**17.871, Political Science Lab**

**Problem set # 1: Using STATA**

Spring 2012

Handed out: Feb. 13

Due: Feb. 27, *at the beginning of class. (Please print before coming to class. Unfortunately, the printers in our classroom are ridiculously slow and loud.)*

Write a do-file that responds to all the parts of this problem set. Turn in the do-file and the log-file that shows that the do-file works. Clearly label each do-file. Make a habit of writing comments in the do-file, to help us and you keep track of things. (You can make non-executed comments in a do-file using the front slash and asterisks as follows:

/\* THIS IS WHAT A COMMENT LOOKS LIKE IN A DO-FILE \*/

**Part I: Golf putting data (6 points for the part, one point for each step)**

Variables

dist                             distance to hole in feet

tries                            number of putting attempts

success                       number of successful puts (one hit only)

1. Open the data set putting.dta from the 17.871 course locker (Examples folder) or off the class website. Paste the code for opening the file into your do file as your answer to question 1. (Don't forget to use the clear command first.)
2. Examine the data:
	1. With small data sets, you can easily see the data with the list command. Try it.
	2. What are the mean, min, and max of each variable? (Hint: summarize)
	3. Use the tabulate command to examine the distribution of each variable one at a time.
	4. Use the tabulate command to examine the distribution of each variable one at a time with one line of code. (Hint: tab1)
3. Create a new variable called success\_rate that is equal the proportion of successes. (Hint: generate)
4. Label your new variable "Put success rate (proportion)." (Hint: label variable)
5. Create a scatter plot of success rate (y-axis) by distance (x-axis). (Hint: scatter)
6. Which is the dependent variable and which is the independent variable in step 5? Why?

     (Don't forget to save your do file.)

**Part II: Getting data into STATA (five points for the part)**

Data comes in many forms. Here's one way to get data into Stata. Using a text editor (such as EMACS), type the text from Exhibit 1 in the handout “How to Use the *STATA*infile and infix Commands” into Athena and save it in a file named scores.dat on your home directory.

Write the code that will create a STATA data set from this raw data and save it as a file called “scores.dta”. Use the list command to see your data.

**Part III: Speed-dating data (16 points for the part, one point for each step, except for point 12)**

Speed-dating data from studies conducted in New York City by Ray Fisman and Sheena Iyengar, an economist and a psychologist at Columbia University. If you're interested, they summarize their findings in [this paper](http://www2.gsb.columbia.edu/faculty/rfisman/datingFULL-EK1.pdf). You'll need to familiarize yourself with the codebook (see the class website). Here's the abstract for the paper:

We study dating behavior using data from a Speed Dating experiment where we generate random matching of subjects and create random variation in the number of potential partners. Our design allows us to directly observe individual decisions rather than just final matches. Women put greater weight on the intelligence and the race of partner, while men respond more to physical attractiveness. Moreover, men do not value women's intelligence or ambition when it exceeds their own. Also, we find that women exhibit a preference for men who grew up in affl­uent neighborhoods. Finally, male selectivity is invariant to group size, while female selectivity is strongly increasing in group size.

1.    Open the speed\_dating.dta file from the 17.871 course locker (Examples folder) or off the class website. Paste the code for opening the file into your do file as your answer to question 1. (Don't forget to use the clear command first.)

2.    Get a sense for the data: How many unique subjects participated in the experiments? (Hint: tabulate)

3.    The questions in this section are about variables that are constant across the multiple speed-dating waves, such as a self-reported question about how often students go on dates. To analyze the responses, we need to eliminate the multiple occurrences of participants so that each individual occurs only once in the data set (that is, one row per person). To do so, use the collapse command. With this command, we can take the average of participants’ responses.

