

Autonomous practice with R

Walter Garcia-Fontes

Teaching workshop

Course: Theory - Practice

- Practice has to be done autonomously: hard to assess individual performance and effort
- Final exam cannot assess practice, but has to have sufficient weight to assess individual performance and effort
- Students may skip practice if final exam is enough to pass the course
- How to motivate and assure students do the practical part?

Course organization

Original setup

2 lectures of 2 hours – 10 weeks

40-student seminars – 1 hour
(2 per group)– 10 weeks

Weekly homeworks



Current setup

1 lecture of 2 hours – 10 weeks

20 or less students seminars – 1 hour
(6 per group) – 8 weeks

Autonomous work for extra concepts
and materials – 1 hour per week

Autonomous work to practice
computer software, programming and
data analysis - 3 to 5 hours per week

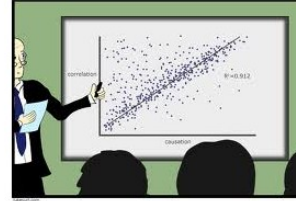
Autonomous practice

- Tutorial
- Model test – with solutions
- Actual short test in seminars
- Provide solutions for actual tests

Typical course week

Week 5



One of the main goals of statistics is to analyze the relations between two or more characteristics to take a decision for a given problem that we have.






This week we have the **third test in seminars** (Grouped data, Data Transformation, Normal Distribution and Files in R)

 [Video of Lecture 5: The Analysis of Two Numerical Variables](#)


Weekly tasks

-  [Autonomous work guidelines 5](#)
-  [Task 5: Correlation and regression](#)

Computer tutorials

-  [Two numerical variables analysis with R](#)
-  [Two numerical variables analysis with Stata](#)
-  [Programming in R: Sequences of Numbers and Vectors](#)



Test resources

-  [Model for the second test in lectures](#)

Seminar activities

-  [Activity of the week 5 seminar](#)
-  [Solutions to activity 1 week 5 seminar](#)
-  [Solutions to activity 2 week 5 seminar](#)

Practice tasks

-  [Practice Autonomous work guidelines 5](#)
-  [Practice Task 5: Correlation and regression](#)

Autonomous work guidelines 5

Introduction

This week we have studied how to summarize numerically the relation between two quantitative variables. Specifically, we have seen how to compute and interpret the coefficient of correlation and linear regression.

We have started to analyze the relation between variables. We have seen that we can use a numerical representation to view the form of the relation between two numerical variables. The graph that we use is called the *scatterplot*.

Example 1. Data Analysis Final Exam

Let us suppose that we try to explain the grades that a group of students got in the final exam of a course. We think there may be a relation between the number of hour that they have studied and the grade in the exam. We can check if these two variables are related by means of a scatterplot, since both are numerical. With the data in the following table, draw a scatterplot. Does it seem to be a relation between the two variables? What type of relation?

Grade	Study hours	Statistics background?
1	3	no
2	3	no
3	2	yes
3.5	5	no
5	5	yes
5.5	7	no
6	9	no
6.5	7	yes
7.5	10	yes
8	12	no
8	7	yes
9	14	yes

Question 1: Does it seem to be any relation between the variables? Which type of relation?

- Yes, there is a quadratic relationship between the two variables.
- No, there is no relation between the two variables.
- Yes, there is a linear inverse relation between the two variables.
- Yes, there is a linear direct relation between the two variables.

You can find this data in a spreadsheet: [HERE](#)

R tutorial

| 0%

| In this tutorial we will see how to analyze the relation between two numerical and one categorical variable, that is, how to introduce a categorical variable into the analysis of the two numerical variables. It consists of separating the data set into groups defined by the categorical variable, and to check the relation between the two numerical variables within each group.

