Homework - Day 2: Simulation I (Observational Data)

Social Analysis and Simulation in R

Yui Nishimura

21 August 2019

Contents

1 Simulation for A Countinuous Outcome

Instruction: In this question, you are asked to simulate an observational data you would want to analyze if you have all the resources and tools to complete the data collection process. First, briefly describe a research question that you are interested in, but have not empirically examined or collected any data about. Specify one continuous random variable of interest and consider the data generating process for that random variable. For example, the continuous random variable can emerge from a normal distribution (with a positive variance) whose expected value is a linear function of several independent variables.

Next, simulate N hypothetical data points for your "dependent variable" under the data generating process, where you first generate N data points for your "independent variables," and then create the continuous variable. Here N must be determined according to your population of interest and consider that you can have a census of the population (i.e., you do not have to consider any form of sampling and just imagine that you can collect all data). During this process, focus on details and make sure that you specify the theoretical range of each variable and the simulated data follows such theoretical bound (i.e., you cannot simulate the value -30 or 120 for a variable denoting percentage). Here create at least one aggregate level variable (e.g., city level median income). Moreover, if you theorize that one or more independent variables are a function of another independent variable, incorporate the perspective into the data generating process.

Finally, provide a set of descriptive statistics of the data both numerically and visually and perform a set of Normal-Linear regression models. Output a coefficient table in LATEX format and create a plot for predicted values of the continuous variable for typical units. Be sure that such typical units must be inside the convex hull of the data, meaning that the prediction must be made for observations the combination of whose covariate values surely exists in the data. To conclude your answer, make some relevant comments on what you would have to account for when you are actually going to collect the observational data.

1.1 Introduction – Research Question and Background

What types of characteristics make states ratify multirateral treaties more quickly than others? In reality, we can observe that some countries ratify agreements very quickly after they became open to ratify, but others takes very long time. In the existing literature, scholars have tried to understand what determines duration until ratification in various cases of international agreements, by assuming that such a variety represents as the willingness of ratify of the state. However, this assumption might not be always true, if there is a systematic element affecting the ratification process at the domestic level. During negotiation process, states bargain about provisions which legally bind themselves in the future. However, once after they reach the consensus, the decision-making to ratify or not is dependent on the domestic political process. Then, for example, it possibly happens that some countries require more time to obtain administrative approval with complicated procedures and various actors, but others just require political leader's signature on a single piece of paper, even if they have a very strong preference to ratify. If so, the differences in terms of duration across countries does not necessarily mean their differences in intentions behind observed behaviors.

Unfortunately, however, there is no assessment to reveal the general ratification pattern of each country. Accordingly, the central motivation of my research is, to identify what states characteristics have an influence on the general mechanism of ratification regardless of treaty types.

1.2 Domestic Political Institutions and Ratification Process

As I mentioned before, the domestic administrative process is the key element. My theoretical expectatin is that democratic countries take more time to ratify treaties. I am interested in how domestic political institutions have an influence on their duration until ratify treaties, comparing to non-democratic regime. There are two phases making ratification process delay in democratic countries. One is the legislative hurdle. Assuming office-seeking legislatures, it is possible that some of them disagree with treaties which can conflict with preferences of their supporters. Or, even if the legislatures do not disagree, they might want to prioritize discussions about domestic policies at the Congress and the Parliament. The other process required in democracies is the executive procedure. Where executive brunch have constraint on the decision making, it becomes more hard to pass potentional new laws related to treaties or to obtain approval for treaties. Comparing to this, in autocracy, decision making is done by a political lader and the small group of relevant actors. Also, leaders do not need to

take accountability to citizens in terms of decisions they made and counsequences of the decisions. Accordingly, I expect that the general time to ratify takes more time in democracies, comparing to autocratic countris.

Before getting in depth about details, I specify what are the "treaty" in this context. Although there are a number of treaty types, I am interested in international multirateral treaties, which are open for all the United Nations member countries. These treaties are differenciated by other treaties, such as bilateral treaties, since every state has the equal opportunity to participate. Regional treaties, for example, only focus on members of the reagion as potential ratifiers. Bilateral treaties are exclusive too. Considering those characteristics, the examination of the UN treaties has an advantage to compare all UN members in terms of their overall ratifying practices.

1.2.1 Sample and Measurement

Based on the discussion above, the analytical unit is country, and all countries are the UN members. The total number of the UN members is 193. To understand the pattern, I measure it by taking a arithmetic mean of years until ratification for all treaties they follow (" *average duration* ", hereafter), as the outcome variable. The first UN treaty became open in 1945. If states do not ratify a treaty yet, it is recognized as still being the ratification process, and the possible latest year is 2019. Therefore, the longest time to ratify treaty is 74 years (= 2019 - 1945). Since I calculate the arismetic mean, so the variable is continuous.

The key independent variable of interest is *regime type*. If countries are currently ruled in democratic political system, it is considered as democracy, otherwise autocracy. For the sake of the requrement of the question, I include the number of *regional institutions* (RIs) as a control variable. This is because the development of RIs can embed members in the regional ties and cooperation based agreements, and peer pressure for ratification emerge. In addition, I include the *regime transition experience* because it might affects the current ruling system and ratification process to some treaties which were done in different ways.

$$Y = \alpha + \beta_1 X_{regime} + \beta_2 X_{RI} + \beta_3 X_{transition}$$

1.3 Simulation

1.3.1 Preparation

First, I clean up the environment.

```
rm(list=ls()) # rm() cleans up your R
gc();gc()     # gc() cleans up your memory
```

Here I simulate the model for the most ideal case, where the every data would be available. To fix the results for the pseudo-random number generation, I also set seed here.

```r
simN <- 193    # the number of total UN members
```

Regime Type

For now, I start to create the independent variables based on substantive knowledge. First, I create the regime variable.

```r
set.seed(7272)
democrat <- rbinom(simN, size = 1, prob = .7)
```

