STVN14: Political Science Methodology

Introduction to R

The R interface is a just that - a basic interface. In order to obtain R, you must first download the software from the link provided here: https://www.r-project.org/. Then, in order to use the software, you must load a dataset and use packages and commands in order to tell the software how to evaluate and handle the data. Each package is open-sourced and contains different tools in order to implement all sorts of data manipulation techniques.

Note: The dataset used in this tutorial and the R Script are online at:
(http://michaelhansenpolitics.org/courses-taught).

Packages, Libraries, Loading Data, and Viewing Data

1. Install packages for use.

Example: The foreign package and library allows you to load SPSS and STATA dataset file formats. You will want to load dependency packages also just in case the package/commands needs the support of others to work.

Command:

```
install.packages("foreign", dependencies=TRUE)
```

2. Loading a package for use.

Command:

```
library(foreign)
```

3. Loading the dataset.

Example: Loading the 2014 European Social Survey Sweden dataset in STATA format (.dta format) that is provided on my webpage (http://michaelhansenpolitics.org/courses-taught). When loading the dataset, the code here specified that variables should remain numerical rather than convert them to factors. Therefore, if a variable was say gender (i.e. men and women), the variable will be coded as a number, such as 1 for men 2 for women. Note: I named the dataset "dat" here.

Command:

```
dat <- read.dta(file.choose(), convert.factors=FALSE)
```

4. Additional examples for different data formats:

Example: Reading an SPSS file (.sav format) into R

Command:

```
dat <- read.spss(file.choose(), convert.factors=FALSE)
```

Example: If you want to read an excel file (.xls or .xlsx format) into R you will need a different package. The code is below.

Command:

```
install.packages(gdata, dependencies=TRUE)
library(gdata)
dat <- read.xls(file.choose())
```

5. Exploring the Dataset.

Example: Looking at the names of the variables in the dataset.

Command:

```
names(dat)
```

Example: If we wanted to know the number of rows we have in the dataset (i.e. observations) we could use the "nrow" function. Here, we have 1791 Swedish Survey Respondents.

Command:

```
nrow(dat)
```

Example: If we wanted to get a quick idea of the structure of the variables. We could simply summarize the entire dataset.

Command:

```
summary(dat)
```

Additional Note: Help with a Library, Command, or Function

Example: If you would like to find out what function a command performs. A help file can be accessed by typing ?? and the code. If we wanted to know what function the "names" command performed, we could type the command below.

Command:

```
??names
```

Understanding the Structure of Variables and Recoding

1. Basic Exploration of Variables

Variables are almost never in a readily useable format. Usually, there are missing values and people that did not answer questions. Therefore, the first step to any analysis is exploring the structure of the variables you want to use and making sure that they are accurately depicting the measurement scales intended.

Example: Let us pretend that the thing I ultimately want to explore is how Swedes view themselves ideologically (left - 0 to right - 10). In the survey, the codebook indicates that the variable "lrscale" provides this information. I can view a summary of this information by telling R that I want a summary of the lrscale variable in the "dat" dataset that I loaded with the command below. Note: The dataset I am loading is titled "dat" and the variable in the dataset is "lrscale".

Command:

```
summary(dat$lrscale)
```

R Output:

```
> summary(dat$lrscale)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   3.000   5.000   8.277   7.000  88.000
```

Discussion: The output indicates that the minimum value for the variable is 0, median is 5, mean is 8.277 and max is 88. Does this make sense? No. Remember, the left-right ideology variable is measured from 0-10. Therefore, we need to continue our exploration.

Example: You could also make a table of the variable in order to get a better view at the categories of the variable.

Command:

```
table(dat$lrscale)
```

R Output:

```
> table(dat$lrscale)

  0   1   2   3   4   5   6   7   8   9  10  77  88
 85  41 114 242 195 329 185 244 182  48  56   2  68
```

Discussion: As you can see, there are two categories "77" and "88" that do not fit in our 0-10 scale. If we look in the codebook, 77 indicates "Refusal" and 88 indicates "Don't know". In order to successfully move on and perform an analysis, we will need to recode this variable in order to remove the categories.

