title: “QMSS-5015-Lab-2”
author: “Adrian Varallyay”
date: “October 12, 2020”

Setting working directory and file access, and library

```
setwd("/Users/adrianvarallyay/Downloads")
gss = read.csv("GSS.2006.csv.xls")
library(psych)
library(gmodels)

library(ggplot2)
```

1. Recode 1 *sort of* continuous variable into categories. Tell me what you did and explain the variable(s).

The *sort of* continuous variable I chose to look at was *gasregs.* This variable inquired about respondents’ opinion on requiring manufacturers to make cars and trucks more fuel-efficient. As table 1 shows, the original variable had five ordinal levels ranging from strongly favor (1) and strongly oppose (5). I decided to restructure the data into a binary variable. I chose to set up the dummy variable as “favor” and “oppose”, assigning the strongly favor group as “favor”, and assigning all other levels to “oppose”, as figure 1 indicates. Since “strongly oppose” was the most assertive claim of support, I figured this would be the best cut off point as its reasonable to assume that they had the least misgivings to oppose the initiative.

Table 1:

533) How much do you favor or oppose requiring car makers to make cars and trucks that use less gasoline? Do you strongly favor, favor, neither favor nor oppose, oppose, or strongly oppose such a requirement?

	TOTAL	**%**
1) Strongly favor	645	69.5
2) Favor	197	21.2
3) Neither favor nor oppose	52	5.6
4) Oppose	14	1.5
5) Strongly oppose	9	1.0
8) Don't know	10	1.1
9) No answer	1	0.1

Missing	3582	
TOTAL	928	100.0

Figure 1:

```
## Breaking a variable into categories:
## create a binary variable (category) for favoring or opposing requiring car makers to make car
s and trucks that use less gasoline
gss$gasregs.cat = cut(gss$gasregs, breaks = c(-1, 1, 8), label=c("favor","oppose"), ordered=TRUE
)

## check code
table(gss$gasregs.cat, gss$gasregs)

##
##            1   2   3   4   5
##   favor  645   0   0   0   0
##   oppose   0 197  52  14   9
```

2. Recode 1 other variable and attach value labels. Tell me what you did and explain the variable(s).

The second variable I chose to recode was *sciinfgw.* This variable inquired about respondents' opinion on how much influence environmental scientists should have in deciding on the approaches that should be taken to address global warming. As table 2 indicates, the original variable had four ordinal levels ranging from a great deal of influence (1) to none at all (4). I decided to restructure the data by reversing the order to reflect the increase in influence (rather than decrease). As Figure 2 indicates, once I reversed the order, I then transformed the response options into factors and collated them with the ascending order of influence: none, a little, a fair amount, and a great deal. For example, as the table in Figure 2 indicates, the 30 respondents in the none at all group (4) are now assigned to the none group (1). The variable is now structured to reflect the increase of support for environmental scientists, ranging from none to a great deal.

Table 2:

527) How much influence should Environmental scientists have in deciding what to do about global warming? Would you say a great deal of influence, a fair amount, a little influence, or none at all?

	TOTAL	**%**
1) A great deal of influence	438	47.2

2) A fair amount	347	37.4
3) A little influence	73	7.9
4) None at all	30	3.2
8) Don't know	38	4.1
9) No answer	2	0.2
Missing	3582	
TOTAL	928	100.0

Figure 2:

```
## Reverse code a variable and then add labels and make it ordered:
## To reverse code: (highest category + 1) - orginal_variable
gss$r_sciinfgw = 5-gss$sciinfgw

## make into a factor
gss$r_sciinfgw.fact = as.factor(gss$r_sciinfgw)

## make factor variable into an ORDERED factor, with value labels: a great deal of influence, a f
air amount, a little influence, or none at all, reversed so that influence increases from 1 to 4
gss$lab.r_sciinfgw <- ordered(gss$r_sciinfgw, levels = c(1,2,3,4), labels = c("none", "a little", "a f
air amount", "a great deal" ))

## check code
table(gss$lab.r_sciinfgw, gss$r_sciinfgw)

##
##                  1   2   3    4
##   none          30   0   0    0
##   a little       0  73   0    0
##   a fair amount  0   0 347    0
##   a great deal   0   0   0  438

## the original variable was numeric, mean
mean(gss$sciinfgw, na.rm=T)

## [1] 1.656532

## the new variable, lab.r_sciinfgw, is an ordered factor -- telling R to treat it like a number, hen
ce, the "as.numeric"
mean(as.numeric(gss$lab.r_sciinfgw), na.rm=T)
```

```
## [1] 3.343468
```

3. Use one (or both) of your recoded variables to do a cross-tabulation (like last week, with prop.table, doBy, or ddply). Explain your results.

