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.

[Fill in here]：Vicky0129

### 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 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](https://vincentarelbundock.github.io/modelsummary/articles/datasummary.html#datasummary).

library(modelsummary); library(tidyverse)  
GrowthSW$treat <- ifelse(GrowthSW$revolutions == 0, "0", " > 0 ")  
GrowthSW\_table <- datasummary(formula = growth + rgdp60 ~ Mean + Median + SD + Min + Max,  
 by = "treat",  
 data = GrowthSW)  
GrowthSW\_table

|  | Mean | Median | SD | Min | Max |
| --- | --- | --- | --- | --- | --- |
| growth | 1.94 | 1.98 | 1.90 | -2.81 | 7.16 |
| rgdp60 | 3103.78 | 2019.00 | 2512.66 | 367.00 | 9895.00 |

# write your code here

**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.

Based on the provided table, the results of the Welch Two Sample t-test suggest that there is not enough evidence to conclude that the true difference in means between group 0 and group more than 0 is significantly different from zero. The p-value of 0.06871 indicates a relatively high chance of observing the observed difference in means under the null hypothesis of no difference. The 95% confidence interval for the difference in means ranges from -0.06182741 to 1.62566475, encompassing zero, further supporting the lack of a significant difference. The mean growth in group 0 is 2.459985, while in group more than 0, it is 1.678066. The median growth rates in countries with more than 0 revolutions are slightly higher compared to countries with 0 revolutions, further supporting the finding that there is a trend of higher growth in the former group.

### 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 <- t.test(GrowthSW$growth ~ treat, data=GrowthSW)  
t\_test

##   
## Welch Two Sample t-test  
##   
## data: GrowthSW$growth by treat  
## t = -1.8531, df = 61.015, p-value = 0.06871  
## alternative hypothesis: true difference in means between group > 0 and group 0 is not equal to 0  
## 95 percent confidence interval:  
## -1.62566475 0.06182741  
## sample estimates:  
## mean in group > 0 mean in group 0   
## 1.678066 2.459985

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

With a p-value of 0.06871, which is slightly above the commonly used significance level of 0.05, we do not have strong statistical evidence to reject the null hypothesis. This suggests that the observed difference in mean growth rates between the groups with more than 0 revolutions and the group with 0 revolutions could potentially be due to random chance or sampling variability. Therefore, we cannot confidently claim that there is a significant difference in growth rates between these groups based on the available data. We can also control for other factors by including them in a linear model, for example:

**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.

The purpose of including the variable rgdp60 is to control for the initial level of real GDP per capita in 1960. By including this variable in the regression model, we can assess the impact of revolutions or political stability on economic growth while accounting for the initial economic conditions. This allows us to examine whether the effect of revolutions on growth is independent of the initial economic development.

We now want to estimate a stepwise model. Stepwise means that we first estimate a univariate regression , 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, . ~ . + rgdp60, data = GrowthSW)  
model3 <- update(model2, . ~ . + tradeshare, data = GrowthSW)  
model4 <- update(model3, . ~ . + education, data = GrowthSW)

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

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

|  | (1) | (2) | (3) | (4) |
| --- | --- | --- | --- | --- |
| (Intercept) | 1.678\*\*\* | 1.826\*\*\* | 0.425 | -0.119 |
|  | (0.286) | (0.374) | (0.638) | (0.590) |
| treat0 | 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 |
| 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 3.5**: Edit the code chunk above to remove many statistics from the table, but keep only the number of observations , and the statistic.

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

According to the provided model summary, it can be observed that the variable “education” is the primary factor influencing economic growth. In model 4, the coefficient associated with “education” demonstrates high statistical significance (p < 0.001), indicating a strong relationship between education and economic growth.

Furthermore, when comparing the coefficients of determination (R²) among the four models, it is evident that model 4 exhibits the highest value of 0.318. This indicates that approximately 31.8% of the variance in economic growth can be explained by the variables included in model 4. Therefore, model 4 outperforms the other models in terms of its ability to effectively explain and predict economic growth.

These findings emphasize the importance of education as a significant driver of economic growth, as supported by the high significance of its coefficient and the strong explanatory power of model 4.