2. Recoding Variables - Numerical

In most instances, we might want to recode data so that these people do not exist in our analysis or "Not Applicable (NA)". For recoding variables, the "car" package provides us tool to do this type of recoding.

Command:

```
install.packages("car", dependencies=TRUE)
library(car)
```

Example: As indicated before, it is necessary to recode the categories/numbers "77" and "88" as NA so that we can perform meaningful statistical tests on the data. Here, we create a new variable called "ideology" in the "dat" dataset where these values are set to NA. The reason we create a new variable rather than simply recoding the old one is because we do not want to risk overwriting a variable and losing valuable information.

Command:

```
dat$ideology <- recode(dat$lrscale, "77=NA; 88=NA")
```

R Output:

```
> dat$ideology <- recode(dat$lrscale, "77=NA; 88=NA")
> summary(dat$ideology)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  0.000   3.000   5.000   5.047   7.000  10.000      70
```

Discussion: After recoding the variable, the summary of the variable indicates that we now have a minimum value of 0, max of 10, mean of 5.047, and 70 observations have been set to NA. As you can see, the new mean makes much more sense theoretically.

3. Recoding Variables - Categorical

Another situation that might arise is that R might treat a categorical variable as numeric in ways that we do not wish. For example, a variable that explores profession might have several categories that cannot be ranked in a meaningful way numerically. Therefore, the variable should be recoded as a factor.
Note: Variables can take on several forms including factors, strings, vector list, data frames, and matrices.

Example: For simplicity sake, here we are recoding the gender variable ("gndr") to be categorical instead of numeric. In addition, if respondents refused to answer, we could do as we did above and recode other values to NA. The codebook indicates that for the variable a 1 indicates male and a 2 indicates female.

Command:

```
table(dat$gndr)
dat$gender <- recode(dat$gndr, "1='Man'; 2='Woman'")
dat$gender <- as.factor(dat$gender)
table(dat$gender)
```

R Output:

```
> table(dat$gndr)

  1   2
893 898
> dat$gender <- recode(dat$gndr, "1='Man'; 2='Woman'")
> table(dat$gender)

  Man Women
  893   898
```

Discussion: There is now a categorical variable in the "dat" dataset named "gender" with a category for men and women.

Graphically Displaying Variables

1. Bar Chart

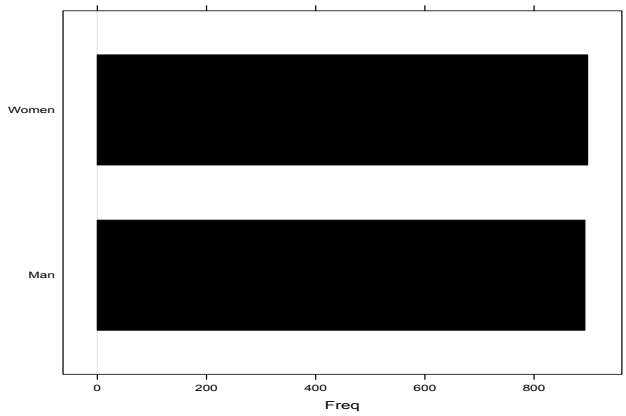
Example: For most variables, the best mode of presentation is usually providing a simple bar chart. Here, we provide a bar chart that presents the frequencies for our newly created gender variable.

Command:

```
barchart(dat$gender, col="black")
```

R Output:

Figure 1: Frequencies: Respondent Gender



2. Density Plot

Example: For continuous numerical variables, a density plot usually provides a better snapshot. Here, we create two density plots. One density plot for political ideology and the other for political ideology while splitting up the sample by gender.