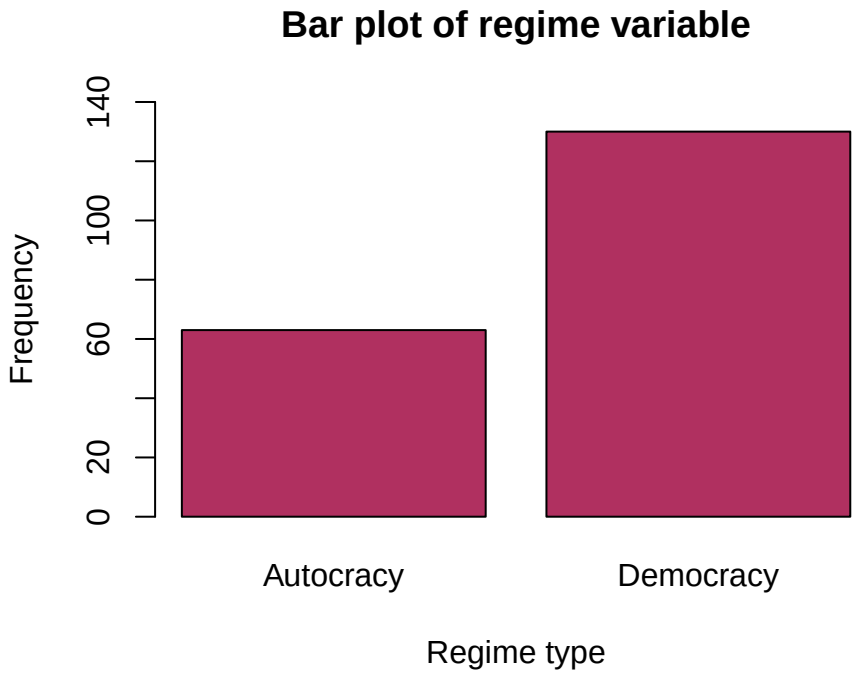
```r
summary(democrat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  1.0000  0.6736  1.0000  1.0000
```

```r
unique(democrat)
```

```
## [1] 1 0
```

```r
# visualization of frequency in barplot - for discrete variable
democrat_name <- ifelse(democrat == 1, "Democracy", "Autocracy")
barplot(table(democrat_name), main = "Bar plot of regime variable",
        xlab = "Regime type", ylab = "Frequency", ylim=c(0,140), col = "maroon")
```



Number of Regional Institutions

Next, I create the variable of RIs. To deffirenciate the region based on this variable, I set the condition based on the general observation of treaty ratification pattern across regions.

```r
# creating an aggregate level variable - regional institutions
set.seed(7272)
region1 <- ifelse(democrat == 1 & runif(simN)<0.8, 6, 0)       # Latino
region2 <- ifelse(democrat == 1 & runif(simN)<0.7, 5, region1) # Europe
region3 <- ifelse(democrat == 1 & runif(simN)<0.3, 4, region2) # North America
region4 <- ifelse(democrat == 0 & runif(simN)<0.5, 3, region3) # Africa
region5 <- ifelse(democrat == 0 & runif(simN) <0.8, 2, region4) # Asia
region <- ifelse(region5==0, 7, region5)                        # Middle East

# check the variable
summary(region)    # no NA
```

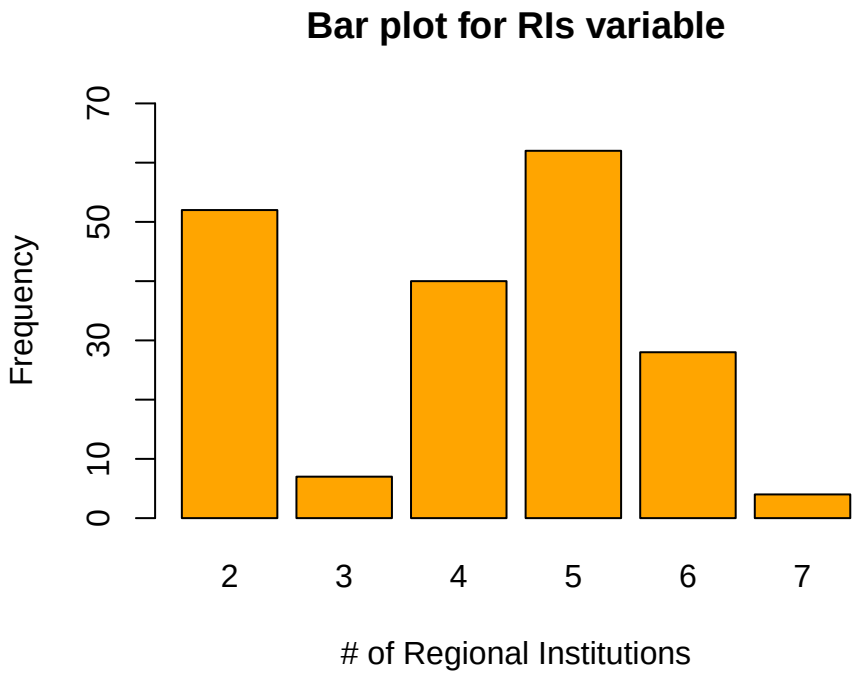
Min. 1st Qu. Median Mean 3rd Qu. Max. 2.000 2.000 4.000 4.098 5.000 7.000

```r
unique(region)     # unique values
```

[1] 6 5 4 2 3 7

This variable is discrete, and thus I use the bar plot instead of histogram.

```r
# visualization of frequency in barplot - for discrete variable
barplot(table(region), main = "Bar plot for RIs variable" ,
        xlab = "# of Regional Institutions", ylab = "Frequency",
        ylim = c(0, 70), col = "orange")
```



Regime Transition

Finally, I create the regime transition dummy variable. This variable is theoretically related both to the regime type and the average duration until ratification. Historically, democrati-

zation happened more than transition into autocracy. If countries change their regime, their trend of ratification process is expected not to be too fast or too late.

```
# creating a transition dummy variable
set.seed(7272)
transit1 <- ifelse(democrat == 1 & runif(simN)<0.3, 1, NA)       # 30% democrat transitined
transit2 <- ifelse(democrat == 0 & runif(simN)<0.1, 1, transit1) # 10% autocrat transitioned
transit <- ifelse(is.na(transit2)==T, 0, 1)      # non-transitioned country

# check the created variable
summary(transit)    # no NA
```

