Quantitative Methods for Business and Social Science

(BPOLO2010E) - Home assignment (UC)

Final Exam

Copenhagen Business School

Pages: 10 (front page and appendix excluded)
Reference system: APA, 7th edition
Examiner: Zoltan Fazekas
Student number:
Characters: 20 732

25th of May 2022

Q 1.1 The main question of this research concerns whether politicians are more willing to learn from the experiences of their own country's government rather than from experiences from other countries' governments where the report in question completed an experimental approach (Butler et al., 2019). Opposed to the experimental approach, the observational research does not contain interventions. An example of an observational study in this case could be to observe how politicians either include statistics from their country or from other countries when carrying out new legislation. Then one might examine whether the proportion of using own country statistics is the biggest. However, in an experiment, the treatment assignment can be randomized which ensures the internal validity of the experiment. This cannot be conducted in the observational study which on the other hand then represents better external validity as the terms of the study is not an experimental setup. This central shortcoming related to the lack for randomization means that there is a possibility of cofounding bias (self-selection bias) which means that findings in the study can be attributed to other factors than the chosen treatment. This could be approached by statistical control such as subclassification of only governmental legislators in Europe concerned with environmental policies in an attempt to make both treatment groups and control groups more similar.

Q 1.2 Firstly, the case states, that the samples from the experiment was based on a correct randomization which enables conclusions in the following without a high probability of bias.
The first part of the experiment contains data from 9 different countries where a different number of municipalities were contacted and a random selection of them were exposed to the treatment of emails concerning the suggestions on addressing climate change. Throughout question one, many of the calculations have been based creation of the data frames in R for the purpose of using the software for the calculations. For this first data frame, the three variables describe the given country, the number of municipalities contacted, and the number of emails opened. Overall, the average proportion of emails opened in the study was 23.57% which was calculated as the total number of emails opened relative to the number of municipalities contacted. Throughout the assignment formulas for calculations will only be introduced once and then appear only in the appendix.

$$p_{emails\ opened} = \frac{\text{emails opened}}{\text{municipalities contacted}} = \frac{799}{3390} = 0.2357 = 23.57\%$$

Moreover, the standard error which reflects the uncertainty in sample distribution. The standard error for proportion can be found as:

$$SE\ for\ proportions = \sqrt{\frac{\bar{X}_n \cdot (1 - \bar{X}_n)}{n}} = \sqrt{\frac{p \cdot (1 - p)}{n}}$$

The standard error of the proportion of emails opened in this case was 0.00729 which is quite small and indicates that the confidence intervals then are quite narrow. The confidence intervals represent the range of values that are likely to contain the true value of a parameter at a given confidence level. The lower and upper critical values of a normal distribution define the given confidence levels that in this case are 90% and 95%. Via the critical values of the distribution ($z_{\frac{\alpha}{2}}$) and the standard error (SE), the confidence interval of a given confidence level (α) is computed by:

$$CI(\alpha\ ) = \left[\bar{X}_n - z_{\frac{\alpha}{2}} \cdot SE, \qquad \bar{X}_n + z_{\frac{\alpha}{2}} \cdot SE\right]$$

Consequently, for this study, the confidence interval of the 90% confidence level was between 0.2237 and 0.2477. This means that during a hypothetical, infinitely repeated data-generating process, 90% of the random samples would contain the true proportion between this interval of 0.2237 and 0.2477. Equivalently, the 95% confidence interval is a little larger (as we would expect) between 0.2214 and 0.2500. Consequently, 95% of the samples of a hypothetical, infinitely repeated data-generating process would contain the true proportion of emails opened within this interval. Thus, the small standard error is related to the findings above that reveal narrow ranges in the confidence intervals that would contain the true value of the proportion of emails opened.

