Applied Biostatistics with R and rk.Teaching

Alfredo Sánchez Alberca (asalber@ceu.es)

Department of Applied Math and Statistics
CEU San Pablo

March 9, 2016

Applied Biostatistics with R

Alfredo Sánchez Alberca (asalber@ceu.es)

License terms

Contents

Practice 1

Introduction to R and RKWard

1 Introduction

The big computing power achieved by computers has made them powerful tools essentials for all those disciplines, such as Statistics, that require managing large volumes of data. Nowadays, nobody addresses a serious statistical study without the help of data analysis software.

R is a powerful programming language that includes a lot of functions for representing and analyzing data. It was developed by Robert Gentleman and Ross Ihaka at the university of Auckland in New Zealand, although now is maintained by a large scientific community all around the world.

The advantages of R with respect to other data analysis software like SPSS, SAS, Matlab, Minitqab or Excel, are multiple:

- It's open source software, what implies that it's free. It can be downloaded from the web http://www.r-project.org/.
- It's multi-platform . There are versions for the main platforms: Windows, Macintosh, Linux, etc.
- It's supported by a huge scientific community that uses the software as an standard for data analysis.
- It has thousands of packages for performing any type of data analysis and graphics representations, from the most common to the most innovative and cutting edge statistical procedures that are not included in other software. The packages are organized and documented in the CRAN (Comprehensive R Archive Network) repository, from where they can be downloaded for free. In Spain there is a mirror of this repository in the web http://cran.es.r-project.org/.
- It's programmable, what means that the user can create easily their own functions or packages for specific data analysis.
- There are a lot of books, manuals and tutorials for free, that allow the learning even for advances uses in all kind of disciplines like Mathematics, Physics, Chemistry, Biology, Psychology, Medicine, etc.

By default the working environment of R is the command line, what means that the computations are done with commands manually typed in a text window. However, there exists some graphical user interfaces (GUI) that ease its use, particularly for newbies. The GUI that we are going to use along these practices is *RKWard*, developed by Thomas Friedrichsmeier, and the package rk.Teaching developed

by Alfredo Sánchez Alberca in the CEU San Pablo University and specifically designed for teaching Statistics.

The main goal of this chapter is to introduce the student to the use of RKWard and rk.Teaching, showing him the basic operations to enter and manipulate data.

2 Installation

2.1 Installation of R

Linux In the Debian distribution and derivatives (Ubuntu, Kubuntu, etc.) is as easy as typing the following command in the command prompt

```
> sudo apt-get install r-base-html r-cran-rcmdr r-cran-rodbc r-doc-html r
  -recommended
```

Windows Download from http://cran.es.r-project.org/bin/windows/base/release.htm the installation program, execute it and follow the instructions.

2.2 Intallation of RKWard and rk.Teaching

The GUI RKWard can be downloaded from the web http://rkward.sourceforge.net/ where there are instructions for the different platforms.

For Windows it is recommended to install the full package that contains R, RKWard and rk.Teaching available in http://aprendeconalf.es/rkteaching/

By default, the installation of R includes the basic packages for the most common operations or analysis. However, for other analysis is required to install additional packages, like for instance the package rkTeaching, that includes menus and dialogs for most of the statistical procedures teached in these practices.

To install the package rk.Teching, first you have to download it from the web http://asalber.github.io/rkTeaching/, then start R or RKWard and enter the following command in the command console.

```
> setwd("path_to_download")
> install.packages("rk.Teaching",repos=NULL,dep=True)
```

For packages hosted in the CRAN, it's possible to install them from RKWard selecting the menu Settings > Manage R packages and plugins. In the dialog displayed, select the Install/Update/Remove R packages tab, then search or select the desired package and click the button OK. After installing a package, to use it in a work session it must be loaded. For that, in the same menu, select the Load/Unload R packages tab, then select the package, click the button Load and finally the button OK.

3 Solved exercises

1. Create a data set with the data in the sample below and save it with name colesterol.rda

Name	Gender	Weight	Height	Cholesterol
José Luis Martínez Izquierdo	H	85	179	182
Rosa Díaz Díaz	F	65	173	232
Javier García Sánchez	M	71	181	191
Carmen López Pinzón	F	65	170	200
Marisa López Collado	F	51	158	148
Antonio Ruiz Cruz	M	66	174	249

To create a data set:

(a) Select the menu File » New » Dataset.

