R Camp - Day 2 Homework

Emily Elia

8/19/2019

Problem 1 - Continuous Variable

Research Question

For Problem 1, I simulate data to examine if the presence of corruption scandals in a country's national legislature impacts foreign direct investment (FDI) flows. I hypothesize that a greater number of corruption scandals should be correlated with lower FDI flows because these scandals signal potential government instability. Scandals can upset constituents who may then punish politicians electorally. Scandals can also cause protests for resignations, government "get clean" overhauls, etc., which would be unattractive for companies interested in FDI because these companies are seeking a stable base to conduct business. I do think an argument can be made that more corruption may increase FDI because businesses think they can cut a good deal under the table or profit illicitly with a corruption-compliant government. While I think this may be a possibility, I make the assumption that stability will be prioritized, and thus a country rampant with corruption scandals will be an unattractive destination for FDI.

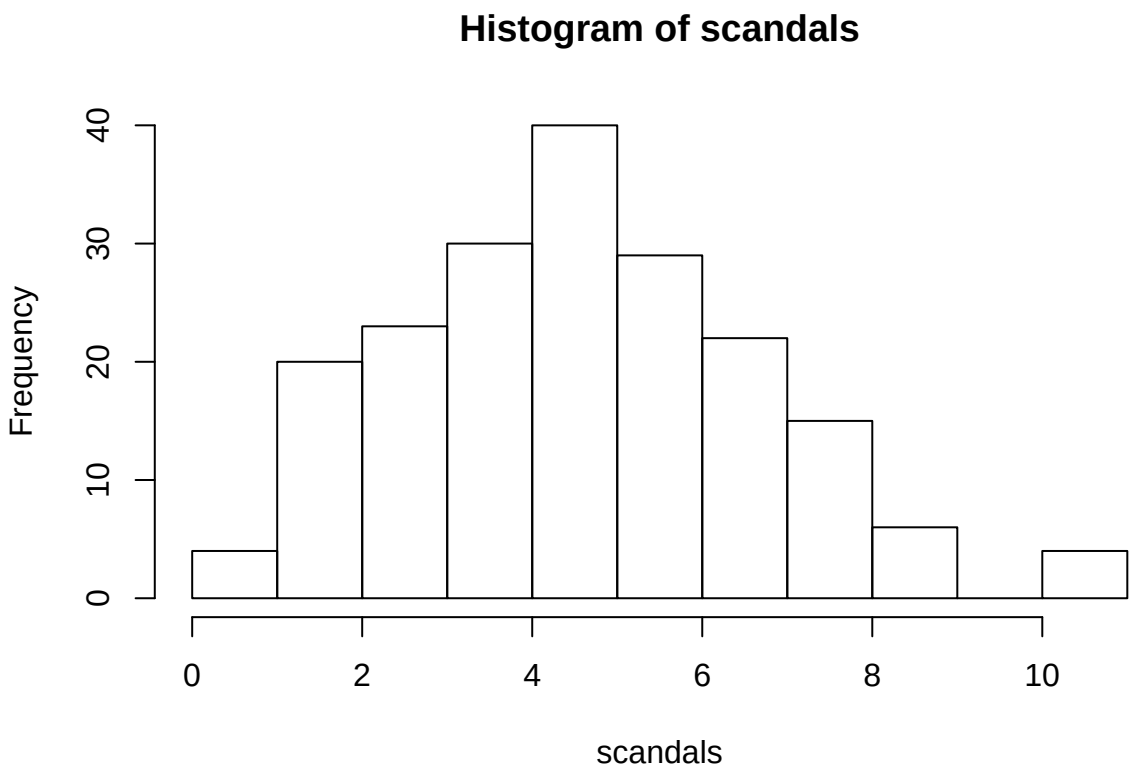
My population is 150 countries. I focus on 150 countries instead of 195, the world total, because I purposefully exclude states that are too weak and/or near-failed states to house any FDI. I consider these states to be irrational choices for FDI and so exclude measuring them. According to the Fragile States Index, an index by the US-based think tank Fund for Peace that measures state strength, there are 45 nations in the world that rank as Alert, High ALert, and Very High Alert. I consider these states to be inhabitable for FDI, and so I settle on 150 countries as my population (world number (195) - alert+ states (45)).

I simulate all my variables by drawing from distributions. My independent variable is "scandals," which refers to corruption scandals that invovled national legilsators. This variable must be 0 or greater. I simulate this variable with a poisson distribution where lamba = 5. This variable could technically have no maximum, but my random draw generates a set of variables that range from 0 to 13 scandals. My dependent variable, "fdi," represents FDI flows in and out of the country in millions. I generate this variable with a normal distribution as flows could be negative due to FDI lost. The minimum value for fdi is -1.963 million, and the maximum is 4.052 million. I also generate a GDP variable as a control. I use an exponential distribution for this variable because I do not allow GDP to be negative and because I expect values for GDP to grow in an exponential-like fashion in the sense that the gains that rich countries experience will likely always be gretaer tahn the gains a poorer country experience. Also, I intended for the range of possible GDP values to be much larger as country GDPs differ greatly from one another. For example, the US GDP is around 22 trillion, sitting as top GDP in the world, while Germany, only 3 spots below in the 4th GDP rank, has a GDP of 5.2 trillion. My minimum GDP value is 0.011 trillion, and my maximum is 23.046 trillion. Lastly, I create an aggregate variable that represents the world average FDI flow, which I set as 1.8 million given World Bank data available about estimated recent world FDI flow numbers.

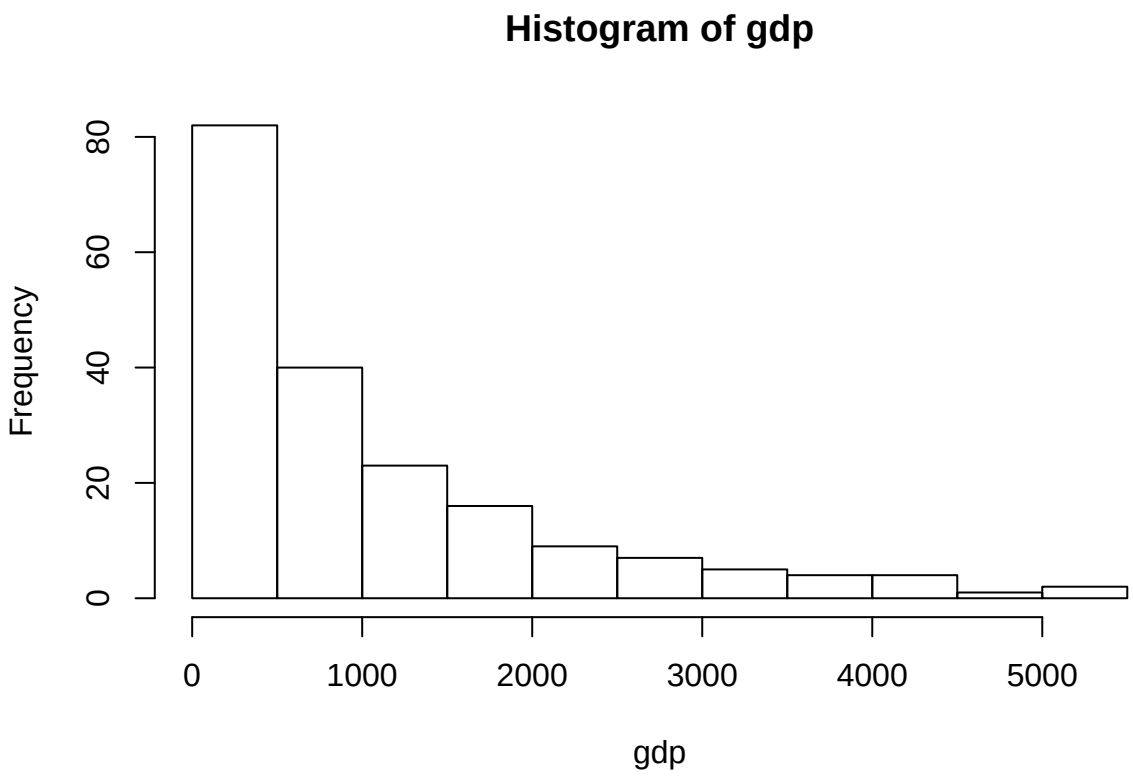
Simulating Data

```
# Setting the seed and defining N #

set.seed(11232017) # Set seed for pseudo-random number generator

N <- 193      # NUMBER OF COUNTRIES
```

```r
# building coefficients
N = 193
a= 0.05
b1 = -2.0
b2 = 1.0
error = rnorm(n=N, mean=0, sd=0.1)

# Drawing from distributions in order to create co-variates

# number of corruption scandals involving national legislators (IV)
scandals <- rpois(n=N, lambda=5)
hist(scandals)
```




```r
# GDP, in billions
gdp_mean = 1000 #theoretical mean, in billions
gdp = rexp(n=N, rate=1/gdp_mean)
hist(gdp)
```