Q 1.3 To find whether the emails opened is statistically different than 0.5 in the Netherlands, the following will conduct a hypothesis test. Firstly, the hypothesis test relies on both a hypothesis (H_1) and a null hypothesis (H_0). In this case, the null hypothesis is that the proportion of emails opened is equal to 0.5 such that we can find whether the proportion of emails is different from 0.5 which is the alternative hypothesis. As the test is based on proportions, the z-score is computed in the following way:

$$z - score = \frac{\bar{X}_n - E(X)}{SE\ of\ the\ \bar{X}_n} = \frac{p - H_0}{SE}$$

Further, the z-score is used to calculate the p-value which will indicate the significance of the hypothesis given the level of test at α = 5%. With the use of R, I find the p-value of 0.611, which indicates that the null hypothesis should be retained as the p-value is above the $\alpha = 0.05$. This does not provide evidence for the alternative hypothesis of the Netherlands' proportion of emails opened to statistically different from 0.5. Additionally, changing the level of test to $\alpha = 0.01$ would not change the conclusion to retain the null hypothesis in this case. Retaining the null in this case in supported when looking at the 95% confidence interval which is between 0.430 and 0.541 and then contains the value 0.5.

Q 1.4 Next, another hypothesis test should be carried out based on a difference-in-proportions. For this case, the null hypothesis is based on the idea that the proportion of emails opened is the same for both Hungary and Sweden. This means that the difference in proportions will be 0 and the null hypothesis will be equal to 0 as well. Then by using doing a two-sample test that is *one-sided*, we can make a test based on the alternative hypothesis of the proportions of emails opened being smaller in Hungary than in Sweden. Unlike before, the standard error is computed based on the difference in proportions:

$$SE\ for\ difference - in - proportions = \sqrt{\hat{p} \cdot (1 - \hat{p}) \cdot \left(\frac{1}{n_0}\right) + \left(\frac{1}{n_1}\right)}$$

The $\hat{p}$ refers to the overall sample proportion and the n_0 and n_1 refers to the sample size of Sweden and Hungary respectively. The z-score is computed as:

$$z - score\ for\ difference\ in\ proportions = \frac{difference\ in\ proportions}{SE}$$

Using software, the p-value of the one-sided test is calculated as 0.104. Whether I chose a level of test (α) to be at 0.05 or 0.01, the p-value is above either one which suggest that the null hypothesis should be retained. Subsequently, the test does not necessarily support the hypothesis that the proportions of emails opened in Hungary is significantly smaller than in Sweden.

Q 1.5 The sample average treatment effect (SATE) is given as the difference in average outcome of the group that received the treatment and the control group:

$$\widehat{\text{SATE}} = \text{average of treated} - \text{average of control}$$

Which in this case is translated to the difference between the proportion of emails answered from own country (treatment) and from another other country (control):

$$\widehat{\text{SATE}} = \text{proportion of own country} - \text{proportion of other country}$$

I find that the SATE is equal to 0.0155. To find whether this treatment effect is statistically significant, a hypothesis test can be computed based under the null hypothesis that the sample average treatment effect is 0. Then the alternative hypothesis is that there is a treatment effect which is different from 0. As before by computing the SE based on the difference in proportions and then then z-score, I find that the p-value under the given null hypothesis is 0.295. With the α-level of 0.05, the null hypothesis is then rejected through this test. Consequently, the findings do not support the hypothesis that the treatment and hence the SATE is statistically significant effect. Overall, this also relates to the general question of the research, as this test alone does not support the idea that politicians in general are more willing to learn from experiences from their own country rather than experiences from other countries. However, just because one cannot find statistical significance through a single test, the relationship between treatment and outcome should not necessarily be ignored. The following will investigate the overall research question further.

Q 1.6 To answer question 1.5 using regression, I have first computed a data frame in R that contains two binary variables. The first variable contains a repetition of the condition that the email received is from either own country (1) or another country (0). The other variable contains the information of whether the link was clicked (1) or not (0). The equation for the linear regression model would look like the following:

$$Y = \alpha + \beta X = \alpha + \beta \cdot condition$$

Y being proportion of links clicked, x being the condition whether treatment is own country or other country. Then using the R-software to regress whether the link was clicked on the condition variable, I created the regression model where the intercept, α, is equal to 0.203 and the slope, β, is 0.0155. In words, this means that proportion of links clicked when the condition was 0 (other country) is 0.203 and the added effect on links clicked of going from other to own country is 0.0155. In relation to question 1.5, this supports the earlier findings regarding the significance of the treatment effect as the regression shows that the treatment increases the proportion of clicks.

In addition, the values of the parameter estimates, $\hat{\alpha},$ and $\hat{\beta}$, based on Q1.5 should be the value of the proportion of emailed opened from other countries (0.205) and then the value of the SATE (0.0155) as it reflects the additional treatment effect of going from other to own country. As expected, these estimates fit the values of the coefficients found by the regression model as seen above.

