Lab 5: Correlation and Regression in R

This lab covers the following skills:

- How to conduct a correlation in R using `cor.test()` and interpret the output
- How to conduct a regression in R using the `lm()` command and `summary()` command

In addition, you should know how to input data from R, visualize data using a histogram, and filter your data to remove outliers. We covered this in a previous lab.

In the lab below, make sure you use code to complete all the bolded **steps** below and create annotations indicating what you are doing. In addition, make sure you answer the numbered questions in your R notebook. Once you are finished, upload your completed R notebook.

Overview

Correlation and regression are ways that we examine association claims, or whether two variables are related. This lab will start to cover these techniques. When examining whether one variable predicts another variable, correlation and regression are both ways that we can examine the relationship.

Generally, people do correlation when they want to measure whether two variables are associated. Regression also accomplishes this, but gives us the materials to use one variable to predict another. In addition, regression can be extended to multiple regression to examine multiple predictors. This can allow us to control for other variables.

Regression in R is very easy, at least when compared to doing it by hand. If we do regression in R, we can see what the relationship between variables are and find other important bits of information which would be hard to calculate, such as the significance of each of the predictor variables.

Inputting and visualizing data

Step 1: The data in this lab are available on canvas as lab5.csv. Before you get started on the lab, you need to do the following: 1. Create a new folder for this lab on your computer 2. Download the data for this project in that folder 3. Create a new R notebook in that same folder and save it as Lab5Notebook.Rmd (or whatever name you want to give it)

This lab uses real data I collected from a section of PSY 101 in order to assess how studying affects grades. Here are the variables:

- student: number that identifies each student
- daysstudying: number of days a person studied. This was defined as how many days before the exam that a student first looked at a study guide on Moodle. For instance, if they first looked at the exam study guide 3 days before the exam, this would be 3.
- examgrade: grade on the exam out of 100 points
- shortanswer: grade on the short answer part of the exam
- mc: grade on the multiple choice questions

Step 2: Now, create two new code chunks. In the first code chunk, load the tidyverse library. In the second code chunk, input your data using the `read.csv()` command to the data frame `lab5`.

You should use your R notebook to do the analyses below.

Step 3: The first thing we do is visualize our data to see if there are outliers and what we should do about them. Do a histogram of the exam scores (variable “examgrade”). You can use the `hist()` command or use ggplot like you did in Lab 4.

1. Are there any outliers? Answer this question as an annotation.

2. View the data by clicking on it in the enviroment tab of RStudio. When you visualize your data, who is the outlier?

Now we are going to create a new dataframe that removes the outlier for sectiong. To do this, we are going to use the `filter()` command from tidyverse. For instance, we can create a subset that selects only students who got above a 70 on their exam by typing the following: `filter(lab5, examgrade > 70)`.

Step 4: Use the filter command to remove the outlier and create a new dataframe called lab5clean which removes the outlier.

Step 5: Create a code chunk to visualize do a histogram of examscores in lab5clean and see if the outlier is there.

Correlations in R

One aspect I may be interested in is whether students' performance in the multiple choice section is related to their performance in the short answer section. We will look at this using a correlation. A correlation is an appropriate way to test this because there is not a clear predictor variable.

To conduct this correlation, we will use the `cor.test()` command. This command takes two vectors (lists of numbers) as two inputs. Since correlation does not have a predictor or outcome variable, the order does not matter.

Step 6: To look at this correlation, we will do the following. Type the following into R as a code chunk.

```
cor.test(lab5$shortanswer, lab5$mc)
```

```
##
##  Pearson's product-moment correlation
##
## data:  lab5$shortanswer and lab5$mc
## t = 2.1596, df = 27, p-value = 0.03985
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.02010695 0.65776886
## sample estimates:
##       cor
## 0.3837844
```

There are three important numbers in the output above. The first is the correlation itself. That is the very last number, after the word "cor". In this case, the correlation is .38. The second number is the degrees of freedom. This is listed as "df = 27" on the second row of the output. The third important number is the p-value, which is listed as the last number on the second row. In this case, the p-value is .03985. Usually I round this to 3 decimal places (or two significant figures), so I would report p = .040.