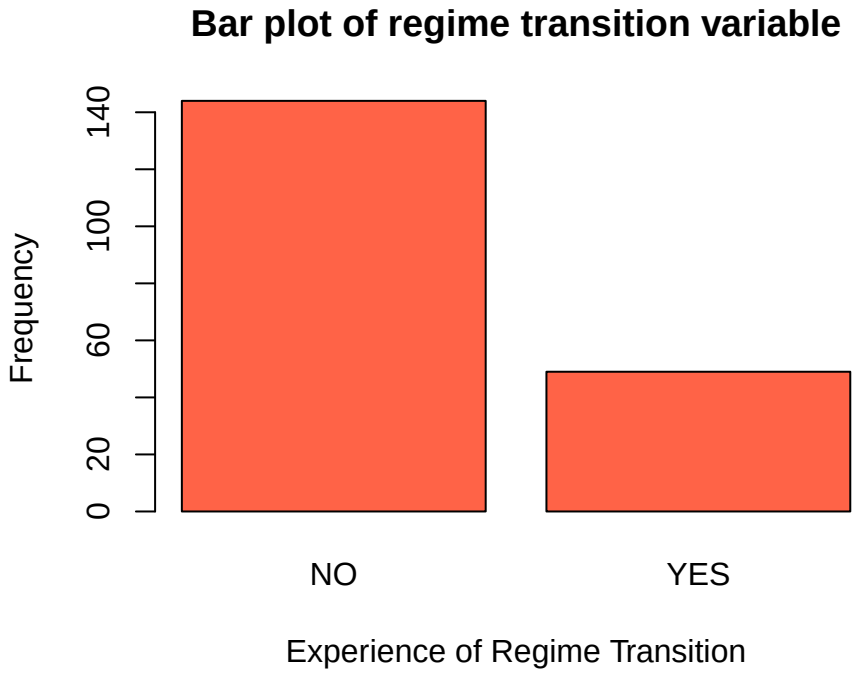
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.2539  1.0000  1.0000
```

```
unique(transit)     # unique values
```

```
## [1] 0 1
```

This variable is discrete variable, and thus I use the bar plot instead of histogram.

```
# visualization of frequency in barplot - for discrete variable
transit_name <- ifelse(transit == 1, "YES", "NO")
barplot(table(transit_name), main = "Bar plot of regime transition variable",
        xlab = "Experience of Regime Transition", ylab = "Frequency",
        col = "tomato")
```



Average Duration

Finally, I simulate values of the outcome variable, which is continuous and normally dis-

tributed and produced by the independent variables above. As I discussed, the possible values the outcome variable take is from 0 to 74. According to the theoretical expectation, the systematic component of the model is specified as follows:

$$\text{Average Duration}_i \sim N(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta_1 X_{regime} + \beta_2 X_{RI} + \beta_3 X_{transition}$$

```
set.seed(7272)                          # set seed

a <- 37       # intercept
b.d <- 5      # coefficient of democracy (regime type)
b.r <- 1.4    # coefficient of RIs
b.t <- -2     # coefficient of regime transition
sigma <- 4    # homoskedastic standard deviation
```

After storing the simulated values of error and coefficients, I first check whether the simulation worked as I supposed. I checked the mean of simulated results and confirmed whether there is any value violatinglower and upper limits of this variable. Here, I could successfully create the acceptable values for the duration variable.

```
# data generating process
set.seed(7272)# set seed
duration <- rnorm(
  n = simN,
  mean = a + b.d*democrat + b.r*region + b.t*transit,
  sd = sigma)

# checking the results of simulation
mean(duration)                          # it should be around 37
```

```
## [1] 45.61342
```

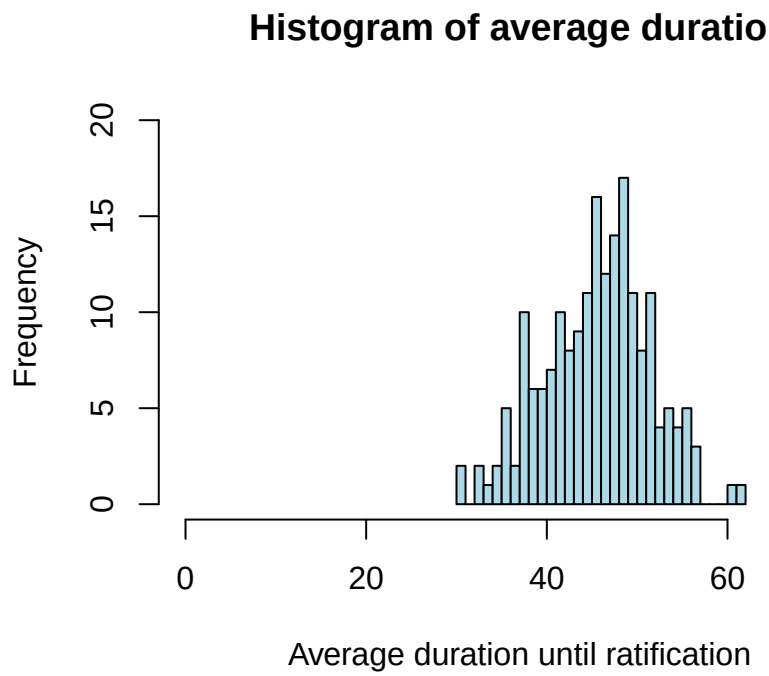
```
unique(ifelse(duration > 0, 0, 1))    # returns 1 if violating of the lower limit
```