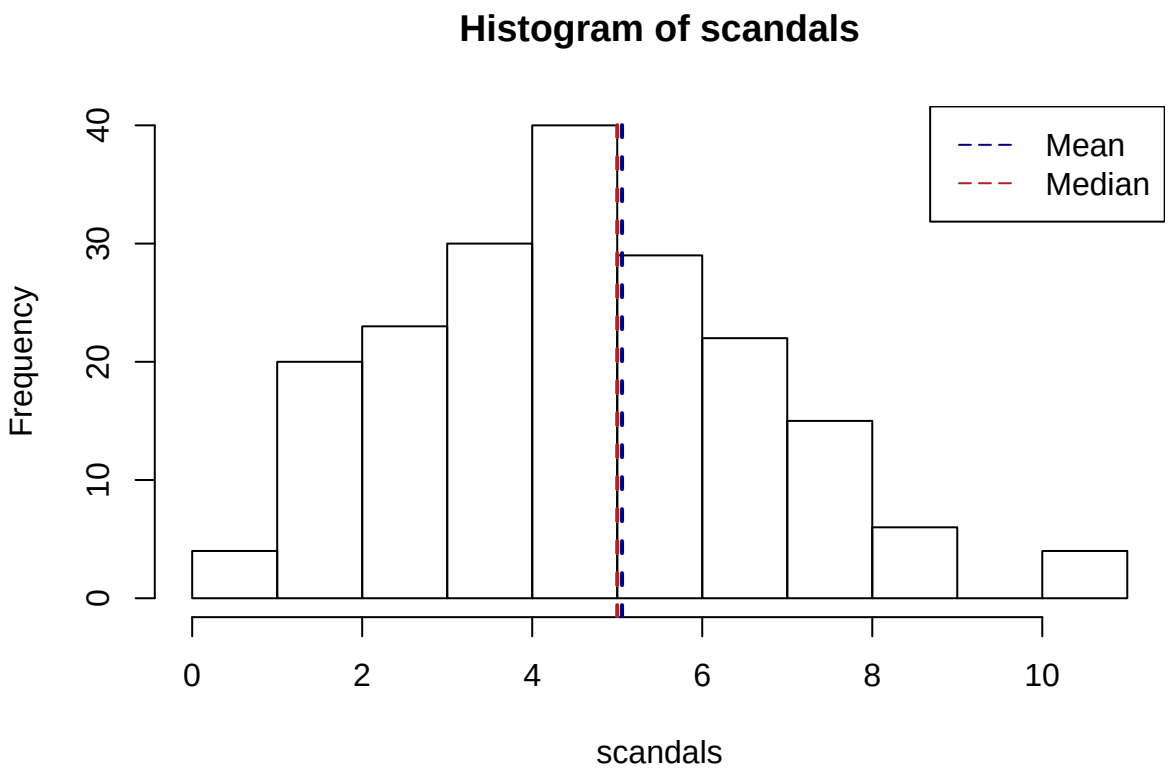

```
# aggregate level variable - world FDI flow average, in millions
world_FDI <- 1.8

# using variables and coefficients to build FDI (DV)
fdi <- a + b1*scandals + b2*gdp + error
```

Visual Data Description

Scandals

scandals	
Mean	5.233
Median	5.00
Min	0.00
Max	13.00

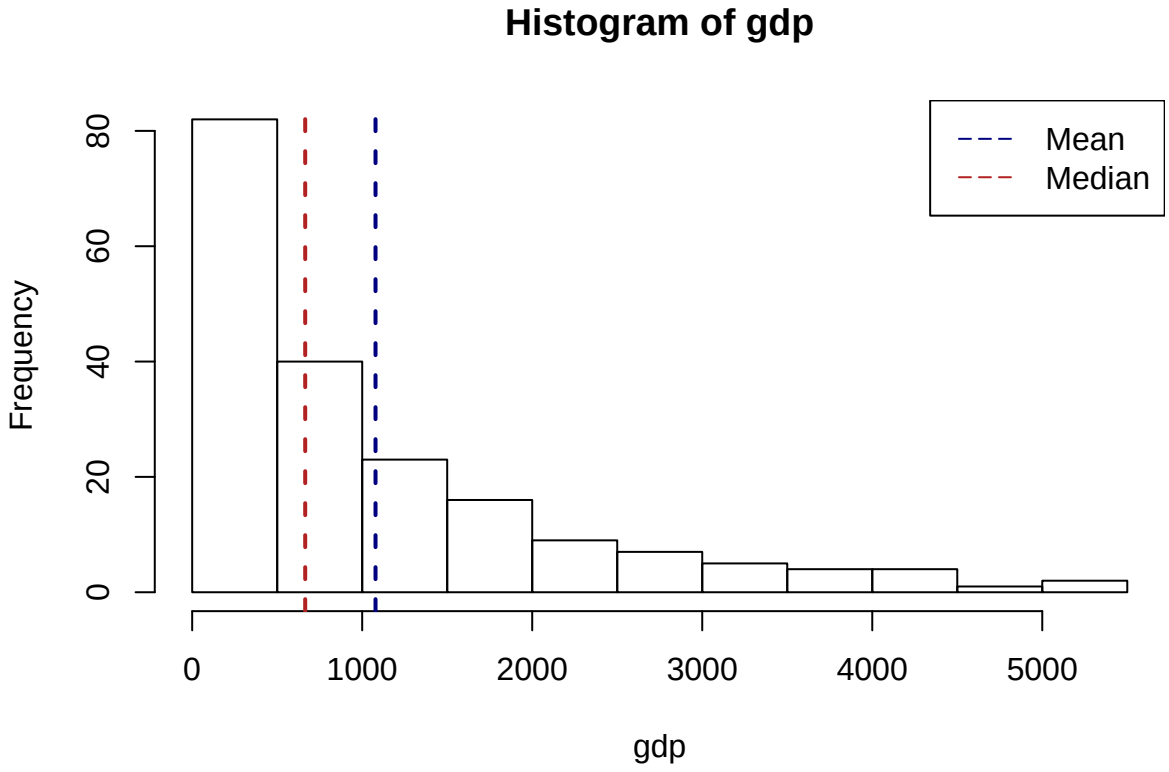
```
#Scandals

hist(scandals)
abline(v=mean(scandals), col="navy", lwd="2", lty=2)
abline(v=median(scandals), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"), col=c("navy", "firebrick"), lty=c(5,5))
```



