Final Paper

Anahi Rodriguez

April 19, 2021

Question of Thesis & Motivation for Research

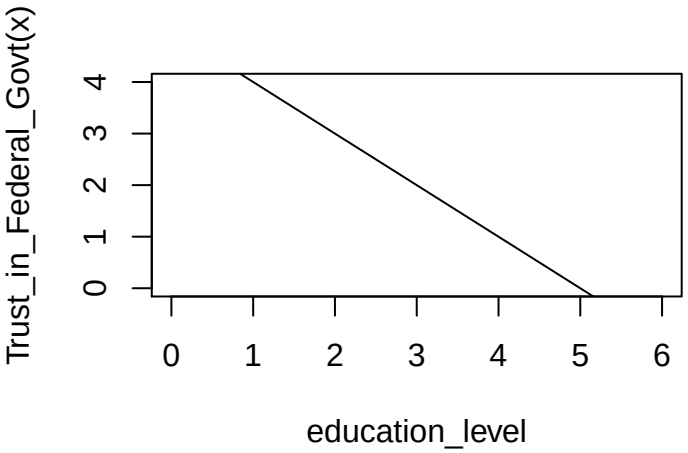
I am planning to address the relationship between amt. of trust in the federal government to do what is right (outcome variable) and education level (explanatory variable). I am interested in seeing if as people become more educated, do they begin to trust the federal government to do what is right more or less of the time. I am also interested in viewing this relationship with two control variables: income and year. I would like to see how holding different groups (5 levels) of income constant over time affects this relationship in either direction.

This research is important becuase it can give us insight into how levels of trust in the federal government change with education and time, but also becuase politicians or other individuals/groups can use these trends to understand their constituents belief's and worries'. Additionally, these trends could help us create new studies to gain even further insight such as for example, trying to figure out what specifically it is about becoming more educated that makes you trust the federal government either less or more.

Expectations / Theory & Hypotheses

We should expect to see a linear relationship between trust in the federal government to do what is right and education level when controlling for year and income. I predict that, holding income and year constant, as people become more educated, they will tend to trust the federal government to do what is right less and less of the time, creating a negative relationship. I believe that this is the relationship between these two variables because I know that generally when people become more educated, they tend to move left ideologically, and I also am aware that in more recent years/decades, polarization within the US has increased. I use Nathaniel Persily's definition of polirization in his book: Solutions to Political Polirization in America, "ideological convergence within parties and divergence between parties – what we might call 'hyperpar- tisanship' "(4 Persily). Therefore, people tend to identify with one side or the other but stay far from the center more and more every year - indicating that we should also control for year - , and that education would only push this already occurring phenomena farther out to either end depending on an individuals education level. If this is the case, I would presume that it would be more difficult for the government to satisfy the wants of these individuals who are moving farther from the center, causing those individuals to believe that they can trust the government to do what is right less and less as time passes. Income will also be controlled for here becuase income level not only affects an individuals ability to recieve an education but also becuase there usually is a relationship between individuals in different income brackets and different political ideologies

- **H__O** : There is no relationship between education level and the amount of trust someone has for the federal government when holding income and year constant
- **H__A** : There exists a relationship between education level and the amount of trust someone has for the federal government when holding income and year constant



Data Description

Dataset

The data was collected from 1948 to 2016 and attempts to ask the same question, using different wording over the years. Questions are asked during interviews; some of which are asked in two separate interviews: one pre-presidential election and one post-presidential election. Questions vary in scope with some being open to interpretation or relating to important political figures that are relevant during the year or general time period while others are always relevant to all respondents such as total income level. These numbers and data can be trusted because according the ANES FAQs, "These materials are based on work supported by the National Science Foundation under grant numbers SES 1444721, 2014-2017, the University of Michigan, and Stanford University", backing their work to reliable sources. This data set contains 59944 observations/cases. Each of these is an individual's answers during interview questions over their own identity, public opinions/attitudes, and questions regarding voting/election behavior for that individual.

