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

Giovanna Tullume Carrion
g.e.tullumecarrion@student.uu.nl
http://www.github.com/GioviEli

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.

[VendelLantos]

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.

treat		mean	median	sd	min	max
no revolution	growth	2.46	2.29	1.28	0.42	6.65
	rgdp60	5283.32	5393.00	2439.39	1374.00	9895.00
revolution	growth	1.68	1.92	2.11	−2.81	7.16
	rgdp60	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)

# write your code here
GrowthSW$treat <- if_else(GrowthSW$revolutions > 0, "revolution", "no revolution")
data_summary <- datasummary(
  treat * (growth + rgdp60) ~ mean + median + sd + min + max,
  data=GrowthSW
)
data_summary
```

Designated place: Both the mean and median growth was higher for countries with 0 revolutions, than the countries with more than 0.

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.

```
# write t test here
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 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 is 0.06871. This means that under the assumption that the null hypothesis is true(which states that the true difference in means between group 0 Revolutions and group More than 0 Revolutions is not equal to 0) , there is a 6.871% chance of obtaining a test statistic as extreme as the one observed. The

	(1)	(2)	(3)	(4)
(Intercept)	2.460***	2.854***	0.839	−0.050
	(0.400)	(0.751)	(1.045)	(0.967)
GrowthSW$treatrevolution	−0.782	−1.028	−0.415	−0.069
	(0.491)	(0.633)	(0.647)	(0.589)
GrowthSW$rgdp60		0.000	0.000	0.000*
		(0.000)	(0.000)	(0.000)
GrowthSW$tradeshare			2.233*	1.813*
			(0.842)	(0.765)
GrowthSW$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

p-value is above 0.05 so we do not have sufficient evidence to reject the null hypothesis. The 95% confidence interval provides a range of plausible values for the true difference in means between the 2 groups. It is from -0.06182741 to 1.62566475. It suggests that with 95% confidence, the true difference in means falls within this interval.

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 shows the GDP of the country in 1960. Growth shows the growth in GDP. rgdp60 is added to know what is the initial value of the GDP, because only a growth does not say to much if you do not know where they originilally came from.

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(GrowthSW$growth ~ GrowthSW$treat)
model2 <- update(model1, . ~ . + GrowthSW$rgdp60)
model3 <- update(model2, . ~ . + GrowthSW$tradeshare)
model4 <- update(model3, . ~ . + GrowthSW$education)
```

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

```r
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.

	(1)	(2)	(3)	(4)
(Intercept)	2.460***	2.854***	0.839	−0.050
	s.e.=0.400	s.e.=0.751	s.e.=1.045	s.e.=0.967
GrowthSW$treatrevolution	−0.782	−1.028	−0.415	−0.069
	s.e.=0.491	s.e.=0.633	s.e.=0.647	s.e.=0.589
GrowthSW$rgdp60		0.000	0.000	0.000*
		s.e.=0.000	s.e.=0.000	s.e.=0.000
GrowthSW$tradeshare			2.233*	1.813*
			s.e.=0.842	s.e.=0.765
GrowthSW$education				0.564***
				s.e.=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?

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)
summary_table_formatted <- list(model1, model2, model3, model4) |>
  modelsummary(
    stars=T,
    gof_map = c("nobs", "r.squared"),
    statistic=c("s.e.={std.error}")
  ) |>
  row_spec(3,color='white',background='red') |>
  row_spec(4,color='white',background='red')
summary_table_formatted
```

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),
  gof_map=c("nobs", "r.squared"),
  title="Regression Table",
  output='table_1.docx'
)
```

The End