```
## [1] 0
```

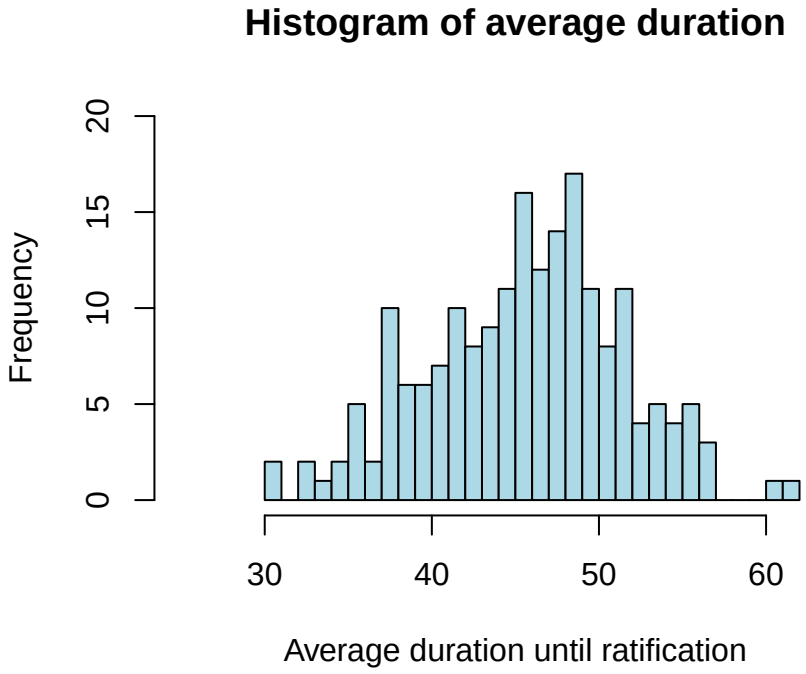
```
unique(ifelse(duration < 74, 0, 1))  # returns 1 if violating of the upper limit
```

```
## [1] 0
```

```
# histogram
hist(duration, breaks = 25,  xlim = c(0,74), ylim = c(0,20),
     col = "lightblue", xlab = "Average duration until ratification",
     main = paste("Histogram of average duration" ))
```




```
# fixing the range
hist(duration, breaks = 25,  xlim = c(25,65), ylim = c(0,20),
     col = "lightblue",xlab = "Average duration until ratification",
     main = paste("Histogram of average duration" ))
```




```
id <- rep(1:simN)  # creating id
sim.data <- data.frame(duration,  democrat, region, transit, id)
```

```r
# check the data
library("xtable")
xtable(head(sim.data[,2:5]))
```

% latex table generated in R 3.6.1 by xtable 1.8-4 package % Wed Aug 21 16:59:21 2019

	democrat	region	transit	id
1	1	6.00	0.00	1
2	1	5.00	1.00	2
3	1	5.00	1.00	3
4	1	4.00	0.00	4
5	1	4.00	0.00	5
6	0	2.00	0.00	6

```r
# creating table for summary statistics
library("stargazer")
stargazer(sim.data, type = "latex", title = "Summary Statistics of Simulated Data",
          summary.stat = c("n", "mean", "sd", "min", "max"))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Aug 21, 2019 - 16:59:21

Table 1: Summary Statistics of Simulated Data

Statistic	N	Mean	St. Dev.	Min	Max
duration	193	45.613	5.838	30.559	61.125
democrat	193	0.674	0.470	0	1
region	193	4.098	1.485	2	7
transit	193	0.254	0.436	0	1
id	193	97.000	55.858	1	193

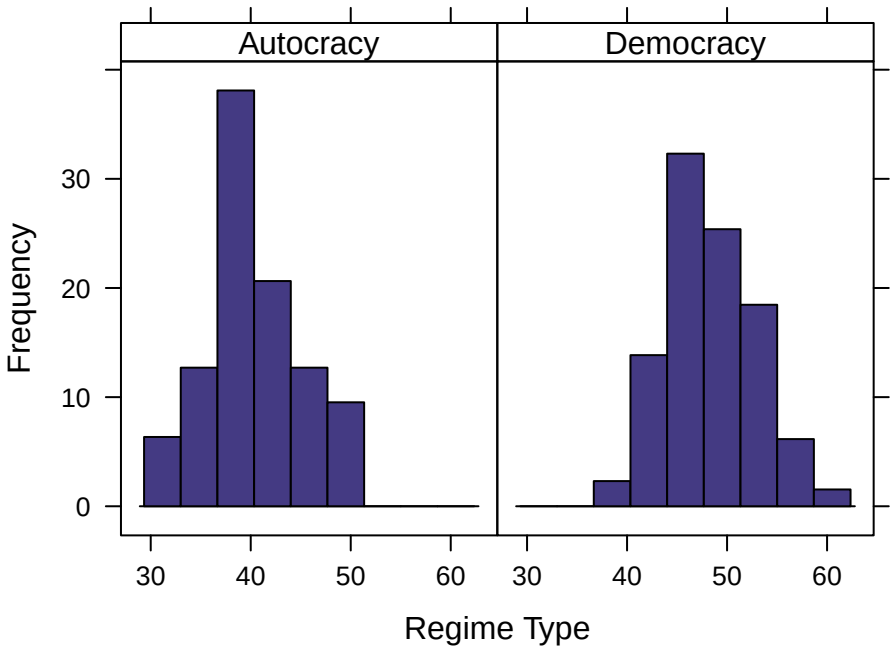
```r
# Correlation between the average duration and democracy
rho <- cor(duration, democrat)
rho
```

```
## [1] 0.6345647
```

Since now I got the dataframe which include both the outcome variable and the key independent variable, the following codes produce the histogram by grouping regime types. There is `lattice` package which is an extentended version of the base plot, and it is useful to use `lattice::histogram()` for this purpose.

```r
library("RColorBrewer")
myColours <- brewer.pal(5,"Blues")

my.settings <- list(
```

```
  superpose.polygon=list(col=myColours[2:5], border="transparent"),
  strip.background=list(col=myColours[6]),
  strip.border=list(col="black")
)

library('lattice')
sim.data$democrat_name <- factor(sim.data$democrat,
                                 labels = paste(c("Autocracy", "Democracy")))
histogram(~ duration | democrat_name, data = sim.data, scales=list(alternating=1),
          auto.key=list(space="top", columns=4,
          points=FALSE, rectangles=TRUE,
          title="District", cex.title=1),
          col = "#443A83FF",
          main = paste("Histograms of average duration grouping by regime"),
          xlab = "Regime Type", ylab = "Frequency",
          par.settings = my.settings)
```



According to this output, the data of autocratic regime are smaller, converting around 42, while the data points of democracy consentrate around 47. Seemingly, it is along with the setting of the data generating process, assigning the longer average duration for democratic countries.