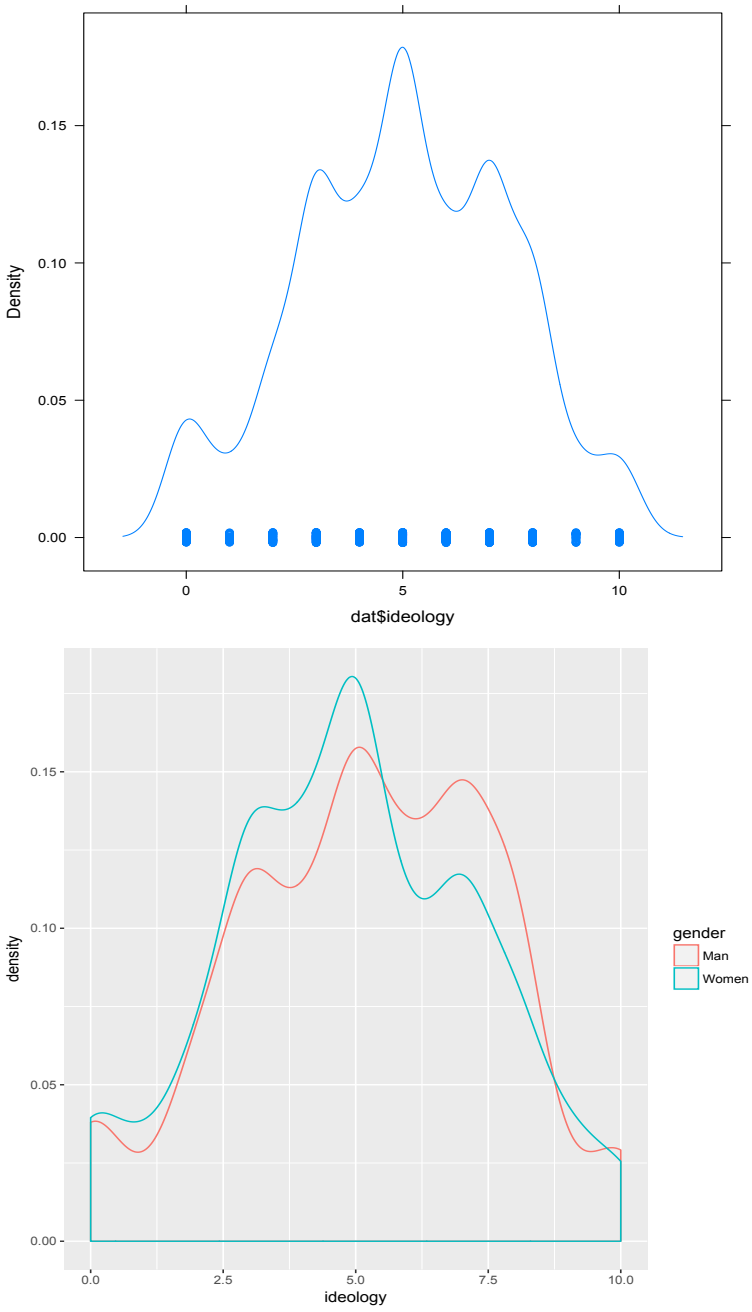
Command:

```
densityplot(dat$ideology)

install.packages("ggplot2", dependencies=TRUE)
library(ggplot2)
ggplot(dat, aes(x=ideology)) + geom_density(aes(group=gender, colour=gender))
```

R Output:

Figure 2: Density Plot: Ideology and Ideology by Gender



Bivariate Analysis

1. Simple Bivariate Correlation

When starting our empirical bivariate analysis, we might simply want to know how correlated two variables are with each other. In other words, we are exploring how much of the variance that exists on one variable is dependent on the variation of another variable. Importantly, to perform a simple test of this nature, we must have two numeric variables.

Example: Let us pretend that I think political ideology in Sweden is dependent at least to some extent on the age of the respondent. I could perform a simple correlation test that explores whether these two variables are correlated.
Note: First, I must recode the age variable using the year born variable "yrbrn".

