72	77.2	3.30	72.6	32.9
1932	3	27.99	62.3	5.17	72.0	3.12	75.8	7.10	72.2	43.0
1933	4	16.76	60.5	1.64	77.8	3.45	76.1	3.01	70.5	40.0
1934	5	11.36	69.5	3.49	77.2	3.85	79.7	2.84	73.4	23.0
1935	6	22.71	55.0	7.00	65.9	3.35	79.4	2.42	73.6	38.4
1936	7	17.91	66.2	2.85	70.1	0.51	83.4	3.48	79.2	20.0
1937	8	23.31	61.8	3.80	69.0	2.63	75.9	3.99	77.8	44.6
1938	9	18.53	59.5	4.67	69.2	4.24	76.5	3.82	75.7	46.3
1939	10	18.56	66.4	5.32	71.4	3.15	76.2	4.72	70.7	52.2
1940	11	12.45	58.4	3.56	71.3	4.57	76.7	6.44	70.7	52.3
1941	12	16.05	66.0	6.20	70.0	2.24	75.1	1.94	75.1	51.0
1942	13	27.10	59.3	5.93	69.7	4.89	74.3	3.17	72.2	59.9
1943	14	19.05	57.5	6.16	71.6	4.56	75.4	5.07	74.0	54.7
1944	15	20.79	64.6	5.88	71.7	3.73	72.6	5.88	71.8	52.0
1945	16	21.88	55.1	4.70	64.1	2.96	72.1	3.43	72.5	43.5
1946	17	20.02	56.5	6.41	69.8	2.45	73.8	3.56	68.9	56.7
1947	18	23.17	55.6	10.39	66.3	1.72	72.8	1.49	80.6	30.5
1948	19	19.15	59.2	3.42	68.6	4.14	75.0	2.54	73.9	60.5
1949	20	18.28	63.5	5.51	72.4	3.47	76.2	2.34	73.0	46.1
1950	21	18.45	59.8	5.70	68.4	4.65	69.7	2.39	67.7	48.2
1951	22	22.00	62.2	6.11	65.2	4.45	72.1	6.21	70.5	43.1
1952	23	19.05	59.6	5.40	74.2	3.84	74.7	4.78	70.0	62.2
1953	24	15.67	60.0	5.31	73.2	3.28	74.6	2.33	73.2	52.9
1954	25	15.92	55.6	6.36	72.9	1.79	77.4	7.10	72.1	53.9
1955	26	16.75	63.6	3.07	67.2	3.29	79.8	1.79	77.2	48.4
1956	27	12.34	62.4	2.56	74.7	4.51	72.7	4.42	73.0	52.8
1957	28	15.82	59.0	4.84	68.9	3.54	77.9	3.76	72.9	62.1
1958	29	15.24	62.5	3.80	66.4	7.55	70.5	2.55	73.0	66.0
1959	30	21.72	62.8	4.11	71.5	2.29	72.3	4.92	76.3	64.2
1960	31	25.08	59.7	4.43	67.4	2.76	72.6	5.36	73.2	63.2
1961	32	17.79	57.4	3.36	69.4	5.51	72.6	3.04	72.4	75.4
1962	33	26.61	66.6	3.12	69.1	6.27	71.6	4.31	72.5	76.0

(a) Fit a Linear Model to estimate the production of corn (in *bu/acre*), using all the other variables as predictors. Comment your results.

(b) Calculate the value of the adjusted R^2. Comment.

(c) Repeat the first question, but now excluding the chronological variable x_1 from the set of predictors. Compare the results of your fits and the residual plots in both cases. Comment.

(d) Test whether the model with all the predictors and the submodel using only those predictors that can be known at the end of the month of June differ significantly. Comment.

(e) Identify a more parsimonious model than the full model, using a backward elimination heuristic based on t-tests ($\alpha = 0.10$). Repeat, but using the `leaps` package/function to carry out a full search of subsets. Comment.