(b) In the dialog displayed enter the name cholesterol and click the button OK.

(c) In the data editor window, define a variable in every column giving a name for the variable in the Name row, a type (Numeric, Factor, String or Logical) in the Type row, and in the case of a factor, defining its levels in the Levels row.

(d) After define the variable, enter the data of every variable in the sample in the corresponding column.

To save the data set:

(a) Select the menu Workspace » Save Workspace.

(b) In the dialog displayed give the name cholesterol.rda to the file, select the folder where to save it and click the button Save.

2. Define a new variable Age with the ages of the individuals in the sample, between the variables Name and Gender.

Name	Age
José Luis Martínez Izquierdo	18
Rosa Díaz Díaz	32
Javier García Sánchez	24
Carmen López Pinzón	35
Marisa López Collado	46
Antonio Ruiz Cruz	68

(a) Click the left tab Workspace.

(b) In the workspace window double-click the data set cholesterol to edit it.

(c) In the data editor window, right-click the column header of the variable Gender and select Insert new variable left.

(d) In the new empty column enter the name and type of the variable Age, and enter the age of every individual.

3. Insert the following data of a new individual:

   Name: Cristóbal Campos Ruiz.
   Age: 44 years.
   Gender: Male.
   Weight: 70 Kg.
   Height: 178 cm.
   Cholesterol: 220 mg/dl.

(a) Insert the data of the new individual in the first empty row.

4. Create a new variable with the body mass index of every individual using the formula

$$\text{bmi} = \frac{\text{Weight (in Kg)}}{\text{Height (in m)}^2}$$

(a) Select the menu Teaching > Data > Compute variable.

(b) In the dialog displayed enter the formula to compute the body mass index in the field Variable computation.

(c) In the field Save as click the button Change.

(d) In the dialog displayed select as a parent object the data set cholesterol and click the button OK.

(e) Enter the name bmi for the new variable and click the button Submit.

5. Recode the body mass index variable into a new categorical variable Obesity according to the following rules:

Less than de 18.5	→	Low weight
From 18.5 to 24.5	→	Healthy
From 24.5 to 30	→	Overweight
Greater than 30	→	Obese

(a) Select the el menu Teaching > Data > Variable recoding.

(b) In the dialog displayed insert the variable bmi in the field Variable to recode.

(c) Enter the recoding rules below in the field Recoding rules:

```
lo:18.5 = 1
18.5:24.5 = 2
24.5:30 = 3
30:hi = 4
```

(d) In the field Save as click the button Change.

(e) In the dialog displayed select as a parent object the data set cholesterol and click the button OK.

(f) Enter the name Obesity for the new variable and click the button Submit.

(g) In the data edition window, enter the levels for the Obesity factor, setting the label "Low weight" for the first category, "Healthy" for the second one, "Overweight" for the third and "Obese" for the fourth.

6. Filter the data set to get a new data set with the data of males.

(a) Select the menu Teaching > Data > Data filtering.

(b) In the dialog displayed insert the data set cholesterol in the field Data set.

(c) Insert the expression `gender=="M"` in the field Selection condition.

(d) Enter the name cholesterol.males for the new data set and click the button Submit.

4 Proposed exercises

1. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

   (a) Load the data set.

1. Click the Workspace tab and double-click the rk.Teaching package to unfold the data sets that it contains.
2. Right-click the data set nenonates and select the menu Copy to .GlobalEnv to copy the data set to the working environment.

(b) Compute the variable APGAR.average as the mean of the variables APGAR1 and APGAR5.

(c) Recode the variable weight into the factor weight.category with two categories corresponding to weights less than and greater than 2.5 Kg.

(d) Recode the variable APGAR1 into the factor APGAR.state with three categories: depressed (APGAR$\leq$ 3), moderately depressed (3 <APGAR$\leq$ 6) and normal (APGAR> 6).

(e) Filter the data set to get a new data set with the neonates of non-smoking mothers with an APGAR score at 1 minute less than or equal to 3. How many neonates are there?

