Adrian Varallyay
October 12, 2020

QMSS-5015-Lab-2

1. Run a simple bivariate regression, and interpret your results. (Did the results fit your expectations? Why? Why not?)

I chose to look at the *knwgw* and *region* variables in the GSS dataset, utilizing the subset of people who responded to *degree* (*degree* will be used in the next regression*)*. I suspect that individuals from the northeast and south would be more informed about global warming, just due to the regional cultures and political values. As Table 1 indicates, *knwgw* is the extent to which respondents felt they were informed regarding global warming. This ran from very informed (1), to very uninformed (5). I reversed the order of the possible responses, labeled them appropriately, and then transformed the variable into a binary option of uninformed (0) and informed (1). I did this to facilitate analysis, and so that the shift from zero to one would be an increase in the level of "being informed".

As table 2 indicates, *region* is the part of the country (USA) where the respondent lived. The regions of the country were grouped and recoded the same way the GSS did. This resulted with *region* being Northeast (1), Midwest (2), South (3), and West (4).

Figure 2 shows two plots (different organization) of the responses, and displays the proportions of those responses. As such, we can see that that the Northeast and the South have similar rates of uninformed and informed respondents (close to an even split) and the west has a higher rate of informed respondents.

Figure 3 presents the results of the bivariate regression. The intercept line refers to the northeast, as it is the reference category for the other regions. Thus, when the respondent's region is the northeast (x=0) the Intercept is 1.61. This result is very statistically significant (P = 4.1e-07)[1]. When holding other regions constant, if the respondent was from the Midwest (X=Midwest), on average, they are 0.09 scale points more informed about global warming. This result was also very statistically significant (P = 0.0004). If the respondent was from the south (X=South), on average they 0.058 scale points more informed about global warming. This result was also statistically significant – though less so compared to the other regions – (p = 0.014). The last result was also very statistically significant (p = 0.0003) with respondents from the west being 0.078 scale points more informed about global warming, on average. The R-squared is 0.01756, meaning that just under 2% of the variance is explained by this model.

I was not expecting the outcomes that the regression produced. I was concerned that these relationships would not be statistically significant, and as such, I was quite surprised to see all the results being statistically significant. However, though I was not expecting the significance levels to be what they are, I did think the coefficients would be larger, that there would be more of an effect. My biggest issue is with the interpretation of the intercept. What exactly does a 1.61 coefficient on being from the Northeast mean when the scale is from zero

[1] How do we write this when the significance is '***' and it's listed in scientific notation?

to one? My initial reaction was to “average” but that still doesn’t make sense when the scale maxes out at 1 and the coefficient is above 1. I suspect it has more to do with the categorical and ordinal nature of these variables than anything else. The only other thing I can think of is if R somehow adjusted my uninformed/informed coding of 0 with 1, and 1 with 2.

Table 1:

585) For each of the following areas, please indicate whether you are very informed, somewhat informed, neither informed nor uninformed, somewhat uniformed, or very uninformed about the issues. D. Global warming (KNWGW)

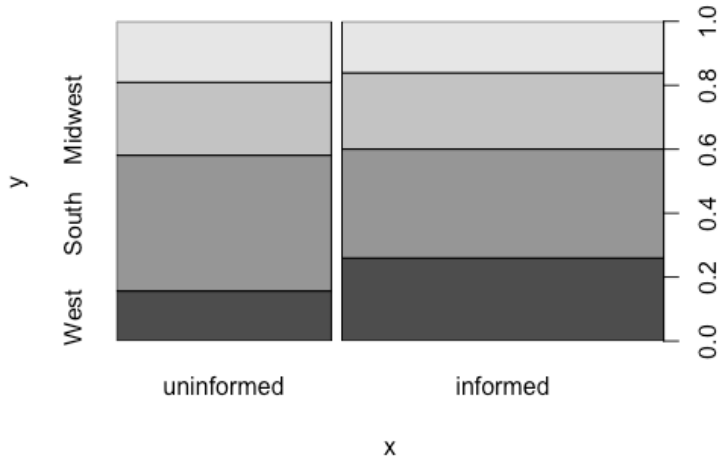
	TOTAL	%
1) Very informed	178	9.5
2) Somewhat informed	923	49.5
3) Neither informed nor uninformed	332	17.8
4) Somewhat uninformed	233	12.5
5) Very uninformed	170	9.1
8) Don't know	25	1.3
9) No answer	3	0.2
Missing	2646	
TOTAL	1864	100.0