An important number to ignore is the *t-value* that `cor.test()` reports. A correlation is not a t-test. This t-value is a relic that R uses to calculate the p-value, and you do not need it. Remember, correlations only range from -1 to 1. They will never be greater than 1.

When you report the correlation, you should do the following for APA format, which gives the correlation, the degrees of freedom, and the p-value. $r(27) = .38$, $p = .040$.

3. Based on what you know about correlations, is this a high correlation? What factors might explain this correlation?

4. Based on what you know about null hypothesis testing, if you assume an alpha of .05, what would you do regarding the null hypothesis?

Step 7: Examine the correlation using the data frame that removed the outlier. Did the correlation change? Note this in your annotation whether you think it changed, and whether this is important.

Regressions in R

Regressions in R use the `lm()` command (which stands for linear model. They use the following formula `outcome~predictor` where you separate the outcome variable from the predictor variable using a tilde.

Our prediction is that studying predicts exam grades. So we will for your data (including the outlier).

Step 8: Type the following into R.

```
x = lm(examgrade~daysstudying, data = lab5)
summary(x)
```

```
## 
## Call:
## lm(formula = examgrade ~ daysstudying, data = lab5)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -22.9332  -6.8787   0.2914   6.3866  21.0668 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)   69.9332     3.3729  20.734   <2e-16 ***
## daysstudying   0.5850     0.2529   2.313   0.0286 *  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 9.354 on 27 degrees of freedom
## Multiple R-squared:  0.1654, Adjusted R-squared:  0.1345 
## F-statistic:  5.35 on 1 and 27 DF,  p-value: 0.02858
```

Make sure you include both lines in the same chunk, so that your regression summary output is included in your RNotebook.

Now you see the regression. The regression output has a lot of sections.

A. Residuals. This is a summary of the residuals of the regression. Residuals are the error in the regression equation, or how far the line of best fit is from the actual data. This gives you the quartiles for residuals, or the minimum, first quartile, median, third quartile, and maximum.

Residuals tell us how far off the regression line is from predicting hte actual values. The residuals should be equally distributed around zero, with roughly half being positive and roughly half being negative. The negative residuals (where the regression line predicts a better grade than the person received in reality reality) are relatively equal to positive residuals (where the line predicts a worse grade than reality). Also, the median should be close to zero.

5. What are the residuals? Are the negative residuals relatively equal to the positive residuals (is the min roughly equal to the max, is the first quartile (1Q) roughly equal to the third quartile (3Q) and the median close to zero?

B. Coefficients: these tell us the regression equation for the intercept and the predictor variable. They have four columns. The first column, titled Estimate, gives us the values for the intercept (69.9332) and the value for the b-value, which is listed in the row for daysstudying (0.5850). The other columns tell us how good of a predictor each value is, by giving each variable a standard error, t statistic, and p-value. We will not discuss the importance of these values in this class, but this p-value would be the same as if we did a correlation with the same data.

6. Based on this output, what would the regression equation be?

C. The rest of this output tells us how good the fit is by giving us the R-squared value and the adjusted R-squared value. The R-squared value tells us conceptually what percent of the variance in a student's grade is explained by how long they studied. This ranges from 0, meaning absolutely none of a student's grade is explained by how long they studied, to 1 which means the student's grade is entirely explained by how long they studied.

7. What is the R-squared value? Do you think this is a high value?

8. Now use your regression equation to answer the following questions:

a. For each extra day studying, how many points should a person expect to increase their exam grade?

b. What would you expect the exam grade to be for a student who studied 5 days?

Step 9: Now run the regression for the lab5clean dataframe, which removed the outlier.