The concepts and how I plan on measuring them

My first concept is level of trust in the federal government to do what is right: `trust_fed_gov`. This variable is measured by interview question results. The interview question has different variations , for different years – asking a respondent to choose the amount of trust they have in the federal government to do what is right. This is a 5-category question where respondents can rate the frequency of how often they believe the government will do what is right or wrong. I believe that this is a good measure of trust in the federal government because respondents can only choose one category and this category would be reflective of their own opinions and feelings towards the federal government. However, one category is "DK; depends" and is the last factor, so it has been recoded to be directly in the center of the 5 factors in order to more accurately use the category variable in order from least to most trust.

- **trust_fed_gov**: Measure of how much of the time an individual believes that they can trust the federal government to do what is right (0: None of the time, 1: Some of the time, 2: Don't know/Depends, 3: Most of the time, and 4: Just about always)

Next, another interview question intended to measure education level: `education_level`, asks different variations, for different years – asking respondents to answer how many years of school they have finished. This is a 7-category question that respondents can place themselves into based on their total level of education. I believe that this is the most straightforward of my concepts in terms of how it is measured because it is not a question about feelings but rather a single number or answer for every individual. Additionally, I believe that this question is not one that respondents would tend to misinterpret as they may others.

- **education_level**: Measure of respondents education level completed (0: 8 grades or less ('grade school'), 1: 9-12 grades ('high school'), no diploma/equivalency, 2: 12 grades, diploma or equivalency, 3:

12 grades, diploma or equivalency plus non-academic, 4: Some college, no degree; junior/community college, 5: BA level degrees, 6: Advanced degrees incl. LLB)

Lastly, another interview question intended to measure total income: `income`, asks different variations , for different years – asking a respondent to make an estimate for their total family income that year and based on this number to place themselves within one of 5 categories. Each respondent's own answers may be subject to their knowledge of how different levels of income vary, so some respondent's answers/choices may not be the most accurate in describing their income level as a range of percentiles.

- **income**: Where respondent places themselves within income ranges for total family income that year (1: 0 to 16 percentile, 2: 17 to 33 percentile, 3: 34 to 67 percentile, 4: 68 to 95 percentile, 5: 96 to 100 percentile)