GDP, in billions

gdp	
Mean	1078.408
Median	664.750
Min	12.544
Max	5075.653

```
hist(gdp)
abline(v=mean(gdp), col="navy", lwd="2", lty=2)
abline(v=median(gdp), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"), col=c("navy", "firebrick"), lty=c(5,5))
```



FDI flows, in millions

fdi	
Mean	1068.353
Median	652.4191
Min	-5.375017
Max	5065.724

```
{ r fdi} hist(fdi)
```

Running A Model

Linear Regression

```
regress1 <-lm(fdi ~ scandals + gdp)
summary(regress1)
```

```
##
## Call:
## lm(formula = fdi ~ scandals + gdp)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.307519 -0.068888  0.000282  0.071045  0.304609
##
## Coefficients:
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept)  0.0832714  0.0208832      3.987 9.52e-05 ***
```

```
## scandals    -2.0048985  0.0035338   -567.348  < 2e-16 ***
## gdp          1.0000002  0.0000068 147057.652  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1051 on 190 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.082e+10 on 2 and 190 DF,  p-value: < 2.2e-16
```

```
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 3.5.2
```

```
xtable(regress1)
```

```
## % latex table generated in R 3.5.1 by xtable 1.8-4 package
## % Thu Aug 22 12:18:15 2019
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & t value & Pr($>$$|$t$|$) \\
##   \hline
## (Intercept) & 0.0833 & 0.0209 & 3.99 & 0.0001 \\
##   scandals & -2.0049 & 0.0035 & -567.35 & 0.0000 \\
##   gdp & 1.0000 & 0.0000 & 147057.65 & 0.0000 \\
##    \hline
## \end{tabular}
## \end{table}
```

% latex table generated in R 3.5.1 by xtable 1.8-4 package % Thu Aug 22 11:17:13 2019

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0833	0.0209	3.99	0.0001
scandals	-2.0049	0.0035	-567.35	0.0000
gdp	1.0000	0.0000	147057.65	0.0000

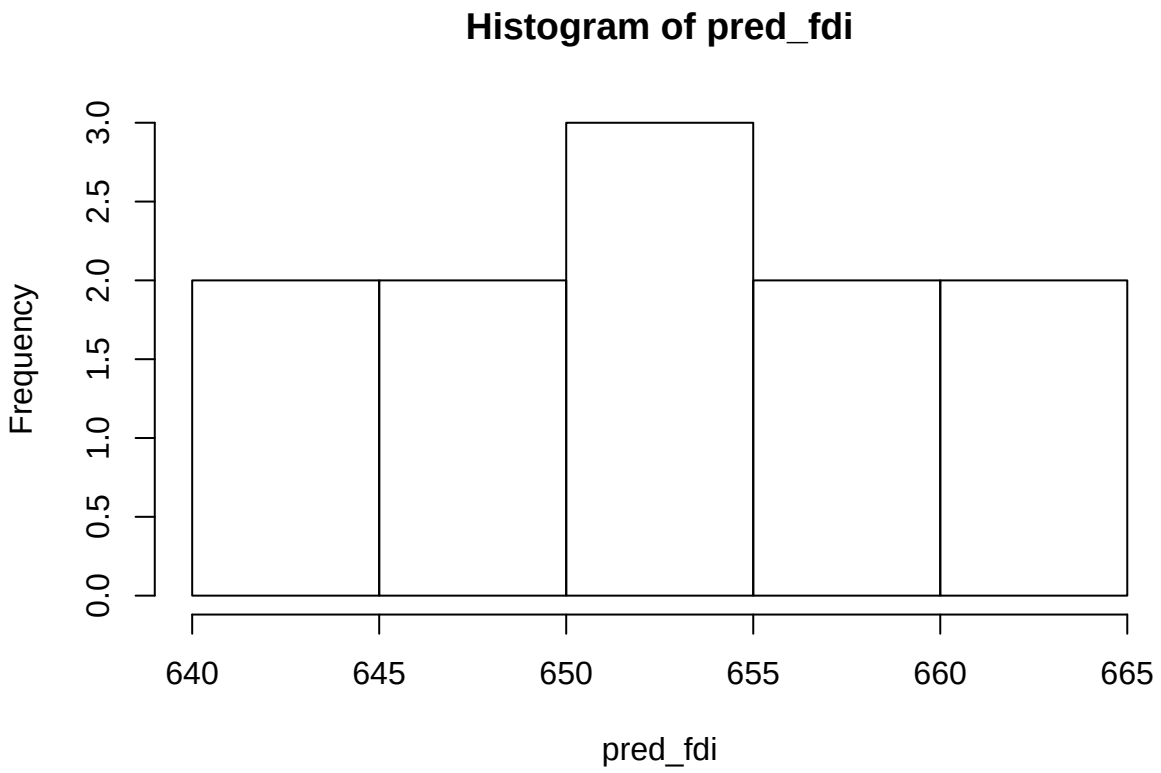
Predicted Probabilities

```
# make a vector for values of scandals to loop over
scandals_p = seq(from=1, to=11, by=1)

# make a variable to store loop results
pred_fdi <- as.numeric()
```

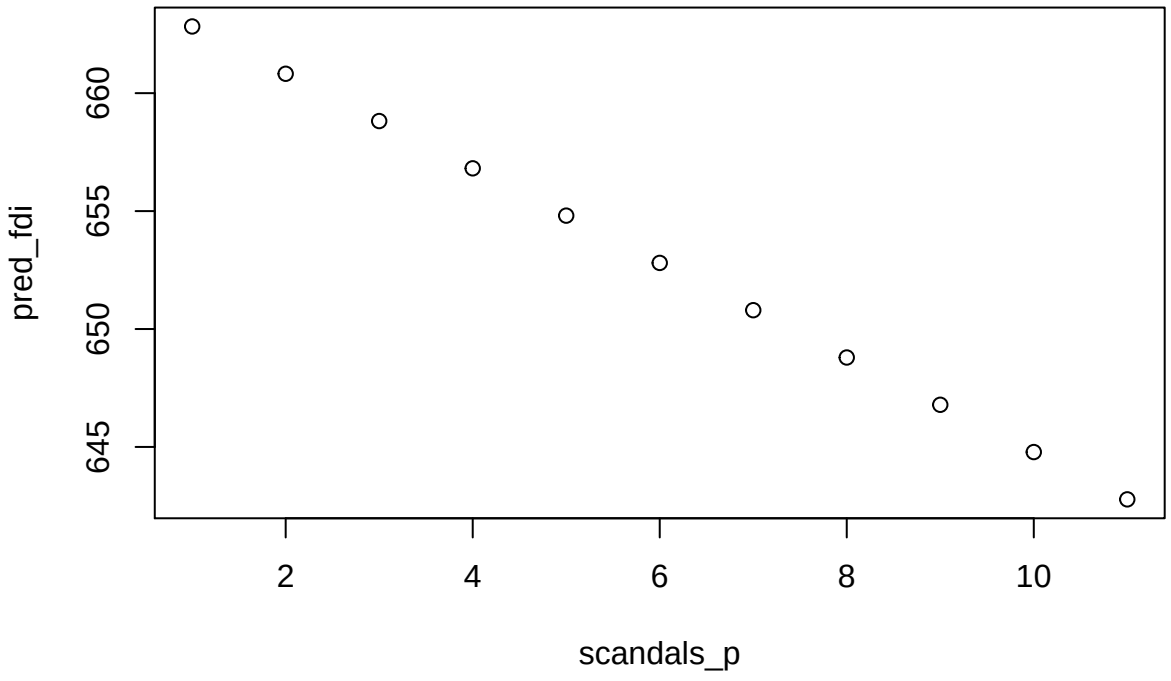
```
# make a loop for generated pred probs

for(i in 1:11){  # alternatively, you can write for(i in fdi)

  pred_fdi[i] = 0.083 + -2.005*scandals_p[i] + 1.00*median(gdp)

}

head(pred_fdi)
```

```
## [1] 662.8277 660.8227 658.8177 656.8127 654.8077 652.8027
```

