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

Maurits de Haan
m.c.p.dehaan@students.uu.nl
https://github.com/MCPdeHaan

20th of June, 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. The username of the class mate I did the project with is: **Wojtek331**

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.

	No revolution					Revolution				
	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, "No revolution", "Revolution"))

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

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.

ANSWER TO DESIGNATED QS: In Maurits table we can see mean, median, sd, max and min values for both growth and rgdp60. I will focus on the analysis on mean and median. The mean of the growth is higher in in the countries that have no revolution = 2.46 than the countries that had revolution = 1.68, and the median value is also higher for the countries where there was no revolution 2.29, as they have slow steady growth, whereas in the revolution one 1.92 there is often a high jump rather than slow steady growth. In both we observe that countries with revolution have lower median and mean as the growth rate can increase a lot during shorter period rather than continues growth.

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 No revolution and group Revolution is
## 95 percent confidence interval:
##  -0.06182741  1.62566475
## sample estimates:
## mean in group No revolution    mean in group Revolution
##                    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 that the t-test gives is **0.06871**. The goal of the p-value is to show what the probability is that we obtain the observed results, while assuming that the null hypothesis (H_0) is true. The general rule is: "*The lower the p-value, the greater the statistical significance of the observed difference.*" (Alwan et. al., 2020). So, the p-value is greater than 0.05 which means that we cannot reject the null hypothesis (H_0) that there is no significant difference between "no revolution" and "revolution" in terms of the growth variable. This suggests that there is not enough evidence to conclude that the presence of a revolution has a significant impact on economic growth, at least based on the available data.

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 purpose of including the variable `rgdp60` in the linear model is to control for the initial level of economic development. This is because the initial level of economic development, as shown by the GDP per capita in 1960, can have a significant impact on economic growth in the future, and this model wants to make sure that we are not just capturing the effect of the revolution on economic growth through the initial level of economic development.

The goal of this is, is to allows us to isolate the effect of the revolution on economic growth from the effect of the initial level of economic development. This is important because we want to know whether the revolution has a causal effect on economic growth, and we can only do this if we control for other factors that could also be affecting 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.

```
library(AER)

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,
  omit = c("AIC", "BIC", "Log-Likelihood", "Adj. R-squared", "Standard errors"),
  keep = c("N", "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?

The main driver of economic growth is `education`, since the coefficient on the variable `education` is positive and statistically significant in all four models. This suggests that an increase in education leads to an

	(1)	(2)	(3)	(4)
(Intercept)	2.460***	2.854***	0.839	−0.050
	(0.400)	(0.751)	(1.045)	(0.967)
treatRevolution	−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
R2 Adj.	0.023	0.014	0.101	0.272
AIC	270.1	271.7	266.6	253.8
BIC	276.7	280.4	277.5	266.9
Log.Lik.	−132.069	−131.867	−128.319	−120.918
F	2.532	1.446	3.403	6.989
RMSE	1.85	1.84	1.74	1.55

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

increase in economic growth. The other variables in the model also have some impact on economic growth, but the impact of education is the strongest. This is likely because education increases the skills of the workforce, which makes children more productive in the future. Overall, the results of this analysis suggest that education is a key driver of economic growth. Policies that promote education are likely to lead to higher rates of economic growth in the long run.

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)

table <- list(model1, model2, model3, model4) |>
  modelsummary(
    gof_map = c("nobs", "r.squared"),
  ) %>%
  row_spec(row = 3:4, background = "red", color = "white")
table
```

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

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

X. . . begin.table..n..centering.n..begin.tabular..t..lcccc..n..toprule.n. . . . .1. . . . .2. . . . .3. . . . .4. . . . . n..midrule.n.Intercept. . . . . nu

The End

	(1)	(2)	(3)	(4)
(Intercept)	2.460	2.854	0.839	−0.050
	(0.400)	(0.751)	(1.045)	(0.967)
treatRevolution	−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

	(1)	(2)	(3)	(4)
(Intercept)	2.460	2.854	0.839	−0.050
	(0.400)	(0.751)	(1.045)	(0.967)
treatRevolution	−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