1.4 Normal-Linear Regression Model

Using this simulated dataset, now I examine the relationship between the average duration and the regime type. The model is normal-linear-homoskedastic.

```
model1 <- lm(duration ~ democrat + region + transit, data=sim.data)
summary(model1)
```

```
##
## Call:
## lm(formula = duration ~ democrat + region + transit, data = sim.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.224 -2.565 -0.086  3.141 10.244
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.7062     0.9461  38.797  < 2e-16 ***
## democrat      4.9448     1.0587   4.671 5.69e-06 ***
## region        1.5384     0.3243   4.743 4.13e-06 ***
## transit      -2.8697     0.7239  -3.964 0.000104 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.122 on 189 degrees of freedom
## Multiple R-squared:  0.5092, Adjusted R-squared:  0.5014
## F-statistic: 65.36 on 3 and 189 DF,  p-value: < 2.2e-16
```

According to the estimated coefficients, it reports close values which I specified before. As I theoretically expected and setted, democratic countries tend to take more time for ratification process in general, and its effect size is almost 5 years. The result is statistically significant. For the region variable, also the result supports as well. Substantively, if the number of regional institutions increase by 1, it makes ratification process be in delay by around one and half years.

To report in LaTeXstyle, `stargazer()` is also useful to report the regression output.

```
stargazer(model1,
          digits = 2, digits.extra = 0, # align = TRUE,
          #star.cutoffs = NA, omit.table.layout = "n", ## this line is important!
          keep.stat = c("n", "adj.rsq", "f"), df = FALSE,
          covariate.labels = c("Regime Type", "Number of RIs", "Regime Transition"),
          dep.var.caption = "Dependent variable",
          dep.var.labels = "The Average Duration",
          title = "Results of Linear Regressions",
          type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Aug 21, 2019 - 16:59:22

Table 2: Results of Linear Regressions

	Dependent variable
	The Average Duration
Regime Type	4.94^{***}
	(1.06)
Number of RIs	1.54^{***}
	(0.32)
Regime Transition	-2.87^{***}
	(0.72)
Constant	36.71^{***}
	(0.95)
Observations	193
Adjusted R^2	0.50
F Statistic	65.36^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

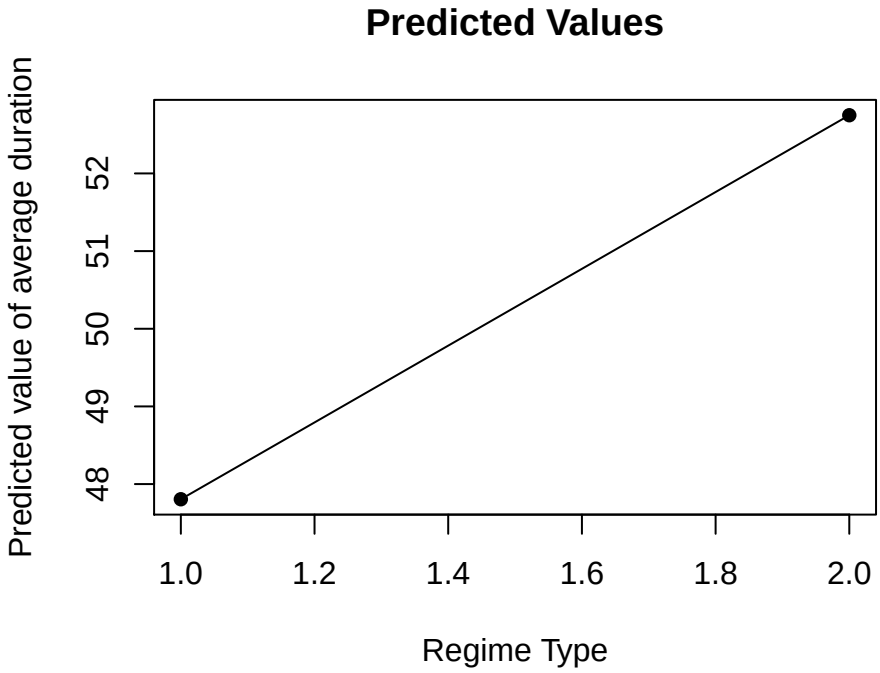
1.4.1 Predicted Values on Plots

To visualize predicted values for typical cases, first we need to extract typical case from the simulated dataset, grap the estimated coefficients according to the results of regression, to substitute them to the linear aggregator function. I determine median as the typical case for each variable. For the extraction of coefficients, I used `coef() function` by specifying the variable.

```
# typical case
m_democ <- median(sim.data[,2])  # regime type
m_region <- median(sim.data[,3])  # RIs
m_transit <- median(sim.data[,4])  # regime transition

# We can extract estimated parameters by using `coef()`
a_hat <- coef(model1)[1]  # intercept
b1_hat <- coef(model1)[2]  # regime type
b2_hat <- coef(model1)[3]  # RIs
b3_hat <- coef(model1)[4]  # regime transition
```

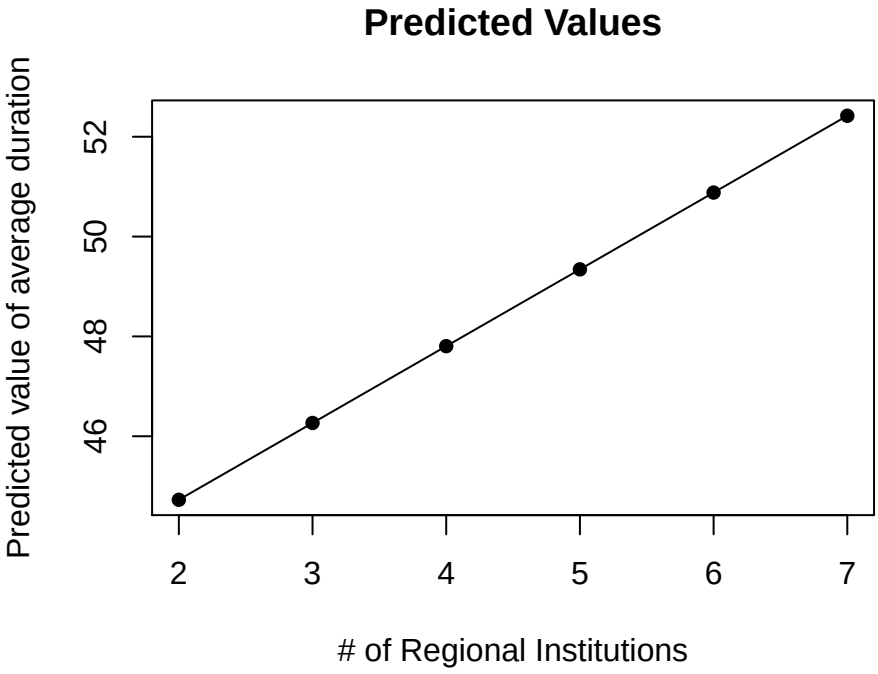
Now I create empty vectors for the predicted values and possible ranges which independent variables can take. By using loop, I storing the predicted values by changing values of one independent variable and by fixing others at median. At the same time, it is useful to store the possible values of each independent variable. Then, by using stored vectors, I create a plot which has data points of predicted values and connects the points.