(f) In the submodel that you chose in the previous question, change the units of measurement as shown below and again fit the model. Comment any observed changes in your results.

$$z^{\circ}\text{F} = \tfrac{5}{9}(z-32)^{\circ}\text{C}$$

Conversions:

$$1\ in = 25.4\ \text{mm}$$

$$1\ \text{bu/acre (corn)} = 0.06277\ \text{t ha}^{-1}$$

20. A study of a tree species seeks to establish a relation between the volume of tree trunks (variable *Volume*, in cubic feet) and trunk height (*Altura*, in feet) and diameter at 1.30 m height (*Diametro*, in inches). Measurements of these three variables were made on $n = 31$ trees. Here are some basic descriptive indicators, as well as the coefficients of linear correlations between pairs of variables:

```
> apply(arvores,2,summary)
       Diametro Altura Volume
Min.       8.30     63  10.20
1st Qu.   11.05     72  19.40
Median    12.90     76  24.20
Mean      13.25     76  30.17
3rd Qu.   15.25     80  37.30
Max.      20.60     87  77.00
```

```
> apply(arvores,2,var)
  Diametro     Altura     Volume
  9.847914  40.600000 270.202796

> cor(arvores)
          Diametro    Altura    Volume
Diametro 1.0000000 0.5192801 0.9671194
Altura   0.5192801 1.0000000 0.5982497
Volume   0.9671194 0.5982497 1.0000000
```

(a) A multiple linear regression model was initially considered, to predict trunk volume from trunk height and diameter. Here are some results:

```
Call: lm(formula = Volume ~ Diametro + Altura)
Residuals:
    Min      1Q  Median      3Q     Max
-6.4065 -2.6493 -0.2876  2.2003  8.4847

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07
Diametro      4.7082     0.2643  17.816  < 2e-16
Altura        0.3393     0.1302   2.607   0.0145

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-Squared: 0.948,      Adjusted R-squared: 0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

i. Carry out a goodness-of-fit test for the model. Discuss the result.

ii. Say whether it is possible to simplify the model, so as to obtain a simple linear regression with goodness-of-fit that is not significantly worse than that of this model. Use the significance levels $\alpha = 0.05$ and $\alpha = 0.01$. Comment.

iii. Regardless of your answer to the previous question, state, for each of the possible simple linear regressions, what would be their Coefficient of Determination and the calculated value of the F statistic in the goodness-of-fit test.

(b) Based on previous experience, it was suggested that the goodness-of-fit could be improved by log-transforming all three variables. Below is the resulting fit.

```
Call: lm(formula = log(Volume) ~ log(Diametro) + log(Altura))
Residuals:
      Min        1Q    Median        3Q       Max
-0.168561 -0.048488  0.002431  0.063637  0.129223

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.63162    0.79979  -8.292 5.06e-09 ***
log(Diametro)  1.98265    0.07501  26.432  < 2e-16 ***
log(Altura)    1.11712    0.20444   5.464 7.81e-06 ***

Residual standard error: 0.08139 on 28 degrees of freedom
Multiple R-Squared: 0.9777,     Adjusted R-squared: 0.9761
F-statistic: 613.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

i. What is the fitted underlying trend, in terms of the original (non log-transformed) variables?

ii. Discuss the following statement: "*the model with the log-transformed data has a better fit, taking into account its larger Coefficient of Determination, its larger value of the F statistic and also the smaller residuals than in the case of the model without logarithmic transformations*".

(c) It was finally decided to try a model without transformation of the variables, but where the variables *Altura* and *Volume* exchange roles, in other words trying to model trunk height from a linear regresssion on diameter and volume. Here are the results:

```
Call: lm(formula = Altura ~ Diametro + Volume)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.2958     9.0866   9.167 6.33e-10
Diametro     -1.8615     1.1567  -1.609   0.1188
Volume        0.5756     0.2208   2.607   0.0145

Residual standard error: 5.056 on 28 degrees of freedom
Multiple R-Squared: 0.4123,     Adjusted R-squared: 0.3703
F-statistic:  9.82 on 2 and 28 DF,  p-value: 0.0005868
```

Test the goodness-of-fit and discuss the result, taking into account the fairly small value of the Coefficient of Determination. How can we explain the fact that the same three variables that were used in the original model give rise to a much worse goodness-of-fit?

21. We know that, given the Linear Regression Model, for any linear combination $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$ we have:

$$\frac{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}} - \vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}}{\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}}} \frown t_{n-(p+1)} \ ,$$

with $\hat{\sigma}_{\vec{\mathbf{a}}^t\hat{\vec{\boldsymbol{\beta}}}} = \sqrt{QMRE \cdot \vec{\mathbf{a}}^t(\mathbf{X}^t\mathbf{X})^{-1}\vec{\mathbf{a}}}$. Based on this result, deduce the expression of a $(1-\alpha)\times 100\%$ confidence interval for a linear combination $\vec{\mathbf{a}}^t\vec{\boldsymbol{\beta}}$.