My second control variable is year

- **year**: Denotes year of interview

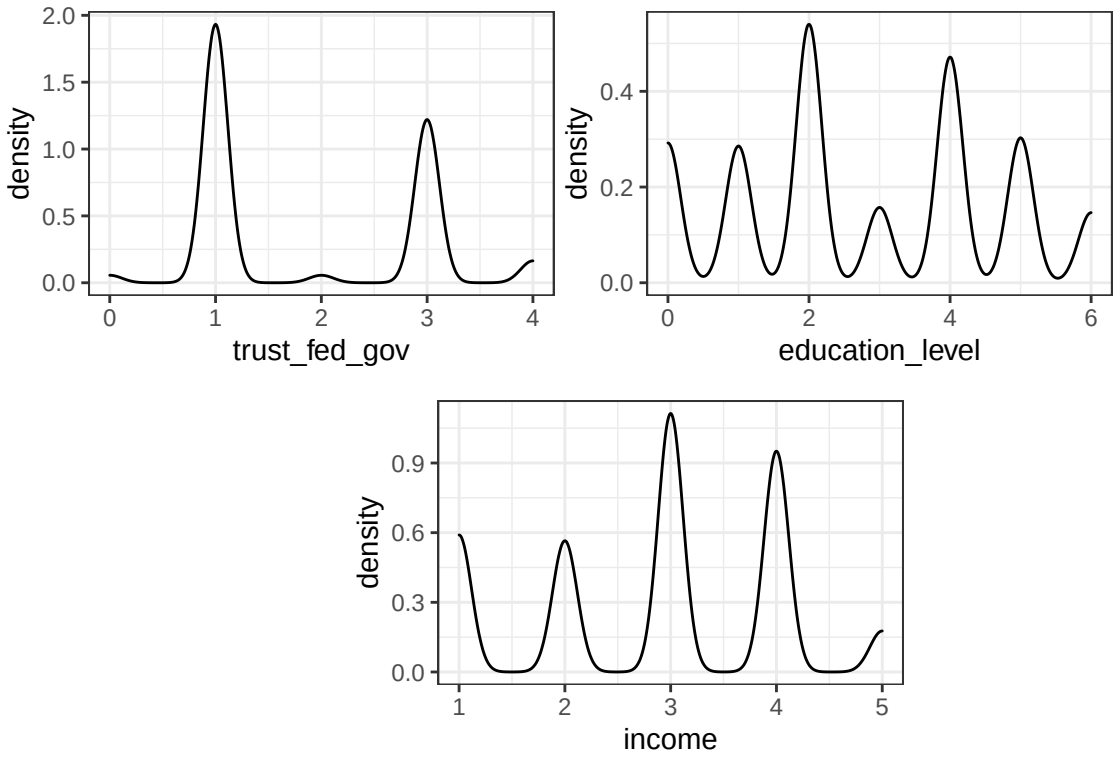


Figure 1: Showing the distribution of key variables

Because all of our variables, besides year, whose distribution was not visualized, have distributions that are not normal, we should use the interquartile range rather than mean to describe the variables.

Table 1: Descriptive Statistics of Levels of Trust in the Federal Government, Education Levles, and Income Levels

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
trust_fed_gov	39,924	1.855	1.075	0.000	1.000	3.000	4.000
education_level	58,696	2.786	1.812	0.000	1.000	4.000	6.000
income	54,495	2.870	1.154	1.000	2.000	4.000	5.000

Analysis of Data

lm1: Control Variable Model

Predicted values from the model

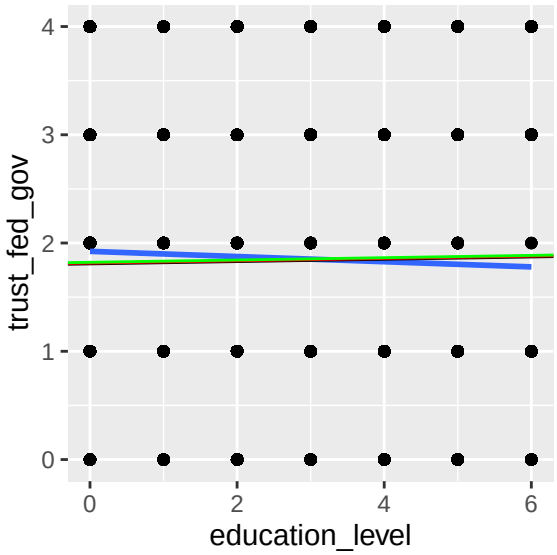
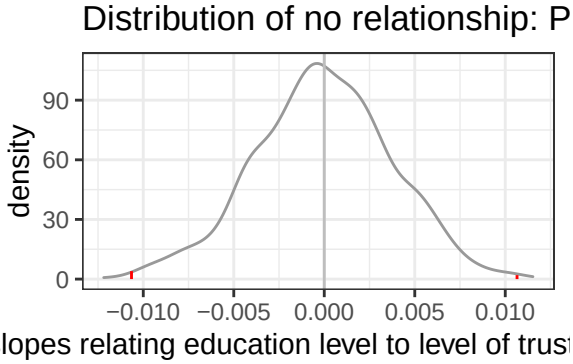


Figure 2: There exists a negative relationship between the education level of an individual and the amount of trust they have in the federal government to do what is right however, with higher levels of income, there exists a positive relationship

For this model, our intercepts based on income level divided into low, mid, and high are 1.807422, 1.814658, and 1.821895 respectfully. We expect that on average with an increase of 1 in `education_level`, there to be an increase of 0.011 in the `trust-fed_gov` response variable. We also see a greater increase in the `trust-fed_gov` variable for higher income groups and show by the slight differences in the intercepts for different income level groups.

pvalue for `education_level` for lm1

This figure shows the hypothesized "no relationship" distribution (sometimes called the "permutation distribution" because we randomly permute the values of the outcome).