Q 1.7 Using the previous formulas concerning the SATE, I find that the SATE for Germany is 0.045 opposed to a value of -0,014 for the Netherlands (difference of -0.059). Firstly, these different values indicate that the treatment effect in Germany is relatively larger than the overall treatment effect computed in Q1.5 where the SATE was 0.0155. This might suggest that politicians in Germany overall are more likely compared to other countries to take on policy experiences from their own country rather than others. On the other hand, the SATE of the Netherlands is a negative value which indicates that the proportion of links clicked was larger for the emails from other countries. Opposed to the values from Germany, this indicates that overall, the politicians of the Netherlands would be more likely to learn from other countries rather than their own compared to politicians of other countries. Then this would indicate that there actually should be no significant treatment effect for the Netherlands when the treatment is emails from own country. However, I do not compute any statistical significance of these expectations and difference in SATE when computing a hypothesis-test based on proportions ($\alpha = 0.05$) for either country's SATE. However, this does not lead to complete rejection of the treatment effect – they might just present weak relationships with the outcome of clicks.

Q 1.8 For the purpose of using the R, I created a data frame with the content of table 3 with three binary variables. The first variable contains countries, Germany (0) and the Netherlands (1), the second contains the condition or email from either other country (1) or own (0) and the third contains whether the link was clicked (1) or not (0). As suggested above, the treatment might have a different effect *depending on* the country which suggests a multiple conditional relationship. Based on this, the following equation describes regression model:

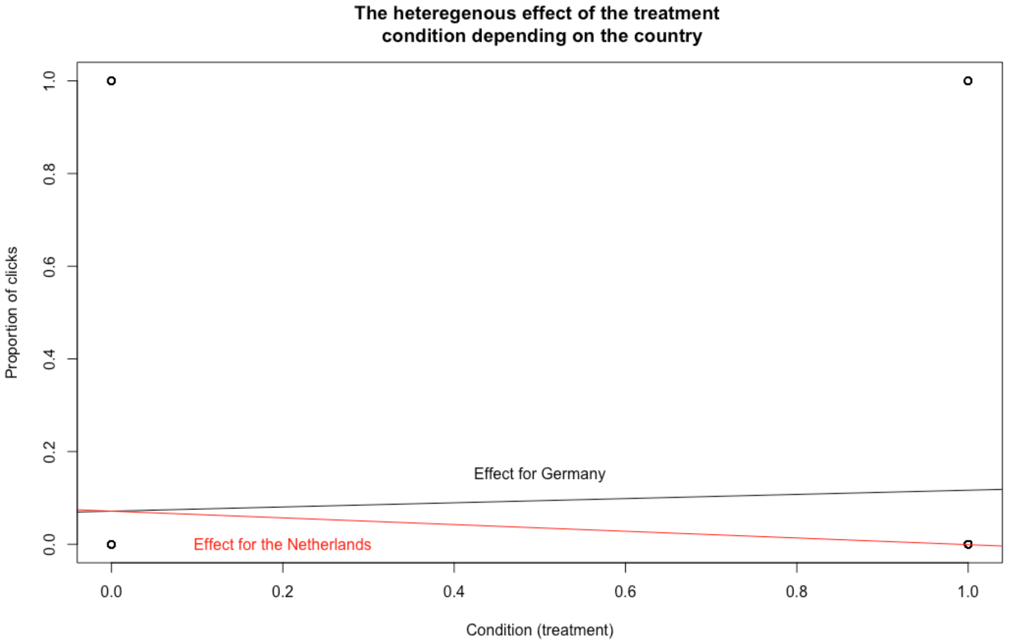
$$Y = \alpha + \beta_1 \cdot condition + \beta_2 \cdot country + \beta_3 \cdot condition \cdot country$$

Based on the regression of this model in R, I find the following coefficients:

Intercept	**Condition**	**Country**	**Condition · Country**
0.0714	0.0450	0.179	-0.0589

