Problem Set #2

DUE IN-CLASS, THURSDAY, MARCH 5

DIRECTIONS

Answer each of the questions below. You should submit your answers in the following form (points will be deducted for improper form):

1. Type up all of your answers, using a reasonable font and margins.
2. Use complete sentences in answering all of the questions.
3. Properly label your problem set with (a) your name, (b) the number of the question you are answering. If it makes sense to do so, you can combine parts of a question together in your answer (i.e. do parts 1a, 1b, and 1c together, etc.). Just be sure to label your answers appropriately.
4. Simply cutting and pasting output from STATA into Word is not enough to answer the question (unless the question just asks you to draw a graph). You should run the analysis in STATA, then pick out the information you need to answer the question and write it up in your problem set. If you are including a table of data in your answers, then you should copy it from STATA into Excel, then properly format it and keep only the information you need to answer the question.

See the example of how to properly submit a problem set.

STATA NOTE: For some of the problems in this problem set, you will need to look carefully at your data to make sure that missing values and such are properly coded. Unfortunately, some of the datasets do not have codebooks associated with them. In that situation, when you need to know what the numeric codes associated with certain categories in a frequency table are, you can use the command

```
tab [insert variable name], nolabel
```

For example, let’s say you want to look at a variable called minority. When you run a frequency table, it tells you that there are 1000 non-minorities and 50 minorities. You do not know, however, if non-minority=1 and minority=2 or, alternately if non-minority=0 and minority=1, etc. So, if you run the command above, it will show you the same frequency table, but use numbers instead of labels.

Question #1

Open the file “marylandcounty.dta” from the course website. Let’s see if there is a relationship between water consumption (variable 6 in the file, “consumptive water use per million gallons (water)”) and the crime rate (variable 2 in the file: “serious crimes per 100,000 (crime)”). In this exercise, we are interested in the pattern of association between crime rates and water use, not cause and effect. Let’s treat crime as the independent variable and water use as the dependent variable. This designation is arbitrary, but for everyone to be on the same page make sure you treat your variables this way: crime is X and water use is Y.

a. What is the unit of analysis in this data?

b. Provide the univariate descriptive statistics on water consumption. Specifically, what is the mean, median, mode, range, and standard deviation are of water consumption?
c. What does this tell us about the distribution of the variable? Indicate what you know about the shape of the data using the mean, median, and mode data. Indicate what you know about how spread out it is and which statistics tell you this.

d. Graph a histogram of water consumption and indicate the shape.

e. What is the correlation between water consumption and crime?

f. Use STATA to draw a scatterplot of water consumption and crime. Keep in mind what we said about the placement of the dependent and independent variables and clearly and fully labeling axes and keeping embellishments to a minimum. Examine the plot. What type of relationship (if any) do you see? That is, is it linear, positive, etc.? How strong is it? Imagine a regression line passing through the “middle” of the points. Would the slope be positive or negative?

g. Use STATA to run a regression. What is the estimated regression coefficient (slope)? Give a statistical and substantive interpretation of this number.

h. What is the estimated regression constant (intercept) for these data? Interpret what this means. Does it make sense in real-world terms?

i. What is the R-squared for this relationship? Give a statistical and substantive interpretation of this measure.

j. If you found a relationship between two variables, what might explain it? Surely, crimes don’t cause people to consume more or less water. Can you think of an explanation? (Tip: There are two ideas to consider. First, where is water most likely to be consumed? And what’s apt to characterize living conditions in those same places? Second, look carefully at the plot. Do you see one or more observations that perhaps do not fit the overall pattern?)

**Question #2**

This question provides you with additional practice recoding variables, creating crosstabs, and making graphs in Excel.

a. Download ANES2004.dta from the course website. First, find the variable that indicates what respondent’s own ideology is. (Right now, that variable has 7 categories. Recode that variable into 3 categories: liberal, moderate, and conservative. Recode it as follows: very liberal, liberal, and slightly liberal=liberal; moderate=moderate; slightly conservative, conservative, and very conservative=conservative. Be sure to label the new categories in your new variable!). Then find the variable that indicates how often the respondent attends church. Then create a crosstab of the relationship between church attendance and your new ideology variable. Show your results and interpret your findings.

b. In Excel, create a clustered bar chart. The x-axis should be ideology. Each cluster should be defined by how often the respondent attends church. The y-axis should show the percent of people who fit into each category.

**Question #3**

Download the dataset “bes2001.dta,” a survey of potential voters in England, from the course conference. Let’s say we want to see if respondents’ level of personal efficacy (“People like me have no say”) has any effect on how interested they are in the election (“Interest in 2001 General Election”).

a. In the question specified above, which is the independent variable and which is the dependent variable?

b. Create a hypothesis for what the relationship between the two variables may be.

c. If we were to run a regression of interest in the election on efficacy, what are other control variables we might need to include? List one possible control variable and explain why it is necessary as a control.

d. List a second possible control variable and explain why it is necessary as a control.
Problem Set #2, Page 3 of 3

e. To learn more about the relationship between the variables, run a regression model. Use “Interest in 2001 General Election” as your dependent variable. Your independent variables will be “People like me have no say,” “Annual Household Income,” and “Gender.” Recode all the variables to exclude categories that should be missing (those labeled “Not app” “don’t know” and “refused”). Run the regression and present your results in a table. Remember not to just cut and paste. Include all the important statistics, ignoring those that are irrelevant to the problem.

f. Interpret your findings from the regression. In particular, interpret what the effect of efficacy seems to be on the dependent variable.

g. Interpret the meaning of the adjusted-R-Squared.

**Question #4**

In this exercise, we will look at the effect of the variable, REGION, on the relationship between senators’ party identification (PARTYRD) and their scores on the ACLU scorecard. Use the data from “senate.dta” on the course conference. In which region does the relationship between party and scorecard score the strongest? The weakest?

a. Describe the relationship between PARTYRD and the ACLU, which you can see by creating a crosstab. Throughout, treat ACLU as the DV. You don’t have to show your results, but interpret the meaning of your results using specific numbers from your table.

b. Regress ACLU on PARTYRD (Note that PARTYRD should be coded as 1 for Dem and 2 for Rep). Show your results (intercept, coefficients, adjusted r-squared, and number of cases). Interpret your findings (don’t forget to interpret the magnitude of the coefficients and the r-squared).

c. Create dummy variables for REGION and regress ACLU on party and the new dummy variables you created for region. Use Midwest as your omitted category. What do you find? Show your results (intercept, coefficients, adjusted r-squared, and number of cases).

d. Interpret your findings, by indicating what the relationship between party and ACLU scores are, and what effect overall region has on the regression results (in comparison to your results in part c). Also indicate how to interpret the results of each of your dummy variables. Be sure to indicate what the comparison group is in your answers.

e. Create a dummy variable for Republicans. Create an interaction term between Northeast and Republican. Then regress ACLU on your dummy variable for Republican, 3 of your dummy variables for region (using Midwest as your omitted category), and the interaction term. What do you find? Show your results (intercept, coefficients, adjusted r-squared, and number of cases).

f. Interpret your findings above, by indicating what the relationship between party and ACLU scores are, and what effect overall region has on the regression results (in comparison to your results in parts c and d).

g. Using the results from part f, indicate how to interpret the results of each of your interaction term. What is the predicted ACLU score of a Northeastern Republican? A Northeastern Democrat? What about a Southern Republican?