"Thanks!" Your friend's voice sounds much calmer than it did when you last

spoke. "My tasks this week are much less scary than engaging with long-tailed

and zero-inflated variables like hours spent helping others and groups." When

you ask what she is up to, she replies, "I do need some help. At a recent

meeting we had a debate about two different statements. Can you help me by

providing answers based in data? We got into a shouting matching and only

stopped when I mentioned the word 'data'. I don't know much

about data. Can you help? Here are the two statements."

> Support for anti-immigration populists like Trump or the UKIP party arises from anti-feminism more than nativism."

> In countries like the UK, education and income are not strongly related."

"I managed to find a survey that is supposed to be representative of people in the UK."

```{r readdata}

# library(foreign) ## For older Stata files

#Before you can load readstata13, you must first install the package, or the loading will not work

install.packages("readstata13")

library(readstata13) # This loads up the package readstata13, which allows R to be able to access older types of datasets, which is what you are trying to access here.

#Next, we can see that the person is loading the dataset that will be used for the experiment and creating a variable to allow for easier recognition of the dataset

bes<-read.dta13("http://jakebowers.org/Data/BES2015/bes\_f2f\_original\_v3.0.dta",convert.factors=FALSE)

```

"And I found a (codebook)[http://jakebowers.org/Data] too. And have learned enough about R to investigate some variables."

#When loading up the dataset, you will be presented with a certain number of variables that provide data information for you. However, the dataset will not show you each of the 477 vectors that you have access to, so it may seem as though you do not have access to this information. In order to rectify this, make sure to have the codebook mentioned above loaded on your computer and use it to look up questions and the associated question letter/number that identifies that question. In doing this, you will be able to see each of the questions whose answers you have access to. At the moment, my group is not sure how to provide access to viewing every one of the variables in the dataset at the same time, but rest assured that the codebook will provide you with the answers you need.

```{r sometabs, results="hide"}

#In order to get a gauge on levels of feminism, the question is asked about how much has been done to ensure equality for women. This question is answered and summarized by using the code below. The code below offers a breakdown of how the question was answered by giving the number of answers for each possible response to the question

table(bes$r03,useNA="ifany") ## Feminism

#The code listed below is used to run a summary of the age of the people who answered this question. The lowest possible answer besides refusing to answer was 18, and the highest answer was 97. By running the table code, you are able to see exactly how many respondents were what particular age, allowing you to be able to isolate particular age groups or simply report on the age breakdown of the respondents

table(bes$Age,useNA="ifany")

#The code listed below is run in order to surmise the level of income for each of the respondents. Respondents could respond with Don't Know, refuse to answer, or answer 1-15, with each of these numbers representing a certain level of income. This variable was likely coded -2-15 instead of true income for multiple reasons; one of which being the ease associated with the coding structure and also to make it easier for the survey takers to respond.

table(bes$y01,useNA="ifany") ## Income

#This code listed below is a bivariate variable that is measuring whether or not the people believe that too many immigrants have been let into their country. This variable is being used in order to get an idea of how many are supportive or against the idea of immigrants being let into their country

table(bes$j05,useNA="ifany") ## Immigration

#The code listed below is the measurement being used for education. The codebook first asks if the respondent has any educational or work related skills, and then follows that up by asking the respondent to pick a number 1-18 that best represents the level of skill that they have acquired. This is supposed to measure education, but in the opinion of our group, this does not seem to be the most accurate measurement of education. Rather, a simplified version seems to be the better way to go, with instead of listing the many different types of degrees one can get, instead there were responses boiled down to extremely similar degrees being coded the dame way

table(bes$education,useNA="ifany")

#The code listed below is answering the question of which party the respondent voted for in the previous general election. The answers listed a great many parties, along with a write in option as well. This question is being asked in order to get a general sense of who each respondent supports

table(bes$b02,useNA="ifany")

#The code listed below was a follow up question regarding who the respondent would have voted for in the election if they had voted. Many respondents to the first question regarding voting had not voted, and the researcher is trying to ascertain who supported or is likely to support the UKIP party in the United Kingdom, so this question is asked to see which of the non-voters would be likely to vote for UKIP if they had voted

table(bes$b04,useNA="ifany")

## 1 if voted or would have voted for ukip

#The code listed below is creating a variable that seeks to identify and separate those who did support or would have supported UKIP in the previous general election from the rest of the respondents to the survey. The code is doing this by creating the variable that takes the respondents who answered 7-UKIP in the question b02 and who answered 8-UKIP in the question b04, both of which were gathering who supported or would have supported UKIP. By separating these respondents from the rest and saving them into their own variable, the researcher can use this variable to measure supporters of anti-immigration policies against other variables, such as income, education, etc.

bes$ukipvoter <- as.numeric(ifelse(is.na(bes$b02),bes$b04==8,bes$b02==7))

#The code listed below is used to just show how many people supported or would have supported the UKIP party in the previous general election and how many people did not. These people were coded in a binary fashion, with a 1 representing support for the UKIP party and a 0 representing the respondent not supporting the UKIP party

table(bes$ukipvoter,useNA="ifany")

#The code listed below is used in order to establish the party identification of the respondents of the survey. There were many different options presented along with an optional write-in section in order to make sure to present the entire spectrum of possibilities in the party ID

table(bes$d01,useNA="ifany") # Party ID

```

"What do you think? Between these variables, and the others in the codebook,

can you provide some evidence for or against those statements?