Practice 2

Frequency distributions and charts

1 Solved exercises

1. The number of children in a sample of 25 families is

   1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.

   Do the following operations:

   (a) Create a data set with the variable children and enter the data.

   (b) Create the frequency table.

   

   1. Select the menu Teaching > Frequency distribution > Frequency table .
   2. In the dialog displayed, select the variable children in the field Variable to tabulate and click the button Submit.

   (c) Create the absolute frequency bar chart.

   

   1. Select the menu Teaching > Charts > Bar chart.
   2. In the dialog displayed, select the variable children in the field Variable and click the button Submit.

   (d) Create also the relative frequency, cumulative absolute frequency and cumulative relative frequency bar charts, with their respective polygons.

   

   Follow the steps above checking, in the Bar options tab, the box Relative frequencies for the relative frequencies bar chart, the box Cumulative frequencies for the cumulative absolute bar chart, and both of them for the cumulative relative frequency bar chart. Check the box Polygon, to plot the corresponding polygon.

2. The number of people treated in the emergency service of a hospital every day of November was

15	23	12	10	28	7	12	17	20	21	18	13	11	12	26
30	6	16	19	22	14	17	21	28	9	16	13	11	16	20

   Do the following operations:

   (a) Create a data set with the variable emergencies and enter the data.

   (b) Create the box plot. Are there some outlier? In that case, remove the outliers and proceed with the next part.

1. Select the menu Teaching > charts > Box plot.
2. In the dialog displayed select the variable emergencies in the field Variables and click the button Submit.
3. In the output windows with the box plot identify the outliers.
4. In the data set edition tab, remove the rows with the outliers right-clicking the row header and selecting Delete this row.

(c) Create the frequency table grouping data into 5 classes.

1. Select the menu Teaching > Frequency distribution > Frequency table.
2. In the dialog displayed select the variable emergencies.
3. In the Classes tab check the box Grouping intervals, check the option Number of intervals, enter the desired number of intervals in the field Suggested intervals and click the button Submit.

(d) Create the absolute frequency histogram.

1. Select the menu Teaching > charts > Histogram.
2. In the dialog displayed select the variable emergencies in the field Variable.
3. In the Classes tab, check the box Grouping intervals, check the box Number of intervals, enter the desired number of intervals in the field Suggested intervals and click the button Submit.

(e) Create also the relative frequency, cumulative absolute frequency and cumulative relative frequency histograms, with their respective polygons.

Follow the steps above checking, in the Histogram options, the box Relative frequencies for the relative frequency histogram, the box Cumulative frequencies for the cumulative absolute frequency histogram, and both of them for the cumulative relative frequency histogram. Check the box Polygon to plot the corresponding polygon.

3. The blood type of a sample of 30 persons are:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, A, AB,
A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

Do the following operations:

(a) Create a data set with the variable blood.type and enter the data.

(b) Create the frequency table.

1. Select the menu Teaching > Frequency distribution > Frequency table.
2. In the dialog displayed, select the variable blood.type in the field Variable to tabulate and click the button Submit.

(c) Create the pie chart.

1. Select the menu Teaching > charts > Pie chart.
2. In the dialog displayed, select the variable blood.type in the field Variables and click the button Submit.

4. The age and the marital status of a sample of 28 persons are:

Marital status	Age								
Single	31	45	35	65	21	38	62	22	31
Married	72	39	62	59	25	44	54		
Widow(er)	80	68	65	40	78	69	75		
Divorced	31	65	59	58	50				

Do the following operations:

(a) Create a data set with the variables marital.status and age and enter the data.

(b) Create the frequency table of the variable age for every marital status.

1. Select the menu Teaching > Frequency distribution > Frequency table.
2. In the dialog displayed, select the variable age in the field Variable to tabulate, check the box Tabulate by groups and select the variable marital.status in the field Grouping variable(s).
3. In the Classes tab, check the box Grouping intervals and click the button Submit.

(c) Create the box plots of age for every marital status. Are there outliers? Which group have more spread in ages?

1. Select the menu Teaching > charts > Box plot.
2. In the dialog displayed, select the variable age in the field Variable(s), check the box Plot by groups, select the variable marital.status in the field Grouping variable(s) and click the button Submit.