The calculated pvalue for the education_level coefficient is 0.002 which is sufficiently low enough based on an alpha value of 0.05 to reject the null and say that there is sufficient evidence to suggest that there is a

relationship between education level of an individual and their level of trust in the federal government to do what is right.

Confidence interval for `education_level` for lm1

Based on our calculated 95% confidence interval of roughly (0.00361664, 0.0177613), we reject the null becuase our null value, 0, is not within the confidence interval so we say that there is sufficient evidence to suggest the alternative hypothesis.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, May 09, 2021 - 14:41:30

Table 2:

	Dependent variable:
	trust_fed_gov
education_level	0.011^{***}
	(0.004)
income	0.004
	(0.005)
year	-0.018^{***}
	(0.0004)
Constant	36.806^{***}
	(0.796)
Observations	35,532
R^2	0.055
Adjusted R^2	0.055
Residual Std. Error	1.046
F Statistic	688.690^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

lm2: Interaction Variable Model

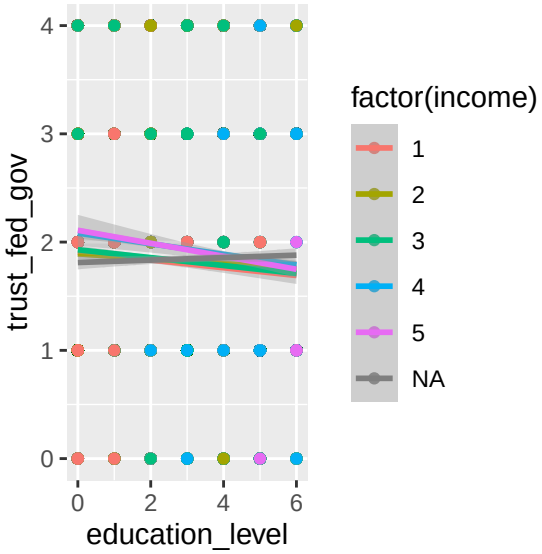
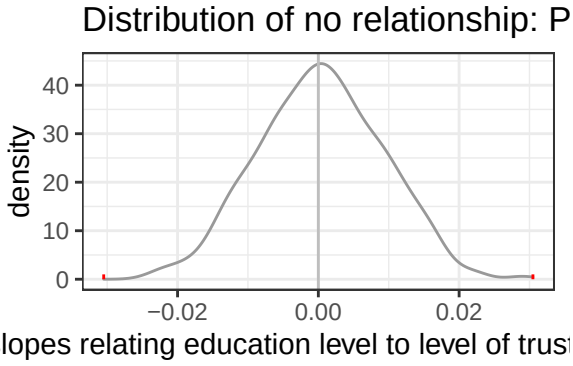


Figure 3: There exists a negative relationship between the education level of an individual and the amount of trust they have in the federal government to do what is right

For this model, our intercepts based on income level divided into low, mid, and high are 1.778096, .822115, and 1.866135 respectfully. We expect that on average with an increase of 1 in `education_level`, there to be an increase of -35.04068 in the `trust-fed_gov` response variable for low income individuals, -35.05437 for middle income, and -35.06807 for high income individuals .

pvalue for `education_level` for lm2

This figure shows the hypothesized "no relationship" distribution (sometimes called the "permutation distribution" because we randomly permute the values of the outcome).



The calculated pvalue for the education_level coefficient is 0.0 which is sufficiently low enough based on an alpha value of 0.05 to reject the null and say that there is sufficient evidence to suggest that there is a relationship between education level of an individual and their level of trust in the federal government to do what is right.