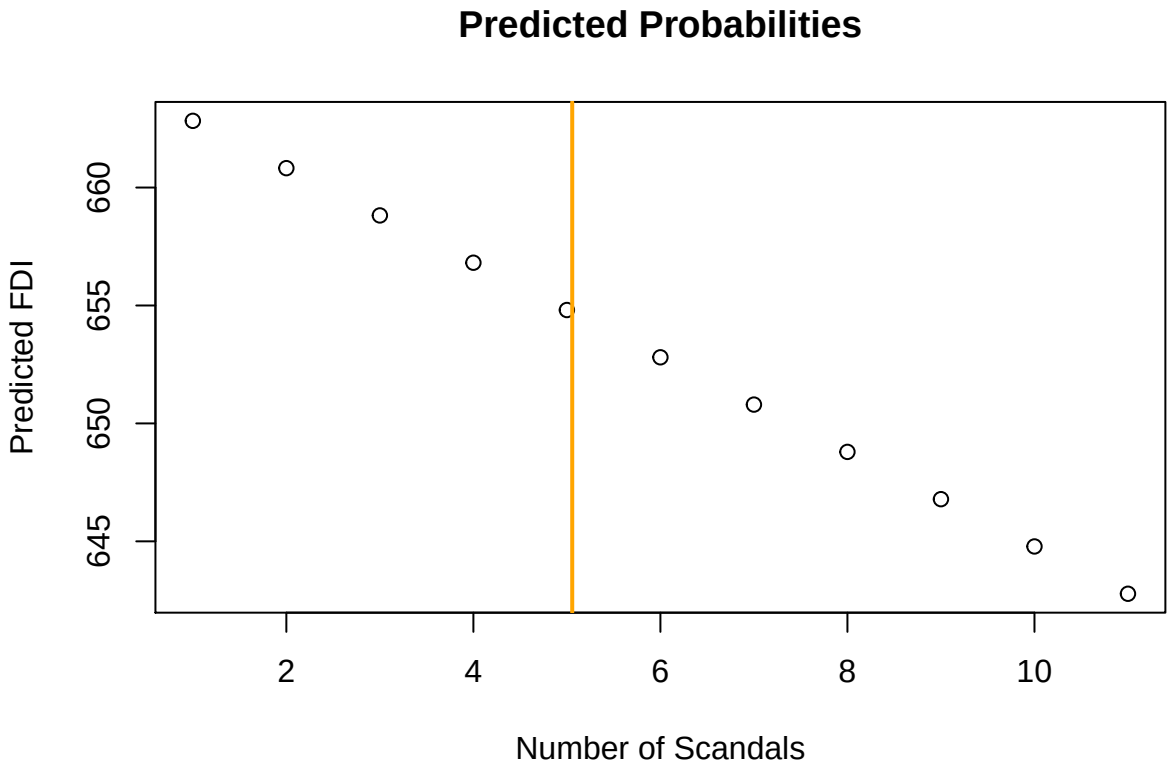
```
hist(pred_fdi)
```




```
# plotting predicted probabilities

plot(pred_fdi ~ scandals_p) #plot with gdp_p because the regular gdp variable has a
```



```r
#different variable length from pred_inf

plot(pred_fdi ~ scandals_p, xlab="Number of Scandals", ylab="Predicted FDI",
     main="Predicted Probabilities")
abline(v=mean(scandals), col="orange", lwd=2)
```



Problem 2

Research Question

For a simulated binary variable, I examine whether the amount of media references to a corruption scandal makes citizens more or less likely to vote. In my simulated data set, I imagine a scenario in which a current mayor has committed some form a corruption. This corruption was caught and made public. Media covered the scandal in various forms, such as newspapers, local newscasts, etc., but not every citizen is guaranteed to consume the same amount of information regarding the corruption scandal. This corruption scandal broke just before an election cycle in which the mayor is running for reelection. In my dataset, I examine if the amount of exposure to the corruption scandal causes citizens to be more likley to go out and vote (whether or the incumbent or opposition) or abstain. Does media exposure to corruption scandals "ignite the fight" or cause voters to withdraw from the political process? I hypothesize that exposure to media references will make citizens less likelt to go out to vote because corruption information often demobilizes the electorate as they withdraw from the political process.

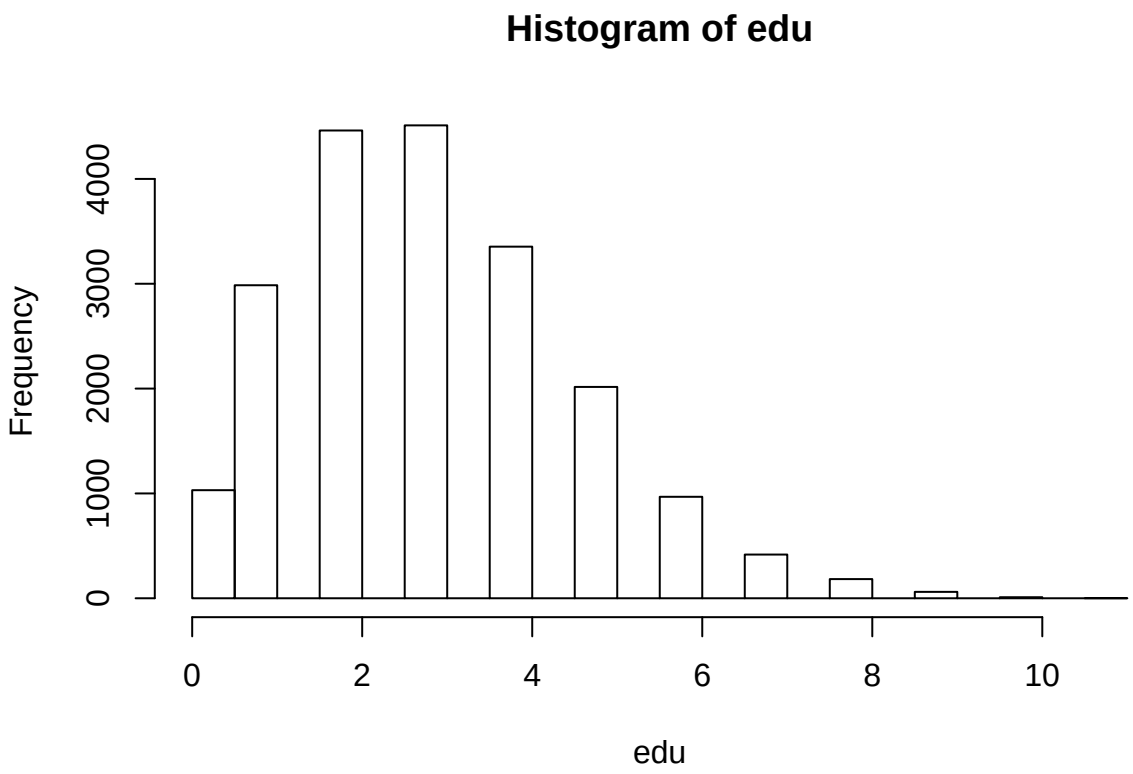
My population is 20,000, an arbitrary municipality citizen-voter age population that this hypotheticla mayor presides over. In my simulated dataset, every citizen in the city has a recorded amount of media consumed that directly discusses the mayor's corruption scandal. Every citizen also has a recorded education level and a recorded income level, as some studies have argued that those who are more educated and more wealthy are more likely to vote, and they are often more likely to have negative perceptions of corruption than less

educated, less afluent citizens. Every citizen is also coded with a binary variable as belonging to the mayor's party or belonging to the opposition party. I include an aggregate level variable of median city income level.