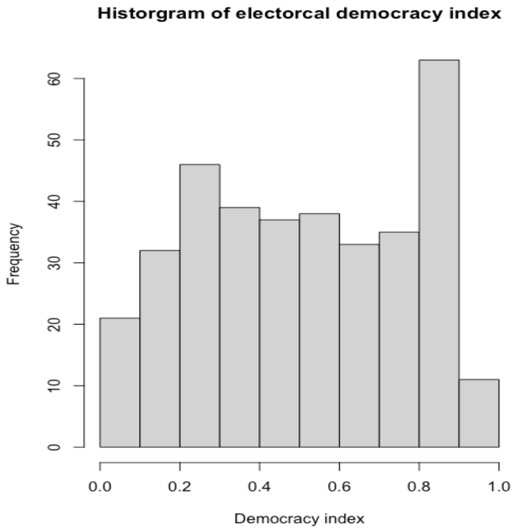
Firstly, when both the condition and country is 0 (other country and Germany), the proportion of links clicked is 0.0714. Then when condition is 1 (going to own country), controlling for country (it stays at 0 for Germany), the proportion of clicks increases with 0.045. Lastly, the coefficient for the interaction term represents the additional effect of the country variable in interaction with the condition. The negative coefficient reflects that the when the country is 1 (the Netherlands) and the condition is 1 (own country) the additional effect of the country is -0.0589. This relates to the earlier finding that the sample average treatment effect in Netherlands was negative (which is visible in the plot below). Finally, the parameter estimate of the interaction term should be the value of the difference in the SATE of Germany and the Netherlands which in Q1.7 was indeed also found to be -0.059. Based on the estimated parameters found via samples average treatments effects for each country, the plot below visualizes two relationships. The plot displays the negative SATE of the

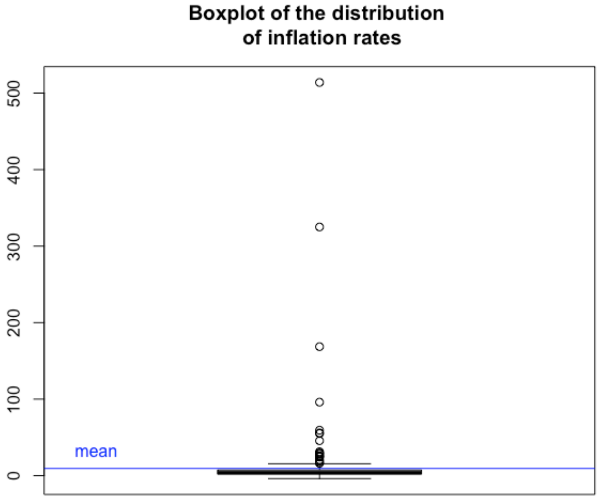
Netherlands and a minor treatment effect in Germany which nonetheless compliments the earlier findings that the treatment effect of Germany was not statistically significant. Both these findings as well as earlier findings relate to the overall research question and not provide any sufficient support for the idea that politicians chose experiences from their local governments rather than from the ones of other countries.



Q 2.1 The following part uses the data frame s151875 of 10 different variables where 9 of them are different potentially explanatory variables of the dependent variable y, the electoral democracy index. The electoral democracy index takes values between 1 and 0 and the distribution of the frequency is displayed in the histogram below where we see a wide and quite even distribution. Furthermore, the binary variable SoMe reflects whether the government ever shut down social media (0 = never, 1= once or more). The frequency of each is reflected in the table below and displays that almost half of the countries have experienced shutdowns of social media by the government. Lastly, the inflation variable is a continuous, interval variable that represents yearly inflation rate. It ranges from -3.85% to 513.9% where the boxplot below indicates that this big range and a rather high mean of 9.40% (despite a median of only 4.00%) can be explained by a few extreme outliers.

Frequency of social media shutdown or not	
Never had shutdown	Shutdowns rarely, sometimes, often
0.589	0.411





Q 2.2 For a mean comparison, a one-sided hypothesis test based on the difference in means is useful. I will use the variable of social media shutdown where the two convenient categories describe whether or not the government ever shut down social media. The expectation is that the means of democracy indexes are statistically significantly higher if social media was never shut down. The null hypothesis is then that the difference-in-means between the democracy index for countries that never shut down social media is equal to the mean for countries that have had government shut down social media (at $\alpha = 0.05$). The standard error for difference-in-means is found by:

$$SE\ for\ differnce - in - means = \sqrt{\frac{var(X_0)}{n_0} + \frac{var(X_1)}{n_1}}$$

And then a t-score is computed:

$$t - score = \frac{difference\ in\ means}{SE}$$

By the use of R software, I find that under the given null hypothesis, the p-value is $2.2 \cdot 10^{-16}$. Consequently, the null should be rejected as the probability of this sampling results given that the null was true is very unlikely. This finding is in favor of the alternative hypothesis which fits expectations as shutdowns of social media might suggest poor levels of democratic freedom. These findings are also supported when including uncertainties. For instance, the 95% confidence interval is [0.322 ; inf] which does not even include 0 and ultimately further supports the alternative hypothesis.