...

|===== | 9%

| We will use a data frame that we have already read for you, called "wagexp". It contains data on the wage of 41 workers, the experience (measured in years) for these workers, and the plant where they work. We want to see if experience has an effect on wage, and if there are any differences of this relation between plants. There are therefore 3 variables, "wage", "exper" and "plant". Enter "wagexp" (without the quotation marks) to check how this data frame looks like.

```
> wagexp
  wage exper plant
1  86818    15    A
2 112316    29    A
3   66252     5    A
4   52927    13    A
5   76868    17    A
6 118042    25    A
7   96676    21    B
8   48283    17    B
9   61815    25    B
10  42743     4    B
```

R tutorial

|=====

| 27%

| As always, it is convenient to start with a scatterplot to see the relation between the two
| numerical variables. We can first plot a scatterplot ignoring the groups defined by the
| categorical variable (the plants). For the scatterplot we use the plot() function as we did in the
| tutorial for two numerical variables. Enter now the appropriate command to obtain the scatterplot.
> plot(wagexp\$wage,wagexp\$exper)

| Give it another try. Or, type info() for more options.

| Enter "plot(wagexp\$wage~wagexp\$exper)" to see the scatterplot.

> plot(wagexp\$wage~wagexp\$exper)

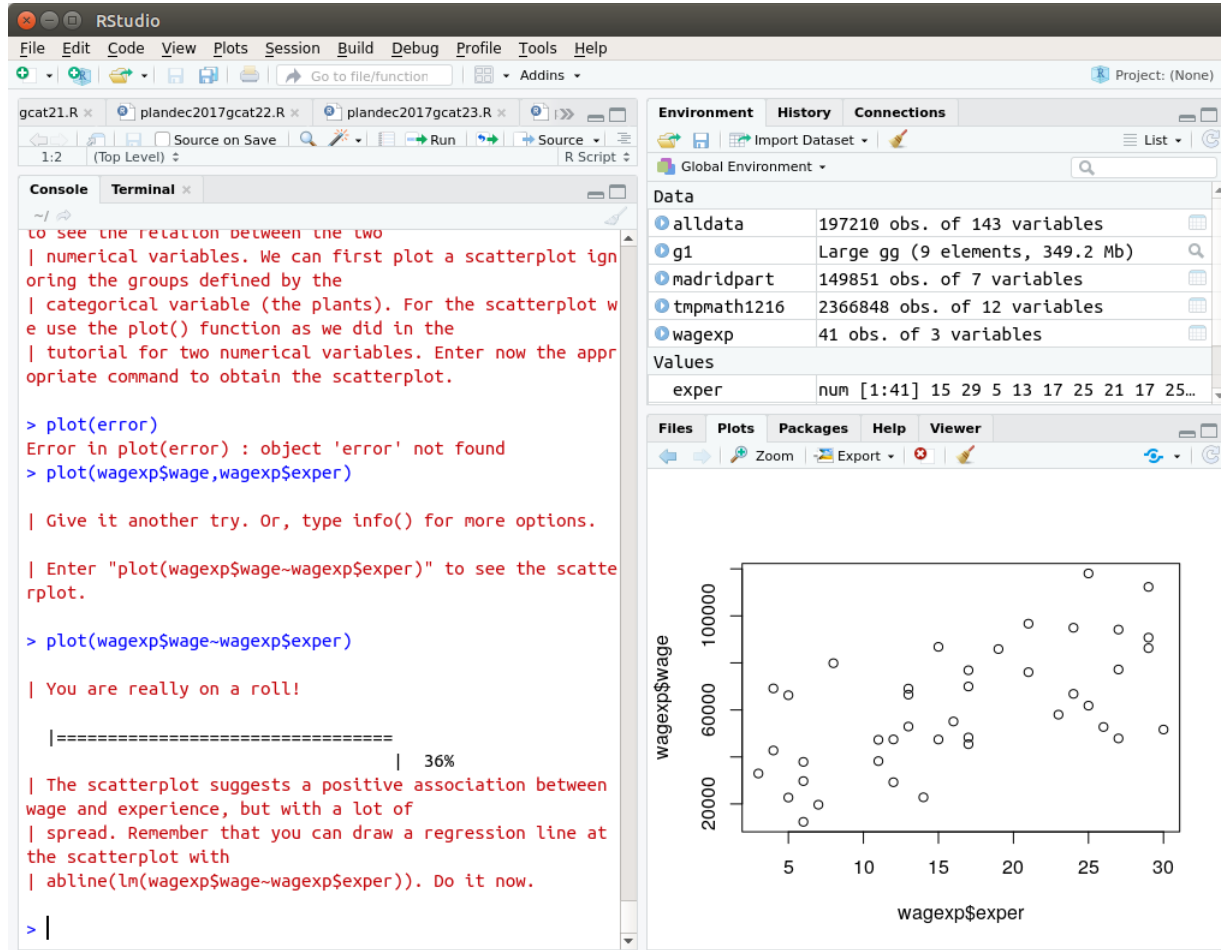
| You are really on a roll!