22. In a study of Royal apples, we seek to relate the diameter of the apples (`Calibre`, in mm) with their weight (`Peso`, in g). Data exists for 1273 fruits with diameters between 53 and 79 mm. A linear regression model was fitted with the following results:

```
Call: lm(formula = Peso ~ Calibre, data = pesocal)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -210.3137     3.8078  -55.23   <2e-16
Calibre        5.1813     0.0577   89.79   <2e-16
---
Residual standard error: 8.525 on 1271 degrees of freedom
Multiple R-squared: 0.8638,Adjusted R-squared: 0.8637
F-statistic:  8063 on 1 and 1271 DF,  p-value: < 2.2e-16
```

(a) What would be the natural intercept for this regression line? Compute a 95% confidence interval and see whether that intercept is admissible, for the fitted model. Comment your conclusions.

(b) A researcher who looked into the residuals of the fitted model claims that there is evidence for some curvature, and that it would be preferable to model weight based on a polynomial of second degree of the diameter. Here is the result.

```
Call: lm(formula = Peso ~ Calibre + I(Calibre^2), data = pesocal)
Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  72.33140   46.76415   1.547   0.1222
Calibre      -3.38747    1.41429  -2.395   0.0168
I(Calibre^2)  0.06469    0.01067   6.064 1.75e-09
---
Residual standard error: 8.408 on 1270 degrees of freedom
Multiple R-squared: 0.8677,Adjusted R-squared: 0.8675
F-statistic:  4163 on 2 and 1270 DF,  p-value: < 2.2e-16
```

i. Write down the equation of the parabola describing the fitted relation.

ii. Do you think that the researcher is right? Justify using an appropriate statistical tool. Comment your results, taking into consideration the R^2 values of each model.

23. Consider a multiple linear regression model with p predictor variables, fitted using n observations.

(a) Describe the model in detail, *using vector/matrix notation.*

(b) Show that the vector of estimators of the model parameters, $\vec{\hat{\beta}}$, can also be written as $\vec{\hat{\beta}} = \vec{\beta} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\vec{\epsilon}$.

(c) Deduce *from the above expression*, the expected vector and the covariance matrix of the vector of estimators, $\vec{\hat{\beta}}$, given the linear regression model.

24. Consider the usual coefficients of determination (R^2) and its adjusted version, R^2_{mod}, in a multiple linear regresstion with p predictors, fitted with n observations.

(a) Prove the relation $R^2_{mod} = 1 - (1 - R^2)\frac{n-1}{n-(p+1)}$.

(b) Show that the F test statistic for the goodness-of-fit test can be written only in terms of R^2 and R^2_{mod}, as: $F_{calc} = \frac{R^2}{R^2 - R^2_{mod}}$.

(c) Show that the adjusted coefficient of determination is negative when $R^2 < \frac{p}{n-1}$. Comment the implications of this condition for the F goodness-of-fit test.

2 Analysis of Variance

1. In the breeding of traditional varieties of tomato, an important characteristic is the resistance of the skin. This characteristic was measured in 6 tomato varieties. For each variety, tomatoes were chosen at random in 3 different plots, and individual observations were taken to be the mean resistance of the skins of tomatoes from a given plot. Measurements were made with a texturometer, in grams-force, *gf*. Here are the values for each plot (which are given in the *data frame* `tomate`), as well as the mean values and variances of all observations and of the observations from each variety:

Variety	Observations			Mean	Variance
18	632.04	629.30	420.59	560.6433	14 713.08
28	253.00	219.34	252.11	241.4833	367.9434
29	223.71	374.48	274.66	290.9500	5881.921
40C	503.51	757.44	856.39	705.7800	33 132.64
Ace	375.18	376.81	379.77	377.2533	5.414433
Roma	333.05	324.82	338.45	332.1067	47.11163

- The overall mean of all observations is $\overline{y}_{\cdot\cdot} = 418.0361$;
- The sample variance of all observations is $s^2_y = 34\,517.82$.

(a) What is the experimental design that was used? Describe the corresponding ANOVA model, specifying all the additional assumptions that are necessary.