Q2.3 To discuss explanatory factors of the democracy index I have created following multiple regression models:

$$Y = \alpha + \beta_1 \cdot inflation + \beta_2 \cdot SoMe$$

$$Y = \alpha + \beta_1 \cdot inflation + \beta_2 \cdot corruption\ \beta_3 \cdot SoMe + \beta_4 \cdot Unis$$

Y is the electoral democracy index value. Overall, I chose to work with the variables describing inflation, corruption control, social media freedom and universities as they all represent variables that are not as dependent on country size (arguably not as correct for universities) and also quite directly linked to democratic factors and values that might have a big effect on the index.

Coefficients for model 1:

Intercept	**Corrupt**	**SoMe**
0.5544	0.1153	-0.2686

Coefficients model 2:

Intercept	**Inflation**	**Corrupt**	**SoMe**	**Unis**
0.06348	$-7.657 \cdot 10^{-5}$	0.11805	-0.2638	$1.018 \cdot 10^{-4}$

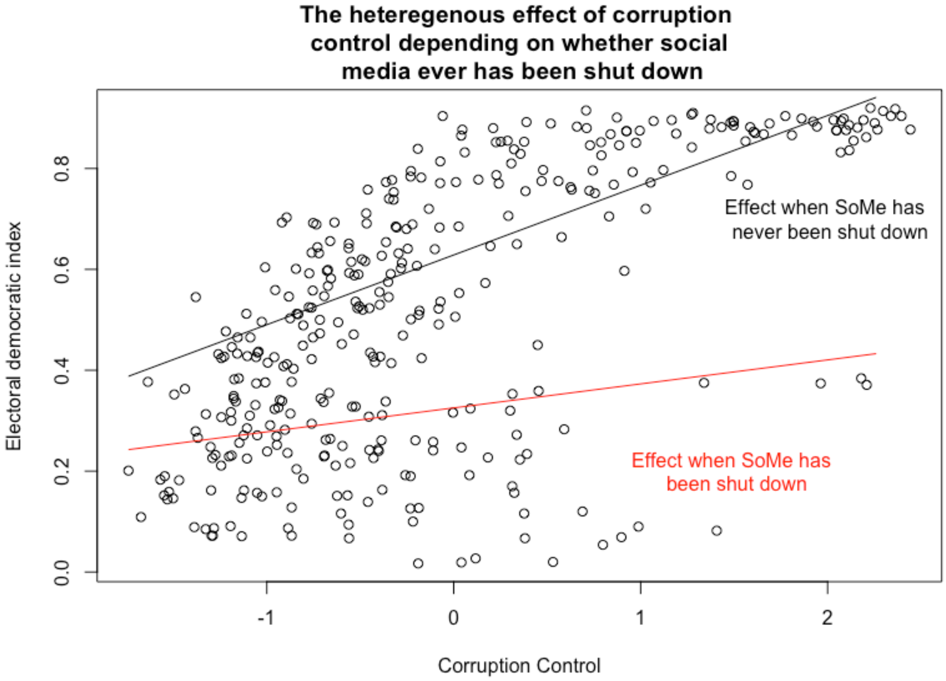
For model 1, I find that controlling for the corruption control, going from cases where social media was never shut down to any case of one shutdown or more decreased the democracy index massively which also fit what we would expect. Also, when controlling for social media freedoms, a one unit increase in corruption controls increases the democracy index with 0.1153. The difference in coefficients between the two can be explained by the difference of scale and type of variable that may make the effect more difficult to compare. For model 2, the same effects of positive and negative influence of increasing corruption control and going from no shutdowns to shutdown of social media respectively, still applies when controlling for inflation and number of universities as well. Additionally, controlling for everything else, a one unit increase of inflation decreases the democracy index with $-7.657 \cdot 10^{-5}$. The impact of an increase in universities is quite minor as well which might be as expected as the one unit increase in these variables of a big scale has little effect. In terms of uncertainty and variation between the models, the R^2 is 0.6066 and 0.624 for model 1 and 2 respectively, which indicates a better fit in model 2 as we control for more predictors. However, there is a probability of overfitting the model, which makes the adjusted R^2 relevant as it restricts the model with the degrees of freedom (values of 0.6043 and 0.619 respectively). Though surprisingly, the values of both R^2 adjusted R^2 for model 1 is nearly as good as for model 2 which suggests a good model fit in model 1 despite there only being two predictors. Consequently, the model suggests that the social media and corruption variables are key explanatory factors.

