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

Morris Trijbits
7317336@students.uu.nl
http://www.github.com/Trijgits

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.

[j-szymborski]

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"))

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

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



datasummary(growth + rgdp60  ~ mean + median + sd + min + max, data = GrowthSW_rev)
```

```
datasummary(growth + rgdp60  ~ mean + median + sd + min + max, data = GrowthSW_non_rev)
```

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.

```
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.

the p-value = 0.06871. meaning that there is a 6.871% chance of obtaining a result that has no significant meaning. or in this case, a 6.871% chance that there is no significant difference in growth between countries that have had or not had revolutions.

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 the value of GDP per capita in 1960, converted to 1960 US dollars. the purpose of this could be to see if a country was already "developed". since a developed country would show less growth from 1965 to 1995 than a country with a lower GDP per capita that is still in devlopment. thus being an explanation for a lower growth rate.

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 ~ 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)


model1
```

```
##
## Call:
## lm(formula = growth ~ education, data = GrowthSW)
##
## Coefficients:
## (Intercept)    education
##      0.9583       0.2470
```

```
model2
```

```
##
## Call:
## lm(formula = growth ~ education + tradeshare, data = GrowthSW)
##
## Coefficients:
## (Intercept)    education   tradeshare
##     -0.3702       0.2500       2.3313
```

```
model3
```

```
##
## Call:
## lm(formula = growth ~ education + tradeshare + treat, data = GrowthSW)
##
## Coefficients:
##        (Intercept)           education          tradeshare  treatRevolutionary
##            -0.9779              0.3038              2.4762              0.4709
```

```
model4
```

```
##
## Call:
## lm(formula = growth ~ education + tradeshare + treat + rgdp60,
##     data = GrowthSW)
##
## Coefficients:
##        (Intercept)           education          tradeshare  treatRevolutionary
##         -0.0498162           0.5641862           1.8129261          -0.0689992
##             rgdp60
##         -0.0003976
```

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")
# edit this to remove the statistics other than R-squared
# and N
)
```

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?

#the main driver of economic growth would be education, because according to the modelsummary it has the most significant effect (statistically)

	(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

	(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

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)
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")
```

```
# use functions from modelsummary to edit this table
```

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