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

Wojciech Deska
w.p.deska@students.uu.nl
https://github.com/Wojtek331

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.

Username of class mate: MCPdeHaan

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.

	revolution = 0					revolution > 0				
	Mean	median	sd	min	max	Mean	median	sd	min	max
growth	2.46	2.29	1.28	0.42	6.65	1.68	1.92	2.11	−2.81	7.16
rgdp60	5283.32	5393.00	2439.39	1374.00	9895.00	1988.67	1259.00	1698.18	367.00	6823.00

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)
GrowthSW <- GrowthSW %>%
  mutate(treat = ifelse(revolutions == 0, " revolution = 0", "revolution > 0"))

datasummary <- datasummary(growth + rgdp60 ~ treat*(Mean + median + sd + min + max), data = GrowthSW)
datasummary
```

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.

If I look at the table of Wojtek, I see the that the value for GDP per capita in 1960 was 5283.32 on average for countries that had now revolutions and over two times smaller for countries that had a revolution (1988.67). This is logical, since the reason why a revolution occurred is often that something (often the economy) underperformed or the revolution had a big effect on the internal and external economy in a country, as shown by the lower average GDP per capita in 1960 for countries that had one or more revolutions.

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.

```
t_test <- t.test(growth ~ treat, data = GrowthSW)
print(t_test)
```

```
##
##  Welch Two Sample t-test
##
## data:  growth by treat
## t = 1.8531, df = 61.015, p-value = 0.06871
## alternative hypothesis: true difference in means between group  revolution = 0 and group revolution >
## 95 percent confidence interval:
##  -0.06182741  1.62566475
## sample estimates:
## mean in group  revolution = 0  mean in group revolution > 0
##                       2.459985                       1.678066
```

Question 3.2: What is the p-value of the test, and what does that mean? Write down your answer below. The p value is: 0.06871 The p value indicates the probability of observing a t statistics at extreme or higher under the assumption of null hypothesis meaning that there is no difference in means between the two groups. futhuremore the p value is greater than the significance level, therefore there is not enough evidence to reject the null hypothesis. The value of the null hypothesis, which states that the true difference in means between group revolution = 0 also lies within the 95% confidence interval which again runs the conclusion that we fail to reject null hypothesis that states that the true difference in means between group is = 0.

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 value means the GDP per capita in 1960, converted to 1960 US dollars, AS you need this as a base level to see how it changes the output in the future for longer period of time. Also USA had the highest GDP in 1960 in USD out of all countries, therefore it is a good estimate and currency to compare and to see the changes as the currency is a pretty stable one for a long period of time and most likely you want to compare to something that is the easiest to compare and as the currency of the leading country was USD it might be beneficial to keep it like that as all of the governmental data in USA is mostly stated in USA. The year is also a pretty good choice to choose and see the changes, as the World Wars were over and the economy and GDP's started to become more stable, as the uncertainty in the world became to be lower and there were no as big and vital events after the Wars, so it is a good time to compare the values to, to see how the economy and in this case especially GDP is changing. Another thing is that GDP or gdp growth is also one of the variables and measures of how economy is performing that might contribute to economic growth, therefore by looking at the values we might find a basic understanding of the potential economic growth or stability or overal performance that might be related to gdp once again by looking whether the values would change over the time will again indicate the overall look of the economy comparing to 1960. Also in the above model it shows that it is one of the factors that contribute to economic growth.

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.

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

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? according to this analysis education is the main driver of economic growth, as it has the biggest impact on the rsquared meaning the bigges correlation value between variables, which is possibly the best estimate as it shows the strength of the value, and mostly making it precise keeping in mind the same number of observations, meaning

	(1)	(2)	(3)	(4)
(Intercept)	2.460***	2.854***	0.839	−0.050
	(0.400)	(0.751)	(1.045)	(0.967)
treatrevolution > 0	−0.782	−1.028	−0.415	−0.069
	(0.491)	(0.633)	(0.647)	(0.589)
rgdp60		0.000	0.000	0.000*
		(0.000)	(0.000)	(0.000)
tradeshare			2.233*	1.813*
			(0.842)	(0.765)
education				0.564***
				(0.144)
Num.Obs.	65	65	65	65
R2	0.039	0.045	0.143	0.318

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

	(1)	(2)	(3)	(4)
(Intercept)	2.460***	2.854***	0.839	−0.050
	(0.400)	(0.751)	(1.045)	(0.967)
treatrevolution > 0	−0.782	−1.028	−0.415	−0.069
	(0.491)	(0.633)	(0.647)	(0.589)
rgdp60		0.000	0.000	0.000*
		(0.000)	(0.000)	(0.000)
tradeshare			2.233*	1.813*
			(0.842)	(0.765)
education				0.564***
				(0.144)
Num.Obs.	65	65	65	65
R2	0.039	0.045	0.143	0.318

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

that the results are based on the same sample size, and as rsquared is the proportion of the variation in the dependent variable that in here is the economic growth that is predictable from the independent variable in this case education shows the greatest impact by more signigicant impact after adding education.

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(kableExtra)
library(modelsummary)
list(model1, model2, model3, model4) |>
  modelsummary(stars=T, gof_map = c("nobs", "r.squared")) |>
  kable_styling() |>
  row_spec(row = 3, background = "red", color = "white") |>
  row_spec(row = 4, background = "red", color ="white")
```

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

```
sjPlot::tab_df(table, file = "tableex3.8.doc")
```

```
## Error in (function (..., row.names = NULL, check.rows = FALSE, check.names = TRUE, : argument is miss
```

The End