Q2.4 To study any possible heterogeneous treatment effects, I have chosen to extend model 1 of Q2.3 which will concern the heterogeneous effect of level of corruption depending on whether the government has ever shut down social media or not. The extended model 3 then takes the form:

$$Y = \alpha + \beta_1 \cdot corruption + \beta_2 \cdot SoMe + \ \beta_3 \cdot corruption \cdot SoMe$$

Intercept	**Corrupt**	**SoMe**	**Corrupt · SoMe**
0.6285	0.1382	−0.3032	−0.0907

Now, the intercept has increased to 0.629 which is the democracy index when social media has never been shut down and corruption control is at 0 (a medium level). However, as we go from no social media shutdowns to 1 or more, the effect of SoMe is both the decrease of -0.3032 as well as the additional effect of -0.0907 when controlling for corruption. Beneath is a visualization of the heterogeneous effect of corruption control depending on whether social media has ever been shut down. It is clear from the plot that the effect of corruption control is much stronger in the cases where social media has never been shut down and vice versa.



Q2.5 In conclusion, most findings concerning this dataset of Q2 corresponds to the expectations related to effects of possible explanatory variables of the electoral democratic index. From the beginning the variables such as important and export were left out as they might need a factor of country sizes to be related to. However, findings in Q2.3 and Q2.4 suggests based on model 1 and 3, that variables such as corruption control and social media (shutdown or not) suggest a causal effect on the democracy index. However, one should be careful with this assumption as determination of causal effect demands elimination of most confounding bias. This limit to the observational study could however be approached by not only controlling for other factors but by subclassification as well. For instance, this could be done by subclassification by regions which would make the cases of comparison more similar and perhaps eliminate some bias related to the causal effects of the variables. Lastly, a potential model where the democracy index would be the predictor could be in an observational study of voting turnout across the world. One could imagine the turnout in percent might be explained by predictors such as the electoral democracy index, GDP per capita and corruption control among others.

Bibliography

Butler, D. M., de Vries, C. E., & Solaz, H. (2019). Studying policy diffusion at the individual level: Experiments on nationalistic biases in information seeking [Article]. *Research & Politics*, *6*(4), 205316801989161. https://doi.org/10.1177/2053168019891619

Imai, K. (2017/2018). *Quantitative Social Science: An Introduction.* Princeton University Press