Command:

```
table(dat$yrbrn)
dat$age <- recode(dat$yrbrn, "7777=NA")
dat$age <- 2014 - dat$age
table(dat$age

cor(dat$ideology, dat$age, use = "pairwise.complete.obs")
```

R Output:

```
> table(dat$yrbrn)

1900 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1930
   1    1    1    1    3    4    1    9    2    6    6    6    7   13   13
1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944 1945
  13    9   11   11   15   21   15   17   21   24   27   28   30   22   34
1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960
  26   34   33   25   26   26   30   22   28   24   28   21   38   27   32
1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975
  24   24   28   24   29   29   38   29   22   25   22   33   29   28   20
1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990
  28   28   21   28   25   24   28   26   21   26   19   24   26   38   31
1991 1992 1993 1994 1995 1996 1997 1998 1999 7777
  24   27   27   22   17   29   25   21    9    1
> dat$age <- recode(dat$yrbrn, "7777=NA")
> dat$age <- 2014 - dat$age
> table(dat$age)

 15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33
  9  21  25  29  17  22  27  27  24  31  38  26  24  19  26  21  26  28  24
 34  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52
 25  28  21  28  28  20  28  29  33  22  25  22  29  38  29  29  24  28  24
 53  54  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71
 24  32  27  38  21  28  24  28  22  30  26  26  25  33  34  26  34  22  30
 72  73  74  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
```

```
 28  27  24  21  17  15  21  15  11  11   9  13  13  13   7   6   6   6   2
 91  92  93  94  95  96  97 114
  9   1   4   3   1   1   1   1
>
> cor(dat$ideology, dat$age, use = "pairwise.complete.obs")
[1] 0.05150771
```

Discussion: The result is .0515 for the correlation test between political ideology and age. The interpretation is that around 5.15% of the variation in political ideology between respondents is explained by age. The correlation relationship is quite weak.

2. T-Test

When we have a continuous variable and a categorical variable, we might want to test whether the continuous variable is statistically different based on the categories of our categorical variable.

Example: We had previously calculated a mean for our political ideology variable. The mean was 5.047. However, we might want to know whether the mean political ideology is different for men and women. A t-test would provide this information.

Command:

```
t.test(dat$ideology~dat$gender)
```

R Output:

```
> t.test(dat$ideology~dat$gender)

Welch Two Sample t-test

data:  dat$ideology by dat$gender
t = 2.4827, df = 1718.7, p-value = 0.01314
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.06030929 0.51410546
sample estimates:
  mean in group Man mean in group Women
           5.190920            4.903712
```

Discussion: Remember, our statistical significance cutoff is a p-value of .05 in most research. The p-value indicates the level of confidence that we have the correct sample. Here, we have a p-value of .01 (rounding). The value indicates that if we took 100 samples we would correctly have the correct sample 99 times. Here, if we guessed that gender had a statistical impact on ideology we would be correct 99/100 times. In particular, we can see that women are statistically more liberal.

3. Cross Tabulation

When we have two categorical variables we might want to test whether the categories of the variables have a relationship with one another. The easiest way to test this relationship is to perform a cross tabulation.

Example: Let us pretend that I think there is a relationship between gender and unemployment in the last 12 months. I could test whether these two categorical variables have a relationship. Note: I need to recode the "uemp12m" variable that asks about unemployment in the last 12 months.

Command:

```
table(dat$uemp12m)
dat$unemp <- recode(dat$uemp12m, "1='Yes'; 2='No'; else=NA")

install.packages("gmodels", dependencies=TRUE)
library(gmodels)
CrossTable(dat$unemp, dat$gender, chisq=TRUE)
```