Confidence interval for `education_level` for lm2

We find a confidence interval of (0.012441,0.04893636). Becuase our null value, 0, is not within this confidence interval, we reject the null and say that there is sufficient evidence to suggest the alternative.

Regression Table for Model 2 - Interaction Model

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, May 09, 2021 - 14:42:47

Table 3:

	Dependent variable:
	trust_fed_gov
education_level	0.030^{***}
	(0.009)
income	0.022^{**}
	(0.009)
year	-0.018^{***}
	(0.0004)
education_level:income	-0.007^{**}
	(0.003)
Constant	36.820^{***}
	(0.796)
Observations	35,532
R^2	0.055
Adjusted R^2	0.055
Residual Std. Error	1.046
F Statistic	518.187^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01

Assessing model fits

Model 1, the control model, has a sum squared residual value of 38898.9 while model 2, the interaction term model, has a sum squared residual value of 38891.93. By this measure, model 2 is a better fit to the data. Model 1 has an R^2 value of 0.05496 while model 2 has an R^2 value of 0.05513, indicating that model 2 also is a better fit to the data because more of the variability can be explained by model 2 than model 1, although both models seem to fit the data poorly based on these values.

Reference

- American National Election Studies. Time Series Cumulative Data File. https://electionstudies.org/data-center/anes-time-series-cumulative-data-file/. Accessed on March 20, 2021.
- Persily, Nathaniel. "Introduction." Solutions to Political Polarization in America, edited by Nathaniel Persily, Cambridge University Press, Cambridge, 2015, pp. 3–14.

