Assignment 4: Collaborating Together
Introduction to Applied Data Science
2022-2023

Aleksandra Tatko
a.tatko@students.uu.nl
https://github.com/otatko

April 2023

Assignment 4: Collaborating Together

Part 1: Contributing to another student's Github repository

In this assignment, you will create a Github repository, containing this document and the .pdf output, which analyzes a dataset individually using some of the tools we have developed.

This time, make sure to not only put your name and student e-mail in your Rmarkdown header, but also your Github account, as I have done myself.

However, you will also pair up with a class mate and contribute to each others' Github repository. Each student is supposed to contribute to another student's work by writing a short interpretation of 1 or 2 sentences at the designated place (this place is marked with **designated place**) in the other student's assignment.

This interpretation will not be graded, but a Github shows the contributors to a certain repository. This way, we can see whether you have contributed to a repository of a class mate.

Question 1.1: Fill in the **github username** of the class mate to whose repository you have contributed.

[Fill in here]

Part 2: Analyzing various linear models

In this part, we will summarize a dataset and create a couple of customized tables. Then, we will compare a couple of linear models to each other, and see which linear model fits the data the best, and yields the most interesting results.

We will use a dataset called `GrowthSW` from the `AER` package. This is a dataset containing 65 observations on 6 variables and investigates the determinants of economic growth. First, we will try to summarize the data using the `modelsummary` package.

```
library(AER)
data(GrowthSW)
```

One of the variables in the dataset is `revolutions`, the number of revolutions, insurrections and coup d'etats in country i from 1965 to 1995.

	Mean	Median	SD	Min	Max
growth	1.68	1.92	2.11	−2.81	7.16
rgdp60	1988.67	1259.00	1698.18	367.00	6823.00

	Mean	Median	SD	Min	Max
growth	2.46	2.29	1.28	0.42	6.65
rgdp60	5283.32	5393.00	2439.39	1374.00	9895.00

Question 2.1: Using the function `datasummary`, summarize the mean, median, sd, min, and max of the variables `growth`, and `rgdp60` between two groups: countries with `revolutions` equal to 0, and countries with more than 0 revolutions. Call this variable `treat`. Make sure to also write the resulting data set to memory. Hint: you can check some examples here.

```
library(modelsummary); library(tidyverse)

library(dplyr)

GrowthSW <- GrowthSW %>%

  mutate(treat = ifelse(GrowthSW$revolutions > 0, "revolutions", "no revolutions"))


Revolution_GrowthSW <- GrowthSW %>%

  filter(treat == "revolutions")

Norevolution_GrowthSW <- GrowthSW %>%

  filter(treat == "no revolutions")

datasummary(growth +rgdp60 ~ Mean + Median + SD + Min + Max, data = Revolution_GrowthSW)
```

```
datasummary(growth +rgdp60 ~ Mean + Median + SD + Min + Max, data = Norevolution_GrowthSW)
```

Designated place: type one or two sentences describing this table of a fellow student below. For example, comment on the mean and median growth of both groups. Then stage, commit and push it to their github repository.

Part 3: Make a table summarizing reressions using modelsummary and kable

In question 2, we have seen that growth rates differ markedly between countries that experienced at least one revolution/episode of political stability and countries that did not.

Question 3.1: Try to make this more precise this by performing a t-test on the variable growth according to the group variable you have created in the previous question.

```
library(tidyverse)

GrowthSW <- GrowthSW %>%
  mutate(treat = ifelse(revolutions > 0, "revolutions", "no revolutions"))
```

```r
Revolution_GrowthSW <- GrowthSW %>%
  filter(treat == "revolutions")

Norevolution_GrowthSW <- GrowthSW %>%
  filter(treat == "no revolutions")

t_test_result <- t.test(Revolution_GrowthSW$growth, Norevolution_GrowthSW$growth)

t_test_result
```

```
##
##  Welch Two Sample t-test
##
## data:  Revolution_GrowthSW$growth and Norevolution_GrowthSW$growth
## t = -1.8531, df = 61.015, p-value = 0.06871
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.62566475  0.06182741
## sample estimates:
## mean of x mean of y
##  1.678066  2.459985
```

Question 3.2: What is the p-value of the test, and what does that mean? Write down your answer below.

p- value is equal to 0.06871. The p-value is a measure that quantifies the probability of observing the observed difference or more extreme, assuming there is no true difference. A smaller p-value suggests stronger evidence against the null hypothesis. If the p-value is below the chosen significance level (e.g., 0.05), we reject the null hypothesis.