R Output:

```
> table(dat$uemp12m)

   1    2    6    8
 148  305 1336    2
> dat$unemp <- recode(dat$uemp12m, "1='Yes'; 2='No'; else=NA")
> table(dat$unemp)

 No Yes
305 148
> CrossTable(dat$unemp, dat$gender, chisq=TRUE)


   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  453


             | dat$gender
   dat$unemp |       Man |     Women | Row Total |
-------------|-----------|-----------|-----------|
          No |       144 |       161 |       305 |
             |     0.953 |     1.001 |           |
             |     0.472 |     0.528 |     0.673 |
             |     0.621 |     0.729 |           |
             |     0.318 |     0.355 |           |
```

```
-------------|-----------|-----------|-----------|
         Yes |        88 |        60 |       148 |
             |     1.965 |     2.062 |           |
             |     0.595 |     0.405 |     0.327 |
             |     0.379 |     0.271 |           |
             |     0.194 |     0.132 |           |
-------------|-----------|-----------|-----------|
Column Total |       232 |       221 |       453 |
             |     0.512 |     0.488 |           |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  5.981257     d.f. =  1     p =  0.01445869

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  5.501155     d.f. =  1     p =  0.01900391
```

Discussion: The p-value for our test indicates that there is a statistically significant relationship between gender and unemployment since the value of .01 is less than .05. In particular, if we explore the percentages in the cells men are more likely than women to be unemployed over the last 12 months.

4. Bivariate Regression analysis

The most useful way to test whether two variables are related is to estimate a regression model.

Example: Let us say I wanted to calculate the precise effect that gender has on political ideology. Since political ideology is theoretically continuous, I could estimate an OLS linear regression model.

Command:

```
mod <- lm(ideology ~ gender, data=dat)
summary(mod)
```

R Output:

```
> mod <- lm(ideology ~ gender, data=dat)
> summary(mod)

Call:
lm(formula = ideology ~ gender, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-5.1909 -1.9037  0.0963  1.8091  5.0963

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.19092    0.08187  63.403   <2e-16 ***
genderWomen -0.28721    0.11568  -2.483   0.0131 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.4 on 1719 degrees of freedom
  (70 observations deleted due to missingness)
Multiple R-squared:  0.003573,Adjusted R-squared:  0.002993
F-statistic: 6.164 on 1 and 1719 DF,  p-value: 0.01313
```

Discussion: There are a few aspects of the model output to explore. First, the intercept. The intercept tells us the value of the dependent variable (political ideology) when all of the independent variables are at a value of zero (gender). Here, the zero category is "man". Therefore, when gender is male the mean political ideology value is 5.19. Next, we can explore the coefficient estimate for the gender variable. Since the p-value is less than .05, the variable is significant. When the gender of a respondent is female there is a decrease of -.28 in political ideology. The drop in political ideology is a little over a 1/4 of a point. Finally, we could look at model fit. How well does our model explain the variation in political ideology between respondents. The R-squared statistic provides this information. The value is .003. The value indicates that our model only explains around .3% of the variation in political ideology.

5. Calculating and Plotting Variable Effects

After we have estimated a regression model, we have the possibility of plotting the precise effect that a significant independent variable has on the dependent variable.

Example: Here, I want to plot the predictions for how gender effectd political ideology.

Command:

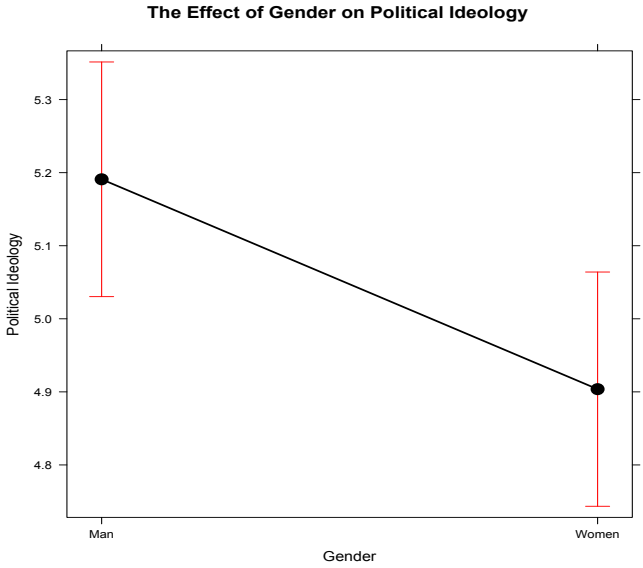
```
install.packages("effects", dependencies=TRUE)
library(effects)

eff <- effect("gender", mod, default.levels=100)

effect <- print(plot(eff, rescale.axis=F, rug=FALSE, xlab="Gender",
ylab="Political Ideology", main=" The Effect of Gender on Political Ideology"))
```