I decided to cross-tabulate *gasregs* from above and *scibstgw.* The Scibstgw variable inquired about respondents' opinion on the extent environmental scientists should have on making policy recommendations about global warming. As table 3 indicates, the original variable for *scibstgw* had a 5-point scale (ordinal levels) ranging from what is best for the country (1) to own narrow interests (4). In other words, this scale reflected the extent of respondents' beliefs on whether an environmental scientist would attempt to influence policy for their own personal interests or the interest of what is best for the country.

Before cross-tabulating, I transformed *scibstgw* into a dummy variable and renamed it *best_for_country*. I split the variable into two bins, collecting the first two respondent groups as best_for_country (1) and the rest of the three groups (3,4, & 5) as own narrow interest (0). I figured splitting in the middle would have been ideal, but not applicable to this variable. As such, I took the first two groups (ostensibly the most conviction in this direction) and made them one, and I included the middle group (3) with other two groups as own narrow interests. Seeing as it would be more difficult to argue the middle groups level of conviction, I thought this was a fair reformation of the variable.

The results of the cross-tabulation can be seen in figure 3. Of the 882 respondents, 269 (30%) believed the own narrow interest narrative, and 613 (70%) believed the best-for-country narrative. On average, regardless of whether a respondent believes an environmental scientist would make policy recommendations by personal or country-based interests, it seems as though they will still support requiring manufacturers to make cars and trucks more fuel-efficient. Of the 269 who believe an environmental scientist would make policy recommendations for their own interest, 173 (64%) favored the more fuel-efficient initiative. Of the 613 who believe an environmental scientist would make policy recommendations for the best for the country, 449 (73%) favored the more fuel-efficient initiative. There doesn't appear to be any systematic relationship here. Essentially, both groups (best for country and own narrow interests) favor/oppose fuel-efficient cars at similar rates.

Table 3:

530) When making policy recommendations about global warming, on a scale of 1 to 5, to what extent do you think the following groups would support what is best for the country as a whole versus what serves their own narrow interests? A. Environmental scientists ()

	TOTAL	**%**

1) What is best for the country	368	39.7
2) 2	245	26.4
3) 3	159	17.1
4) 4	50	5.4
5) Own narrow interests	61	6.6
8) Don't know	43	4.6
9) No answer	2	0.2
Missing	3582	
TOTAL	928	100.0

Figure 3:

```
## Dummy variable for scibstgw: "When making policy recommendations about global warming
, on a scale of 1 to 5, to what extent do you think Environmental scientists would support what i
s best for the country as a whole versus what serves their own narrow interests? best for countr
y coded as 1; 0 otherwise
gss$best_for_country = ifelse((gss$scibstgw<3), 1, 0)

## same var as Q1 above: favoring or opposing requiring car makers to make cars and trucks th
at use less gasoline
gss$gasregs.cat = cut(gss$gasregs, breaks = c(-1, 1, 8), label=c("favor","oppose"), ordered=TRUE
)

CrossTable(gss$gasregs.cat, gss$best_for_country, prop.r=F, prop.c=T, prop.t=F, prop.chisq=F, f
ormat="SPSS")

## 
##    Cell Contents
## |-------------------------|
## |                   Count |
## |          Column Percent |
## |-------------------------|
## 
## Total Observations in Table:  882 
## 
##                 | gss$best_for_country 
## gss$gasregs.cat |        0  |        1  | Row Total | 
## ----------------|-----------|-----------|-----------|
##        favor |     173  |     449  |     622  |
##              |  64.312% |  73.246% |          |
## -------------|-----------|----------- |-----------|
##       oppose |      96  |     164   |     260  |
##              |  35.688% |  26.754% |          |
## -------------|----------- | -----------    | -----------|
##   Column Total |     269  |     613  |     882  |
##              |  30.499% |  69.501% |          |
## ----------------|-----------|-----------|-----------|
##
```

4. Run a linear regression with 1 independent and 1 dependent variable; make all of the recodes necessary to make the model as easy to interpret as possible; and explain your results.