I simulate all my variables by drawing from distributions. My independent variable, "refs," refers to the number of media references about the mayor's corruption that a citizen could consume. I simulate this variable with a poisson distribution where lamba = 2. This variable could technically have no maximum, but my random draw generates a set of variables that range from 0 to 10 references. I generate a binary variable named "sameparty" to indicate that a voter is or isn't of the same political party as the mayor. I draw this variable from a binomial distribution where the probability of being the same party as the mayor is set at 0.45. I then generate two variables to represent education - "edu" - and income level - "income", I draw both of these from poisson distributions where lambda = 3. Both of these variables' medians are seta t 3, which I am considering an average education level, such as completed high school, and a lower-middle class income level. I create a separate, aggregate-level median education level varibable named median_edu, which is set to 3.

Simulating Data

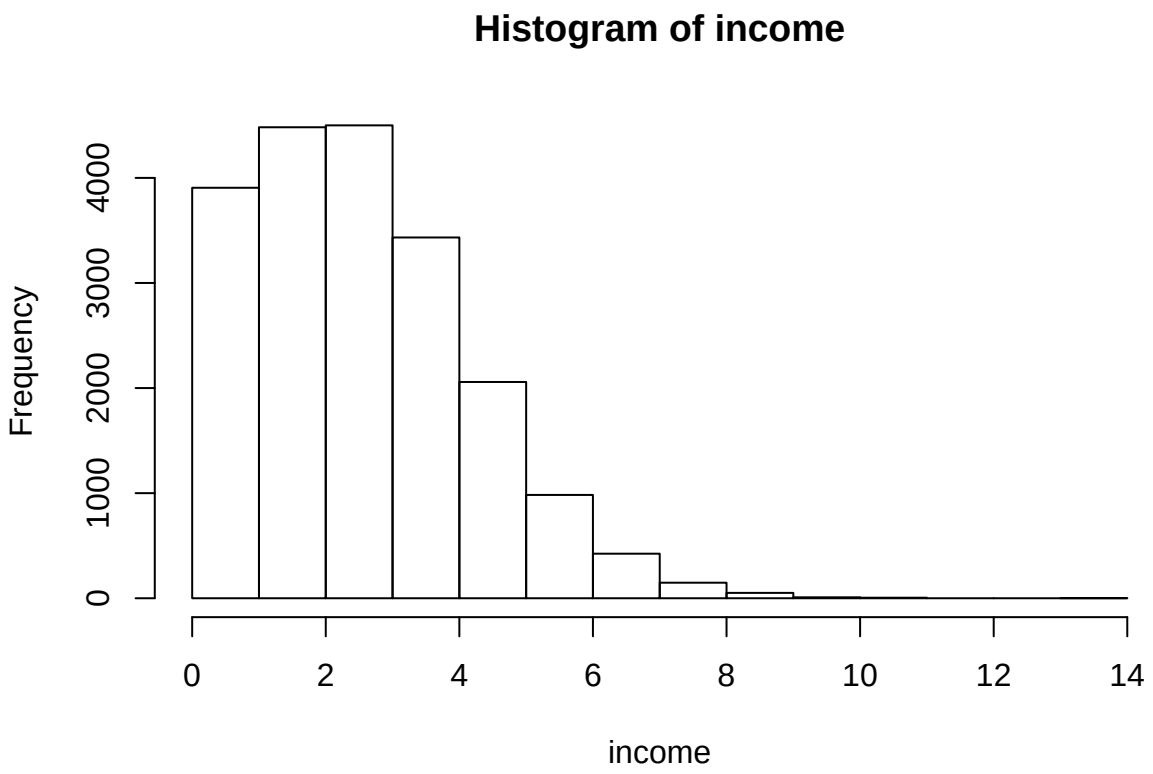
```r
set.seed(11232017) # Set seed for pseudo-random number generator

simN <- 20000      # NUMBER OF TOTAL VOTERS

# Drawing from distributions in order to create variables

# Media references consumed: refs (IV)
refs <- rpois(n=20000, lambda=2)   # Draw values from N(0,1)

# Same party as incumbent: sameparty
sameparty <- rbinom(n=20000, size=1, prob=0.45)

# Education level: edu
edu <- rpois(n=20000, lambda=3)
hist(edu)
```




```
max(edu)
```

```
## [1] 11
```

```
# median education level: median_edu
median_edu <- 3

# Income level: income
income <- rpois(n=20000, lambda=3)
hist(income)
```




```
max(income)
```

```
## [1] 14
```

Hypothetical expectations for dependent variable, voting (Y_0) vs. asbtaining (Y_1)

```
# error: e
e <- rnorm(n=20000, mean=0, sd=1)

#alpha, intercept
a <- 0.5

# Whether or not you will turn out to vote or abstain (DV)
Y_0 <- a + -2*refs + 2*sameparty + -0.7*edu + -0.5*income + e #received refs --> more likely to abstain
Y_1 <- a + 0*refs + 2*sameparty + -0.7*edu + -0.5*income + e #no refs --> more likely to vote
```

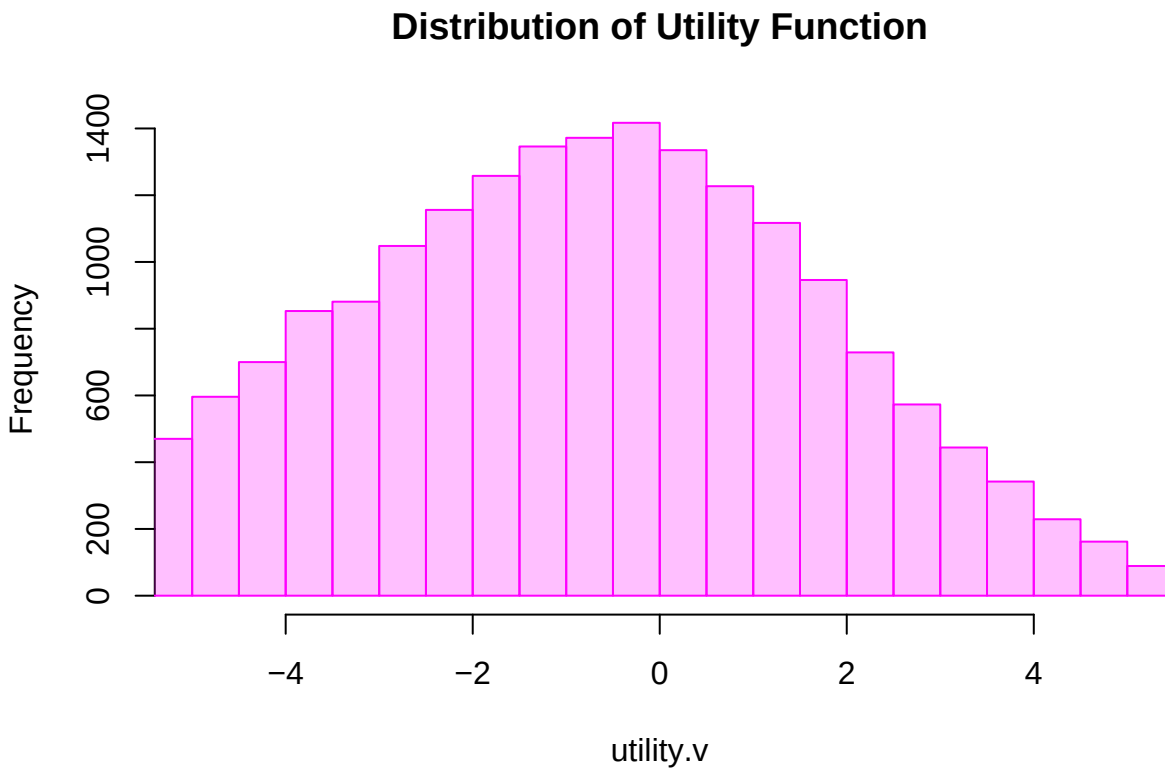
```
pop_data <- data.frame(Y_0, Y_1, refs, sameparty, edu, income) # Population level data
head(pop_data)
```

```
##         Y_0         Y_1 refs sameparty edu income
## 1 -1.945843  0.05415672    1         1   1      3
## 2 -2.763254 -2.76325389    0         0   1      5
## 3 -5.872179  0.12782069    3         1   4      1
## 4 -7.915184 -1.91518426    3         0   2      1
## 5 -9.389238 -3.38923787    3         0   3      6
## 6 -4.356314 -2.35631424    1         0   3      2
```