* 1. To see which variables we are going to analyze in this section, first run the following command: sum wave gender date dec \*1\_1 (Note how the \* acts as a wildcard.)
	2. Eliminate multiple occurrences by running the following command: collapse wave date gender dec \*1\_1, by(iid).
	3. Recode the variable date so that the values roughly correspond with number of dates per year (e.g., once a week = 52) and call this variable dates. Do this with generate and replace. Drop this first variable (drop dates).
	4. Do this recoding again with recode. (Hint: the recode command is structured very differently from generate or replace. Look back at the book or at the Stata help file for examples of how to use it.)
	5. What's the modal category on the dates variable? (Hint: tabulate)
	6. On average, how many dates do participants go on each year?
	7. How many men and how many women participated in the experiments?
	8. Who goes out on dates more often: men or women? (Hint: tabulate gender with sum(dates) as an option.)
	9. In speed dating, are men or women more selective? (Hint: similar to previous question)
	10. In waves 6-9, the experimenters use different scales for the attribute preference questions (e.g., attr1\_1). To simplify, drop waves six through nine. (Hint: drop if wave == 6) You can also save yourself time with the following command: for num 6/9: drop if wave == X.
	11. Do men and women report placing similar weights on traits in potential partners? What's the biggest difference? (Hint: by gender, sort: sum attr1\_1)

4.    Using a non-collapse data set with waves 6-9 restored, determine which participant(s) (iid) sought the most matches? (Hint: first clear and reload the original data set, then create a variable decisions that totals the number decisions to pursue a match by each participant (dec) with the egen command (egen decisions = total(dec), by(iid)).

5.    What was the maximum number of "matches" participants received across the speed dating rounds? (Hint: similar to 4.)

6.    What was the highest success rate observed among participants? (Hint: create a new variable match\_rate with the generate that equals matches divided into decisions.)

7.    The speed-dating data contains a variable that codes the median SAT score for participants’ undergraduate institutions. The variable, however, is not coded in numeric form. What form is it in? Convert it to a numeric variable. (Use describe to determine the variables’ format. Use destring to convert the variable. You will have to use the ignore and replace options.)

8.    SAT terciles I: Create a variable that equals 1 for the bottom third of participants’ undergraduate institutions based on the median SAT variable, 2 for the middle third, and 3 for the top third. Do so with recode using the generate new variable option.  To figure out the 33rd and 66 SAT percentiles for the sample, use tab.

9.    SAT terciles II: Now that you've practiced recoding, show how you can save yourself considerable time in the future by creating this variable again using xtile.

10.    Does a higher SAT tercile predict a higher match\_rate? (Hint: use one of the commands above.)

11.    More practice with the collapse command.

* 1. Create a new data set that contains the average ratings for each self-reported attribute (e.g., attr3\_1) and the average ratings by partners for each participant (e.g., attr\_o). (Hint: use the collapse command with the by option as in question 3.)
	2. Do any variables have missing data?
	3. On what traits do participants’ self reports tend to correspond with those of their partners? On what traits is there no correspondence? (Hint: corr.)
	4. Using this same data set, generate a scatter plot of participants’ own attractiveness ratings by partners’ ratings of the attractiveness of these participants. So that you can see each point, add some randomness to each point with the jitter option. Would you describe the relationship as strong, moderate, or weak? No need to print the scatter plot. (Hint: scatter attr3\_1 attr\_o, jitter(10).)

12.    Look through the codebook and come up with an interesting question that these data can answer. Use STATA to answer the question (five points).

      (Don't forget to save your do file.)

**Part IV: Red versus blue states? (four points total, one point for each step)**

 Is America polarized? To answer this question, we will analyze data from a survey of more than 36,000 Americans conducted by researchers at MIT in 2006. Determine whether states are polarized by partisan identification (Democrat versus Republican). That is, are most states either heavily Democrat or heavily Republican? Alternatively, are most states in the middle, with both Democrats and Republicans. To get started, open CCES.dta from Examples folder in the course locker.

 Using this data, create a histogram that shows the distribution of partisan identification by state. The partisan identification variable is a seven-point scale (pid7).

1. First, recode irrelevant values to missing on pid7.
2. Then, you need to take the average partisan identification for each state. To do so, use the collapse command with by state option. Your new data set should have one row for each state.
3. Create a histogram of average partisan identification.
4. Does partisan identification look polarized at the state level?

**Part V: Research design (ten points total for the part, five points for each step)**

Comment on the research designs of the following two studies. Discuss whether they are designed in a way that would allow the researcher to draw the stated conclusion. State what are the dependent and independent variables in these designs, and what any confounding variables might be. If the research design was insufficient, write a short paragraph indicating why not, and what could, or should, have been done to improve the design.