We can also control for other factors by including them in a linear model, for example:

$$\text{growth}_i = \beta_0 + \beta_1 \cdot \text{treat}_i + \beta_2 \cdot \text{rgdp60}_i + \beta_3 \cdot \text{tradeshare}_i + \beta_4 \cdot \text{education}_i + \epsilon_i$$

Question 3.3: What do you think the purpose of including the variable `rgdp60` is? Look at `?GrowthSW` to find out what the variables mean.

The variable rgdp60 in the GrowthSW dataset represents the Real Gross Domestic Product (GDP) per capita in the year 1960 for each country in the dataset. Real GDP per capita is a measure of economic output per person, adjusted for inflation.

Including the variable rgdp60 in the linear model serves the purpose of controlling for the initial economic conditions or economic development of each country in the analysis. By including rgdp60 as an independent variable in the model, we can examine the effect of revolutions (treat variable) on economic growth (growth variable) while accounting for the differences in initial economic conditions among countries.

We now want to estimate a stepwise model. Stepwise means that we first estimate a univariate regression $\text{growth}_i = \beta_0 + \beta_1 \cdot \text{treat}_i + \epsilon_i$, and in each subsequent model, we add one control variable.

Question 3.4: Write four models, titled `model1`, `model2`, `model3`, `model4` (using the `lm` function) to memory. Hint: you can also use the `update` function to add variables to an already existing specification.

```r
model1 <- lm(growth ~ treat, data = GrowthSW)
model2 <- update(model1, . ~ . + rgdp60)
model3 <- update(model2, . ~ . + education)
model4 <- update(model3, . ~ . + tradeshare)
```

	(1)	(2)	(3)	(4)
(Intercept)	2.460***	2.854***	1.478+	−0.050
	(0.400)	(0.751)	(0.747)	(0.967)
treatrevolutions	−0.782	−1.028	−0.527	−0.069
	(0.491)	(0.633)	(0.577)	(0.589)
rgdp60		0.000	−0.001**	0.000*
		(0.000)	(0.000)	(0.000)
education			0.612***	0.564***
			(0.148)	(0.144)
tradeshare				1.813*
				(0.765)
Num.Obs.	65	65	65	65
R2	0.039	0.045	0.254	0.318

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Now, we put the models in a list, and see what `modelsummary` gives us:

```
list(model1, model2, model3, model4) |>
  modelsummary(stars=T, gof_map = c("nobs", "r.squared"))
```

Question 3.5: Edit the code chunk above to remove many statistics from the table, but keep only the number of observations N, and the R^2 statistic.

Question 3.6: According to this analysis, what is the main driver of economic growth? Why?

Based on this analysis, it appears that the variable education consistently shows a significant positive effect on economic growth across all four models. This suggests that education could be considered as one of the main drivers of economic growth. The coefficient estimate for education consistently has a positive sign and is statistically significant in all models.

Question 3.7: In the code chunk below, edit the table such that the cells (including standard errors) corresponding to the variable `treat` have a red background and white text. Make sure to load the `kableExtra` library beforehand.

```
library(modelsummary)
library(kableExtra)
list(model1, model2, model3, model4) %>%
modelsummary(stars=T, gof_map = c("nobs", "r.squared")) %>%

  kable_styling() %>%

  row_spec(7:8, bold = F, color = "white", background = "red")
```

Question 3.8: Write a piece of code that exports this table (without the formatting) to a Word document.

```
install.packages(flextable)
```

```
## Error in eval(expr, envir, enclos): object 'flextable' not found
```

```
library(flextable)

list(model1, model2, model3, model4) %>%
  modelsummary(stars=T, gof_map = c("nobs", "r.squared"), output = "growth_table.docx")
```

	(1)	(2)	(3)	(4)
(Intercept)	2.460***	2.854***	1.478+	−0.050
	(0.400)	(0.751)	(0.747)	(0.967)
treatrevolutions	−0.782	−1.028	−0.527	−0.069
	(0.491)	(0.633)	(0.577)	(0.589)
rgdp60		0.000	−0.001**	0.000*
		(0.000)	(0.000)	(0.000)
education			0.612***	0.564***
			(0.148)	(0.144)
tradeshare				1.813*
				(0.765)
Num.Obs.	65	65	65	65
R2	0.039	0.045	0.254	0.318

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

The End