Table 2:

1238) I-REGION: Region of interview (Recoded for use with online analysis) (I-REGION)

	TOTAL	%
1) Northeast	711	15.8
2) Midwest	1038	23.0
3) South	1745	38.7
4) West	1016	22.5
TOTAL	4510	100.0

Figure 2: Plots of responses



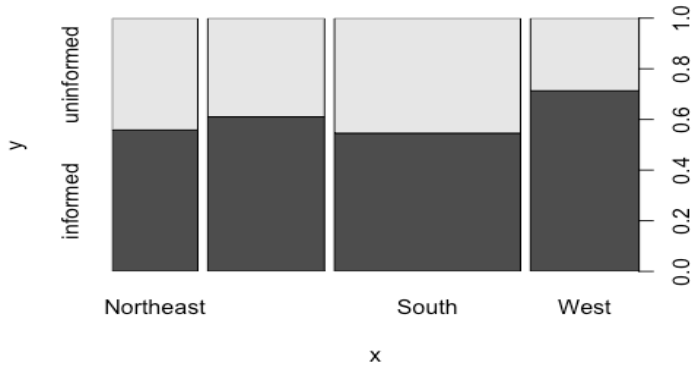


Figure 3: Regression results

```
lm1 = lm(as.numeric(lab.r_knwgw.cat) ~ region.cat, data=gss, subset = !is.na(degree))
summary(lm1)

##
## Call:
## lm(formula = as.numeric(lab.r_knwgw.cat) ~ region.cat, data = gss,
##     subset = !is.na(degree))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7132 -0.5458  0.2868  0.4416  0.4541
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.60691    0.01180 136.190  < 2e-16 ***
## region.cat.L  0.08949    0.02540   3.523 0.000438 ***
## region.cat.Q  0.05776    0.02360   2.448 0.014475 *
## region.cat.C  0.07780    0.02164   3.595 0.000333 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4862 on 1832 degrees of freedom
##   (2671 observations deleted due to missingness)
## Multiple R-squared:  0.01758,    Adjusted R-squared:  0.01597
## F-statistic: 10.93 on 3 and 1832 DF,  p-value: 4.1e-07
```

2. Add an additional variable that might mediate or partly "explain" the initial association from that simple regression above – and explain your results. Did it work out? Yes? No?

I decided to add *degree* to verify if it would behave like a mediating variable. My logic was that the higher level of degree completion, the more you've been exposed to, the more information you have/the more you know, and therefore, you would be more informed about global warming. As Table 1 below indicates, *degree* is the level of degree completed with less than highschool (0), highschool (1), junior college (2), Bachelor's (3), and graduate (4) as categories. Seeing as I suffered some confusion with interpreting the last results, I decided to simplify this variable as either "highschool or less", or as "junior college through graduate". I chose these cut offs, as the depth and span of information students are exposed to is much more in general once entering into post-secondary education.

Figure 2 indicates the difference in density of the responses, with the uninformed group's plot being more densely concentrated, whereas the informed group has more spread to the respective plot. The median *degree* of both groups seems to be from highschool, and there are more, higher degrees in the informed group. An interesting thing to note, is that there is an outlier in the uninformed group that had a graduate degree.

Figure 3 shows that the result of this multiple regression doesn't vary from the previous regression too much. When holding *region* constant, *degree* is also statistically significant (P = 9.07e-11)[2] with a coefficient of 0.062. This implies that if a respondent completed some college (college degree that is, though arguably just some amount of college as I collapsed all degrees into one category), then they would be 0.062 scale points more informed about global warming, on average. When holding *degree* and the other regions constant, region stayed statistically significant except for if the respondent was from the south (P = 0.057). That

[2] Same question as above about p-value display when results are in scientific notation