(d) Create the relative frequency histogram of age for every marital status. Compare the histograms.

1. Select the menu Teaching > charts > Histogram.
2. In the dialog displayed, select the variable age in the field Variable, check the box Plot by groups and select the variable marital.status in the field Grouping variable(s).
3. In the Classes tab, check the box Grouping intervals and click the button Submit.

2 Proposed exercises

1. The number of injuries suffered by the members of a soccer team in a league were

$$0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1$$

Do the following operations:

(a) Construct the frequency table.

(b) Create the relative frequency and the cumulative relative frequency bar charts.

(c) Create the box plot.

2. The heights (in cm) of 30 students are

$$179, 173, 181, 170, 158, 174, 172, 166, 194, 185,$$
$$162, 187, 198, 177, 178, 165, 154, 188, 166, 171,$$
$$175, 182, 167, 169, 172, 186, 172, 176, 168, 187.$$

Do the following operations:

(a) Create the absolute frequency histogram with classes of width 10 cm from 150 to 200 cm. Are there outliers?

3. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

(a) Construct the frequency table of the APGAR score at 1 minute. If an score less than or equal to 3 indicates that the neonate is depressed, what percentage of neonates is depressed in the sample?

(b) Construct the frequency table of the neonate weight, grouping into classes of width 0.5 Kg from 2 to 4.5 Kg. What intervals contains more neonates?

(c) Compare the frequency distribution of the APGAR score at 1 minute for mothers less than 20 years old and for mothers greater than 20 years old. What group has more depressed neonates?

(d) Compare the relative frequency distribution of the neonate weight according to whether the mother smoked or not during the pregnancy. If a weight under 2.5 Kg is considered a low weight, what group has a higher percentage of neonates with low weight?

(e) Compute the prevalence of neonates with low weight for smoking and non-smoking mothers during the pregnancy.

(f) Compute the relative risk of low weight of neonate when the mother smokes vs when then mother doesn't smoke during the pregnancy.

(g) Create the bar chart of the APAGAR score at 1 minute. What is the more common score?

(h) Construct the cumulative relative frequency bar chart of the APGAR score at 1 minute. Below what value is half of the neonates?

(i) Compare the relative frequency distribution bar charts of the APGAR score at 1 minute according to whether the mother smoked or not during the pregnancy. What conclusion can be drawn?

(j) Construct the histogram of the neonates weights with classes of width 0.5 Kg from 2 to 4.5 Kg. What class contains more neonates?

(k) Compare the relative frequency histograms of the neonates weights, with classes of width 0.5 Kg from 2 to 4.5, Kg, according to whether the mother smoked or not during the pregnancy. What group has neonates with less weight?

(l) Compare the relative frequency histograms of the neonates weights, with classes of width 0.5 Kg from 2 to 4.5, Kg, according to whether the mother smoked or not before the pregnancy. What conclusions can be drawn?

(m) Construct the box plot of the neonates weights. What range of weights can be considered normal in the sample? Are there outliers in the sample?

(n) Compare the box plots of the neonates weights according to whether the mother smoked or not during the pregnancy and whether the mother was less than 20 or greater than 20 years old. What group has more central spread? What group has neonates with less weight?

(o) Compare the box plots of the APGAR scores at 1 minute and at 5 minutes. What variable has more central spread?

Practice 3

Sampling statistics

1 Solved exercises

1. The number of children in a sample of 25 families is

$$1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2, 0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.$$

Do the following operations:

(a) Create a data set with the variable children and enter the data.

(b) Compute the arithmetic mean, variance and standard deviation of the number of children. Interpret the statistics.

1. Select the menu Teaching » Descriptive statistics » Statistics.
2. In the dialog displayed insert the variable children in the field Variable.
3. In the Basic statistics tab check the boxes of Arithmetic mean, Variance and Standard deviation, and click the button Submit.

(c) Compute the quartiles, the range, the interquartile range, the third decile and the 68th percentile.

1. Select the menu Teaching » Descriptive statistics » Statistics.
2. In the dialog displayed insert the variable children in the field Variable.
3. In the Basic statistics tab check the boxes of Quartiles, Range, Interquartile range, enter the values 0.3 and 0.68 in the field Percentiles, and click the button Submit.