|=====

| 36%

| The scatterplot suggests a positive association between wage and experience, but with a lot of
| spread. Remember that you can draw a regression line at the scatterplot with
| abline(lm(wagexp\$wage~wagexp\$exper)). Do it now.

R tutorial - RStudio



The screenshot shows the RStudio interface. The console window contains the following text:

```
to see the relation between the two
| numerical variables. We can first plot a scatterplot ignoring the groups defined by the
| categorical variable (the plants). For the scatterplot we use the plot() function as we did in the
| tutorial for two numerical variables. Enter now the appropriate command to obtain the scatterplot.

> plot(error)
Error in plot(error) : object 'error' not found
> plot(wagexp$wage,wagexp$exper)

| Give it another try. Or, type info() for more options.

| Enter "plot(wagexp$wage~wagexp$exper)" to see the scatterplot.

> plot(wagexp$wage~wagexp$exper)

| You are really on a roll!

|=====| 36%

| The scatterplot suggests a positive association between wage and experience, but with a lot of
| spread. Remember that you can draw a regression line at the scatterplot with
| abline(lm(wagexp$wage~wagexp$exper)). Do it now.

> |
```

The Environment pane on the right shows the following data objects:

Data	Details
alldata	197210 obs. of 143 variables
g1	Large gg (9 elements, 349.2 Mb)
madridpart	149851 obs. of 7 variables
tmpmath1216	2366848 obs. of 12 variables
wagexp	41 obs. of 3 variables

The Values pane shows the following data for the 'exper' variable:

Values	Details
exper	num [1:41] 15 29 5 13 17 25 21 17 25...

The Plots pane shows a scatterplot of wage vs experience. The y-axis is labeled 'wagexp\$wage' and ranges from 20,000 to 100,000. The x-axis is labeled 'wagexp\$exper' and ranges from 5 to 30. The plot shows a positive correlation between wage and experience, with a significant amount of scatter.

Stata/Excel tutorial

Analysis of Two Numerical Variables with Stata

Index

To start this tutorial click [here](#). You can also go directly to any of the chapters clicking in the corresponding entry in the index.

1. [1. Scatterplot](#)
2. [2. Correlation and linear regression](#)
3. [3. Prediction](#)
4. [4. Residuals](#)
5. [5. Influential observations](#)
6. [6. Non-linear transformations](#)

Stata/Excel tutorial

[<< Previous](#)

[Index](#)

[Next >>](#)

Analysis of Two Numerical Variables with Stata

2. Correlation and linear regression

The correlation coefficient is a measure of the linear correlation between two numerical variables. We can get the correlation coefficient with the "correl" command in Stata:

```
correl FINAL MIDTERM
```

Stata shows us the correlations between all variables in the list provided to the "correl" command:

```
          |      FINAL  MIDTERM
-----+-----
    FINAL |      1.0000
    MIDTERM |     0.6403      1.0000
```

The diagonal is equal to 1 because by definition the correlation of a variable with itself is 1. The correlation between the FINAL and MIDTERM grade is 0.6403, showing a positive and not too strong linear association between the variables in the midterm and final exam.

The other main numerical summary for the relation between two numerical variables is the regression line. We can compute the coefficients (constant and slope) of the regression line using the "reg" command:

```
reg FINAL MIDTERM
```

We get a lot of information:

Source	SS	df	MS	Number of obs =	21
Model	1604.08089	1	1604.08089	F(1, 19) =	13.20
Residual	2308.58578	19	121.504515	Prob > F =	0.0018
Total	3912.66667	20	195.633333	R-squared =	0.4100
				Adj R-squared =	0.3789
				Root MSE =	11.023