(b) Build the summary-table for the appropriate analysis of variance,

i. first using only a hand calculator, and the information given above;

ii. then, with the R command `summary(aov(res.pel ~ variedade , data=tomate))`.

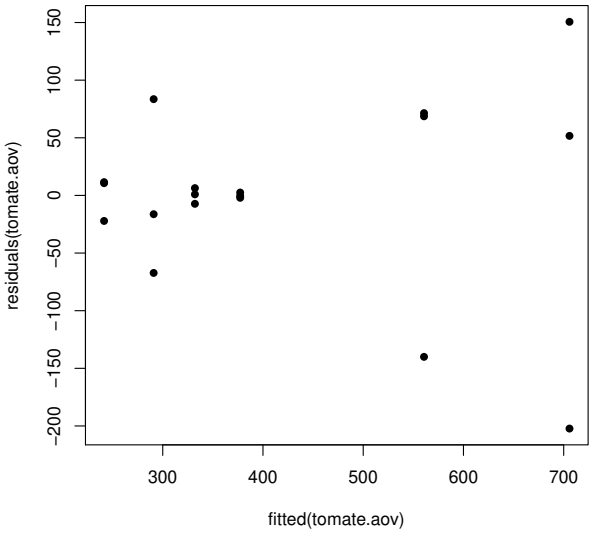
(c) Formalize and perform an appropriate F test for this problem, using a 5% significance level. Can we conclude that not all varieties have the same mean skin resistance?

(d) What is the largest significance level α for which you would change your answer? What is that value called?

(e) Use the command `model.matrix` in R to inspect the nature of the matrix $\mathbf{X}$, in this context.

(f) Use R's `fitted` command to identify the fitted values of the response variable for this Analysis of Variance.

(g) Below is the scatterplot of the (usual) residuals versus the fitted values for this ANOVA model. Discuss the plot and any implications it may have. Identify the observation with the largest magnitude residual.



2. A study of three varieties of coffee, referenced as `CA`, `CL` and `PR`, focused on the lengths of the leaf stomata. From each variety, 12 plants were selected and the mean length of the leaf stomata in each plant was measured in a controlled environment (variable `Comprimento`, in μm). Only the means and variances for the 12 plants of each variety are known:

	CA	CL	PR
Mean	22.85833	19.49333	25.31583
Variance	13.69303	2.725424	9.388936

(a) Describe in detail the ANOVA model appropriate for this problem.

(b) Compute the ANOVA summary-table for the model that you chose.

(c) What is the sample variance for the stomata lengths of all 36 observations?

(d) Is it possible to state that, in the population, the mean stomata length is the same in all three varieties, with an $\alpha = 0.05$ significance level? Provide a detailed answer.

3. It is known that carbon dioxide has a critical effect on the growth of microbial populations. Small quantities of CO_2 may stimulate the growth of some species whereas large concentrations, on the contrary, tend to have an inhibiting effect. The latter effect is commercially used to preserve stored

food. A study was carried out to investigate the effect of different concentrations of CO_2 on the growth rate of *Pseudomonas fragi*; the different concentrations (treatments) were pre-specified and the response variable that was measured was the percentage variation in the mass of the microbial culture after one hour of growth in the specified conditions. The results are given in the following table.

CO_2 concentration				
0.0	.083	.29	.50	.86
62.6	50.9	45.5	29.5	24.9
59.6	44.3	41.1	22.8	17.2
64.5	47.5	29.8	19.2	7.8
59.3	49.5	38.3	20.6	10.5
58.6	48.5	40.2	29.2	17.8
64.6	50.4	38.5	24.1	22.1
50.9	35.2	30.2	22.6	22.6
56.2	49.9	27.0	32.7	16.8
52.3	42.6	40.0	24.4	15.9
62.8	41.6	33.9	29.6	8.8

The data are available in the *data frame* `CO2`, with the CO_2 concentrations repeated in two columns: once as a factor, and then as a numerical variable.

(a) It is suggested that an Analysis of Variance be performed. Test the existence of factor effects of the CO_2 concentrations on the variation of the mass of *Pseudomonas fragi*. Check the validity of the ANOVA model assumptions.

