STAT202 Assignment 1: Review of simple linear regression

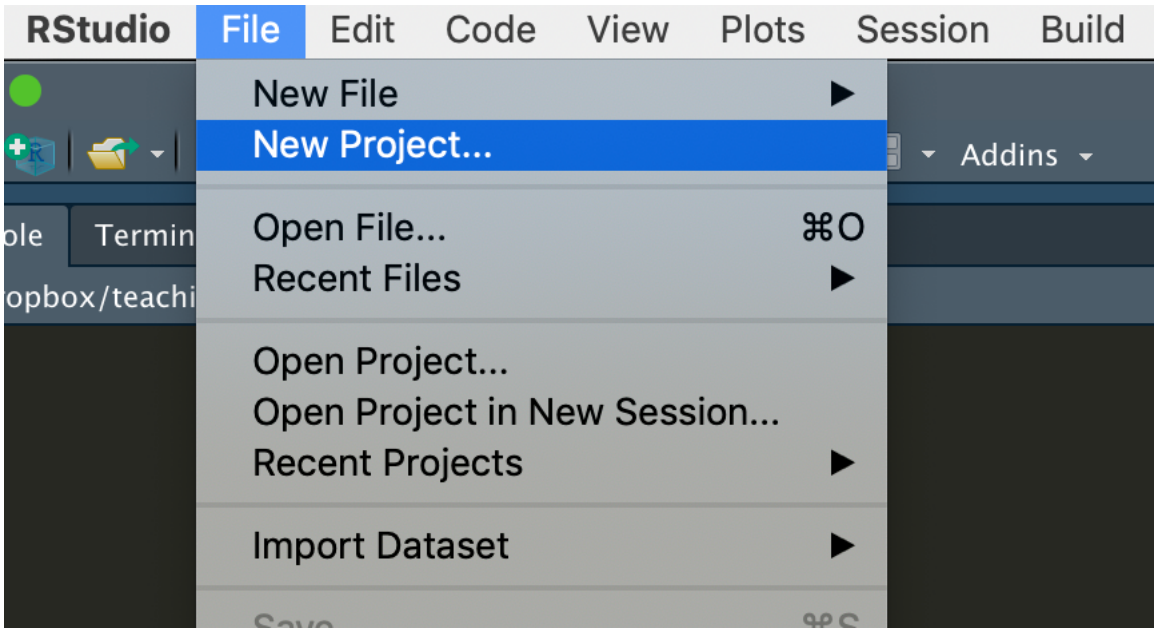
Due on Wednesday 31st July, 1 pm

You will work with a couple of datasets: **starwars** and **cineole.csv**. The first one comes with the `dplyr` package and contains vital statistics of Star Wars characters. The second dataset has data from a *Eucalyptus bosistoana* trial, with `cineole` production (micrograms/l), `month_number` since the start of the study and `leaf_type` (juvenile or mature).

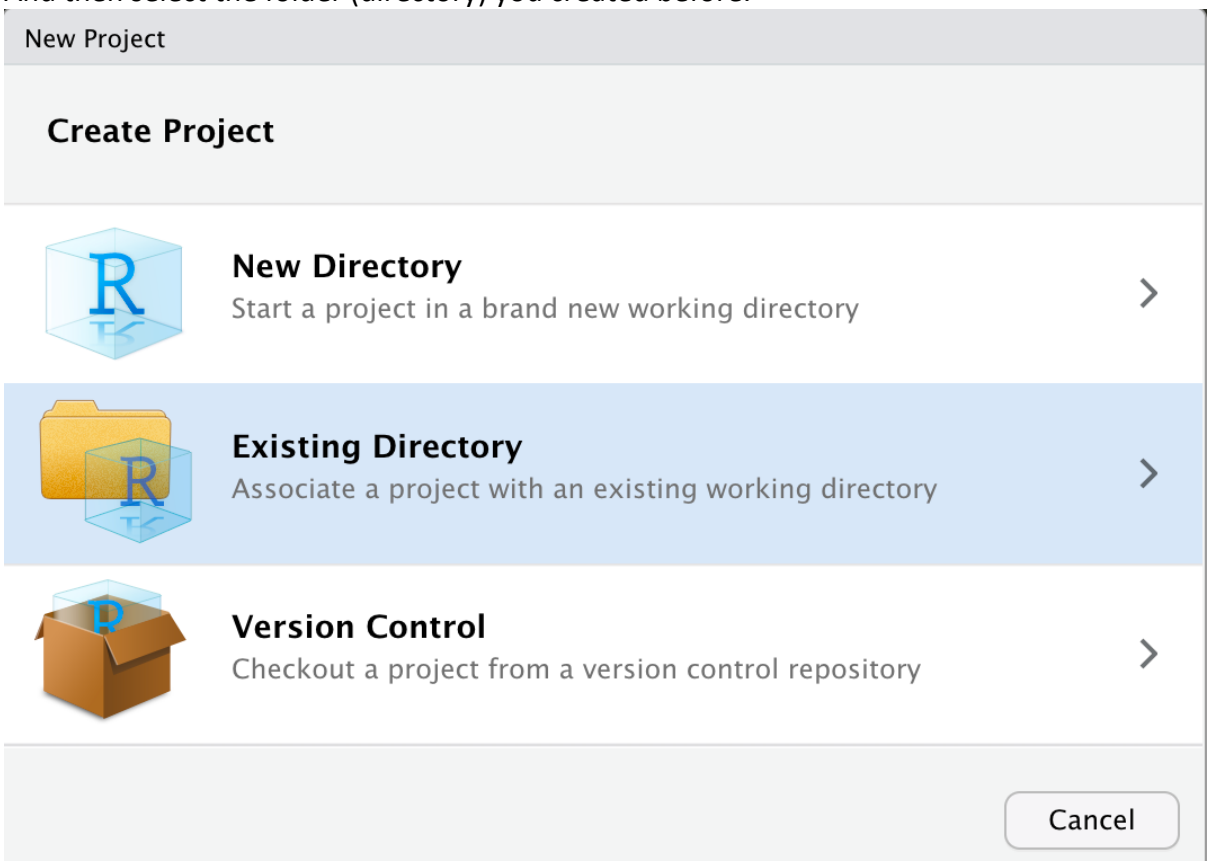
Each of you will use random subsets of the original datasets, depending on your student ID number. This means that results for the assignment are slightly (or not so slightly) different for each student.

