

STAT202 Assignment 1: Review of simple linear regression

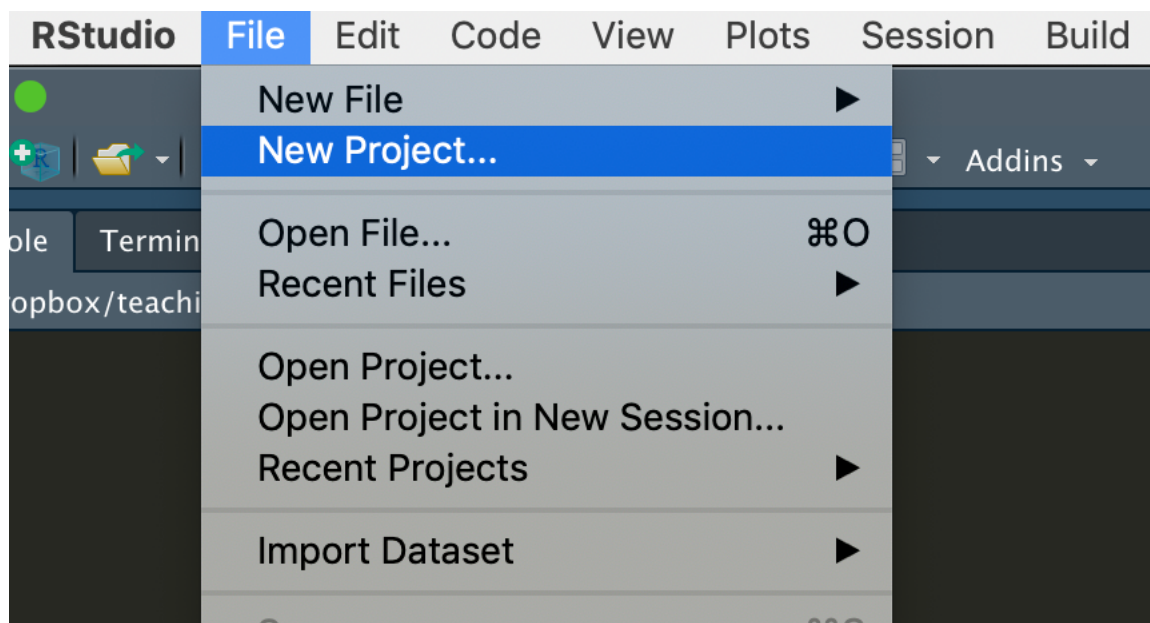
Due on 4th August 12 noon

You will work with a subset of the **euc_tricarpa.csv** dataset, which contains data from a *Eucalyptus tricarpa* trial in New Zealand, part of a programme looking for naturally durable wood to replace CCA treated pine in orchards. The data contains the following variables: **MOE** (modulus of elasticity, a measure of wood stiffness), **acoustic_velocity** (velocity of propagation of sound waves in the wood, km/s) and **dry_density** (mass of wood per cubic metre, kg/m³).

Each of you will have a random subset of the original dataset, which depends on your student code (that's the number in your student ID). This means that results for the assignment are slightly different for each student.

For each assignment you will a- create a folder (as explained in lectures), b- download the data for the assignment in that folder (**euc_tricarpa.csv** in this case), and c- create an RStudio project based on that folder.


Launch RStudio and follow the screenshots





And then select the folder (directory) you created before.

New Project

Create Project

**New Directory**
Start a project in a brand new working directory >

**Existing Directory**
Associate a project with an existing working directory >

**Version Control**
Checkout a project from a version control repository >

Cancel

For your reference we will use the following functions in the assignment:

library(tidyverse)	Makes available functions of the tidyverse, a group of easier to use R functions (read_csv, write_csv, %>%, ggplot, filter).
<-	Assign objects to a name (greater than followed by hyphen)
read_csv()	Read a comma separated values (CSV) file.
write_csv()	Write a comma separated values (CSV) file.
%>%	Read as 'then' or 'pass to'. This lets you chain (apply consecutively) multiple functions without naming each individual step. It passes the results of a function to the next one.
filter(condition)	Filters data for one or more conditions
mutate(var = some function)	Create new variables or modify variables
ggplot(dataset, aes(x, y)) + geom_point()	Create advanced plots
lm(response ~ predictors, data = dataset)	Fits linear models. One usually assigns this function to a name, as in model_1 <- lm(y ~ x, data = my_data)
summary(object_name)	Provides a summary of the object. Examples: summary(mydata) summary(model_1)
plot(object_name)	For models produces a handy plot of residuals

1. Read the **euc_tricarpa.csv** data file into RStudio, and take a random sample based on your student code (e.g. 99999999), so you keep only the data for yourself. Call that data file **eucalyptus**.

```
# Read Eucalyptus tricarpa data set
# The file below should be in your project folder
eucalyptus <- read_csv('euc_tricarpa.csv')

# Use YOUR student user code to select the observations
# you have to keep. For example, for 99999999
set.seed(99999999)
my_eucs <- eucalyptus %>% sample_n(500)
```

2. Plot **MOE on acoustic_velocity** (this means y-axis is MOE and x-axis is acoustic velocity). Before running any analyses, describe the relationship between the two variables in no more than 20 words.
3. Fit a linear regression of **MOE on acoustic_velocity**, and call it model1. By fit I mean perform the estimation of regression coefficients and related information. Write down the regression coefficients, the standard error of the residuals, multiple R^2 and the adjusted- R^2 . Explain the meaning of the intercept and the slope in your own words and in the context of the problem.
4. Using the data from question 3 now center **acoustic_velocity** (that is, express it as deviation from its mean value). You will create a new variable (let's call it **cent_velocity**) that equals **acoustic_velocity - mean(acoustic_velocity)** and use it to repeat step 3, call it model2. Write down the new coefficients and compare them with the non-centred analyses. Explain any differences in the estimates; What's now the meaning of the intercept?
5. Add **dry_density** to the regression in question 3 (predictors are now acoustic_velocity + dry_density) and call it model3. Write down the regression coefficients, the standard error of the residuals, the multiple R^2 and the adjusted- R^2 . Has the model fit improved? Use both the standard error of the residuals and adjusted- R^2 to justify your answer.
6. Plot the residuals for the model in question 5. Comment how well your residuals meet the normality and equal variance assumptions.
7. Save **my_eucs** as a csv file, and read it in Excel, create a scatterplot and add a regression line as for question 3. Display the regression coefficients (intercept and slope) produced by Excel. Based on the coefficients, add a column to the file in which you calculate the predicted value for each observation, and another column where you calculate the residuals. Check that you got the same results as in

R and save the file using an Excel format (if you don't save the file as .xlsx it will not count as correct)¹.

8. Write down all your answers in a Word document and upload it together with your Excel file to Learn, before 4 August 10 am.

¹ Rationale: Funnily enough, many statistical models are fitted using statistical software, but when used by industry and government they end up implemented in Excel. This question aims to show how statistical models look like in Excel.