R Output:

Figure 3: Predicted Probabilities: Effect of Gender on Political Ideology



Multivariate Analysis

1. Multivariate OLS Linear Regression

It is most assuredly unreasonable to assume that only one independent variable impacts a dependent variable of interest. In fact, there are usually multiple variables that impact one dependent variable. In order to account for this type of relationship, we can estimate a model with several independent variables much like how we did prior with the bivariate OLS linear regression.

Example: Say we want to know which variables have a statistical relationship with political ideology in Sweden.
Note: This example requires the use of our previously created age and gender variables, and the recoding of education and income variables.

Command:

```
summary(dat$edulvlb)
dat$education <- recode(dat$edulvlb, '0 = 1; 113 = 2; 129 = 3; 212 = 4; 213 = 5;
221 = 6; 222 = 7; 223 = 8; 229 = 9; 311 = 10; 312 = 11; 313 = 12; 321 = 13; 322 = 14;
323 = 15; 412 = 16; 413 = 17; 421 = 18; 422 = 19; 423 = 20; 510 = 21; 520 = 22;
610 = 23; 620 = 24; 710 = 25; 720 = 26; 800 = 27; else=NA')
summary(dat$education)

summary(dat$hinctnt)
dat$income <- recode(dat$hinctnt, '77=NA; 88=NA; 99=NA')
summary(dat$income)

mod1 <- lm(ideology ~ age + gender + income + education, data=dat)
summary(mod1)
```

R Output:

```
> mod1 <- lm(ideology ~ age + gender + income + education, data=dat)
> summary(mod1)

Call:
lm(formula = ideology ~ age + gender + income + education, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7694 -1.7189  0.1049  1.6841  5.7350

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.509142   0.268672  13.061  < 2e-16 ***
age          0.011422   0.003211   3.557 0.000386 ***
genderWomen -0.252831   0.120327  -2.101 0.035781 *
income       0.150045   0.022056   6.803 1.45e-11 ***
education    0.007091   0.008488   0.835 0.403598
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.363 on 1586 degrees of freedom
  (200 observations deleted due to missingness)
Multiple R-squared:  0.03981,Adjusted R-squared:  0.03739
F-statistic: 16.44 on 4 and 1586 DF,  p-value: 3.326e-13
```

Discussion: If you look at the p-values for our variables you will notice that there are three statistically significant variables that have a relationship with political ideology (age, gender, income). We already talked about the impact of gender. However, age and income have the opposite relationship. In particular, as you increase in age and income you identify at higher values of political ideology (meaning, you are more conservative). Specifically, a one unit increase in income categories leads to a .15 increase in conservatism. Notice, our R-squared statistic indicates that this model explains about 4% of the variation of our dependent variable.

2. Multivariate Logit Regression

The example above was regression estimation for when the dependent variable of interest is a continuous numerical variable. However, many times in social science our variable of interest is a binary categorical variable. In order to explore this type of dependent variable, we must estimate a logistic regression.

Example: Let me pretend I have a strong theory to assume that Sweden Democrat voters are completely different from all other voters. Therefore, I might want to estimate a model that simply explores the variables that impact voting for the Sweden Democrats over all other parties. I would recode the vote choice variable ("prtvtbse") so that voting for SD is a 1 and voting for all other parties is a 0.