For the assignment you will a- create a folder (as explained in lectures and lab 0), b- download the data for the assignment in that folder (in this case cineole.csv), 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.



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

`library(tidyverse)`
 	
`library(performance)`
	Makes available functions of the tidyverse, a group of easier to use R functions (read_csv, write_csv, %>%, ggplot, filter).
	
	Functions to analyse model fit and performance

`<-`	Assign objects to a name (less 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
`drop_na()`	Drop rows with missing values
`mutate(var = some function)`	Create new variables or modify variables
`ggplot(dataset, aes(x, y)) + geom_point()`	Create advanced plots
`lm(response ~ predictors,`	Fits linear models. One usually assigns this function to a name, as in

`data = dataset)`	model_1 <- lm(y ~ x, data = my_data)
`summary(object_name)`	Provides a summary of the object. Examples: summary(mydata) summary(model_1)
`check_model(object_name)`	For models produces a handy plot of residuals

1. Load the `starwars` data set into R. As `starwars` is a dataset that comes with `dplyr` (part of the `tidyverse`), it is available from the time you call the `tidyverse` package with the `data()` function as in.

```
# Load starwars data
data(starwars)
```

2. Produce a scatterplot of `mass` (in galactic mass units) on `height` (in cm). This means y-axis is mass and x-axis is height. Before running any analyses, **describe** the relationship between the two variables using poor/weak/moderate/strong and positive/negative in no more than 20 words.
3. You may not have noticed, but there are a few rows in the data that have missing `mass` or `height`. We could see that using:

```
starwars |> select(mass, height) |> summary()
```

4. Using R code drop (remove) the rows that contain any missing values for `mass` and/or `height` (hint: `drop_na()` and identify which character is the outlier observation (hint: `filter()` will be useful). Create a subset of the data that **does not** contain missing values nor the outlier, name it `starno`.
5. Now take a random sample of 50 observations based on your student code (e.g. if it were 9999999), so you keep only the data for yourself. Call that data set `my_starno`. All the following instructions apply to your work with `mystar_no`.

```
# Use YOUR student ID number to select observations
# you have to keep. For example, for 9999999
set.seed(9999999)
my_starno <- starno |> sample_n(50)
```

6. Questions 6 through 9 use the `mystar_no` dataset. Plot again `mass` on `height`. Before running any analyses, describe the relationship between the two variables using weak/moderate/strong and positive/negative in no more than 20 words.

7. Fit a linear regression of `mass` on `height`, and call it `model_1`. 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**.

8. Now center `height` (that is, express it as deviation from its mean value). You will create a new variable (using `mutate()`, call it `cent_height`) that equals `height - mean(height)` and use it to repeat step 7, naming the fit `model_2`.

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?

9. Plot the residuals for `model_2` using `check_model()`; this function comes from the `performance` package. Comment on how well your residuals meet the normality and equal variance assumptions.

10. Now you are going to work with the `cineole.csv` data set (cineole is an essential oil found in eucalypts). Reset the random number seed using your student code, read the data set and create a random sample of 500 observations called `my_cineole`.

11. Create a scatterplot of `cineole` production vs `month_number`. There is a lot of overplot (points on top of others). Try using alpha = 0.5 in `geom_point()` or `geom_jitter(width = 0.2)` to get a better idea of the number of observations. Explain the observed relationship in 50 words or so.

12. Create a new variable called `new_month` in the cineole data set. This variable is `month_number` divided by 12, which will help us to account for the periodicity of oil production. Repeat the plot from the previous question.

13. Now fit a model, call it `oil_1`, with cineole as the response and `new_month` as the predictor. Plot the residuals vs fitted values and comment on the linearity of the relationship.

14. Now fit model `oil_2`, with cineole as the response and `sin(2*pi*new_month) + cos(2*pi*new_month)` as the predictors. This is one way to fit a regular cyclical component. Write down the regression coefficients and compare `oil_1` and `oil_2` for residual standard error and how well the model fits. Explain the change in 100 words.

15. Plot the residuals vs fitted values for model `oil_2` and comment on the linearity of the relationship. No more than 50 words.

16. Write down all your answers in a Word document (either directly or using Rmarkdown), showing the code used for each question, followed by your comments. Upload it to Learn.