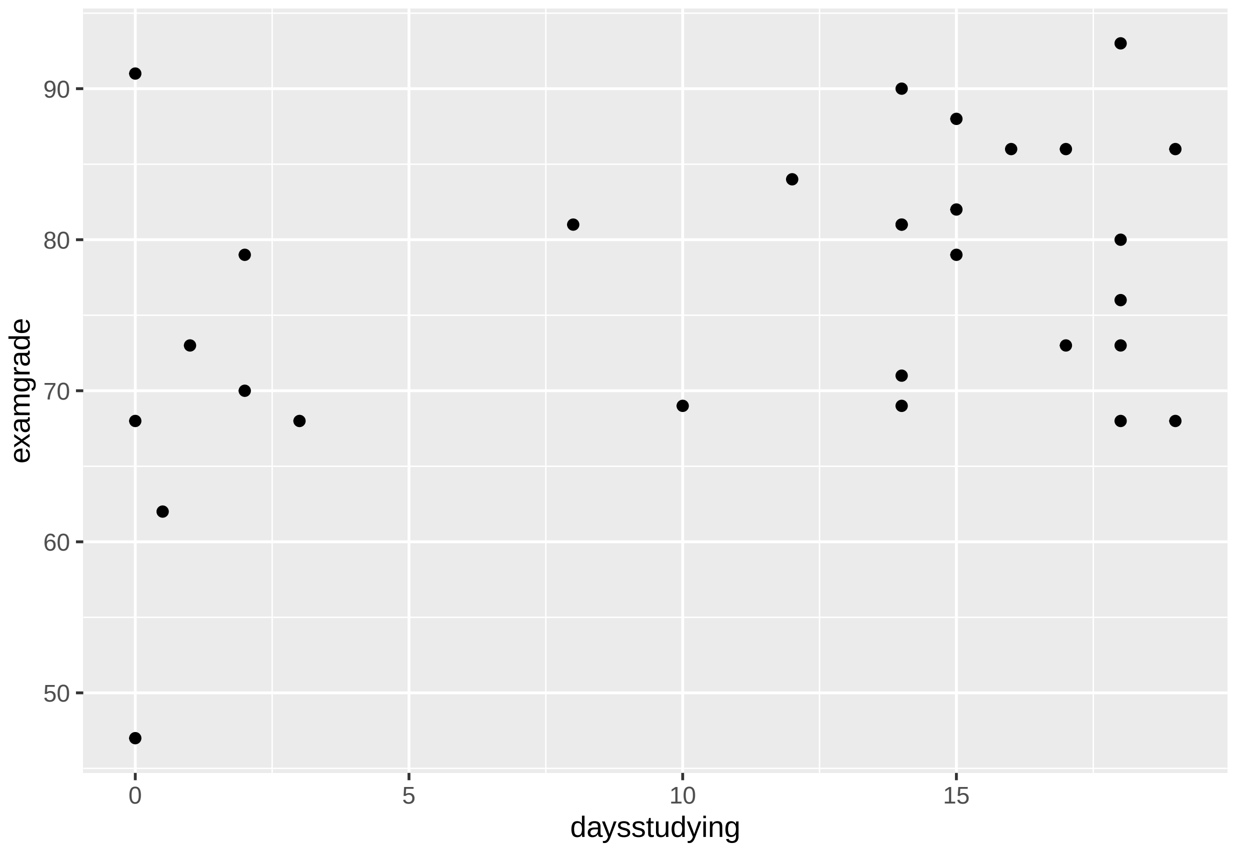
9. Write the equation you get below. Is it very different than the equation with the outlier? Do you think we should remove the outlier from the analysis?

Visualizing your data using scatterplots

It is best to examine whether the data are distributed well using a scatterplot. ggplot allows you to do this by using the geom_point() function. To plot a scatterplot of days studying as the x variable and exam grades as the y variable, you would type the following.

