17.871, Political Science Lab
Spring 2012
Problem set # 2

Handed out: February 29
Due: March 7

Submit

1. A paper copy of your answers to each question, your graphs, and calculations.
2. Your log file. Please use one do file for the entire problem set.

You may work on these together, but you must write them up separately.

Part I: Distributions (2 points for each variable, 6 points total)

Consider the following three variables. Complete the following exercises with respect to each:

1. Create a dataset of each of the variables described below. (That is, create three datasets, one for each variable.) The datasets should include the variable mentioned along with the case identifier(s) mentioned (congressional district, state, etc.) Find the mean, standard deviation, skewness, and kurtosis of the variables. Also produce a histogram, or other appropriate graph, that shows the distribution of the variables. (Produce a do-file that reads in the data, saves a Stata file in .dta format, and then performs the statistical procedures requested.)
2. In order to find the answer to step 1, find the original sources for the data you are describing. (By *original source,* I mean a source that produces the data in reasonably raw form. I also mean you must avoid pointing out a dataset in a spreadsheet that someone has posted on the web, although you may find that someone has already done the data entry for you for this exercise.) Give the citations to how to locate the data, either a traditional bibliographic citation (author, title, etc.) of a book or a URL of an electronic data source.

Here are the variables:

1. The percentage of votes received by each Republican candidate for the U.S. Congress from California in the 2010 election.
2. The average SAT score in each state in 2011.
3. Military spending of countries in 2005 as a percentage of GDP.

Part II: Golf putting data and regression (six points total, one point for each individual question)

Variables

dist	distance to hole in feet
tries	number of putting attempts
success	number of successful puts (one hit only)

1. Use your code from the previous problem set to create the variable success_rate that is equal the proportion of successes (data file is **putting.dta**).
2. Using the regression command, estimate the effect of distance on success rate.
3. Interpret the coefficient estimate for distance.
4. Interpret the confidence interval.
5. Interpret the Standard Error of Regression (SER, Stata calls it Root MSE).
6. Create a scatter plot of success rate (y-axis) by distance (x-axis), **plus** add the regression line (see lecture slides for code).

Part III: Scatter plots and ecological inference (nine points total, three points each)

Use the **CCES.dta** file in the Examples folder of the course locker. The CCES is a survey about the 2006 election conducted at MIT by Stephen Ansolabehere, who is now, alas, at Harvard. It interviewed over 30,000 individuals.

1. Create a publishable scatter plot of average party identification (pid7) by income *at the individual-level*. You should have one data point for average partisanship (y-axis) for each income level (x-axis) in your scatter plot. (Hint: Use the collapse command to average partisanship by income. By publishable, I mean
   a) Recode income to meaningful values (and code irrelevant values to missing).
   b) Code irrelevant values to missing on party identification.
   c) Label both variables.

   Note: To see how variables are coded, use tabulate, e.g., `tab income`. The variables in this data set have value labels, so tabulate does not reveal the actual values of the variable. To see the actual values, use tabulate with the no label option, e.g., `tab income, nol`.
2. Create a publishable scatter plot of average state partisan identification by average state income *at the state level*. You should have one data point for each state in your scatter plot, with a state's average partisanship on the y-axis and a state's average income on the x-axis. Before doing so, drop the District of Columbia. On the scatter plot, label the data points with state abbreviations. (Hint: Use the collapse command to average partisanship and income by state.)
3. Does the relationship between partisan identification and income differ between the state level and individual level? (If it doesn't, you have made a mistake.) Briefly suggest an explanation for any difference.

Part IV: Interpreting regression coefficients (eight points total, two points each)

Using the data set **quartet.dta** in the Examples folder of the course locker:

1. Regress each y on its corresponding x (e.g., y1 on x1, y2 on x2). Present the results in a table with four columns, one for each regression. The rows of the table should be the slopes, the constant, the confidence intervals, and the Standard Error of Regression (Root MSE). You can create this table in Word, Excel, Open office, etc. No need to recode variables before running these regressions.
2. Interpret the coefficients and the Standard Error of Regression.
3. Do you believe these estimates? Explain.
4. What should you conclude about the use of regression (and other fancy statistical procedures/predictions) from this example?

Part V: Practice generating alternative explanations (three points total, one point each)

1. In observational studies, what are the two main problems researchers face with internal validity?
2. Why do experiments overcome these two problems?
3. For each of the following examples, write a sentence or two that explains how these two problems could or could not explain the finding. Make sure to label your alternative explanation as either one of the two types from Question 1.
    a. A study finds a positive association between watching Fox news and supporting the Newt Gingrich for President. It concludes that Fox news exposure induced support for Gingrich.
    b. A study finds that regions of Afghanistan with a stronger military presence experience higher insurgent activity. It concludes that the military generally fail to suppress the insurgency, and may even inflame it.
    c. Numerous studies find that breast-fed children have higher IQs than bottle-fed children. They conclude that breast-feeding increases IQ.