Appendix

```
#---------------------Q1.2----------------------
# create a data frame
country <- c("Austria", "Belgium", "Estonia", "Finland", "Germany",
           "Hungary", "Italy", "Netherlands", "Sweden")
munis_contacted <- c(70, 305, 17, 65, 1353, 84, 1014, 313, 169)
email_opened <- c(20, 115, 6, 11, 286, 26, 117, 152, 66)
maildata1 <- data.frame(country, munis_contacted, email_opened)
#sum of emails opened
summail <- sum(maildata1$email_opened)

#sum of munis contacted
summuni <- sum(maildata1$munis_contacted)

#prop of mails opened
prop_mail <- summail/summuni
#23.6 %

#the standard error, SE
se <- sqrt((prop_mail*(1-prop_mail))/summuni)
#summuni = n

#the CI find the critical values:
```

```
cv90 <- qnorm(0.95, 0, 1)
cv95 <- qnorm(0.975, 0, 1)

#Find the confidence interval of these critical values
CI90 <- c(prop_mail - cv90 * se , prop_mail + cv90 * se)
CI95 <- c(prop_mail - cv95 * se , prop_mail + cv95 * se)

#---------------------Q1.3----------------------
#The H0
H0 <- 0.5

#finding the probability of emails opened in NL
sum_open_NL <- maildata1$email_opened[maildata1$country == "Netherlands"]
sum_mails_NL <- maildata1$munis_contacted[maildata1$country == "Netherlands"]
prop.NL <- sum_open_NL/sum_mails_NL

#SE
se.NL <- sqrt((H0*(1-H0))/sum_mails_NL)

#z-score
z.score.NL <- (prop.NL - H0)/se.NL

#P-value
p_val_NL <- 2*pnorm(z.score.NL)

#Critical value
cv95 <- qnorm(0.975, 0, 1)
#Find the confidence interval of these critical values
CI95 <- c(prop.NL - cv95 * se.NL , prop.NL + cv95 * se.NL)

#---------------------Q1.4----------------------
#H0: the difference in proportions is zero
```

```
H0_2 <- 0

#n for each
n.Swed <- maildata1$munis_contacted[maildata1$country == "Sweden"]
n.Hung <- maildata1$munis_contacted[maildata1$country == "Hungary"]

sum_mails <- n.Swed + n.Hung

#sum of opened emails
sum_open <- maildata1$email_opened[maildata1$country == "Sweden"]
+ maildata1$email_opened[maildata1$country == "Hungary"]

#proportion is the overall proportion for the two because the variance is the same
prop <- sum_open / sum_mails

#for difference in proportions
prop.se <- 66/169
prop.hu <- 26/84

#SE of difference in proportions
se_dif_in_prop <- sqrt(prop * (1 - prop) * (1 / n.Swed + 1 / n.Hung))

#z-score
z.score <- (prop.se - prop.hu)/se_dif_in_prop

#the p-value (one-sided calculation)
p_val1 <- pnorm(-abs(z.score))

#the prop.test gives the same p-value
prop.test(table(data.h.s$country1, data.h.s$mails_open),
        alternative = "greater", correct = FALSE)
```

```
#--------------------Q1.5---------------------
#creating the data frame
condition <- c("Other country","Own Country")
opened <- c(400, 399)
clicked <- c(81, 87)
linksdata <- data.frame(condition, opened, clicked)

#The SATE
prop.other <- linksdata$clicked[linksdata$condition == "Other country"] /
  linksdata$opened[linksdata$condition == "Other country"]
prop.own <- linksdata$clicked[linksdata$condition == "Own Country"] /
  linksdata$opened[linksdata$condition == "Own Country"]

sate <- prop.own - prop.other
#test for the significance of the treatment effect. the H0: the SATE is 0 => difference in prop is 0.
h0 <- 0

#over-all proportion
prop_all <- (87+81) / (400+399)

#SE computed by inserting values of n from table 2
se_dif_in_prop <- sqrt(prop_all * (1 - prop_all) * (1 / 399 + 1 / 400))

#z.score calculated as before
Z.score <- (prop.own - prop.other)/se_dif_in_prop

#the p-value (one-sided calculation)
p_val2 <- pnorm(-abs(Z.score))

#--------------------Q1.6---------------------
# creating new data, condition is either other country (test) or own (treatment), Email is either 1
# (linked clicked), or 0 (mail open, link not clicked)
```

```
condition1 <- rep(c(0, 1), times = c(400, 399))
mails <- rep(c(1,0,1,0), times = c(81, 319, 87, 312))
maildata2 <- data.frame(condition1, mails)

#regress the outcome variable (number of clicks) on the condition
reg.treat <- lm(mails ~ condition1, data = maildata2)
coef(reg.treat)
summary(reg.treat)

#the parameter estimates, alpha \hat and beta \hat are values of prop.other and SATE respectively
prop.other <- linksdata$clicked[linksdata$condition == "Other country"] /
  linksdata$opened[linksdata$condition == "Other country"]