I used *best_for_country* from above and *colsci* as my variables to run in a linear regression. As table 4 indicates, *colsci* reflects whether respondents ever took a college level science course. I recoded this variable as r_colsci, and made it binary with yes as 1 and no as 0. I did this so that the "taken science college course" variable would increase if the individual took at least one science course. The y-intercept is 0.67, meaning that if a respondent did not take a college science course, about 67% of the time they still lean towards believing that environmental scientists would make policy recommendations based on the country's interest[1]. The coefficient for *r_colsci* (0.0499) shows that taking a college science course increases the extent to which an individual believes environmental scientist would make policy recommendations by what is best for the country. However, the *r_colsci* variable was statistically insignificant meaning we cannot reject the null hypothesis – taking a college science course has no seeming effect greater than zero on whether an individual believes environmental scientist would make policy recommendations by own personal interests or best-for-country interests). The R-squared (.002) explains 0.2% of the variance in the model. The predictive power of the model does not seem particularly good.

Table 4:

517) Have you ever taken any college-level science courses?

	TOTAL	**%**
1) Yes	824	44.2

[1] I'm a little iffy on this interpretation. If not percentage of time, would it simply be 0.67 scale points (out of 1)? Since the variables are not continuous this doesn't really make sense to me. Thus, I went with a percentage interpretation.

2) No	1035	55.5
8) Don't know	4	0.2
9) No answer	1	0.1
Missing	2646	
TOTAL	1864	100.0

Figure 4:

```
## recode colsci: 517) Have you ever taken any college-level science courses? yes coded as 1, 0 o
therwise
gss$r_colsci = ifelse(gss$colsci==1, 1, 0)

## Dummy variable for scibstgw (same as above): "When making policy recommendations abou
t global warming, on a scale of 1 to 5, to what extent do you think Environmental scientists woul
d support what is best for the country as a whole versus what serves their own narrow interests
? best for country coded as 1; 0 otherwise
gss$best_for_country = ifelse((gss$scibstgw<3), 1, 0)

lm1 = lm(best_for_country ~ r_colsci , data=gss)
summary(lm1)
## Call:
## lm(formula = best_for_country ~ r_colsci, data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7207 -0.6708  0.2793  0.3292  0.3292
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.67083    0.02104  31.889   <2e-16 ***
## r_colsci     0.04986    0.03118   1.599     0.11
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4609 on 879 degrees of freedom
##   (3629 observations deleted due to missingness)
## Multiple R-squared:  0.002901,   Adjusted R-squared:  0.001767
## F-statistic: 2.558 on 1 and 879 DF,  p-value: 0.1101
```

5. Plot two variables, either as a scatter plot or boxplot; add in trend/regression lines; and explain your results.

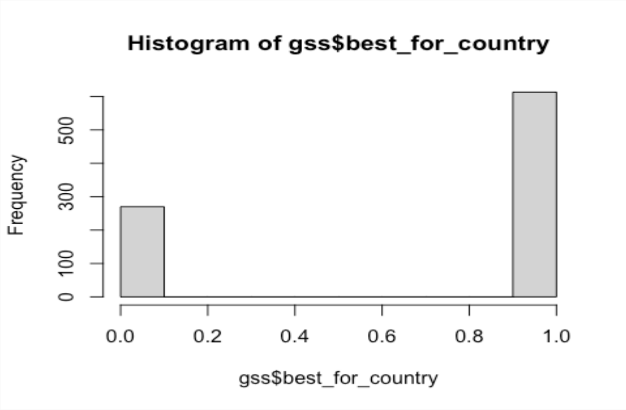
I continued with my two previous variables from above, *best_for_country* and *colsci.* Figure 5 displays the frequencies and densities of both variables. The majority of respondents reported opinions aligned with best-for-country in the *best_for_country* variable (best-for-country or own interests), and reports of having not taken a college level science course in the *r_colsci* variable (no college level science course, or at least one college level science course).