2. The number of people treated in the emergency service of a hospital every day of November was

15	23	12	10	28	7	12	17	20	21	18	13	11	12	26
30	6	16	19	22	14	17	21	28	9	16	13	11	16	20

Do the following operations:

(a) Create a data set with the variable emergencies and enter the data.

(b) Compute the arithmetic mean, variance, standard deviation and coefficient of variation of the number of emergencies. Interpret the statistics.

1. Select the menu Teaching » Descriptive statistics » Statistics.
2. In the dialog displayed insert the variable emergencies in the field Variable.
3. In the Basic statistics tab check the boxes of Arithmetic mean, Variance, Standard deviation and Coefficient of variation, and click the button Submit.

(c) Compute the coefficients of skewness and kurtosis and interpret the statistics.

1. Select the menu Teaching > Descriptive statistics > Statistics.
2. In the dialog displayed insert the variable emergencies in the field Variable.
3. In the Basic statistics tab check the boxes of Coefficient of skewness and Coefficient of kurtosis and click the button Submit.

3. In a group of 20 students the grades in Mathematics were

SS, AP, SS, AP, AP, NT, NT, AP, SB, SS
SB, SS, AP, AP, NT, AP, SS, NT, SS, NT

Do the following operations:

(a) Create a data set course with the variable grades and enter the data.

(b) Recode the grades into scores assigning 2.5 to SS, 6 to AP, 8 to NT and 9.5 to SB.

1. Select the menu Teaching > Data > Variable recoding.
2. In the dialog displayed insert the grades in the field Variable to recode.
3. Enter the following recoding rules in the field Recoding rules:

```
"SS" = 2.5
"AP" = 6
"NT" = 8
"SB" = 9.5
```

4. In the Save new variable click the button Change.
5. In the dialog displayed select as parent object the data set course and click the button Accept.
6. Enter the name score for the new variable, uncheck the box Convert in a factor and click the button Submit.

(c) Compute the median and the interquartile range.

1. Select the menu Teaching > Descriptive statistics > Statistics.
2. In the dialog displayed select the variable score in the field Variable.
3. In the Basic statistics tab check the boxes of Median and Interquartile range and click the button Submit.

4. The heights (in cm) of 30 students are

Females:	173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168.
Males:	179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187.

Do the following operations:

(a) Create a data set with the variables height and gender and enter the data.

(b) Compute the arithmetic mean, median, variance, standard deviation and quartiles according to the gender. Interpret the statistics.

1. Select the menu Teaching > Descriptive statistics > Statistics.
2. In the dialog displayed insert the variable height in the field Variable, check the box Statistics by groups and insert the variable gender in the field Grouping variable(s).
3. In the Basic statistics tab check the boxes of Arithmetic mean, Median, Variance, Standard deviation and Quartiles, and click the button Submit.

2 Proposed exercises

1. The number of injuries suffered by the members of a soccer team in a league were

   0, 1, 2, 1, 3, 0, 1, 0, 1, 2, 0, 1, 1, 1, 2, 0, 1, 3, 2, 1, 2, 1, 0, 1

   Do the following operations:

   (a) Compute la arithmetic mean, median, variance and standard deviation of the number of injuries and interpret them.

   (b) Compute the coefficients of skewness and kurtosis.

   (c) Compute the fourth and the eighth deciles and interpret them.

2. We want to compare the reliability of two blood pressure monitors, an arm monitor and a wrist monitor. For that purpose we have performed 8 repeated measures of the blood pressure of the same person with both moniors. The measurements (in mmHg) were:

Arm monitor:	111, 109, 112, 111, 113, 113, 114, 111
Wrist monitor:	115, 113, 117, 116, 112, 112, 117, 112

   Which monitor is more reliable?

3. The age and the marital status of a sample of 28 persons are:

Marital status	Age								
Single	31	45	35	65	21	38	62	22	31
Married	72	39	62	59	25	44	54		
Widow(er)	80	68	65	40	78	69	75		
Divorced	31	65	59	58	50				

   Do the following operations:

   (a) Compute the arithmetic mean and the standard deviation of the age according to the marital status and interpret them.