Code Appendix

```
#######################
## Data and Variables Section



  ##fed_trust_gov has different descriptions attached to numbers 1 - 5
  ##recode to have just the single number without description
  levels(df$trust_fed_gov)
  summary(df$trust_fed_gov)

  df$trust_fed_gov <- recode_factor(df$trust_fed_gov, "1. None of the time (VOLUNTEERED); almost never

  ## 1 originally named "1. None of the time (VOLUNTEERED); almost never (1966"
  ## 2 originally named "2. Some of the time"
  ## 3 originally named "3. Most of the time"
  ## 4 originally named "4. Just about always"
  ## 5 originally named "9. DK; depends"


  ##rearranging 3,4, and 5 for each other because those observations in trust coded as 5
  ##fit in the scale better at the center instead of after 4 because that indicates a
  ##higher level of trust when it is not (see old names above)
  ##then rearranging these from 0-4 instead to easily interpret intercept in models later

  levels(df$trust_fed_gov)
  table(df$trust_fed_gov)

  summary(df$trust_fed_gov)
  df$trust_fed_gov<- as.numeric(as.character(df$trust_fed_gov))



  ##education_level has different descriptions attached to numbers 1 - 7
  ##recode to have just the single number without description
  levels(df$education_level)
  summary(df$education_level)

  ## 1 originally named "1. 8 grades or less ('grade school')"
  ## 2 originally named "2. 9-12 grades ('high school'), no diploma/equivalency"
  ## 3 originally named "3. 12 grades, diploma or equivalency"
  ## 4 originally named "4. 12 grades, diploma or equivalency plus non-academic"
  ## 5 originally named "5. Some college, no degree; junior/community college"
  ## 6 originally named "6. BA level degrees"
  ## 7 originally named " 7. Advanced degrees incl. LLB"
```

```
  ##then rearranging these from 0-6 instead to easily interpret intercept in models later

  df$education_level <- recode_factor(df$education_level, "1. 8 grades or less ('grade school')" = "0",

  levels(df$education_level)
  table(df$education_level)


  summary(df$education_level)
  df$education_level<- as.numeric(as.character(df$education_level))







  ##income has different descriptions attached to numbers 1 - 5
  ##recode to have just the single number without description
  levels(df$income)
  summary(df$income)

  ## 1 originally named "1. 0 to 16 percentile"
  ## 2 originally named "2. 17 to 33 percentile"
  ## 3 originally named "3. 34 to 67 percentile"
  ## 4 originally named "4. 68 to 95 percentile"
  ## 5 originally named "5. 96 to 100 percentile"

  df$income <- recode_factor(df$income, "1. 0 to 16 percentile" = "0",   "2. 17 to 33 percentile" = "1"

  levels(df$income)
  table(df$income)




  ##need to change levels to numeric instead of factor
  df$income <- as.numeric(df$income)
  summary(df$income)
  df$income<- as.numeric(as.character(df$income))



ggplot(df) + geom_density(aes(trust_fed_gov)) + theme_bw()
ggplot(df) + geom_density(aes(education_level)) + theme_bw()




ggplot(df) + geom_density(aes(income)) + theme_bw()


stargazer::stargazer(df[,c("trust_fed_gov", "education_level", "income")], title="Descriptive Statistics
                     header=F)
```

```r
#do i need to change the iqr range to be 68% instead of the 50% that is the default? how do i do this?

########################
## Expectations Section
#prediction function
Trust_in_Federal_Govt <- function(Education_level){
    5.0 - 1.0 * Education_level
}

# Based on the prediction function, you can plot your expectation:
xname = "education_level"
curve(Trust_in_Federal_Govt,from=0,to=50,xlim=c(0,6),ylim=c(0,4), xlab = xname)



########################
## Analysis and Results Section
lm1 = lm(trust_fed_gov~education_level + income + year, data = df)
coef(lm1)




 ggplot(df, aes(x=education_level, y=trust_fed_gov)) + geom_point() +
  geom_smooth(method="lm", formula="y~x") +
  geom_abline(slope = 0.010650017 , intercept = 1.807422) +
  geom_abline(slope = 0.010650017, intercept = 1.814658, color = "red") +
  geom_abline(slope = 0.010650017, intercept = 1.821895, color = "green")




slope_mod1 = coef(lm1)[2]
slope_mod1

low_income_intercept = coef(lm1)[1] + coef(lm1)[3]*1 + coef(lm1)[4]*mean(df$year, na.rm = TRUE)
