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

Marija Tuneska
m.tuneska@students.uu.nl
http://www.github.com/marija-tuneska

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.

Answer: I have contributed to a github repository **marija-tuneska-yahoo**. This is my second github username, my primary one is **marija-tuneska**. Both contain repository of same content.

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.

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)

treat_yes_no <- if_else(GrowthSW$revolutions>0, 'yes', 'no')
treat <- if_else(GrowthSW$revolutions>0, 1, 0)
data_set <- mutate(GrowthSW, treat_yes_no, treat)
datasummary(treat_yes_no * (growth + rgdp60) ~ mean + median + sd + min + max, data = data_set)
```

treat_yes_no		mean	median	sd	min	max
no	growth	2.46	2.29	1.28	0.42	6.65
	rgdp60	5283.32	5393.00	2439.39	1374.00	9895.00
yes	growth	1.68	1.92	2.11	-2.81	7.16
	rgdp60	1988.67	1259.00	1698.18	367.00	6823.00

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.

The mean and the median growth in countries with no revolutions is higher then in the countries with revolution, while the standard deviation of the growth is lower. So, the countries without revolutions with stable democratic system are expected to have higher growth, then the countries with 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.

I create two new data sets `treat_no` and `treat_yes`, containing data of countries with no revolutions, and with revolution, respectively. Then I perform t-test, testing the alternative hypothesis that mean growth of countries without revolution is bigger then the mean growth of countries with revolution, against the null hypothesis that they are equal. I assume that the variance in growth in those two sets of data is equal.

```
treat_no <- filter(data_set, treat==0)
treat_yes <- filter(data_set, treat==1)
t.test(treat_no$growth, treat_yes$growth, var.equal = TRUE, alternative = "greater")
```

```
##
##  Two Sample t-test
##
## data:  treat_no$growth and treat_yes$growth
## t = 1.5911, df = 63, p-value = 0.0583
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -0.03849276         Inf
## sample estimates:
## mean of x mean of y
##  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.

Answer: The p-value of the test is 0.0583, meaning that based on the data we have for the growth, the probability that the mean growth in countries with no revolution will be bigger that the mean growth in countries with revolution, is equal to 0.0583.

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.

Answer: Variable `rgdp60` is the value of GDP per capita in 1960, converted to 1960 US dollars. It is included in order to allow comparisons between countries at that time period. It adjusts for inflation and exchange rate fluctuations over time. This means that the data can be compared across countries and over time without being affected by changes in currency values or inflation rates.

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 = data_set)
model2 <- update(model1, . ~ . + rgdp60)
model3 <- update(model2, . ~ . + tradeshare)
model4 <- update(model3, . ~ . + education)
```

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

```
list(model1, model2, model3, model4) |>
  modelsummary()
```

	(1)	(2)	(3)	(4)
(Intercept)	2.460	2.854	0.839	-0.050
	(0.400)	(0.751)	(1.045)	(0.967)
treat	-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

	(1)	(2)	(3)	(4)
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

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.

Answer: Edited code chunk from Question 3.4 is

```
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
)
```

	(1)	(2)	(3)	(4)
(Intercept)	2.460***	2.854***	0.839	-0.050
	(0.400)	(0.751)	(1.045)	(0.967)
treat	-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

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

Answer: According to this analysis, the main driver of economic growth is tradeshare because in the linear model the coefficient for the variable `tradeshare` is the highest positive one (1.813). In the dataset `GrowthSW` the variable `tradeshare` is calculated as the average share of trade in the economy from 1960 to 1995, measured as the sum of exports (X) plus imports (M), divided by GDP; that is, the average value of (X + M)/GDP from 1960 to 1995.

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)
```

```
## Error: package or namespace load failed for 'kableExtra':
##  .onLoad failed in loadNamespace() for 'kableExtra', details:
##   call: !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output %in%
##   error: 'length = 2' in coercion to 'logical(1)'
```

```
list(model1, model2, model3, model4) |>
    modelsummary(stars=T, gof_map = c("nobs", "r.squared")) %>%
row_spec(c(7,8), background  = "red", color = "white")
```

```
## Error in row_spec(., c(7, 8), background = "red", color = "white"): could not find function "row_spe
```

```
# 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.

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

The End