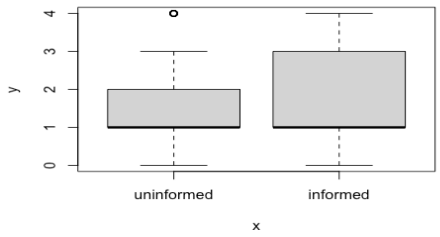
coefficient dropped from 0.058 to 0.045 from the first to the second model. In fact, all coefficients dropped slightly except for responses from the Midwest, which increased slightly from 0.089 to 0.095. The R-squared for the first model (0.018) was a decent amount lower than the second model (0.040), as such, it seems the model with the *degree* explains about twice as much variation than the first model. This is also supported by the reported F statistics (10.93 for model one and 19.01 for model two) which are both highly statistically significant ($p < 0.01$)[3]. Though the effects of the variables didn't change too much, and though the significance levels were only slightly different, the second model seems to be slightly better. Thus, I suppose I could state that it worked out considering the statistical results.

Table 1:

40) Highest educational degree earned by respondent (DEGREE)

	TOTAL	%
0) Less than high school	691	15.3
1) High school	2273	50.4
2) Associate/Junior college	377	8.4
3) Bachelor's	763	16.9
4) Graduate	403	8.9
9) No answer	3	0.1
TOTAL	4510	100.0

Figure 1: Plot of "whether informed about global warming"



[3] I suspect that this may be an inappropriate mode of comparison and therefore not relevant to the discussion.

Figure 2: Regression results

```
lm2 = lm(as.numeric(lab.r_knwgw.cat) ~ region.cat + degree, data=gss)
summary(lm2)

##
## Call:
## lm(formula = as.numeric(lab.r_knwgw.cat) ~ region.cat + degree,
##     data = gss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8542 -0.5126  0.3021  0.4259  0.5547
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.503857   0.019646  76.549  < 2e-16 ***
## region.cat.L 0.094747   0.025134   3.770 0.000169 ***
## region.cat.Q 0.044640   0.023422   1.906 0.056817 .
## region.cat.C 0.077428   0.021400   3.618 0.000305 ***
## degree       0.061790   0.009477   6.520 9.07e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4808 on 1831 degrees of freedom
##   (2674 observations deleted due to missingness)
## Multiple R-squared:  0.03987,    Adjusted R-squared:  0.03777
## F-statistic: 19.01 on 4 and 1831 DF,  p-value: 2.492e-15
```

Figure 3: Model comparison

```
library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

stargazer(lm1, lm2, type = "text")

##
## ===============================================================
##                         Dependent variable:
```

```
##                 ------------------------------------------------
##                           as.numeric(lab.r_knwgw.cat)
##                          (1)                 (2)
## ------------------------------------------------------------------
## region.cat.L              0.089***            0.095***
##                          (0.025)             (0.025)
##
## region.cat.Q              0.058**             0.045*
##                          (0.024)             (0.023)
##
## region.cat.C              0.078***            0.077***
##                          (0.022)             (0.021)
##
## degree                                        0.062***
##                                              (0.009)
##
## Constant                  1.607***            1.504***
##                          (0.012)             (0.020)
##
## ------------------------------------------------------------------
## Observations               1,836               1,836
## R2                        0.018               0.040
## Adjusted R2               0.016               0.038
## Residual Std. Error    0.486 (df = 1832)       0.481 (df = 1831)
## F Statistic        10.928*** (df = 3; 1832) 19.009*** (df = 4; 1831)
## ==================================================================
## Note:                                *p<0.1; **p<0.05; ***p<0.01
```

3. Run another multiple regression. Tell me how you expect your dependent variable to be affected by the independent variables. Interpret your results.

I chose to look at *V131, V80, V238*, variables in the WVS dataset. I renamed the variables *tax_rich*, *Serious_problems*, and *se_class* respectively. I also include the variables Professor Eirich used: *V242* (age), *V57* (married), and *V240* (female). I wasn't exactly sure how the variables would affect the dependent variable (DV). I suspected that younger respondents would have an increase in opinion that taxing the rich and subsidizing the poor was essential for democracy. However, I wasn't sure how gender or marriage status would affect the DV.