low_income_intercept

middle_income_intercept = coef(lm1)[1] + coef(lm1)[3]*3 + coef(lm1)[4]*mean(df$year, na.rm = TRUE)
middle_income_intercept

high_income_intercept = coef(lm1)[1] + coef(lm1)[3]*5 + coef(lm1)[4]*mean(df$year, na.rm = TRUE)
high_income_intercept


stargazer::stargazer(lm1, df = F)



set.seed(210418)
nsamps<-1000
lm1.perm<-do(nsamps)*lm(shuffle(trust_fed_gov)~education_level + income + year , data=df)
coef(lm1)
```

```r
## Remember that you are doing a permutation test of the education level term
lm1.observed <- coef(lm1)[2] #education level coefficient

(thep.lessthan <- sum(lm1.perm$education_level<=lm1.observed)/nrow(lm1.perm))
(thep.greaterthan <-sum(lm1.perm$education_level>=lm1.observed)/nrow(lm1.perm))
p_value_education <- (ptwosided<-2*min(thep.lessthan, thep.greaterthan))
p_value_education

summary(lm1)



library(ggplot2)
ggplot() + geom_density(aes(lm1.perm$education_level), color="gray60") + theme_bw() +
  labs(x="Difference in slopes relating education level to level of trust in the federal gov't",
     title = "Distribution of no relationship: Permutation") +
  geom_vline(xintercept = 0, color="gray") +
  geom_segment(aes(x=coef(lm1)[2], xend=coef(lm1)[2], y=0, yend=2),color="red") +
  geom_segment(aes(x=-coef(lm1)[2], xend=-coef(lm1)[2], y=0, yend=4),color="red")



mylm.fn<-function(){ ## naming the function
  thelm = lm(trust_fed_gov~education_level + income + year,data=resample(df))## running several regress
  thecoef<-coef(thelm) ## record the coefficients
  return(thecoef) ## Tell us the coefficients
}


lm2.bs <- do(500)*mylm.fn()
confint(lm2.bs$education_level, 0.95)
(mySE <- sd(lm2.bs$education_level))
mySE




lm2 = lm(trust_fed_gov~education_level*income + year, data = df)
coef(lm2)


ggplot(df, aes(x=education_level, y=trust_fed_gov, color = factor(income))) + geom_point() +
  geom_smooth(method="lm", formula="y~x")

coef(lm2)
slope_mod2_low = coef(lm2)[2] + coef(lm2)[5] * 1 + coef(lm2)[4]*mean(df$year, na.rm = TRUE)
slope_mod2_low

slope_mod2_mid = coef(lm2)[2] + coef(lm2)[5] * 3 + coef(lm2)[4]*mean(df$year, na.rm = TRUE)
slope_mod2_mid

slope_mod2_high = coef(lm2)[2] + coef(lm2)[5] * 5 + coef(lm2)[4]*mean(df$year, na.rm = TRUE)
slope_mod2_high
```

```r
intercept_mod2_low = coef(lm2)[1] + coef(lm2)[3] * 1 + coef(lm2)[4]*mean(df$year, na.rm = TRUE)
intercept_mod2_low

intercept_mod2_mid = coef(lm2)[1] + coef(lm2)[3] * 3 + coef(lm2)[4]*mean(df$year, na.rm = TRUE)
intercept_mod2_mid

intercept_mod2_high = coef(lm2)[1] + coef(lm2)[3] * 5 + coef(lm2)[4]*mean(df$year, na.rm = TRUE)
intercept_mod2_high


stargazer::stargazer(lm2, df = F)



set.seed(210418)
nsamps<-1000 #more times might give a number other than 0
lm2.perm<-do(nsamps)*lm(shuffle(trust_fed_gov)~education_level * income + year , data=df)
coef(lm2)

## Remember that you are doing a permutation test of the education level term
lm2.observed <- coef(lm2)[2] #education level coefficient

(thep.lessthan2 <- sum(lm2.perm$education_level<=lm2.observed)/nrow(lm2.perm))
(thep.greaterthan2 <-sum(lm2.perm$education_level>=lm2.observed)/nrow(lm2.perm))
p_value_education2 <- (ptwosided<-2*min(thep.lessthan2, thep.greaterthan2))
p_value_education2




library(ggplot2)
ggplot() + geom_density(aes(lm2.perm$education_level), color="gray60") + theme_bw() +
  labs(x="Difference in slopes relating education level to level of trust in the federal gov't",
     title = "Distribution of no relationship: Permutation") +
  geom_vline(xintercept = 0, color="gray") +
  geom_segment(aes(x=coef(lm2)[2], xend=coef(lm2)[2], y=0, yend=1),color="red") +
  geom_segment(aes(x= - coef(lm2)[2], xend= - coef(lm2)[2], y=0, yend=1),color="red")



mylm.fn<-function(){ ## naming the function
  thelm = lm(trust_fed_gov~education_level * income + year,data=resample(df))## running several regress
  thecoef<-coef(thelm) ## record the coefficients
  return(thecoef) ## Tell us the coefficients
}


lm2.bs <- do(500)*mylm.fn()
confint(lm2.bs$education_level, 0.95)
(mySE <- sd(lm2.bs$education_level.income))
mySE
```

```r
ssr.mod.fn<-function(mod){ sum(residuals(mod)^2) }

ssr_lm1 = ssr.mod.fn(lm1)
ssr_lm1

ssr_lm2 = ssr.mod.fn(lm2)
ssr_lm2

summary(lm1)
summary(lm2)
```