```
# create the empty vector
# same as as.numeric()
pred_dur <- NA      # predicted average duration
democlim <- NA      # possible value of regime type
regionlim <- NA     # possible value of RIs
transitlim <- NA    # possible value of regime transition

# Regime type
for(i in 1:2) {
  pred_dur[i] <- a_hat + b1_hat* i + b2_hat*m_region + b3_hat*m_transit
  democlim[i] <- i
}

plot(democlim, pred_dur, ylab = "Predicted value of average duration", pch=16,
     xlab="Regime Type")
lines(pred_dur~democlim,type="l")
title("Predicted Values")
```

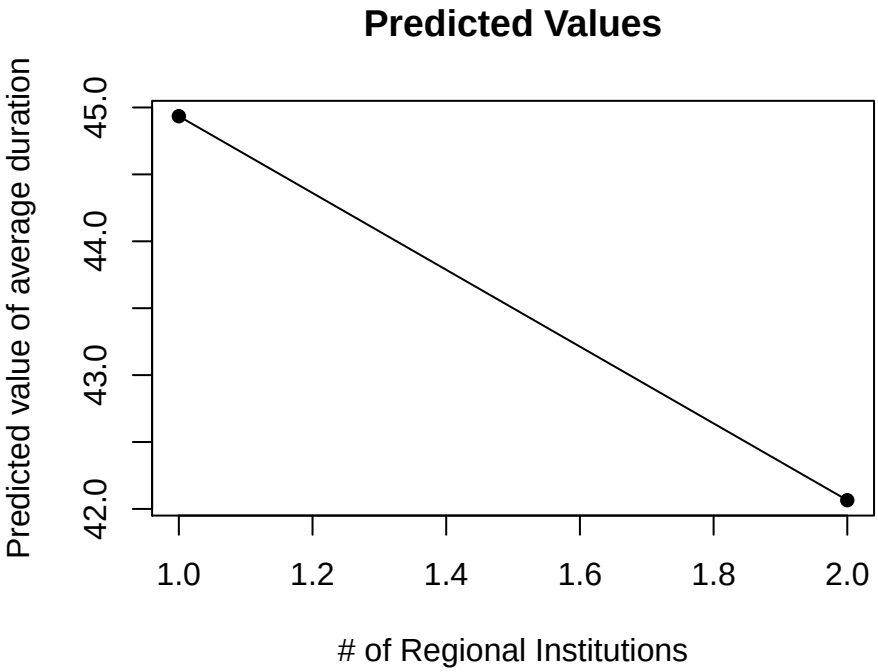



```
# # of Regional Institutions
pred_dur <- NA      # creating an empty vector again
```

```r
for(i in 2:7) {
  pred_dur[i] <- a_hat + b1_hat*m_democ + b2_hat*i + b3_hat*m_transit
  regionlim[i] <- i
}

plot(regionlim, pred_dur, ylab = "Predicted value of average duration", pch=16,
     xlab="# of Regional Institutions")
lines(pred_dur~regionlim,type="l")
title("Predicted Values")
```




```r
# Experience of regime transition
pred_dur <- NA       # creating an empty vector again
for(i in 1:2) {
  pred_dur[i] <- a_hat + b1_hat*m_democ + b2_hat*m_region + b3_hat*i
  transitlim[i] <- i
}

plot(transitlim, pred_dur, ylab = "Predicted value of average duration", pch=16,
     xlab="# of Regional Institutions")
lines(pred_dur~transitlim,type="l")
title("Predicted Values")
```



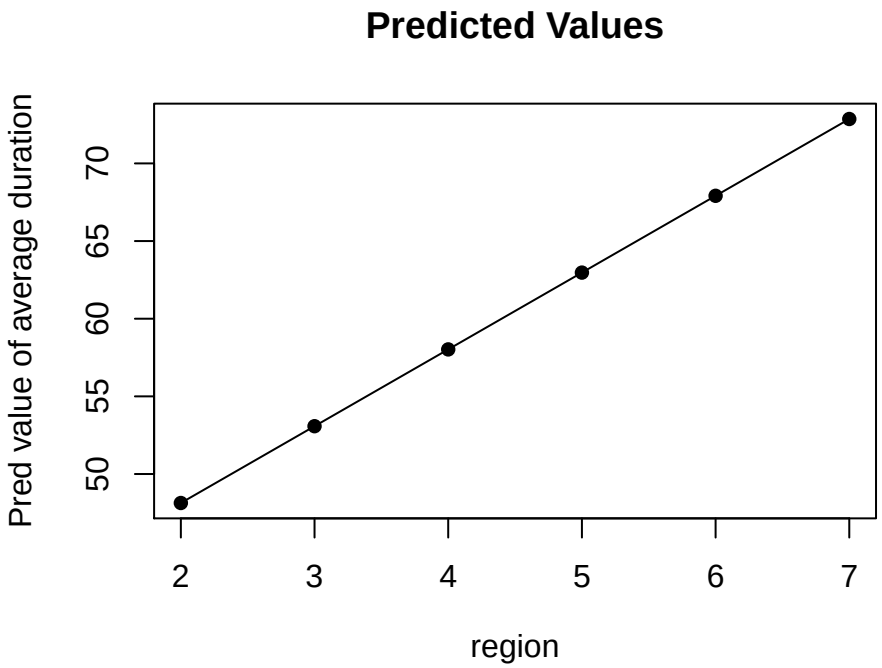
1.4.2 Predicted Values on Plots

According to Day 3 R codes, I create the function which returns predicted values of a model, by specifying the model, the independent variable of interest,, and its range of interest, and how to break the range.