(b) Given the nature of the predictor variable, a linear regression of the growth rates on the carbon dioxide concentrations, viewed as a numerical variable, can also be considered. Using the data frame's column `CO2` with the concentrations as numerical values (that is, column `CO2.numerico`), fit this simple linear regression model. Compare the results of the F goodness-of-fit tests in both contexts resulting from changing the nature of the predictor `CO2`. Explain the resulting differences.

4. Yields obtained with four different wheat varieties are to be compared. Thirteen fields with different soil characteristics were chosen because they are to be used in the future. Each field was divided into four plots of equal size. Within each field, one plot is assigned at random to each of the four varieties. After the harvests, the yields obtained (in t/ha) were registered in the table below (and are given in the data frame `terrenos`).

(a) The sample means for each variety suggest that some varieties may have a better yields. But are they significant differences? Reply using an appropriate Analysis of Variance. Compute the summary table and discuss your results.

(b) Test whether there are significant differences between fields, as could be expected. Comment.

	Variety			
Field	A	B	C	D
I	1.800	2.457	0.722	0.789
II	1.709	1.839	1.546	1.304
III	1.277	1.293	1.515	1.273
IV	1.675	1.745	0.800	0.846
V	1.814	1.833	1.678	1.732
VI	1.896	1.203	1.192	1.580
VII	1.078	1.689	1.583	1.168
VIII	1.740	1.518	1.050	1.305
IX	1.200	1.133	0.778	1.033
X	1.500	0.722	0.636	0.925
XI	1.932	1.700	1.203	0.850
XII	1.169	1.209	1.112	0.986
XIII	1.438	1.577	1.355	1.525
Mean	1.556	1.532	1.167	1.178
Variance	0.0879	0.1855	0.1266	0.0934

5. In a study on the growth characteristics of the stone pine (*Pinus pinea*), carried out in Sines and Tavira by the Portuguese National Institute of Agrarian and Veterinarian Research (INIAV), it was assessed whether the mean height of pines of five different origins (Morrocco, Greece, Portugal and two different sites in Italy), at two years of age. Both in Sines and in Tavira, six plots were planted with trees of each origin, thus generating $n=60$ height values (variable `alt2`, in cm), whose sample variance is $s^2=34.49584$. Here are the resulting means.

```
 prov                                                    local
Grecia Italia-1 Italia-2 Marrocos Portugal              Sines Tavira
 28.81    32.75    30.23    35.13    31.90              28.14  35.38

 prov:local
local                                                   Grand mean
 prov        Sines  Tavira                               31.76298
   Grecia    22.52   35.10
   Italia-1  31.03   34.46
   Italia-2  26.91   33.56
   Marrocos  31.16   39.09
   Portugal  29.09   34.70
```

   (a) Identify the exeprimental design and appropriate ANOVA model. Describe the model in detail.

   (b) Knowing that the Residual Mean Square is 16.59 and that the Sum of Squares associated with the five different origins is 280.61, compute the summary table of the appropriate ANOVA model.

   (c) Use an F test to assess whether there are effects of the pine trees' origin. Comment your results. Briefly state what other kinds of effects should be considered significant. Consider $\alpha = 0.05$.

6. An agricultural engineer wants to select four wheat cultivars for the four agricultural estates for which she is responsible, and which are located in Elvas, Évora, Benavila and Revilheira. In each estate, 36 plots were demarcated, and four plots were associated at random to each of the nine cultivars: Celta, Helvio, TE9006, TE9007, TE9008, TE9110, TE9115, TE9204 and Trovador. Yields were measured in each plot, in kg/ha. The overall variance of all observed yields was $s^2 = 1\,714\,242$.

   (a) Specify the experimental design that was used and describe in detail the ANOVA model suited to this experiment.

(b) An ANOVA model was fitted using R. Partial results from this fit are shown below.

```
                     Df    Sum Sq  Mean Sq  F value
localidade          ??? 183759916      ??? 234.9531
cultivar            ???       ???   964060      ???
localidade:cultivar ???       ???      ???   4.0768
Residuals           ???  28156076   260704
```