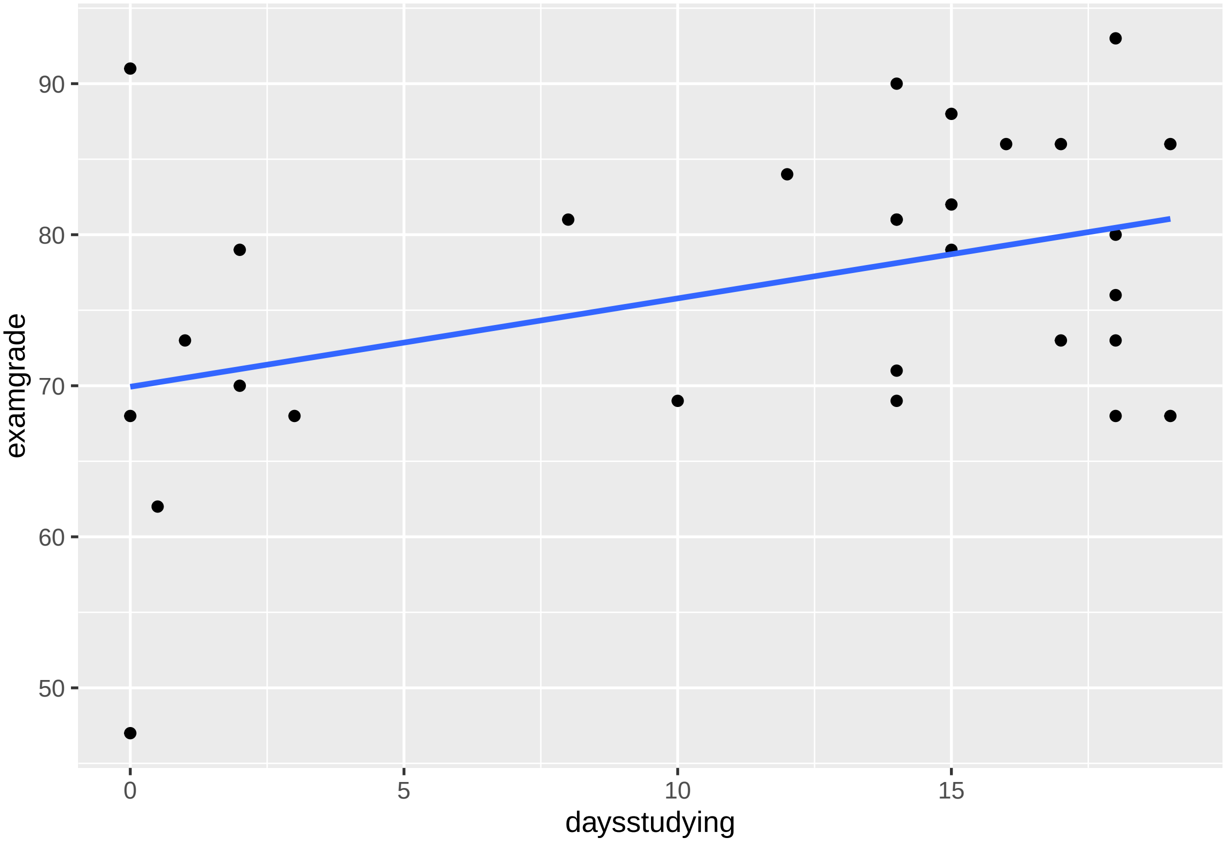
Step 10: Type the following in your R notebook to produce a scatterplot.

```r
ggplot(data = lab5) +
  geom_point(aes(x = daysstudying, y = examgrade))
```



We can also add a line of best fit to the plot which corresponds to our equation. This requires adding a second geom to our plot. We take the code from the first plot and then add a second geom, like as follows.

```
ggplot(data = lab5) +
  geom_point(aes(x = daysstudying, y = examgrade)) +
  geom_smooth(aes(x = daysstudying, y = examgrade), method = 'lm', se = F)
```



Notice that the `geom_smooth()` function had an additional few arguments. The `method = 'lm'` part says to use a linear model. The `se = F` removes the standard error. Try plotting the plot when removing each of these options. What happens?

Step 11: Now, create four scatterplots: 1. The first plot should be a scatterplot looking at the relationship between daysstudying on the y-axis and and multiple choice scores (variable mc) on the x-axis 2. Now create a plot like the one above in step 1 with a line of best fit added 3. Create a scatterplot with no line of best fit looking at the relationship between daysstudying on the y-axis and short answer scores (variable shortanswer) on the x-axis 4. Create a scatterplot just like in step 3, adding a line of best fit to

Does studying predict multiple choice grade or short answer better?

We can use a linear model to examine whether studying is a better predictor of grades on the multiple choice or short answer part of the exam better.

Step 12: Create two regressions, one looking at daysstudying as a predictor of mc and a second looking at daysstudying as a predictor of shortanswer. Make sure the output is visible in your RNotebook by including the summary() function afterward.

10. Based on these two regression, which variable is better predicted by days studying, multiple choice or short answer? Does this reflect what you found in your graphs?