#One way to provide evidence for these statements is to run tests that evaluate the relationship between the variables being mentioned in the statement. One method of doing this would be to do the least squares method. It is called the least squares method because it first creates a best fitting line that shows the relationship between the two variables. Then, after creating the best fitting line, the least squares method attempts to minimize the sum of all the squares of the observed responses in the dataset and the sum of the squares that were predicted by the best fitting lines. By having a smaller sum of the squares, we can see that there is a stronger relationship because that means the points of the dataset are closer to the best fitting line, meaning that the relationship is strong. If the sum is higher, the relationship is weaker because the points from the dataset are fairly far away from the best fitting line, meaning that the relationship is weaker. There is no computation or existing relationship called most squares, but there is a way called the least absolute deviations. You should care about least squares compared to least absolute deviations because it does have some advantages over the other method. For instance, least squares will always provide you with a stable solution, and there will always be one solution for you. This will be helpful in times that you need a solid answer given to you and you want to be able to trust the answer given to you.

#To run the least squares computation regarding the first statement about support for anti-Immigration, you need to run two different computations and compare the results. First, you need to run a linear model computation between the feminist variable and the support for ukip variable in order to measure the relationship between these two variables. That computation would look like this

lm(bes$r03 ~ bes$ukipvoter)

#Running this will return you an intercept and a coefficient measuring the relationship between the two variables. When running this function, we find the coefficient to be -.05757, which means that the more feminist the respondents were, the less likely they were to vote for the UKIP, and vice versa. While there is a relationship between these two variables, it is not a very strong coefficient. The next step is to run the relationship test between the variable measuring feelings on immigration and the support for the UKIP, which would look like this

lm(bes$j05 ~ bes$ukipvoter)

#When running this computation, we find the intercept to be -.09586, which, although not as significant as one would like, is almost twice as strong as the relationship between anti-feminism and voting for UKIP. What this seems to suggest is that the statement being made about anti-feminism being more of a tell than nativism in support for anti-immigration politicians seems to be false. In fact, nativism/anti-immigration feelings in voters is almost twice as much of a predictor in voting for the UKIP than anti-feminist feelings.

#Next, we want to run this same least squares regression between education and income in order to provide a basis for the relationship between the two variables. To run this regression, we will use the code

lm(bes$education ~ bes$y01)

#Here we find that the coefficient given is .022, which would indicate that there is a positive relationship between education and income. While this relationship is positive, the coefficient is only .022, which is a very low coefficient. This provides evidence speaking to the idea that there is actually a fairly weak relationship between these two variables, giving support to the statement that was made up above.

#There are both strengths and weaknesses that are associated with running a least squares regression in order to measure the strength of the relationship between two variables. As far as the strengths associated with running a least squares regression, there are two that stand out. First of all, it is the easiest way to identify the best fitting regression line, which then allows you to run the least squares or even any other method of regression in order to test the relationship. Another strength of running this method is that is guarantees to only give you one best fitting line. Other regression models will sometimes give you multiple regression best fitting lines, which some may see as a positive since it gives you a choice, but we argue that having one line is better because it does not leave decisions up to the researcher. There may be times that different researchers us the same dataset but when running a regression some researchers may choose different lines. By having the one line, it guarantees that researchers will be on the same page. There are some weaknesses regarding using this method though, one of which is the way that this method handles outliers. In using the least squares regression model, an extreme outlier can severely change the best fitting line, thereby throwing off the least squares response regarding the coefficient. This means that if you are using a dataset that you know has extreme outliers, it would be best not to run the least squares method unless you want to take out all of the extreme outliers from the dataset first.

#Another of the ways that you can test the relationship between variables is the correlation function cor(). The correlation function will take two different variables and provide you with a coefficient that determines the strength of the relationship between the two variables. With regards to the first question mentioning support for Trump and the UKIP related to either anti-feminism or nativism, the first thing to do is find the variables that would represent each of the ideas. Regarding the first statement measuring support for anti-Immigration candidates like Trump or the UKIP, we were not able to run the correlation function. We believe this is because the ukipvoter vector is coded as a row instead of a column, and the correlation function measures the relationship between two columns. We tried to remedy this situation by removing the NA values from each of the vectors, which we believed would allow us to run the computation. However, RStudio would not create a new variable that took the NA values out of the ukipvoter vector, thus disabling us from running the correlation computation on the two variables. However, if you are able to fix this situation and run the correlation computation on the two variables, you would be given a coefficient that gives a measurement of the two variables, allowing you to see what type of relationship between two variables and whether it is a significant value. That code, when fixed, would look something like this

cor(bes$r03, bes$ukipvoter)

cor(bes$j05, bes$ukipvoter)

#The second statement regarding the relationship between income and education can be ran using the correlation function already. The function would go as follows

cor(bes$education, bes$y01)

#When running this correlation function, you are given the coefficient of .0216, which would indicate that there is a very small and insignificant relationship between these two variables. This would seem to suggest support for the statement made above, which claimed that there was not a strong relationship between income and education in countries like the UK.