Support for tutorials

Question on drawing a histogram		WALTER ALFREDO GARCIA FONTES BADANIAN	0	Tue, 17 Oct 2017, 3:38 PM WALTER ALFREDO GARCIA FONTES BADANIAN Sun, 8 Oct 2017, 12:46 PM	<input checked="" type="checkbox"/>
R Studio, Excel, Stata		JACOB ANDO	2	JACOB ANDO Sun, 8 Oct 2017, 11:57 AM	<input checked="" type="checkbox"/>
Question 7 of Autonomous Work 1		WALTER ALFREDO GARCIA FONTES BADANIAN	0	WALTER ALFREDO GARCIA FONTES BADANIAN Sun, 8 Oct 2017, 11:18 AM	<input checked="" type="checkbox"/>
Question on empty stems in the stemplot diagram		WALTER ALFREDO GARCIA FONTES BADANIAN	0	WALTER ALFREDO GARCIA FONTES BADANIAN Sun, 8 Oct 2017, 11:04 AM	<input checked="" type="checkbox"/>
Found solution to QDAP Package issues		ERÉNDIRA LEON SALVADOR	2	ERÉNDIRA LEON SALVADOR Fri, 6 Oct 2017, 10:05 AM	<input checked="" type="checkbox"/>
Problems with qdap		NEREA PÉREZ BENÍTEZ	8	WALTER ALFREDO GARCIA FONTES BADANIAN Fri, 6 Oct 2017, 6:36 AM	<input checked="" type="checkbox"/>
Issues building a table frequency in R Studio		NÚRIA LÓPEZ I GABALDÀ	5	WALTER ALFREDO GARCIA FONTES BADANIAN Fri, 6 Oct 2017, 6:30 AM	<input checked="" type="checkbox"/>
Can't use Qdap		JAVIER GRACIA NAVARRO	15	WALTER ALFREDO GARCIA FONTES BADANIAN Thu, 5 Oct 2017, 4:17 PM	<input checked="" type="checkbox"/>
Problems with QDAP Package		ERÉNDIRA LEON SALVADOR	2	ERÉNDIRA LEON SALVADOR Tue, 3 Oct 2017, 11:07 PM	<input checked="" type="checkbox"/>
RStudio		NEUS MARTÍ TRULL	1	WALTER ALFREDO GARCIA FONTES BADANIAN Mon 2 Oct 2017 7:21 AM	<input checked="" type="checkbox"/>

Support for tutorials

Support forum

Found solution to QDAP Package issues

📧 Subscribed

◀ Problems with qdap

Question on empty stems in the stemplot diagram ▶

Export whole discussion to portfolio

Display replies flat, with oldest first ▾

Move this discussion to ... ▾

Move

Pin



Found solution to QDAP Package issues

by ERÉNDIRA LEON SALVADOR - Thursday, 5 October 2017, 7:49 PM

Hello,

I've tried this and it solved it :) Hope for you too.

https://www.r-statistics.com/2012/08/how-to-load-the-r-java-package-after-the-error-java_home-cannot-be-determined-from-the-registry/

P.S: You have to delete the package qdap and to download again so it can restart and use the functions required :D

[Permalink](#) | [Edit](#) | [Delete](#) | [Reply](#) | [Export to portfolio](#)



Re: Found solution to QDAP Package issues

by WALTER ALFREDO GARCIA FONTES BADANIAN - Friday, 6 October 2017, 6:35 AM

Dear Erendida, thank you very much! Let's hope this helps other students, I posted your solution also in the Catalan/Spanish groups of Data Analysis.

[Permalink](#) | [Show parent](#) | [Edit](#) | [Split](#) | [Delete](#) | [Reply](#) | [Export to portfolio](#)



Re: Found solution to QDAP Package issues

by ERÉNDIRA LEON SALVADOR - Friday, 6 October 2017, 10:05 AM

Happy to help!

Let's hope this can solve it for everyone with the same issue :D

[Permalink](#) | [Show parent](#) | [Edit](#) | [Split](#) | [Delete](#) | [Reply](#) | [Export to portfolio](#)

Model tests

Question 1

Not yet answered
Marked out of
1.00

Flag question

Edit question

In the following file you will find data on the consumption of natural gas and the hours a family has heating on (variability in consumption is explained because the temperature can be set up or down):

[Heating and gas consumption](#)

Enter the data into R to analyze them and answer the following questions.

