

Assignment 4: Collaborating Together

Introduction to Applied Data Science

2022-2023

Jodi Tay Seow Xuan
j.joditayseowxuan@students.uu.nl
<http://www.github.com/joditay>

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.

[Fill in here]

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'états 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)

GrowthSW$treat <- ifelse(GrowthSW$revolutions > 0, "yes", "no")

datasummary((mean + median + sd + min + max) ~ (growth+rgdp60) * ((revolutions > 0) + (revolutions == 0)))
```

	growth / revolutions > 0	growth / revolutions == 0
## 1 mean	1.68	2.46
## 2 median	1.92	2.29
## 3 sd	2.11	1.28
## 4 min	-2.81	0.42
## 5 max	7.16	6.65
	rgdp60 / revolutions > 0	rgdp60 / revolutions == 0
## 1	1988.67	5283.32
## 2	1259.00	5393.00
## 3	1698.18	2439.39
## 4	367.00	1374.00
## 5	6823.00	9895.00

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.

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.

```
GrowthSW <- GrowthSW %>%
  mutate(revno = ifelse(revolutions == 0, growth, NA),
         revyes = ifelse(revolutions > 0, growth, NA))

t.test(na.omit(GrowthSW$revno), na.omit(GrowthSW$revyes))
```

```
##
## Welch Two Sample t-test
##
## data:  na.omit(GrowthSW$revno) and na.omit(GrowthSW$revyes)
## t = 1.8531, df = 61.015, p-value = 0.06871
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
## -0.06182741  1.62566475
## sample estimates:
## mean of x mean of y
## 2.459985  1.678066
```

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

Question 3.2: What is the p -value of the test, and what does that mean? Write down your answer below.

The p -value of the test is 0.06871. For a test at the 5% level of significance, the result is statistically insignificant as the p -value is more than 0.05.

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. `rgdp60` can be used as a predictor or explanatory variable in regression models to explore the determinants of 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.

```
model1 <- lm(growth ~ treat, data = GrowthSW)
model2 <- update(model1, growth ~ treat + rgdp60)
model3 <- update(model2, growth ~ treat + rgdp60 + tradeshare)
model4 <- update(model3, growth ~ treat + rgdp60 + tradeshare + education)
```

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

```
list(model1, model2, model3, model4) |>
  modelsummary(stars=T, statistics = c("R-squared", "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.

	(1)	(2)	(3)	(4)
(Intercept)	2.460*** (0.400)	2.854*** (0.751)	0.839 (1.045)	-0.050 (0.967)
treatyes	-0.782 (0.491)	-1.028 (0.633)	-0.415 (0.647)	-0.069 (0.589)
rgdp60		0.000 (0.000)	0.000 (0.000)	0.000* (0.000)
tradeshare			2.233* (0.842)	1.813* (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?

The main driver of economic growth is education. It has the highest r-squared value, and thus has the highest level of correlation with the growth variable.

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")) |>
  row_spec(3, color = "white", background = "red") |>
  row_spec(4, 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, statistics = c("R-squared", "N"), output = "table.docx")
```

```
## Error: Package 'flextable' required for this function to work.
## Please install it by running 'install.packages('flextable')'.
```

The End