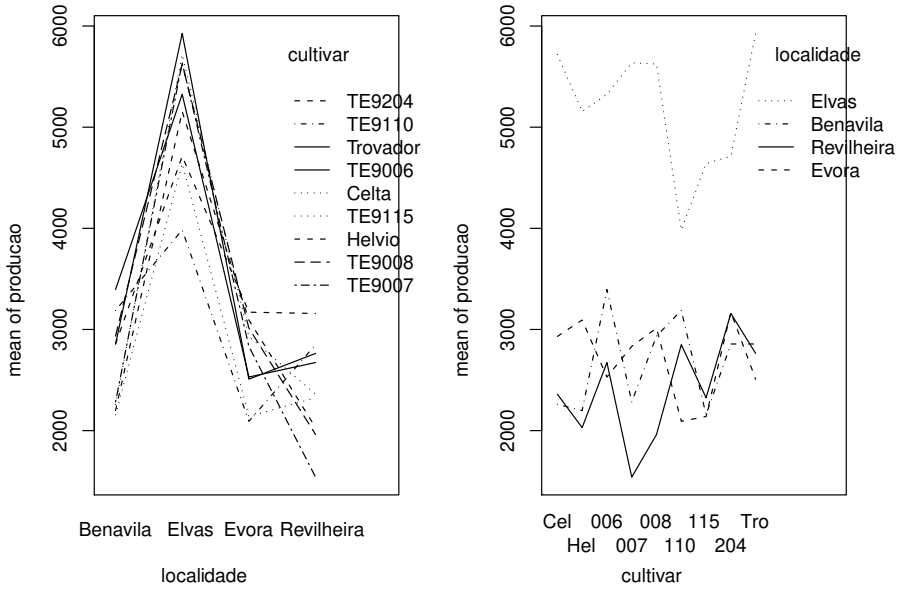
i. Complete the summary table, indicating how each of the missing values is obtained.

ii. What is the estimated value of the variance of the model's random errors, and what are the corresponding units of measurement?

iii. Formally test (for the $\alpha = 0.01$ significance level) which types of effects should be considered significant.

iv. Discuss the effects of changing the units of measurement of the response variable, from kg/ha to tons per hectare. Which values in the summary table change, and which remain the same? What are the effects of this change in units on the conclusions of the F tests?

v. The interaction plots associated with this experiment are shown below. Comment and relate these plots with the results of the previous questions.



7. With the purpose of assessing the changes in tannin content in the pulp of the sapodilla fruits (*Manilkara achras*) resulting from storage at two different temperatures (high/low) during four different time periods (0, 3, 6 or 9 days) a study was carried out with the following results:

	Time							
Temperature	0 days		3 days		6 days		9 days	
high	20.8	19.7	26.5	27.5	26.5	26.4	26.5	26.9
	18.0	19.5	27.0	26.4	27.0	24.0	25.9	26.3
low	32.3	34.1	20.8	20.5	16.4	15.7	10.3	9.7
	30.7	31.8	21.0	20.9	15.9	16.0	7.8	9.8

The overall mean and variance of all 32 observations are 22.14375 and 47.83222, respectively. The means associated with each storage time, each temperature and each combination of time and temperature, are:

```
Tables of means
 tempo
     0      3      6      9
25.862 23.825 20.987 17.900

 temperatura
  alta  baixa
24.681 19.606
```

```
tempo:temperatura
      temperatura
tempo  alta baixa
    0 19.50 32.23
    3 26.85 20.80
    6 25.97 16.00
    9 26.40  9.40
```

(a) Identify the experimental design that was used and describe in detail the ANOVA model best suited for the study.

(b) Knowing that the Residual Sum of Squares was 20.72 and that the Mean Square associated with the different storage times was 96.01, build the summary table of the ANOVA associated with this experiment.

(c) Can it be said that different storage times influence the tannin content in the pulp of these fruits? Answer with an appropriate hypothesis test.

8. [Warning: ignore this exercise, which uses a nested design and was not studied in class.]

9. Show that the sum of residuals is zero:

   (a) for each factor level in a one-way ANOVA;

   (b) for each cell, in a two-way ANOVA with interaction effects.

3 Analyis of Covariance

1. Consider the vineleaves measurements discussed in the Linear Regression Exercise 18 (`videiras` data frame).

   (a) Draw the scatterplot of the main vein lengths (variable *NP*), on the horizontal axis, and right lateral vein lengths (variable *NLdir*) on the vertical axis, but using different colours to identify the leaves from each variety (factor *Casta*). Comment.