Command:

```
table(dat$prtvtbse)
dat$sdvote <- recode(dat$prtvtbse, "10=1; 66=NA; 77=NA; 88=NA; else=0")

mod2 <- glm(sdvote ~ age + gender + income + education, data=dat, family=binomial)
summary(mod2)
```

R Output:

```
> mod2 <- glm(sdvote ~ age + gender + income + education, data=dat, family=binomial)
> summary(mod2)

Call:
glm(formula = sdvote ~ age + gender + income + education, family = binomial,
    data = dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.5586  -0.3552  -0.2853  -0.2335   2.7525

Coefficients:
```

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.487180   0.568188  -2.617  0.00886 **
age         -0.003918   0.007017  -0.558  0.57664
genderWomen -0.497841   0.263685  -1.888  0.05902 .
income      -0.050019   0.045982  -1.088  0.27668
education   -0.049576   0.018484  -2.682  0.00732 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 534.16  on 1360  degrees of freedom
Residual deviance: 516.93  on 1356  degrees of freedom
  (430 observations deleted due to missingness)
AIC: 526.93

Number of Fisher Scoring iterations: 6
```

Discussion: It is important to note that regression coefficients are uninterpretable when running logistic regression in terms of substantive numerical values. However, statistical significance and coefficient directionality can provide us with useful insights into findings. The only variable in our model that is statistically significant is the education variable. In particular, Sweden Democrat voters have a statistically lower level of education than all other voters. Since the coefficient direction is negative, it means that increases in education lead to decreases in the probability of voting for Swedish Democrats.

3. Calculating and Plotting Variable Effects

Since coefficients are difficult to interpret when estimating logistic regression, plotting predicted probabilities provides us with a useful way of exploring the substantive impact that an independent variable has on the dependent variable.

Example: Using the last multivariate model that was estimated, let us plot the effect that education has on the probability of voting for the Sweden Democrats.

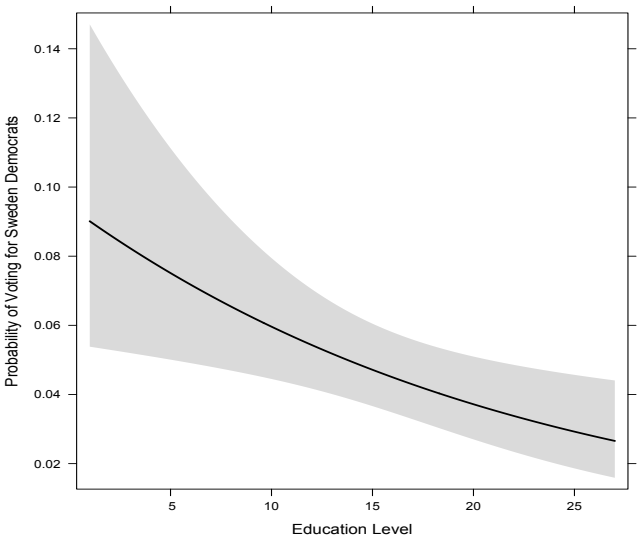
Command:

```
eff2 <- effect("education", mod2, default.levels=100)

effect2 <- print(plot(eff2, rescale.axis=F, rug=FALSE, xlab="Education Level",
ylab="Probability of Voting for Sweden Democrats", main=""))
```

R Output:

Figure 4: Predicted Probabilities: Effect of Education on Voting SD



Where to Find Additional Help

Online help

Stack overflow: https://stackoverflow.com/

R Project: https://www.r-project.org/help.html

R Bloggers: https://www.r-bloggers.com/getting-help-with-r-programming-useful-survival-skills/

R Programming: http://rprogramming.net/

R Cookbook (good for graphing): http://www.cookbook-r.com/

Books

Fox, John and Sanford Weisberg. 2011. *An R Companion to Applied Regression.* Sage Publishing.

Crowley, Michael J. 2015. *Statistics: An Introduction Using R: Second Edition.* John Wiley and Sons, Ltd.

Hothorn, Torsten, and Brian S. Everitt. 2014. *A Handbook of Statistical Analyses using R: Third Edition.* CRC Press.