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

Candela Milikowsky
m.c.milikowskyfernandez.@students.uu.nl
http://www.github.com/CandelaMilikowsky

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.

[julka-dybala]

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, "Revolutionary", "Non-revolutionary"))


RevGrowthSW <- GrowthSW |>
  filter(treat == "Revolutionary")

UnrevgrowthSW <- GrowthSW |>
  filter(treat == "Non-revolutionary")



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

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

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.

Julia: There are two tables: one represents the number of countries that experienced revolution and the other that has not. The growth mean for the first one was equal to 1.68 and was smaller than second one 2.46.

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.

```r
t.test(growth ~ treat, data = GrowthSW)
```

```
##
##  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 Non-revolutionary and group Revolutio
## 95 percent confidence interval:
##  -0.06182741  1.62566475
## sample estimates:
## mean in group Non-revolutionary     mean in group Revolutionary
##                        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.

In this case the p value is 0.06871, this explains that there is a 6.871% of probability that there is no significant difference between revolutionary and non-revolutionary groups in terms of the variable "growth".

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.

rgdp60 is a variable that lists the value of GDP per capita in 1960, converted to 1960 US dollars. By including this variable in the linear model, we can examine the effect of the initial economic conditions of countries in 1960, by knowing this initial conditions, we can explore the study the impact of other independent variables, such as the variable "treat", "tradeshare", and "education" on the economic growth. As conclusion, the use of this variable in the linear model, helps us calculate the total growth of a country as we know their GDP per capita in the 60's, and to help isolate the effect of this different variables on growth, separate from the influence of the initial economic situation.

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 ~ education, data = GrowthSW)

model2 <- lm(growth ~ education + tradeshare, data = GrowthSW)

model3 <- lm(growth ~ education + tradeshare + treat, data = GrowthSW)

model4 <- lm(growth ~ education + tradeshare + treat + rgdp60, data = GrowthSW)
```

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

```r
list(model1, model2, model3, model4) |>
  modelsummary(stars=T,
# edit this to remove the statistics other than R-squared
# and N
)
```

	(1)	(2)	(3)	(4)
(Intercept)	0.958*	−0.370	−0.978	−0.050
	(0.418)	(0.570)	(0.935)	(0.967)
education	0.247**	0.250**	0.304**	0.564***
	(0.089)	(0.083)	(0.106)	(0.144)
tradeshare		2.331**	2.476**	1.813*
		(0.728)	(0.751)	(0.765)
treatRevolutionary			0.471	−0.069
			(0.573)	(0.589)
rgdp60				0.000*
				(0.000)
Num.Obs.	65	65	65	65
R2	0.110	0.236	0.244	0.318
R2 Adj.	0.095	0.211	0.207	0.272
AIC	265.2	257.2	258.5	253.8
BIC	271.7	265.9	269.4	266.9
Log.Lik.	−129.578	−124.605	−124.247	−120.918
F	7.752	9.571	6.572	6.989
RMSE	1.78	1.65	1.64	1.55

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

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.

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

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

Education is the most significant, because it shows a statistically significant positive coefficient (all the coefficient estimates of education are statistically significant at the 0.01 level in the four models) , suggesting that education is the main driver of economic growth, meaning that higher levels of education are associated with higher economic growth rates.

Question 3.7: In the code chunk below, edit the table such that the cells (including standard errors)

	(1)	(2)	(3)	(4)
(Intercept)	0.958*	−0.370	−0.978	−0.050
	(0.418)	(0.570)	(0.935)	(0.967)
education	0.247**	0.250**	0.304**	0.564***
	(0.089)	(0.083)	(0.106)	(0.144)
tradeshare		2.331**	2.476**	1.813*
		(0.728)	(0.751)	(0.765)
treatRevolutionary			0.471	−0.069
			(0.573)	(0.589)
rgdp60				0.000*
				(0.000)
Num.Obs.	65	65	65	65
R2	0.110	0.236	0.244	0.318

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

	(1)	(2)	(3)	(4)
(Intercept)	0.958*	−0.370	−0.978	−0.050
	(0.418)	(0.570)	(0.935)	(0.967)
education	0.247**	0.250**	0.304**	0.564***
	(0.089)	(0.083)	(0.106)	(0.144)
tradeshare		2.331**	2.476**	1.813*
		(0.728)	(0.751)	(0.765)
treatRevolutionary			0.471	−0.069
			(0.573)	(0.589)
rgdp60				0.000*
				(0.000)
Num.Obs.	65	65	65	65
R2	0.110	0.236	0.244	0.318

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

corresponding to the variable `treat` have a red background and white text. Make sure to load the `kableExtra` library beforehand.

```
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.

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

The End