Creating variables coefficients and utility function for voting

```
# Creating coefficients for all variables

  error      <- rnorm(n=simN, mean=-1, sd=1)    # HOMOSKEDASTIC ERROR
  b.refs     <- runif(n=simN, min=-1.7, max=-1.7) # COEF OF REFS: START FROM A SIMPLE MODEL
```

```r
  b.sameparty <- runif(n=simN, min=1.3, max=1.3) # COEF OF PARTYID: START FROM A SIMPLE MODEL
  b.edu <- runif(n=simN, min=0.7, max=0.7) # COEF OF EDU
  b.income <- runif(n=simN, min=0.2, max=0.2) #COEF OF INCOME

# Creating a utility function for voting vs. for abstaining

utility.v <- b.refs*refs + b.sameparty*sameparty + b.edu*edu + b.income*income + error  # UTILITY FUNCT

  hist(utility.v, breaks=40, col = "#ff00ff40", border = "#ff00ff", xlim=c(-5,5),
     main="Distribution of Utility Function")
```

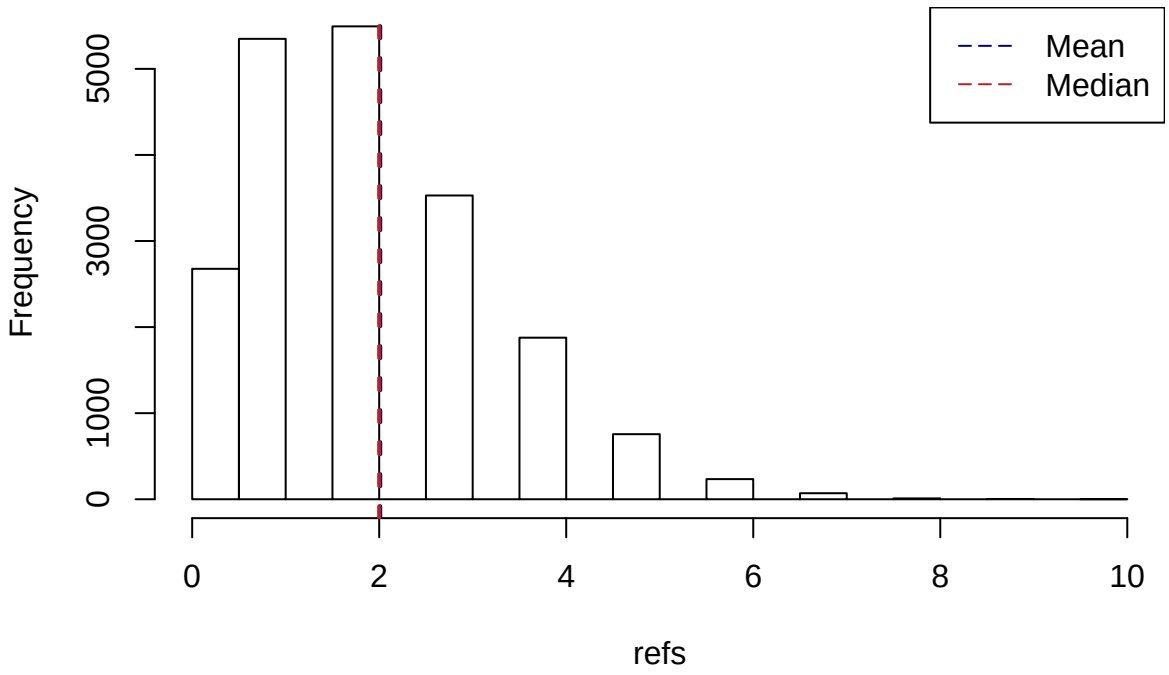



```r
vote_yes <- ifelse(utility.v > 0, 1, 0)        # Vote if Utility > 0
pop_data$vote_yes <- vote_yes
```

Visual Data Description

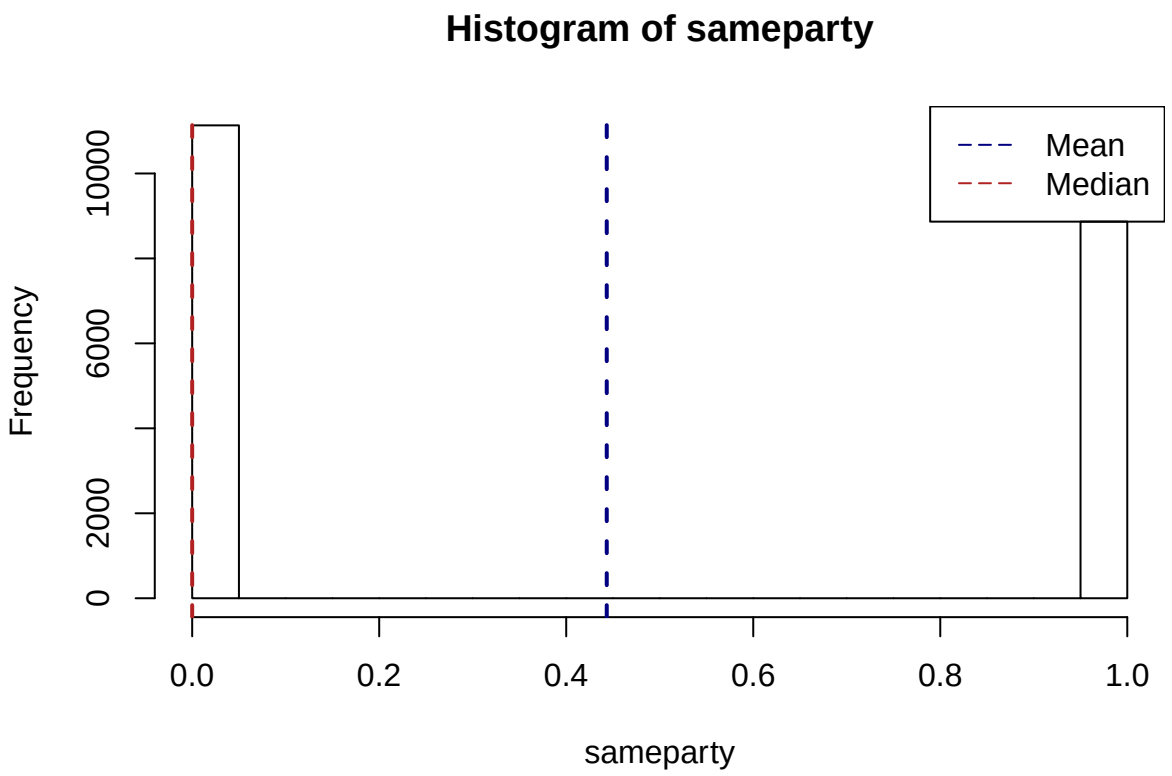
References refs | ——|—— Mean | 5.233 Median | 5.00 Min | 0.00 Max | 13.00

```r
# References
hist(refs)
abline(v=mean(refs), col="navy", lwd="2", lty=2)
abline(v=median(refs), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"), col=c("navy", "firebrick"), lty=c(5,5))
```