sate <- prop.own - prop.other
#---------------------Q1.7----------------------
#Proportions for germany
prop.other.g <- 10/140
prop.own.g <- 17/146
sate.g <- prop.own.g - prop.other.g

#Proportions for Netherlands
prop.other.n <- 20/80
prop.own.n <- 17/72
sate.n <- prop.own.n - prop.other.n
#difference in SATE
dif.sate <- sate.n - sate.g

#new dataframe for GERMANY
country <- rep(c(0), times = c(140+146))
condition <- rep(c(0, 1), times = c(140, 146))
clicks <- rep(c(1,0,1,0), times = c(10, 130, 17, 129))
data.ger <- data.frame(country, condition, clicks)
```

```
#new dataframe for Netherlands
country <- rep(c(1), times = c(80+72))
condition <- rep(c(0, 1), times = c(80, 72))
clicks <- rep(c(1,0,1,0), times = c(20, 60, 17, 55))
data.net <- data.frame(country, condition, clicks)

#prop.tests
prop.test(table(data.ger$condition, data.ger$clicks), alternative = "greater")
prop.test(table(data.net$condition, data.net$clicks), alternative = "less")

#---------------------Q1.8----------------------
#merging the data sets
data.ger.net <- merge(data.ger, data.net, by = c("country", "condition", "clicks"), all = TRUE)
#regression model
coef(model2)
model2 <- lm(clicks ~ condition * country, data = data.ger.net)
summary(model2)

#plot
plot(data.ger.net$clicks ~ data.ger.net$condition,
    main = "The heteregenous effect of the treatment \n condition depending on the country",
    xlab = "Condition (treatment)",
    ylab = "Proportion of clicks")
#lines creates based on the SATEs
abline(0.0714, 0.045)
slope_n <- sate.n + dif.sate
abline(0.0714, slope_n, col = "red")
text(0.2, 0.0, "Effect for the Netherlands", col = "red")
text(0.5, 0.15, "Effect for Germany")
#---------------------Q2.1----------------------
#visual and tabular univariate summary of democracy variable (former y)
dim(s151875)
```

```
nrow(s151875)
range(s151875$democ)
table(s151875$democ)
summary(s151875$democ)
boxplot(s151875$democ)
hist(s151875$democ,
    main = "Historgram of electorcal democracy index",
    xlab = "Democracy index")

#SoMe
prop.table(table(s151875$SoMe))
#summary and boxplot of inflation
summary(s151875$inflation)
boxplot(s151875$inflation, main = "Boxplot of the distribution \n of inflation rates")
abline(h = mean(s151875$inflation, na.rm = TRUE), col = "blue")
text(text("mean", x = 0.5, y = 30, pos = 4, col = "blue"))
range(s151875$inflation, na.rm = TRUE)

#--------------------Q2.2----------------------
#t.test used for test based on difference in means and computing 95% CI
t.test(s151875$democ[s151875$SoMe == 0],
      s151875$democ[s151875$SoMe == 1], alternative = "greater")

#--------------------Q2.3----------------------
#multiple regression models
m01 <- lm(democ ~ corrupt + SoMe, data = s151875)
m02 <- lm(democ ~ inflation + corrupt + SoMe + unis, data = s151875)
#coefficiants
coef(m01)
coef(m02)
#The R^2 and adjusted R^2 as well as significance values for the variables
summary(m01)
summary(m02)

#---------------------Q2.4----------------------
#the conditional regression model
m03 <- lm(democ ~ corrupt * SoMe, data = s151875)
coef(m03)

#New data frames
never_SoMe <- data.frame(corrupt = seq(min(s151875$corrupt, na.rm = TRUE),
                                       max(s151875$corrupt, na.rm = TRUE),
                                       by = 1),
                         SoMe = 0)

has_shut_SoMe <- data.frame(corrupt = seq(min(s151875$corrupt, na.rm = TRUE),
                                     max(s151875$corrupt, na.rm = TRUE),
                                     by = 1),
                       SoMe = 1, na.rm = TRUE)
#predictions
preds_never_SoMe <- predict(m03, newdata = never_SoMe)
preds_has_shut_SoMe <- predict(m03, newdata = has_shut_SoMe)

#plotting and prediction lines
plot(s151875$democ ~ s151875$corrupt,
    main = "The heteregenous effect of corruption \n control depending on whether social \n media
ever has been shut down",
    xlab = "Corruption Control",
    ylab = "Electoral democratic index")
lines(never_SoMe$corrupt, preds_never_SoMe)
lines(has_shut_SoMe$corrupt, preds_has_shut_SoMe, col = "red")
text(2, 0.7, "Effect when SoMe has \n never been shut down")
text(1.5, 0.2, "Effect when SoMe has \n been shut down", col = "red")
```