As table one indicates, *Tax_rich* asked respondents which characteristics of democracy are essential, in this case, government taxing the rich and subsidizing the poor. This scale runs from "not an essential characteristic of democracy (1), to an essential characteristic of democracy (10). I turned this variable into a binary variable to aid in analysis by splitting the

scale at five, those above 5 are responses of essential characteristic and those that are five and below, I assigned to not essential characteristics.

Table 2 displays *Serious_problems*, which reported respondents' opinions of which problems they consider the most serious. The categories provided are, "People living in poverty and need" (1), "Discrimination against girls and women" (2), "Poor sanitation and infectious diseases" (3), "Inadequate education" (4), and "Environmental pollution" (5). Table 3 indicates that *Se_class* asked respondents how they would describe themselves in terms of working/social class. I reversed this variable to aid in analysis so that as class increases the more "class" increases. The new order is lower (1), working (2), lower-middle (3), upper-middle (4), and upper (5).

Figure one shows the regression predicting opinions on essential characteristics of democracy as a function of *age, sex,* and *marital status*, utilizing the subset of responses that were from the USA and all those who responded to the *se_class* variable. The intercept[4] is 0.45 meaning that when X (age, sex, and marriage status) is zero, the average response to *tax_rich* (essential tax) is .45 on the scale. *Age* was the only variable that was not statistically significant (p = 0.49), though interestingly, it had a negative effect. On average, when age goes up, responses would be 0.0005 points lower on the scale. Again, this was not statistically significant, and the effect is minimal. As such, overall, age doesn't have an effect on opinions regarding essential tax.

Sex was statistically significant (p = 0.017) with a coefficient of 0.051 meaning that, on average, if you are female then responses to *tax_rich* would be 0.051 points higher on the essential tax scale than males. *Married* was very statistically significant with a coefficient of -0.089, showing that respondents who are married, on average, would be 0.089 scale points lower regarding essential taxes. The R-squared is 0.012 meaning that this model explains about one percent of the variation in responses.

As we can see, my prediction for age was incorrect; I was surprised *married* had the impact it did, and that being female also had a statistically significant result. I suspect there may be some interaction or collinearity between *married* and *sex,* though I'm not sure.

Table 1:

148) Many things are desirable, but not all of them are essential characteristics of democracy. Please tell me for each of the following things how essential you think it is as a characteristic of democracy. Use this scale where 1 means 'not at all an essential characteristic of democracy' and 10 means it definitely is 'an essential characteristic of democracy': Government tax the rich and subsidize the poor (V131)

[4] I'm not certain, but is the intercept in this regression utilizing a reference category? Like, is it the average male response? I was not clear on this as I can't recall how R deals with that. If it is male, then this result is statistically significant (p = 2e-16) the 0.45 coefficient indicates that on average if your male, then you are 0.45 scale points higher on the essential tax binary scale.

	TOTAL	%
-5) AM, DE, SE: Inapplicable; RU: Inappropriate response; Missing (Inappropriate)	22	0.0
-2) No answer	763	1.0
-1) Don't know	2489	3.4
1) Not an essential characteristic of democracy	7392	10.0
2) 2	3132	4.2
3) 3	4205	5.7
4) 4	4180	5.6
5) 5	8664	11.7
6) 6	5861	7.9
7) 7	7615	10.3
8) 8	9491	12.8
9) 9	6023	8.1
10) An essential characteristic of democracy	14205	19.2
TOTAL	74042	100.0

Table 2:

82) I'm going to read out some problems. Please indicate which of the following problems you consider the most serious one for the world as a whole? (V80)

	TOTAL	%
-5) Missing; SE: Inapplicable; RU: Inappropriate response; BH: Missing	10	0.0
-2) No answer	369	0.5
-1) Don't know	669	0.9
1) People living in poverty and need	41640	56.2
2) Discrimination against girls and women	4999	6.8

3) Poor sanitation and infectious diseases	7743	10.5
4) Inadequate education	8465	11.4
5) Environmental pollution	10147	13.7
TOTAL	74042	100.0

Table 3:

291) People sometimes describe themselves as belonging to the working class, the middle class, or the upper or lower class. Would you describe yourself as belonging to the: (V238)