Using the `lm()` function, the regression line shows the following coefficients:

Select one:

- a. Constant = -0.5622, Slope = 0.3020
- b. Constant = -0.2936, Slope = 0.5400
- c. None of the other options is correct
- d. Constant = -0.6216, Slope = 0.7020
- e. Constant = -0.4648, Slope = 0.4700

Question 2

Not yet answered
Marked out of
1.00

Flag question

Edit question

The relation that we observe in use of heating and gas consumption is

Select one:

- a. an indirect or negative linear association
- b. None of the other options is correct
- c. a skewed distribution
- d. a direct or positive linear association
- e. a symmetrical distribution

Question 3

Not yet answered
Marked out of
1.00

Flag question

Edit question

Using the `cor()` function, compute the correlation coefficient between use of heating and gas consumption. The value that R shows is equal to

Select one:

- a. 0.8320391
- b. None of the other options is correct
- c. 0.9780921
- d. 0.8920203
- e. 0.9441597

Question 4

Not yet answered
Marked out of
1.00

Flag question

Edit question

Using the `plot()` and `abline()` functions draw a scatterplot with a regression line. Based on what you observe in the scatterplot

Select one:

- a. the residuals are all equal to 0
- b. the residuals become larger in absolute value for more than 15 hours of heating use
- c. the residuals become smaller in absolute value for more than 15 hours of heating use
- d. None of the other options is correct
- e. there is no special pattern in the residuals

Tests solutions

Question 1

Not answered

Marked out of 1.00

Flag question

Edit question

In the following file you will find data on the consumption of natural gas and the hours a family has heating on (variability in consumption is explained because the temperature can be set up or down):

Heating and gas consumption

Enter the data into R to analyze them and answer the following questions.

Using the `lm()` function, the regression line shows the following coefficients:

Select one:

- a. Constant = -0.5622, Slope = 0.3020
- b. Constant = -0.2936, Slope = 0.5400
- c. None of the other options is correct
- d. Constant = -0.6216, Slope = 0.7020
- e. Constant = -0.4648, Slope = 0.4700

Your answer is incorrect.

The correct answer is: Constant = -0.4648, Slope = 0.4700

Question 2

Not answered

Marked out of 1.00

Flag question

Edit question

The relation that we observe in use of heating and gas consumption is

Select one:

- a. an indirect or negative linear association
- b. None of the other options is correct
- c. a skewed distribution
- d. a direct or positive linear association
- e. a symmetrical distribution

Your answer is incorrect.

The correct answer is: a direct or positive linear association

Actual tests

Test creation



- ✓ Test 4 for the seminar
Hidden from students
- ✓ Test 4 for the seminar
Hidden from students
- ✓ Test 4 for the seminar
Hidden from students
- ✓ Test 4 for the seminar
Hidden from students
- ✓ Test 4 for the seminar
Hidden from students
- ✓ Test 4 for the seminar
Hidden from students
- ✓ Test 4 for the seminar
Hidden from students
- ✓ Test 4 for the seminar
Hidden from students



Each student gets a random test



Restrict access

Access restrictions

Student match the following

Group ×



















Activity completion

Actual tests

Make test visible only for the time of the student's seminar and protect with password

- [Edit settings](#)
- **Group overrides**
- [User overrides](#)

Test 4 for the seminar

Group	Overrides		Action
Franja A	Require password	Enabled	  
Franja B	Require password	Enabled	  
Franja C	Require password	Enabled	  
Franja D	Require password	Enabled	  
Franja E	Require password	Enabled	  
Todos	Require password	Enabled	  

Assessment

Activity grading		
Final grading	Final exam	40 points, minimum 16 of the final grade (or 60 points, minimum 24, of the grades if the lecture tests do not improve the average with the final exam)
Continuous grading	Activities	
	Class participation	5 points of the final grade
	Weekly task assignments	5 points of the final grade
	Lecture tests	20 points of the final grade (only taken into account if they raise the final grade)
	Seminar tests	15 points of the final grade
	Team project	15 points of the final grade
Total points to be earned		100 points (A minimum of 60 points and more than 40% in the final are needed)

Thank you!