```
PredValue <- function(mymodel, xvar, xstart, xend, by){  # Five arguments

#   xvar <- enquote(xvar)
   coefs <- coef(mymodel)                              # Coefficient list
   indnames <- attr(mymodel$terms, "term.labels")  # Names of variables
   xmat <- mymodel$model[,-1]                          # Matrix for predictors


   # Median values for all variables except for xvar
   typical <- apply(xmat[-which(names(xmat)==deparse(substitute(xvar)))], 2, median)

   xseq <- seq(from=xstart, to=xend, by=by)    # Variable of interest
   intercept <- rep(1, length(xseq))             # Intercept
   typicalmat <- matrix(rep(typical, length(xseq)), ncol=length(typical), byrow=TRUE)

   typical_dat <- data.frame(intercept, xseq, typicalmat) # Typical data frame
   colnames(typical_dat) <- c("Intercept", indnames)      # Typical data frame name

   typical_dat <- as.matrix(typical_dat)         # Typical data matrix (10 by 6)
   coefs2 <- as.matrix(coefs)                     # Coefficient matrix (6 by 1)
```

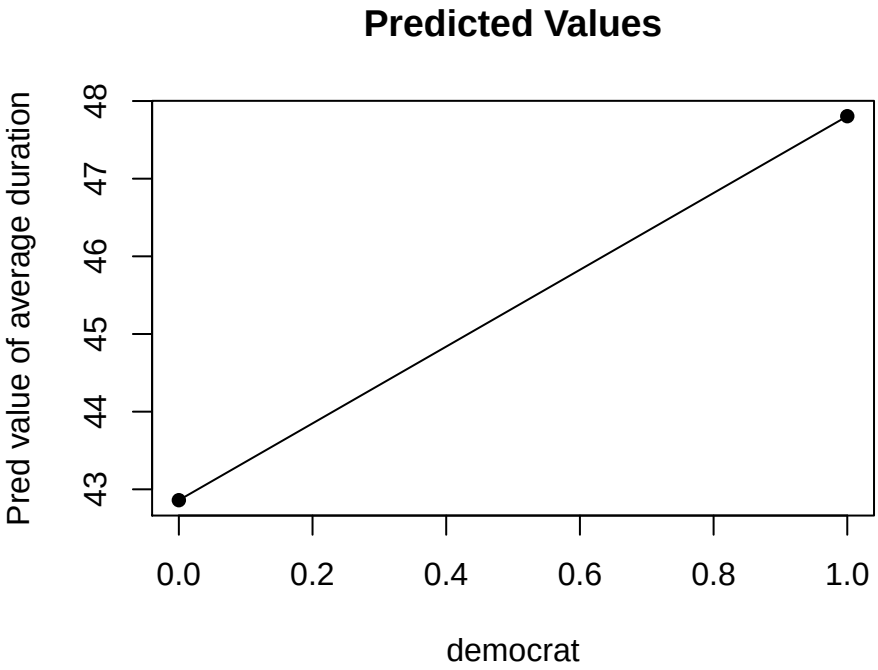
```
  linagg <-  typical_dat %*% coefs2              # Linear aggregators

  pred <- linagg                                 # Predicted probabilities

  # Provide a visualization of predicted probabilities
  plot(linagg ~ xseq, ylab="Pred value of average duration", pch=16,
       xlab=substitute(xvar))
  lines(pred ~ xseq, type="l")
  title("Predicted Values")

  return(linagg)
 }
```

```
# checking the median as the tipical case
PredValue(model1, region, 2, 7, 1)
```



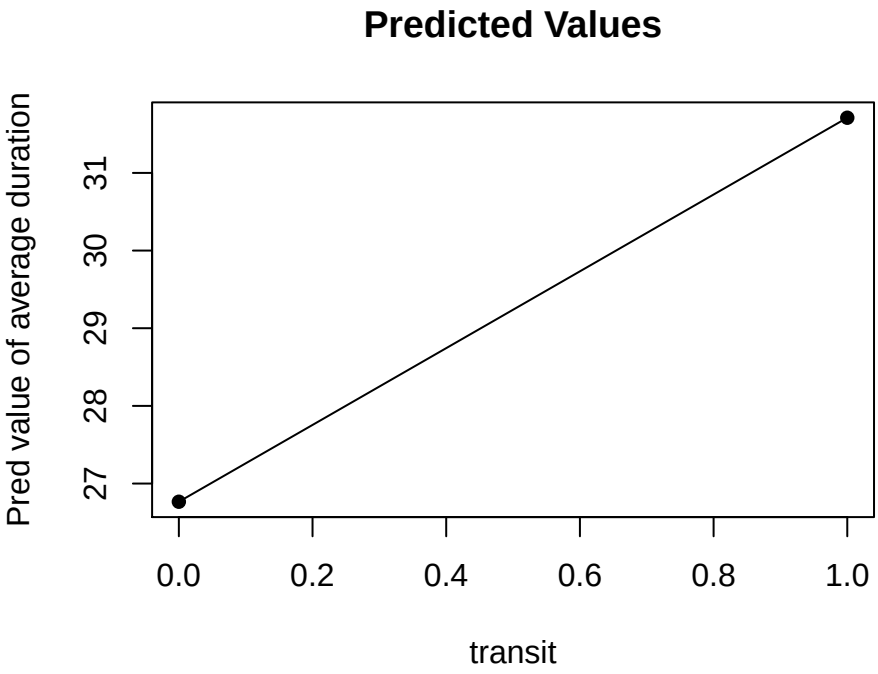

```
##          [,1]
## [1,] 48.13411
## [2,] 53.07886
## [3,] 58.02361
## [4,] 62.96837
## [5,] 67.91312
## [6,] 72.85787
```

```
PredValue(model1, democrat, 0, 1, 1)
```




```
##           [,1]
## [1,] 42.85989
## [2,] 47.80464
```

```
PredValue(model1, transit, 0, 1, 1)
```




```
##           [,1]
## [1,] 26.76575
## [2,] 31.71050
```

1.5 Agregated Level Analysis

Since this practice includes the use of aggregate level variable, I create the aggregate level. I already define the number of region level institutions, but here I consider the region as the aggregate level.