As can be seen in Figure 6, the regression line has a very subtle rising slope, this relates to the *r_colsci* coefficient (0.0499) from above. At first glance, this means that if an individual took at least one science course in college, their opinion regarding environmental scientists making policy recommendations based off of what is good for the country increases by about .05 scale points. However, this result was not statistically significant so we cannot say there is an effect from taking at least one science course.

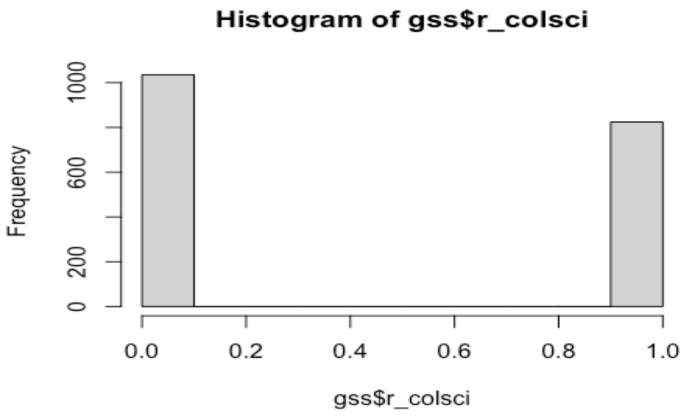
What I am having trouble deciphering is the placement of the plot points in Figure 6. They're coordinates seem to be visually very similar (regardless of the X, the Ys look similar) which I suppose makes sense when we recall that the bulk of responses to the *best_for_country* categories (country or own interest), 613 out of 928 respondents (about 2/3^{rds}), were concentrated in the best-for-country response. Thus, seeing as most respondents – independent of college courses taken, felt policy recommendations would be made with the interest of the country, it seems logical that both Ys could reflect that when we include *r_colsci* in the model, given that the data drives it (remember in the linear regression, the role of *r_colsci* was statisitically not significant). So, if college science course taken was zero, one, or more, it is no more of a predictor of opinion for environmental scientists' making policy recommendations with consideration for what's best for the country.

Figure 5: Collection of frequencies and densities

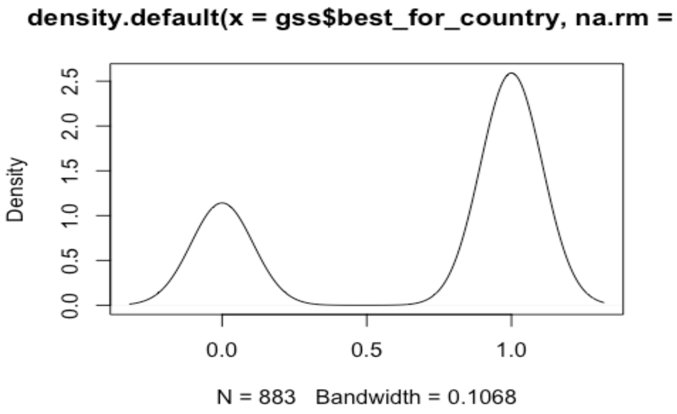
```r
hist(gss$best_for_country)
```



```r
hist(gss$r_colsci)
```



```r
dense <- density(gss$best_for_country, na.rm=T) # returns the density data
plot(dense)
```



```
dense <- density(gss$r_colsci, na.rm=T) # returns the density data
plot(dense)
```