   (b) Fit a single regression line to predict right lateral vein lengths from main vein lengths, using all $n = 600$ observed leaves and ignoring the varieties of origin. Draw that line on the scatterplot from the previous question. Discuss the goodness-of-fit of this simple linear regression.

   (c) Fit an ANCOVA model to all $n = 600$ observations, which envisages the possibility that the leaves from each variety have a different regression line. Draw the three resulting lines, using a colour code that matches that which you used for the points in the scatterplot. Discuss your results.

   (d) Formally test whether the ANCOVA model in the previous question and the single regression line fitted in question 1b) differ significantly. Comment the conclusions of your test.

   (e) Fit simple linear regressions of *NLdir* over *NP*, for each of the following subsets of $n_i = 200$ $(i = 1, 2, 3)$ observations:

      i. the n_1 observations from the Água Santa variety;

      ii. the n_2 observations from the Fernão Pires variety;

      iii. the n_3 observations from the Vital variety.

      Comment your results. In particular, compare the Coefficients of Determination in each of these three models with the Coefficient of Determination in the ANCOVA model fitted in question 1c).

(f) Inspect the model matrix $\mathbf{X}$ used by `R` to fit each of the models discussed in this Exercise (they can be obtained with the `model.matrix` command, when applied to each fitted `lm` object).

2. Consider the vineleaves measurements from Linear Regression Exercise 18 (`videiras` data frame).

   (a) Draw the scatterplot of main vein lengths (variable *NP*, horizontal axis), and leaf surface area (variable *Area*, vertical axis), using different colours or symbols to represent the leaves from each variety (factor *Casta*). Comment.

   (b) Repeat the above question, but using the logarithmic transformations of the variables *NP* and *Area*. Comment.

   (c) Fit a single regression line to model the logarithms of leaf area, based on the logarithms of main vein lengths, regardless of the varieties. Comment the model's goodness-of-fit.

   (d) Fit a new model for the logarithm of leaf surface areas, but crossing the linear relation over *log-NP* with the factor *Casta*. Comment the new model's goodness-of-fit.

   (e) Discuss the significance of the ANCOVA model from the previous question, with differentiated fits for each variety, *in terms of the non-log-transformed variables*.

   (f) Formally test whether the linearized model which allows for variety-specific equations has a significantly better fit.

   (g) Regardless of your answer to the previous question, draw the following lines on the scatterplot obtained in question 2b:

      i. the regression line from the model that ignored the varieties of each leaf;
      ii. the three variety-specific regression lines (use different colours for each line).

   (h) On the scatterplot for the original (non-log-transformed) variables that you obtained in question 2a, draw the following curves (resulting from your fitted linear regressions):

      i. the curve associated with the relation between leaf surface area and main vein length, regardless of the leaves' variety of origin.
      ii. the three curves associated with the non-linear relations between leaf surface area and main vein length, for each variety.

   Compare your results with those of the previous question and comment.

3. Consider the *iris* data measurements (data frame `iris`).

   (a) Draw the scatterplot for the measurements of petal length (horizontal axis) and width (vertical axis), but identifying the species of origin of each observation. Comment.

   (b) Fit a simple linear regression of petal width over petal length, for all $n = 150$ observations. Comment your results.

   (c) Now fit an ANCOVA model for petal width, but crossing the simple linear regression with the *Species* factor. In particular,

      i. Draw the regression lines obtained for each species on the scatterplot of question 3a).
      ii. Compare the value of the coefficient of determination now obtained with the R^2 value for the model with a single regression line, regardless of species.
      iii. The information available suggests that the population regression lines for the *setosa* and *virginica* species may be parallel. Formally test this hypothesis.

   (d) Now fit the 3 regression lines of petal width over length, for each species separately. Compare the coefficients of determination obtained for the data from each species with the overall coefficient of determination of the ANCOVA model in question 3c). What is the reason for the discrepancy between the R^2 value in the ANCOVA model and those of the separate models? Additionally, discuss the value of the single simple linear regression model fitted with all $n = 150$ iris flowers. Taking into account the low values of the R_i^2 $(i = 1, 2, 3)$ for each

species-specific model when fitted separately, how can the high value of R^2 in the single linear regresssion fitted with all 150 observations be explained? Comment the implications of this situation.

(e) Calculate the Sums of Squares for each of the models fitted in the previous question and confirm the formula given in class relating each kind of Sums of Squares with the coefficients of determination.