```
# creating region ID
sim.data$id <- NA
for (i in seq_along(sim.data$region)) {
  sim.data$id[i] <- sim.data$region[i] - 1
}

 # creating empty vectors
 # region level average duation, % of democ, % of transit
 agg.duration<-c()
 agg.democ<-c()
 agg.transit<-c()
 temp<-c()
 agg.total<-c()

 # creating temporary vectors to make agg. variables
 rid  <- sim.data$id
 rdur <- sim.data$duration
 rdem <- sim.data$democrat
 rtra <- sim.data$transit

 #  loop to store aggregated level information
 for(i in 1:6) {
   agg.total[i] <- length(rid[sim.data$id==i])

   # calculating arithmetic mean of average duration
   agg.duration[i] <- sum(rdur[sim.data$id==i])/length(rdur[sim.data$id==i])

   # % of democratic country in each region
   agg.democ[i] <- sum(rdem[sim.data$id==i])/length(rdem[sim.data$id==i])

   # % of states experienced regime transition
   agg.transit[i] <- sum(rtra[sim.data$id==i])/length(rtra[sim.data$id==i])
 }

mean(agg.duration)     # it should be around 37
```

```
## [1] 45.30931
```

```
# reporting 1 if violating the upper or lower limit
unique(ifelse(agg.duration > 0 | agg.duration < 74, 0, 1))
```

```
## [1] 0
```

```
mean(agg.democ)
```

```
## [1] 0.5
```

```
# reporting 1 if violating the upper or lower limit
unique(ifelse(agg.democ >= 0 | agg.democ <= 1, 0, 1))
```

```
## [1] 0
```

```
mean(agg.transit)
```

```
## [1] 0.1956384
```

```
# unique value: 1 shows violation of the lower limit
unique(ifelse(agg.transit >= 0 | agg.transit <= 1, 0, 1))
```

```
## [1] 0
```

From now on, I'm going to creat a new dataset consists of region as a unit of analysis.

```
# creating region level data
regid <- 1:6

# the majority of regional members are democracy or not
d.democ <- ifelse(agg.democ > .5, 1, 0)
d.democ
```

```
## [1] 0 0 1 1 1 0
```

```
# storing all variables into a data frame
agg.data <- data.frame(regid = regid, r.democ = agg.democ,
                       r.transit = agg.transit, r.duration = agg.duration,
                       d.democ)
stargazer(agg.data, type = "latex", title = "Summary Statistics of Aggregated Data",
          summary.stat = c("n", "mean", "sd", "min", "max"))
```

```
##
## % Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac
## % Date and time: Wed, Aug 21, 2019 - 16:59:24
## \begin{table}[!htbp] \centering
##   \caption{Summary Statistics of Aggregated Data}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev
## \hline \\[-1.8ex]
## regid & 6 & 3.500 & 1.871 & 1 & 6 \\
## r.democ & 6 & 0.500 & 0.548 & 0 & 1 \\
```

```
## r.transit & 6 & 0.196 & 0.168 & 0.000 & 0.403 \\
## r.duration & 6 & 45.309 & 4.515 & 39.832 & 50.982 \\
## d.democ & 6 & 0.500 & 0.548 & 0 & 1 \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

```
# agg.data
# write.table(agg.data, "data/sim_precinct.csv", sep=",", row.names=F) # Output the data
```

```
 # Simple Analysis
model2 <- lm(r.duration ~ r.democ + r.transit, data=agg.data)
model3 <- lm(r.duration ~ d.democ + r.transit, data=agg.data)
summary(model2) # Compare with the individual level model
```

```
##
## Call:
## lm(formula = r.duration ~ r.democ + r.transit, data = agg.data)
##
## Residuals:
##       1       2       3       4       5       6
## -2.7952 -0.1342 -0.7550  1.3127 -0.5577  2.9294
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   43.680      1.686  25.904 0.000126 ***
## r.democ       13.430      4.341   3.094 0.053561 .
## r.transit    -25.992     14.140  -1.838 0.163333
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.518 on 3 degrees of freedom
## Multiple R-squared:  0.8134, Adjusted R-squared:  0.689
## F-statistic: 6.538 on 2 and 3 DF,  p-value: 0.08062
```

```
stargazer(model2, model3,
          digits = 2, digits.extra = 0, # align = TRUE,
          star.cutoffs = NA, omit.table.layout = "n", ## this line is important!
          keep.stat = c("n", "adj.rsq", "f"), df = FALSE,
          covariate.labels = c("Regime Type", "Regime Transition"),
          dep.var.caption = "Dependent variable",
          dep.var.labels = "The Average Duration",
          title = "Results of Linear Regressions (Region Level Data)",
          type = "latex")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Wed, Aug 21, 2019 - 16:59:24

Table 3: Results of Linear Regressions (Region Level Data)

	Dependent variable	
	The Average Duration	
	(1)	(2)
Regime Type	13.43	
	(4.34)	
Regime Transition		13.43
		(4.34)
r.transit	−25.99	−25.99
	(14.14)	(14.14)
Constant	43.68	43.68
	(1.69)	(1.69)
Observations	6	6
Adjusted R^2	0.69	0.69
F Statistic	6.54	6.54

1.6 Comments on Data Collection

The advantage of simulation was that I could imaginally create possible values and input and use them. However, in the data collection, I need to pay more attention to definition in the first place. For example, independence from the colonial governance or the former goverment has a special process to success the international law. Also, it is important whether we should exclude or include the case without ratification at this moment. Although I assumed that those countries are still in the ratification process, they never commit to the international agreement without any intentions. Therefore, substantively, I have to elaborate the proxy what I really want to capture, and technically, definition should be specified.

2 Simulation for A Binary Outcome

Instruction: Repeat the same simulation for a binary random variable of your interest. Here use a Bernoulli-Logistic model.

2.1 A Binary Variable of Interest

2.2 Simulation

2.3 Bernoulli-Logistic Model