17.871, Political Science Lab
Spring 2009
Problem set # 1: Using STATA

Handed out: Feb. 9, 2009
Due: Feb. 23, 2009, *at the beginning of class.*

For Parts I-III and V, turn in a "log" file produced from running the do file.

Part I: Golf putting data (one point each)

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.
2. Examine the data:
   a. With small data sets, you can easily see the data with the `list` command.
   b. What are the mean, min, and max of each variable? (Hint: `summarize`) Use the underlined characters as a command shortcut.
   c. Use the `tabulate` command to examine the distribution of each variable one at a time.
   d. 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? Why?

Part II: Speed-dating data (one point each)

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. 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.
2. 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.
    a. 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.)
    b. Eliminate multiple occurrences by running the following command: `collapse wave date gender dec *1_1, by(iid)`
    c. Recode the variable `date` (see page 4 of the codebook) 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`).
    d. Do this recoding again with `recode`.
    e. What's the modal category on the `dates` variable? (Hint: `tabulate`)
    f. How many dates does the average participant go on each year? (Hint: `summarize`)
    g. How many men and how many women participated in the experiments?
    h. Who goes out on dates more often: men or women? (Hint: `tabulate gender` with `sum(dates)` as an option.)
    i. In speed dating, are men or women more selective? (Hint: similar to previous question)
    j. 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 .`
    k. 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`) These questions appear on pages 5-6 of the codebook.
4. Which participant(s) (`iid`) sought the most matches? (Hint: first 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).`) Before answering this question, reopen the survey to restore waves 6-9 and to undo the collapse command.
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. First do so with `recode` using the generate new variable option.
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.
    a. 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.)
    b. Do any variables have missing data?
    c. On what traits do participants' self reports tend to correspond with those of their partners? On what traits is there no correspondence? (Hint: `corr`.)
    d. 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).

Part III: Getting data into STATA (five points)

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 document "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 IV: Research design (five points each)

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 V: Finding and merging data (one point each)

1. To examine incarceration rates at the state level, find and download from HMDC (http://libraries.mit.edu/get/hmdc) 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 in 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 these variables called?
2. Find and download a file with state median household income from the 2000 U.S. Census (http://www.census.gov/)
    a. Utilize the American Fact Finder system (linked from the left)
    b. Under Get Detailed Data, select Decennial Census
    c. Income information is only asked of selected households and is located in the file Census 2000 Summary File 3 (SF 3) - Sample Data
    d. Create a Geographic Comparison Table
    e. Select as the geographic type "nation" and for the table format "United States -- States; and Puerto Rico"
    f. Download the table to your Athena locker as a .csv file called `income`.
3. Prepare the first file for merging: Open ICPSR Study 4456 and drop data for all years except 1999 (Hint: use to keep command). Drop all variables except State name and incarceration rate, which is per hundred thousand people (`STATE_NA INC_RATE`). Sort by state name. Save this file as `incarceration.DTA`.
4. Prepare the second file for merging: To get the income data into Stata, use the insheet command (delete the irrelevant top rows, delete irrelevant variables, and rename the relevant variables, which are state name and median household income). 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. 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.