1. MIT faculty members were interested in determining whether ending spring-term freshman Pass/No Record had been a success. They decided to answer this question by comparing the GPA of spring-term freshmen before and after the change in Pass/No Record grading had taken effect. The average freshman GPA in the spring of 2002 is 4.0; the average freshman GPA in the spring of 2003 is 4.4. The faculty concluded that the change was a success. (Note the obvious: these are made-up data.)
2. Researchers were interested in determining whether postcards sent to registered voters encouraging them to vote actually worked. The researchers took the list of registered voters in a town (about 100,000 individuals) and randomly assigned them to one of two samples—T, a sample of voters who were sent the get-out-the-vote postcard, and C, a sample of voters who were not sent the get-out-the-vote postcard. After the election, the researchers went to the town clerk to see who voted. They discovered that 70% of the T group voted, whereas 59% of the C group voted, a highly significant difference, a highly statistically significant difference. The researchers concluded that the “causal effect” of the postcards is to increase turnout by 70%-59% = 11%.

**Part VI: Finding and merging data (twenty points for the part, one point for each step, plus a bonus for getting through the whole thing)**

1. To examine incarceration rates at the state level, find and download from ICPSR (<http://www.icpsr.umich.edu/icpsrweb/ICPSR/>) the following study: IMPACT OF STATE SENTENCING POLICIES ON INCARCERATION RATES IN THE UNITED STATES, 1975-2002 (ICPSR: 4456). Make sure you download the Stata file (note: Primary Data) and save it to your Athena directory. We want to merge this file with another data file also at the state level. To do so, we need to have the same unique identifier for states in both files (e.g., state names with the same capitalization, etc.). What unique state identifiers are available in this file? What are the names of these variables?
2. Find and download a file with state median household income from the 2010 U.S. Census (<http://www.census.gov/>)
	1. Utilize the American FactFinder system (linked from the bottom left, under the FIND DATA section.
	2. On the left-hand of the American FactFinder page, you’ll want to use the **Topics** option to search for data about “median household income.” Under **Geographies** you’ll want to select “All states of the United States.”
	3. You will be given a list of data tables to choose from. It is likely that the one you want will be at the top, labeled something like “median household income in the past 12 months” or “selected economic characteristics.” Click on the link to bring you to the table.
	4. You will notice that the table will likely be pretty big, and include a lot more information than you need. You will also discover that the American FactFinder interface is very inefficient for slicing down tables. Therefore, the best thing to do is to download the table in Excel format, and then use the spreadsheet to delete everything you don’t need.
	5. It is likely that the dataset you download will have the variables along the rows and the observations (state) along the columns. You will want these reversed. Once you have selected only the variable you want, the best way to reverse the rows and columns in Excel is to copy the data you want to another worksheet, using the “Copy special” command. You want to select the option to transpose the data matrix.
	6. Once you have processed the data in Excel, output the data into a .csv file called income.
3. Prepare the first file for merging:
	1. Open ICPSR Study 4456.
	2. Drop data for all years except 1999 (Hint: use the keep command).
	3. Drop all variables except State name and incarceration rate, which is per hundred thousand people (STATE\_NA INC\_RATE). Notice that state names are truncated after 10 characters; this will have to be addressed before merging (see below).
	4. Sort by state name. Save this file as incarceration.dta.
4. Prepare the second file for merging:
	1. To get the income data into Stata, use the insheet command. If you previously saved a clean version of the dataset (with only the data and perhaps labels at the top of each column with variable names), you should be all set. If not, use a text editor (not the spreadsheet) to edit out any extraneous information in the .csv file.
	2. Notice whether the income data have been imported as a numeric or string variable. Because of the commas in some Census Bureau datasets, the income data may be non-numeric. If so, convert income to a numeric variable with the destring command.
	3. You may also notice a few blank rows in your data set. Drop them.
	4. To merge state income with the ICPSR Study 4456, you need to name the state-names variable the same as the state-name variable in the ICPSR study.
	5. To merge the data sets, the state names variable not only needs to have the same name in both data sets, but also has to have the same values for each state. To address the problem with truncated state names in the incarceration data by truncating the names in the income data. Use the substr command to extract the first 10 characters, e.g., replace STATE\_NA =substr(STATE\_NA ,1, 10) .
	6. To show that you have successfully saved these data into Stata, run the list command. Sort by state name. Save this file as income.dta.
5. Merge the two data sets. Save the merged dataset.
6. Label the income and incarceration rate variables.
7. Test that you have successfully merged your data by tabulating \_merge.
8. Are your merged variables actually related? Check with the scatter command.