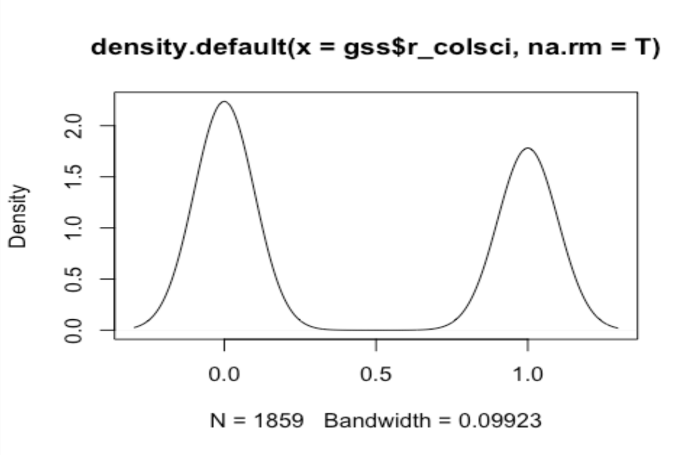
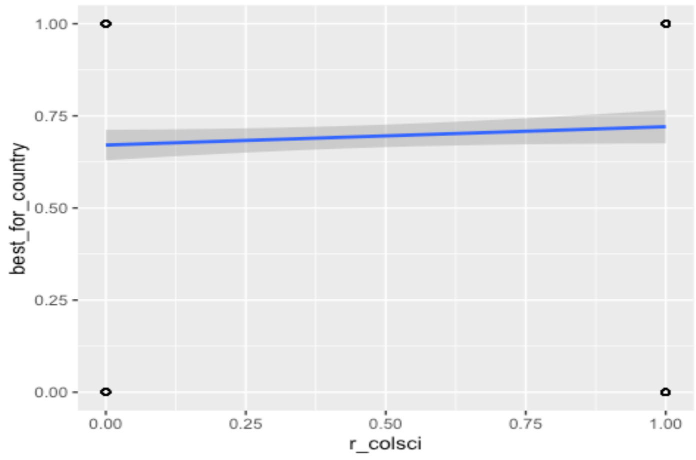


Figure 6:

```
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':
##
##     %+%, alpha

ggplot(gss, aes(x=r_colsci, y=best_for_country)) + ## scatter plot
  geom_point(shape=1)      +   # Use hollow circles
  geom_smooth(method=lm)   # Add linear regression line

## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 3629 rows containing non-finite values (stat_smooth).

## Warning: Removed 3629 rows containing missing values (geom_point).
```



6. Tell me two theories/ideas you might want to test in this course. Do you have a dataset for these ideas/theories already? Do you have it in R-readable format already? What is your main independent variable? What is your main dependent variable? Write it here, but also please send me your proposal as an email with the subject "Independent Project Ideas - [your name]" to gme2101@columbia.edu

To be honest, my initial idea seems unlikely to work, at least at the original scale. I met with you at the beginning of the semester and expressed my interest in using the data that I was working with, when I was with Professor Subiaul. His work has investigated cognitive imitation, and my original plan was to use the entire set of studies he has done over the past decade or so. The hope for this approach was to build up to what he and I wanted to explore: How/what demographic variables/factors predict performance on the various tasks he used in his studies. As such, there would be no specific main variable, at least there may be more than one operationalization of it. Broadly speaking, the independent variable would be the various demographic factors and the dependent variable would be performance and fidelity on the task. I know one specific line of interest Professor Subiaul had, regarded exposure to media/technology (like tablets) and how that predicted performance, I still need to think about this a little and really contemplate the variables that are in the dataset.

The data may not exactly be in R-readable format per se, though all of the data is in excel and already organized (though I may need to discard some observations). If I proceed with this line, I decided I will decrease the amount of studies I'll include, looking at only 2-4 studies that are the most alike. This would make it more manageable and still allow me to practice what we learn on data/a study I was a part of. Hopefully I'll be able to extend the experience to the rest of the studies later.

My second Idea is nascent, and I have no real organization to it yet. But essentially, I would use the GSS or the WVS to explore variables related to climate change. I still need to familiarize myself more intimately with these two datasets so I can offer a cohesive line of thinking.

Obviously the Subiaul line is more developed than the other, and I should probably just go with that data. However, I'm concerned that I may not fully understand the variables or theory behind Subiaul's work and I don't know how that will impact my analysis.

I'm still churning away, and in some senses, I think I'm within a reasonable frame, but at times it also feels I'm quite behind and need to hurry up to understand what I'm doing. Don't know if this is normal or not.