**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':  
## loadNamespace()里算'kableExtra'时.onLoad失败了，详细内容：  
## 调用: !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output %in%   
## 错误: 'length = 2' in coercion to 'logical(1)'

table <- list(model1, model2, model3, model4) |>  
 modelsummary(stars=T, gof\_map = c("nobs", "r.squared")) %>%   
 row\_spec(3:4, bold = TRUE, color = "white", background = "red")

## Error in row\_spec(., 3:4, bold = TRUE, color = "white", background = "red"): 没有"row\_spec"这个函数

print(table)

## function (..., exclude = if (useNA == "no") c(NA, NaN), useNA = c("no",   
## "ifany", "always"), dnn = list.names(...), deparse.level = 1)   
## {  
## list.names <- function(...) {  
## l <- as.list(substitute(list(...)))[-1L]  
## if (length(l) == 1L && is.list(..1) && !is.null(nm <- names(..1)))   
## return(nm)  
## nm <- names(l)  
## fixup <- if (is.null(nm))   
## seq\_along(l)  
## else nm == ""  
## dep <- vapply(l[fixup], function(x) switch(deparse.level +   
## 1, "", if (is.symbol(x)) as.character(x) else "",   
## deparse(x, nlines = 1)[1L]), "")  
## if (is.null(nm))   
## dep  
## else {  
## nm[fixup] <- dep  
## nm  
## }  
## }  
## miss.use <- missing(useNA)  
## miss.exc <- missing(exclude)  
## useNA <- if (miss.use && !miss.exc && !match(NA, exclude,   
## nomatch = 0L))   
## "ifany"  
## else match.arg(useNA)  
## doNA <- useNA != "no"  
## if (!miss.use && !miss.exc && doNA && match(NA, exclude,   
## nomatch = 0L))   
## warning("'exclude' containing NA and 'useNA' != \"no\"' are a bit contradicting")  
## args <- list(...)  
## if (length(args) == 1L && is.list(args[[1L]])) {  
## args <- args[[1L]]  
## if (length(dnn) != length(args))   
## dnn <- paste(dnn[1L], seq\_along(args), sep = ".")  
## }  
## if (!length(args))   
## stop("nothing to tabulate")  
## bin <- 0L  
## lens <- NULL  
## dims <- integer()  
## pd <- 1L  
## dn <- NULL  
## for (a in args) {  
## if (is.null(lens))   
## lens <- length(a)  
## else if (length(a) != lens)   
## stop("all arguments must have the same length")  
## fact.a <- is.factor(a)  
## if (doNA)   
## aNA <- anyNA(a)  
## if (!fact.a) {  
## a0 <- a  
## op <- options(warn = 2)  
## on.exit(options(op))  
## a <- factor(a, exclude = exclude)  
## options(op)  
## }  
## add.na <- doNA  
## if (add.na) {  
## ifany <- (useNA == "ifany")  
## anNAc <- anyNA(a)  
## add.na <- if (!ifany || anNAc) {  
## ll <- levels(a)  
## if (add.ll <- !anyNA(ll)) {  
## ll <- c(ll, NA)  
## TRUE  
## }  
## else if (!ifany && !anNAc)   
## FALSE  
## else TRUE  
## }  
## else FALSE  
## }  
## if (add.na)   
## a <- factor(a, levels = ll, exclude = NULL)  
## else ll <- levels(a)  
## a <- as.integer(a)  
## if (fact.a && !miss.exc) {  
## ll <- ll[keep <- which(match(ll, exclude, nomatch = 0L) ==   
## 0L)]  
## a <- match(a, keep)  
## }  
## else if (!fact.a && add.na) {  
## if (ifany && !aNA && add.ll) {  
## ll <- ll[!is.na(ll)]  
## is.na(a) <- match(a0, c(exclude, NA), nomatch = 0L) >   
## 0L  
## }  
## else {  
## is.na(a) <- match(a0, exclude, nomatch = 0L) >   
## 0L  
## }  
## }  
## nl <- length(ll)  
## dims <- c(dims, nl)  
## if (prod(dims) > .Machine$integer.max)   
## stop("attempt to make a table with >= 2^31 elements")  
## dn <- c(dn, list(ll))  
## bin <- bin + pd \* (a - 1L)  
## pd <- pd \* nl  
## }  
## names(dn) <- dnn  
## bin <- bin[!is.na(bin)]  
## if (length(bin))   
## bin <- bin + 1L  
## y <- array(tabulate(bin, pd), dims, dimnames = dn)  
## class(y) <- "table"  
## y  
## }  
## <bytecode: 0x000001db085a62c8>  
## <environment: namespace:base>

# 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 = "table.docx")

## The End