#A strength of using the correlation function to find the relationship between variables is that it is a good starting point that will enable you to get a first glance at the relationship between variables. While it only provides you with a single coefficient, it can be a good initial test to first tell whether or not there is even a somewhat significant relationship between a couple of variables. However, a weakness of using this computation is that it does not help with establishing causation. Running a correlation test does indeed tell you whether or not a relationship exists, but it leaves the researcher to decipher whether or not there is causation, which can make it much more difficult for the researcher.

#A third and final way to measure the relationship between variables would be to run the least absolute deviation regression. This regression is similar to the least squares regression in that it attempts to formulate a best fitting line from the dataset, and it then attempts to minimize the absolute values between the dataset points and the points on the best fitting lines.

#In order to run this regression, you must first install a package called quantreg. To do this you would use the code

install.packages("quantreg")

#Next you want to load the package you just installed, using the library function

library(quantreg)

#Finally, we want to run two different regressions first between anti-feminism and voting for UKIP and then between nativism/anti-immigration and voting for UKIP. After we do this, we can measure the coefficients given by the computation and see which one is stronger. To run the regression between anti-feminism and voting for the UKIP, we would use the code

rq(bes$r03 ~ bes$ukipvoter)

#This returns you with an intercept of 3 and a coefficient of 0. We believe that this returned a zero because it rounded the actual value of the coefficient to the nearest integer, and since the value was so low it returned a zero. The next step is to run the computation between nativism and voting for the UKIP, for which you would use the code

rq(bes$j05 ~ bes$ukipvoter)

#This code will return an intercept of 1 and a coefficient of 0 as well. Again, our group believes that this happened because the computation rounded the results to the nearest integer. This variable makes it very difficult to measure the relationships, especially due to the fact that they returned the same integer with their coefficients. Because of this, it is the opinions of our group that this measurement, while it works, is not a very good indication of measuring the difference.

#Regarding the second statement, we need to run the same computation regarding the two variables that represent the education and income of the respondents. The code to do this is

rq(bes$education ~ bes$y01)

#When running the computation, the function returns an intercept of 3.22222 and a coefficient of .111111 which would indicate a positive but mostly not significant relationship. This regression provided a stronger coefficient than the other tests did, but even with the increased size of the coefficient, we find that the statement made regarding the weak relationship between income and education holds up with the empirical data supporting it.

#As with the other methods, there are strengths and weaknesses associated with this method. One of the strengths of using this model is that it gives less importance to the extreme outliers in the dataset. By giving equal weight to each of the points from the dataset, the outliers are not going to skew the coefficient near as much as it would in other methods, such as the least squares method. While there may be times that one would want to emphasize the outliers in the data, many times a researcher would not want their results skewed by the outliers, so this method can be helpful in that way. One of the bigger weaknesses of using this method would be that it can return multiple different best fitting lines with your dataset. This means that the researcher will need to determine which best fitting line would be better, but this can be tougher for replication and would most likely be a weakness.

Since this was such a heated discussion in our staff meeting, I'd like to see at least three ways to describe the relationships including that least squares stuff that people always talk about. And I'll need you to justify the weaknesses and strengths of the tactics you choose to use. And, especially for the least squares description, can you please help me understand why it is called least squares and why I should care about least squares versus most squares or least something else? Right now, we just want to describe relationships so $p$-values or confidence intervals would not be helpful."

#In the opinion of this group, we think that the best method for measuring the relationship between the variables mentioned would be using the least squares method. Although it can be skewed by extreme outliers in the dataset, there are ways to eliminate the outliers from the dataset and then run the least function squares regression. When running these computations, we found support for the statement regarding the relationship between education and income. There was a very small relationship that existed, but due to the weakness we can say that the relationship is not significant, just as the statement above said. However, with regards to the first statement, we found that the opposite was true. While there was a relationship between anti-feminism and voting for anti-immigration candidates, the least squares regression coefficient was nearly twice as large with the nativism variable, meaning that it was much more likely that you were influenced by being anti-immigration rather than being anti-feminist. This makes sense because we are measuring support for anti-immigration candidates. While there could be some relationship between anti-feminism and support for the candidates, it would make much more sense for the people that don't like immigrants coming into their country to support candidates who are likely to cater to that political stance.