   (b) What group has the most representative mean?

4. A study wants to determine if there are relations between the blood pressure and the tobacco and drink. The values observed in a sample of 25 persons were:

Smokes	yes	no	yes	yes	yes	no	no	yes	no	yes	no	yes	no
Drinks	no	no	yes	yes	no	no	yes	yes	no	yes	no	yes	yes
Blood pressure	80	92	75	56	89	93	101	67	89	63	98	58	91

Smokes	yes	no	no	yes	no	no	no	yes	no	yes	no	yes
Drink	yes	no	yes	yes	no	no	yes	yes	yes	no	yes	no
Blood pressure	71	52	98	104	57	89	70	93	69	82	70	49

   (a) Compute the arithmetic mean, the standard deviation and the coefficients of skewness and kurtosis of the blood pressure for smokers and non-smokers, and interpret them.

   (b) Compute the same statistics for drinkers and non-drinkers. Interpret the statistics.

   (c) Compute the same statistics for smokers and drinkers, smokers and non-drinkers, non-smokers and drinkers, and non-smoker and non-drinkers. Interpret the statistics

Practice 4

Linear regression

1 Solved exercises

1. The values of two variables X and Y measured in a sample of 10 individuals are:

X	0	1	2	3	4	5	6	7	8	9
Y	2	5	8	11	14	17	20	23	26	29

Do the following operations:

(a) Create a data set with the variables X and Y and enter the data.

(b) Construct the scatter plot of X and Y.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed, select the variable Y in the field Y variable, and the variable X in the field X variable, and click the button Submit.

According to the point cloud, what type of regression model explains better the relation between X and Y?

(c) Compute the regression line of Y on X.

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed, insert the variable Y in the field Dependent variable and the variable X in the field Independent variable(s), and click the button Submit.

(d) Plot the regression line on the scatter plot.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed, insert the variable Y in the field Y variable and the variable X in the field X variable.
3. In the Fitted line tab, check the box Linear and click the button Submit.

(e) Compute la regression line of X on Y and plot it on the scatter plot.

Repeat the steps of the previous part but inserting the variable X in the field Dependent variable and the variable Y in the field Independent variable(s).

(f) How are the residuals? Comment the results.

2. A study pretends to determine the relation between the daily hours of study and the number of failed subjects in a course. The values of these variables in a sample of 30 students were:

Study hours	Failed subjects	Study hours	Failed subjects	Study hours	Failed subjects
3.5	1	2.2	2	1.3	4
0.6	5	3.3	0	3.1	0
2.8	1	1.7	3	2.3	2
2.5	3	1.1	3	3.2	2
2.6	1	2.0	3	0.9	4
3.9	0	3.5	0	1.7	2
1.5	3	2.1	2	0.2	5
0.7	3	1.8	2	2.9	1
3.6	1	1.1	4	1.0	3
3.7	1	0.7	4	2.3	2

Do the following operations:

(a) Create a data set with the variables study.hours and failed.subjects and enter the data of the sample.

(b) Construct the two-dimensional frequency table of the variables study.hours and failed.subjects.

1. Select the menu Teaching » Frequency distribution » Two-dimensional frequency table.
2. In the dialog displayed insert the variable study.hours in the field Variable to tabulate in rows and the variable failed.subjects in the field Variable to tabulate in columns.
3. In the Row classes tab, check the box Grouping intervals for the row variable, and click the button Submit.

(c) Compute la regression line of failed.subjects on study.hours and plot it.

To compute the regression line:

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable failed.subjects in the field Dependent variable and the variable study.hours in the field Independent variable(s), check the box Save the model, enter the name linear.model.failed.subjects.on.study.hours for the regression model and click the button Submit.

To plot he regression line:

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed insert the variable failed.subjects in the field Y variable and the variable study.hours in the field X variable.
3. In the Fitted line tab, check the box Linear and click the button Submit.

(d) What is the regression coefficient of the failed subjects on the daily hours of study? Interpret it.

The regression coefficient is the slope of the regression line.

(e) The linear relation is stronger or weaker than in the previous exercise? Answer the question comparing the residuals in both linear models.