	TOTAL	%
-5) SE, Inapplicable; RU: Inappropriate response; BH: Missing (Inappropriate)	34	0.0
-3) Not applicable	3	0.0
-2) No answer	767	1.0
-1) Don't know	1149	1.6
1) Upper class	1445	2.0
2) Upper middle class	15623	21.1
3) Lower middle class	26284	35.5
4) Working class	20722	28.0
5) Lower class	8015	10.8
TOTAL	74042	100.0

Figure 1:

```
lm3 = lm(as.numeric(essential.tax) ~ age + female + married, data = wvs, subset=V2==840 & !is.na(lab.se_class))  ## This is for USA
summary(lm3)

##
## Call:
## lm(formula = as.numeric(essential.tax) ~ age + female + married,
##     data = wvs, subset = V2 == 840 & !is.na(lab.se_class))
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -0.4963 -0.3995 -0.3394  0.5673  0.6736
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.4537319  0.0349432  12.985  < 2e-16 ***
## age         -0.0004478  0.0006428  -0.697   0.4861
## female       0.0505855  0.0210977   2.398   0.0166 *
## married     -0.0892617  0.0220249  -4.053 5.24e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4884 on 2144 degrees of freedom
##   (33 observations deleted due to missingness)
## Multiple R-squared:  0.0118, Adjusted R-squared:  0.01042
## F-statistic: 8.534 on 3 and 2144 DF,  p-value: 1.239e-05
```

4. Now add another independent variable to that model in Question 3, preferably a set of dummy variables. Tell me why you added that new set of variables and what effect you expected them to have. Did they have an effect? Interpret that new model.

Similar to the example from Professor Eirich, I moved *se-class* from the subsetting command and placed it as another X variable. I only retained respondents from the USA in the subsetting command. I thought that there would be a statistically significant effect from including the type of working/social class to the model. Specifically, I thought lower- or working-class would have had a positive and significant effect.

The intercept is 0.44, and this is for the reference category of "Lower class"[5]. This result was very statistically significant (P = 2e-16) and means that, on average, if you're lower-class then you would be 0.44 points higher on the essential tax scale. *Age* remains statistically insignificant and its effect is marginally less (-0.00043 compared to the previous -0.00045). *Sex* (being female) remains statistically significant (P = 0.016)[6] with a coefficient of 0.051, meaning on average, being female is associated with a 0.051 scale point increase on opinions regarding *tax_rich*. *Married* remains very statistically significant (P = 0 .000176, compared to the previous P = 5.24e-05). Though still very significant, the level of significance dropped from the first regression to the second for this variable. All other *se_class* categories are statistically insignificant and have the following P-values for the remaining four categories: working (P =

[5] Similar to the last footnote, is this reference category the combination of male and lower class? Or does this model just not report the male statistics?

[6] NOTE: P-value is same as previous model due to rounding

0.16), lower-middle (P= 0.67), upper-middle (P = 0.41), and upper (P = 0.66). It's interesting to note, that all coefficients for the remaining working/social classes have negative effects. Though these effects are not statistically significant[7], the coefficients are -0.092, -0.023, -0.03, and -0.01, for working, lower-middle, upper-middle, and upper classes respectively. The effects of these variable on the responses to *tax_rich* are slight, and as they are not significant results there's no real effect greater than zero occurring. The R-squared of this model is 0.013 which is 0.001 larger than the previous model. Essentially only a slight amount more of variation is explained by this model compared to the first. I would say that the adding the *se_class* did have an effect, though slight, and my prediction for *se_class* seems to be partially accounted for.

Figure 1: Regression results

```
lm4 = lm(as.numeric(essential.tax) ~ age + female + married + lab.se_class, data = wvs, subset=
V2==840)
summary(lm4)

##
## Call:
## lm(formula = as.numeric(essential.tax) ~ age + female + married +
##     lab.se_class, data = wvs, subset = V2 == 840)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5366 -0.4036 -0.3372  0.5663  0.7520
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.4395957  0.0388343  11.320  < 2e-16 ***
## age            -0.0004275  0.0006461  -0.662 0.508273
## female          0.0511300  0.0211104   2.422 0.015517 *
## married        -0.0841782  0.0224009  -3.758 0.000176 ***
## lab.se_class.L -0.0915189  0.0648847  -1.410 0.158542
## lab.se_class.Q -0.0237546  0.0555935  -0.427 0.669210
## lab.se_class.C -0.0302897  0.0364665  -0.831 0.406283
## lab.se_class^4 -0.0097192  0.0221388  -0.439 0.660697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4886 on 2140 degrees of freedom
##   (84 observations deleted due to missingness)
```