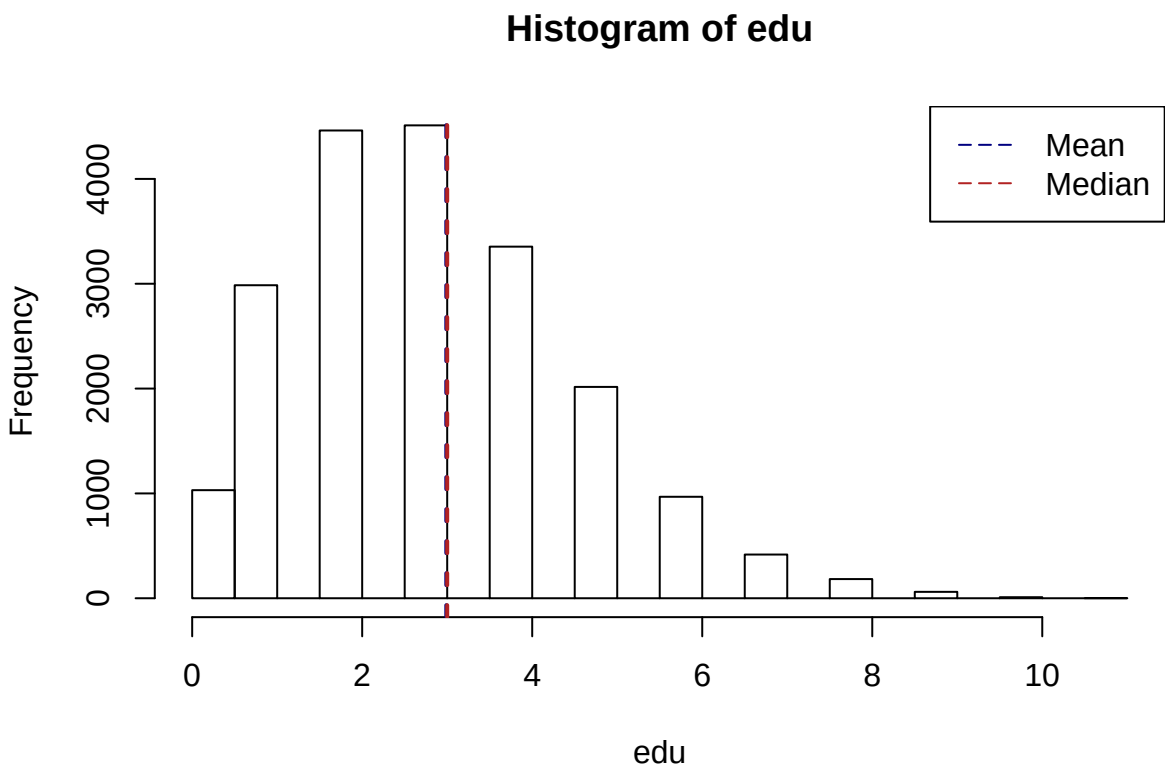
Same Party sameparty | ———-|——— Mean | 0.452 Median | 0.00 Min | 0.00 Max | 1.00

```
# Same party as mayor
hist(sameparty)
abline(v=mean(sameparty), col="navy", lwd="2", lty=2)
abline(v=median(sameparty), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"), col=c("navy", "firebrick"), lty=c(5,5))
```



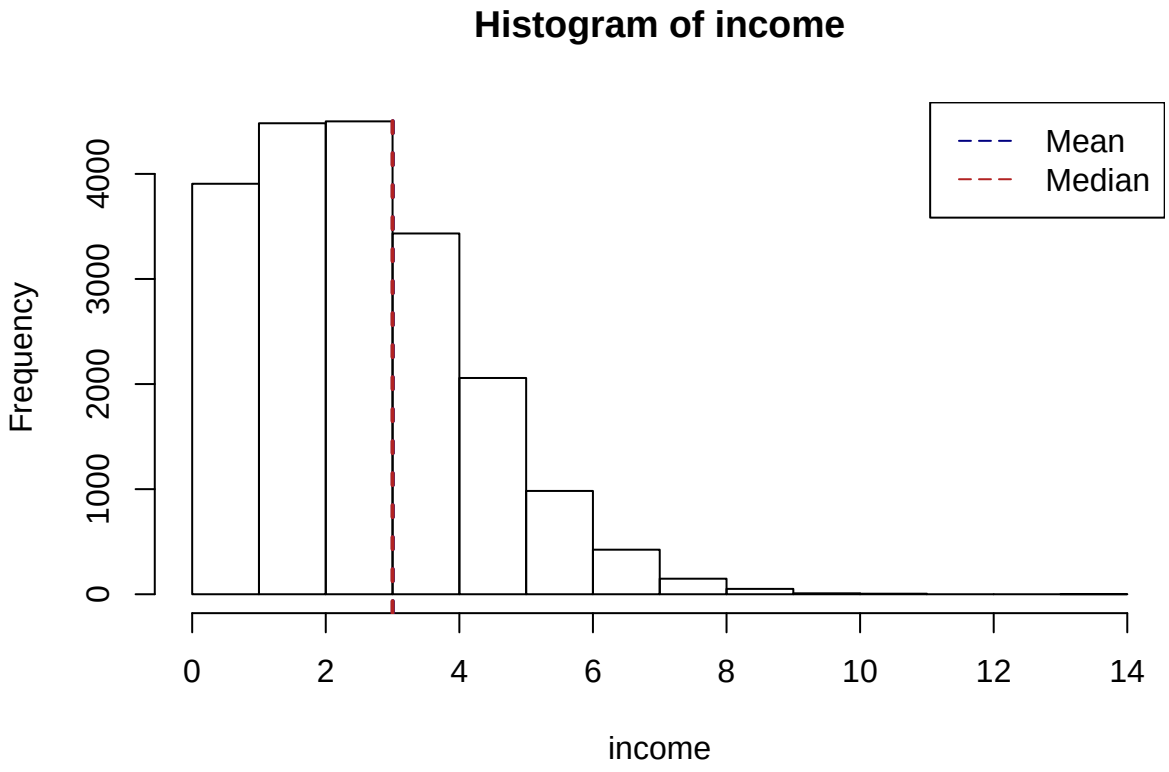
Education Level edu | —-|——— Mean | 3.012 Median | 3.00 Min | 0.00 Max | 12.00

```
# Education Level
hist(edu)
abline(v=mean(edu), col="navy", lwd="2", lty=2)
abline(v=median(edu), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"), col=c("navy", "firebrick"), lty=c(5,5))
```



Income Level income | ——-|—— Mean | 2.983 Median | 3 Min | 0.00 Max | 12.00

```
# Income Level
hist(income)
abline(v=mean(income), col="navy", lwd="2", lty=2)
abline(v=median(income), col="firebrick", lwd="2", lty=2)
legend("topright", legend=c("Mean", "Median"), col=c("navy", "firebrick"), lty=c(5,5))
```