(f) Compute the linear coefficient of determination and the correlation coefficient. Is the linear model a good model to explain the relation between the failed subjects and the daily hours

of study? What percentage of the variability of the failed subjects is explained by the linear model?

The coefficient of determination is showed as R^2 in output window, and the correlation coefficient is the square root.

(g) Use the linear model to predict the expected number of failed subjects for a student that studies 3 hours a day. Is this prediction reliable?

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.model.failed.subjects.on.study.hours in the field Regression model, enter the value 3 in the in field Predictions for and click the button Submit.

(h) According to the linear model, how many hours of study are required at least to pass all the subjects?

To compute the regression line:

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable study.hours in the field Dependent variable and the variable failed.subjects in the field Independent variable(s), check the box Save the model, enter the name linear.model.study.hours.on.failed.subjects for the regression model and click the button Submit.

To make the prediction:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.model.study.hours.on.failed.subjects in the field Regression model, enter the value 0 in the in field Predictions for and click the button Submit.

3. To determine how an organism metabolizes the alcohol, an experiment was conducted where we measured the alcohol in blood every half an hour after drinking a liter of wine. The data of the experiment are below.

Time (min)	30	60	90	120	150	180	210
Alcohol (gr/l)	1.6	1.7	1.5	1.1	0.7	0.2	2.1

Do the following operations:

(a) Create a data set with the variables time and alcohol and enter the data of the sample.

(b) Compute the linear correlation coefficient of the alcohol and the time and interpret it. Is the linear model a good model to explain the metabolization of alcohol?

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable alcohol in the field Dependent variable, the variable time in the field Independent variable(s), and click the button Submit.

(c) Plot the regression line of alcohol on time. Are there some point with a big residual? In such a case, remove the point from the sample and compute again the linear correlation coefficient. Has the model improved?

1. Select the menu Teaching > Charts > Scatter plot.
2. In the dialog displayed insert the variable alcohol in the field Y variable, the variable time in the field X variable.
3. In the Fitted line tab click the box Linear and click the button Submit.

It is observed that the point $(210, 2.1)$ has a huge residual compared to the others, what means that it's an outlier. To remove the outlier in the data edition windows, right-click the header row corresponding to the point and select Delete this row.

(d) If, according to the law, the maximum concentration of alcohol in blood to drive is 0.3 g/l, how much time must wait this person to drive after drinking a liter of wine? Is this prediction reliable?

To compute the regression line:

1. Select the menu Teaching > Regression > Linear regression.
2. In the dialog displayed insert the variable time in the field Dependent variable and the variable alcohol in the field Independent variable(s).
3. Check the box Save the model, enter the name linear.model.time.on.alcohol for the linear model and click the button Submit.

To make the prediction:

1. Select the menu Teaching > Regression > Predictions.
2. In the dialog displayed insert the model linear.model.time.on.alcohol in the field Regression model, enter the value 0.3 in the field Predictions for and click the button Submit.

4. The data set age.height of the package rk.Teaching contains the age and the height of 30 persons. Do the following operations:

(a) Load the data set age.height from the package rk.Teaching.

(b) Compute la regression line of the height on the age. Is the linear model a good model to explain the relation between the height and the age?

1. Select the menu Teaching > Regression > Linear regression.
2. In the dialog displayed insert the variable height in the field Dependent variable, the variable age in the field Independent variable(s), and click the button Submit.

(c) Create the scatter plot of the height on the age. Around which age changes the tendency?

1. Select the menu Teaching > Charts > Scatter plot.
2. In the dialog displayed insert the variable height in the field Y variable, the variable age in the field X variable and click the button Submit.

(d) Recode the variable age into the categorical variable age.group with two categories for younger and older than 20 years.

1. Select the menu Teaching > Data > Variable recoding.
2. In the dialog displayed insert the variable age in the field Variable to recode.
3. In the field Recoding rules enter the following rules:

```
lo:20 = "younger"
20:hi = "older"
```

4. In the field Save new variable click the button Change.
5. In the dialog displayed select as parent object the data set age.height and click the button OK.
6. Enter the name age.group for the new variable and click the button Submit.

(e) Compute the regression line of the height on the age for every age group. In which group the linear model explains better the relation between the height and the age? Justify the answer.