[7] I know usually we don't report insignificant results, but are we supposed to talk about the coefficients of insignificant results for our assignments?

```
## Multiple R-squared:  0.01294,    Adjusted R-squared:  0.009708
## F-statistic: 4.007 on 7 and 2140 DF,  p-value: 0.0002294
```

5. Now run a partial F test comparing the model in Question 3 to the model in Question 4. Does the F test support the idea of adding those new variables? Why? Why not?

I initially ran an anova, but was so confused as to how to interpret the results, that I switched back to stargazer. Now I'm wondering if I understand how to work with F statistic in general and whether stargazer is even appropriate to compare the models.

Nevertheless, figure 1 reports a statistically insignificant F stat of 0.6159 (P = 0.6512). I suppose that with the insignificant P-value reported that the second model is not better than the first and that essentially, I could have left out the *se_class.* According to the results in figure 2, if I understand correctly, the second model with the added *se_class* is also not better than the first. My reasoning is that the recurring variables either stay the same or diminish (married and constant) in the stargazer results. In addition, the adjusted R-squared remains the same (though the regular R-squared increased by 0.001). Both F statistics are statistically significant, with the second models F stat being about half. I am not confident of this interpretation of the stargazer results.

However, when I return to examine the results behind the anova, because the F is not larger than P *and* the result was not anywhere near statistically significant, the results do not support the use of the new model with the addition of *se_class.* Not much more is explained, which is seemingly consistent with the fact that in general, *se_class* was statistically insignificant in the second model.

Figure 1: anova results

```
anova(lm3, lm4)
```

```
## Analysis of Variance Table
##
## Model 1: as.numeric(essential.tax) ~ age + female + married
## Model 2: as.numeric(essential.tax) ~ age + female + married + lab.se_class
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   2144 511.52
## 2   2140 510.93  4   0.58822 0.6159 0.6512
```

Figure 2: Stargazer results

```
library(stargazer)
stargazer(lm3, lm4, type = "text")
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     -------------------------------------------
```

```
##                                as.numeric(essential.tax)
##                               (1)                    (2)
## ---------------------------------------------------------------------
## age                          -0.0004                -0.0004
##                              (0.001)                (0.001)
##
## female                       0.051**                0.051**
##                              (0.021)                (0.021)
##
## married                     -0.089***              -0.084***
##                              (0.022)                (0.022)
##
## lab.se_class.L                                      -0.092
##                                                     (0.065)
##
## lab.se_class.Q                                      -0.024
##                                                     (0.056)
##
## lab.se_class.C                                      -0.030
##                                                     (0.036)
##
## lab.se_class4                                       -0.010
##                                                     (0.022)
##
## Constant                    0.454***               0.440***
##                              (0.035)                (0.039)
##
## ---------------------------------------------------------------------
## Observations                  2,148                  2,148
## R2                            0.012                  0.013
## Adjusted R2                   0.010                  0.010
## Residual Std. Error    0.488 (df = 2144)      0.489 (df = 2140)
## F Statistic         8.534*** (df = 3; 2144) 4.007*** (df = 7; 2140)
## =====================================================================
## Note:                                    *p<0.1; **p<0.05; ***p<0.01
```

Appendix A (codes)