Running A Model

```
# Running a logit with simulated data

m_logit <- glm(vote_yes ~ refs + sameparty + edu + income, family=binomial(link="logit"))

summary(m_logit)
```

```
##
## Call:
## glm(formula = vote_yes ~ refs + sameparty + edu + income, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0809  -0.2793  -0.0296   0.2270   3.4320
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.84177    0.08133  -22.65   <2e-16 ***
## refs        -3.05009    0.04971  -61.36   <2e-16 ***
## sameparty    2.31659    0.06070   38.16   <2e-16 ***
## edu          1.27340    0.02383   53.44   <2e-16 ***
## income       0.35769    0.01578   22.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 26282.7  on 19999  degrees of freedom
## Residual deviance:  9717.8  on 19995  degrees of freedom
## AIC: 9727.8
##
## Number of Fisher Scoring iterations: 7
```

```
library(xtable)
xtable(m_logit)
```

```
## % latex table generated in R 3.5.1 by xtable 1.8-4 package
## % Thu Aug 22 12:18:20 2019
## \begin{table}[ht]
## \centering
## \begin{tabular}{rrrrr}
##   \hline
##  & Estimate & Std. Error & z value & Pr($>$$|$z$|$) \\
##   \hline
## (Intercept) & -1.8418 & 0.0813 & -22.65 & 0.0000 \\
##   refs & -3.0501 & 0.0497 & -61.36 & 0.0000 \\
##   sameparty & 2.3166 & 0.0607 & 38.16 & 0.0000 \\
##   edu & 1.2734 & 0.0238 & 53.44 & 0.0000 \\
##   income & 0.3577 & 0.0158 & 22.66 & 0.0000 \\
##    \hline
## \end{tabular}
## \end{table}
```

```
head(pop_data)
```

```
##         Y_0         Y_1 refs sameparty edu income vote_yes
## 1 -1.945843  0.05415672    1         1   1      3        0
## 2 -2.763254 -2.76325389    0         0   1      5        1
## 3 -5.872179  0.12782069    3         1   4      1        0
## 4 -7.915184 -1.91518426    3         0   2      1        0
## 5 -9.389238 -3.38923787    3         0   3      6        0
## 6 -4.356314 -2.35631424    1         0   3      2        0
```

% latex table generated in R 3.5.1 by xtable 1.8-4 package % Thu Aug 22 11:19:12 2019

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.8418	0.0813	-22.65	0.0000
refs	-3.0501	0.0497	-61.36	0.0000
sameparty	2.3166	0.0607	38.16	0.0000
edu	1.2734	0.0238	53.44	0.0000
income	0.3577	0.0158	22.66	0.0000

Predicted Probabilities

```
range(pop_data$refs)
```

```
## [1]  0 10
```

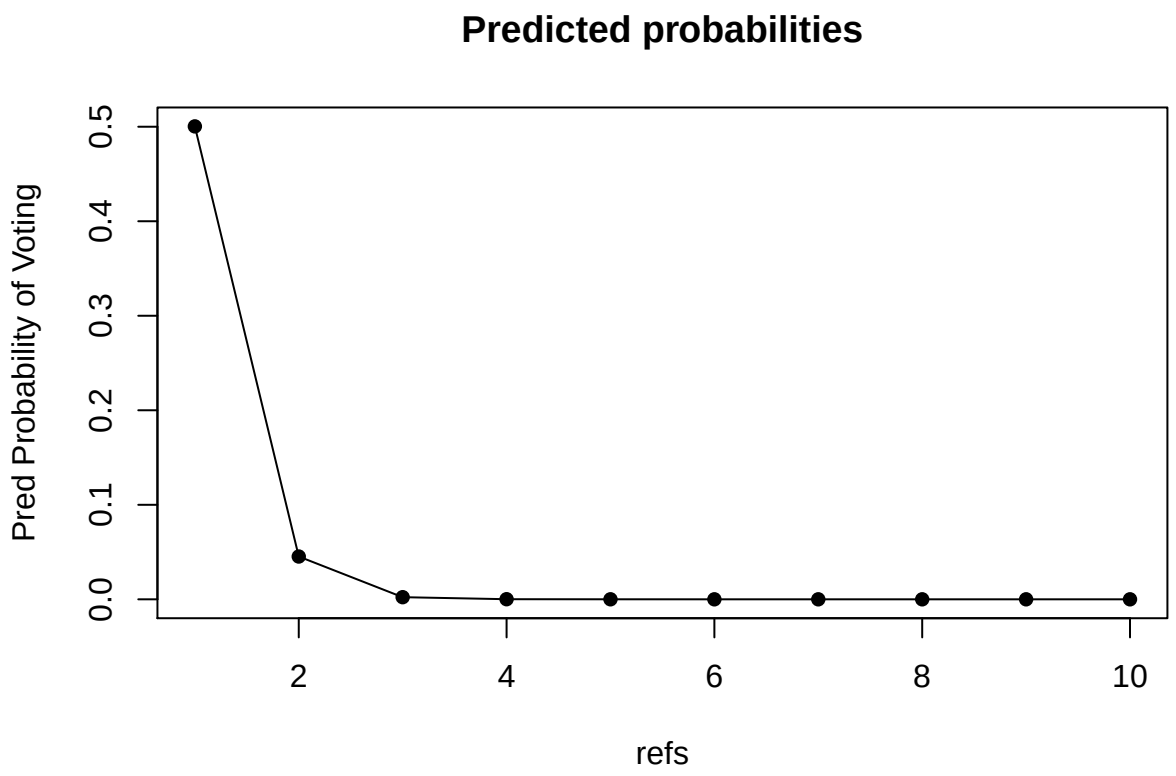
```
xrange <- seq(from=0, 10, by=1)

PredProb <- function(model, xvar, xstart, xend, xind){  # Five arguments
```

```r
  #   xvar <- enquote(xvar)
  coefs <- coef(model)                            # Coefficient list
  indnames <- attr(model$terms, "term.labels")    # Names of variables
  xmat <- model$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=xind)      # 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)
  coefs <- as.matrix(coefs)                       # Coefficient matrix (6 by 1)
  linagg <-  typical_dat %*% coefs                # Linear aggregators

  pred <- exp(linagg)/(1 + exp(linagg))           # Predicted probabilities

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

  return(pred)
}

# Let's use our function here
summary(m_logit)
```

```
##
## Call:
## glm(formula = vote_yes ~ refs + sameparty + edu + income, family = binomial(link = "logit"))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0809  -0.2793  -0.0296   0.2270   3.4320
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.84177    0.08133  -22.65   <2e-16 ***
## refs        -3.05009    0.04971  -61.36   <2e-16 ***
## sameparty    2.31659    0.06070   38.16   <2e-16 ***
## edu          1.27340    0.02383   53.44   <2e-16 ***
## income       0.35769    0.01578   22.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26282.7  on 19999  degrees of freedom
```

```
## Residual deviance:  9717.8  on 19995  degrees of freedom
## AIC: 9727.8
##
## Number of Fisher Scoring iterations: 7
```

```
PredProb(m_logit, refs, 1, 10, 1)
```




```
##               [,1]
##  [1,] 5.003538e-01
##  [2,] 4.527474e-02
##  [3,] 2.240610e-03
##  [4,] 1.063303e-04
##  [5,] 5.035747e-06
##  [6,] 2.384673e-07
##  [7,] 1.129254e-08
##  [8,] 5.347548e-10
##  [9,] 2.532314e-11
## [10,] 1.199169e-12
```