1. Select the menu Teaching » Regression » Linear regression.
2. In the dialog displayed insert the variable height in the field Dependent variable and la variable age in the field Independent variable(s).
3. Check the box Regression by groups and insert the variable age.group in the field Grouping variable(s).
4. Check the box Save the model, enter the name linear.model.height.on.age for the linear model and click the button Submit.

(f) Plot the regression lines of the previous part.

1. Select the menu Teaching » Charts » Scatter plot.
2. In the dialog displayed insert the variable height in the field Y variable and the variable age in the field X variable.
3. Check the box Plot by groups and insert the variable age.group in the field Grouping variable(s).
4. In the Fitted line tab, check the box Linear and click the button Submit.

(g) According to the linear model, what is the expected height for a 14 years old person? And for a 38 years old person?

To predict the height of the 14 years old person:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.regression.height.on.age.younger in the field Regression model, enter the value 14 in the field Predictions for and click the button Submit.

To predict the height of the 38 years old person:

1. Select the menu Teaching » Regression » Predictions.
2. In the dialog displayed insert the model linear.regression.height.on.age.older in the field Regression model, enter the value 38 in the field Predictions for and click the button Submit.

2 Proposed exercises

1. A research study has been conducted to determine the loss of activity of a drug. The table below shows the results of the experiment.

Time (in years)	1	2	3	4	5
Activity (%)	96	84	70	58	52

Do the following operations:

(a) Compute the regression line of the drug activity on time.

(b) What percentage decreases the drug activity every year?

(c) How much time must pass for the drug to have an activity of 80Are these predictions reliable?

2.

3. In an study about the effect of different doses of a medicament, 2 patients got 2 mg and took 5 days to cure, 4 patients got 2 mg and took 6 days to cure, 2 patients got 3 mg ant took 3 days to cure, 4 patients got 3 mg and took 5 days to cure, 1 patient got 3 mg and took 6 days to cure, 5 patients got 4 mg and took 3 days to cure and 2 patients got 4 mg and took 5 days to cure. Do the following operations:

   (a) Compute the regression line of the days to cure on the dose.

   (b) Compute the regression coefficient of the days to cure on the dose and interpret it.

   (c) Compute the correlation coefficient and interpret it.

   (d) Determine the expected time required to cure with a 5 mg dose. Is this prediction reliable? Justify the answer.

   (e) What dose must be applied to last 4 days to cure? Is this prediction reliable? Justify the answer.

4. The data set heights.weights.students of the package rk.Teaching, contains the height, the weight and the gender of a sample of students. Do the following operations:

   (a) Load the data set heights.weights.students from the package rk.Teaching.

   (b) Compute the regression line of weight on height and plot it.

   (c) Compute the regression lines of weight on height for males and females and plot them.

   (d) Compute the coefficients of determination for both models. Which model explains better the relation between weight and height, the males or the females one? Justify the answer.

   (e) What is the expected weight for a man 170 cm tall? And for a women of the same height?

5. The data set neonates of the package rk.Teaching, contains information about a sample of 320 newborns that meet the normal gestation time in a hospital during one year. Do the following operations:

   (a) Construct the two-dimensional frequency table of the APGAR score at 1 minute and whether the mother smoked or not during the pregnancy. What conclusions can you draw?

   (b) Construct the two-dimensional frequency table of te weight and the age of the mother. What conclusions can you draw?

   (c) Compute the regression line of the weight on the daily number of cigarettes smoked by the mother during the pregnancy. Is there a strong linear relation between the variables?

   (d) Plot the regression line of the previous part. Why the regression line doesn't fit well the point cloud?

   (e) Compute the regression line of the weight on the daily cigarettes smoked by the mother during the pregnancy in the group of smoking mothers. Is this regression model better or worse than the previous one? According to this model, how much decreases the weight of newborns for every daily cigarette smoked by the mother?

   (f) According to the previous linear model, what will be the expected weight of a nenonate with a mother that smokes 5 daily cigarettes during the pregnancy? And for a mother that smokes 30 daily cigarettes? Are these predictions reliable?

(g) Are there the same linear relation between the weight and the daily cigarettes somoked by the mother for mothers younger than 20 years and mothers older than 20 years? What conclusions can you draw?