Figure 1:

```
## Y-VAR
#reverse order
gss$r_knwgw <- 6-gss$knwgw
# turn into factor
gss$r_knwgw.fact <- as.factor(gss$r_knwgw)
```

```
#label the reverse order
gss$lab.r_knwgw = ordered(gss$r_knwgw, levels = c(1,2,3,4,5), labels = c("very uninformed", "uninformed", "neither", "informed", "very informed"))
# make binary for regression
gss$lab.r_knwgw.cat = cut(as.numeric(gss$lab.r_knwgw), breaks = c(-1, 3, 8), label=c("uninformed", "informed"), ordered=TRUE)
## check code
table(gss$lab.r_knwgw.cat, gss$knwgw)

##
##                1   2   3   4   5
##   uninformed   0   0 332 233 170
##   informed   178 923   0   0   0

table(gss$lab.r_knwgw.cat)

##
## uninformed   informed
##       735       1101

## X-VAR
# recode and label region (just like the gss did)
gss$region.cat = cut(gss$region, breaks = c(-1, 2, 4, 7, 9), label=c("Northeast","Midwest", "South", "West"), ordered=TRUE)
## check code
table(gss$region.cat)

##
## Northeast   Midwest     South      West
##       711      1038      1745      1016

# Plot and Bivariate regression
plot(as.factor(gss$lab.r_knwgw.cat), gss$region.cat)

plot (gss$region.cat, as.factor (gss$lab.r_knwgw.cat))
```

Figure 2:

```
# recode and label education
gss$degree.cat = cut(gss$degree, breaks = c(-1, 1, 9), label=c("highschool or less","junior college through graduate"), ordered=TRUE)
# test code
table(gss$degree.cat)
```

```
##
##           highschool or less junior college through graduate
##                         2964                         1543

# Plot and 2 IV  regression
plot(as.factor(gss$lab.r_knwgw.cat), gss$degree)
```

```
library(plyr)
wvs = read.csv("WVS.csv") ## choose the WVS.csv from Lab 3 ## gss <- read.csv("GSS.2006.csv.x
ls")

## Here is a question about "Many things are desirable, but not all of them are essential charact
eristics of democracy. Please tell me for each of the following things how essential you think it i
s as a characteristic of democracy. Use this scale where 1 means 'not at all an essential characte
ristic of democracy' and 10 means it definitely is 'an essential characteristic of democracy': Gov
ernment tax the rich and subsidize the poor  (V131)" with higher scores meaning essential for d
emocracy

# rename Y-VAR 131 and add labels
wvs = rename(wvs, c("V131"="tax_rich_sub_poor"))
wvs$r_tax_rich.lab <- ordered(wvs$tax_rich_sub_poor, levels = c(1,2,3,4,5,6,7,8,9,10), labels = c
("Not an essential characteristic of democracy", "2", "3", "4", "5", "6", "7", "8","9", "An essential
characteristic of democracy"))
wvs$essential.tax = ifelse((wvs$r_tax_rich.lab>5), 1, 0)
#check code
table(wvs$essential.tax)

##
##    0    1
## 27573 43195

# I'm going to read out some problems. Please indicate which of the following problems you co
nsider the most serious one for the world as a whole? (V80) 1)People living in poverty and need
, 2)Discrimination against girls and women, 3)Poor sanitation and infectious diseases, 4)Inadequ
ate education, 5)Environmental pollution

# X-VARs
wvs = rename(wvs, c("V80"="serious_problems"))
wvs$r_serious.lab <- ordered(wvs$serious_problems, levels = c(1,2,3,4,5), labels = c("People livi
ng in poverty and need", "Discrimination against girls and women", "Poor sanitation and infecti
ous diseases", "Inadequate education", "Environmental pollution"))
#check code
table(wvs$r_serious.lab)
```

```
##
##      People living in poverty and need  Discrimination against girls and women
##                                  41640                                   4999
## Poor sanitation and infectious diseases                 Inadequate education
##                                   7743                                   8465
##                 Environmental pollution
##                                  10147
```

```
wvs = rename(wvs, c("V242"="age"))
wvs$married=ifelse(wvs$V57==1, 1,0)
wvs$female = ifelse(wvs$V240==2, 1, 0)
```

```
## People sometimes describe themselves as belonging to the working class, the middle class, o
r the upper or lower class. How would you describe yourself: (V238)
wvs = rename(wvs, c("V238"="se_class"))
# reverse
wvs$se_class <- 6-wvs$se_class
#turn into factor
wvs$se_class.fact <- as.factor(wvs$se_class)
#label the reverse order
wvs$lab.se_class = ordered(wvs$se_class, levels = c(1,2,3,4,5), labels = c("lower", "working", "lo
wer-middle", "upper middle", "upper"))
# check code
table(wvs$lab.se_class)
```

```
##
##       lower     working lower-middle upper middle       upper
##        8015       20722       26284       15623        1445
```