17.871, Political Science Lab
Spring 2015
Problem set # 3

Handed out: March 11
Due: March 18

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: Placement of political parties (12 points, 2 points for each part)

In /mit/17.871/Examples you will find cces12_common_subset, which is a subset of the 2012 Cooperative Congressional Election Study. You will use that dataset for this question.

1. Using the regression command, estimate the effect of one's own ideological self-placement on how positively a respondent views the U.S. Supreme Court. (Note, the proper weighting variable to use in this study is variable V103.)
2. Interpret the slope coefficient.
3. Interpret the confidence interval of the slope coefficient.
4. Interpret the Standard Error of Regression (SER, Stata calls it Root MSE).
5. Draw a graph that plots the estimated regression line against the independent variable. There is no need to plot the original data. (Why?)
6. Write a paragraph or two in which you explain, in substantive terms, what the slope coefficient means, in terms of how voters view the Supreme Court.

Part II: Scatter plots and ecological inference (14 points total, two points each)

Use the cces12_common_subset dataset that was used previously.

1. Create a publishable scatter plot of party identification (pid7) by education *at the individual-level*. By publishable, I mean
    a) Recode education to meaningful values (and code irrelevant values to missing).
    b) Code irrelevant values to missing on party identification.
    c) Label both variables.
    d) You may use the "jitter" subcommand, if you think it helps.
2. Add a best-fit linear regression line to this graph.
3. Conduct the regression implied by # 2 and interpret the coefficients.

Note: To see how variables are coded, use tabulate, e.g., `tab educ`. The variables in this data set have value labels, so tabulate does not reveal the actual values of the variable.

4. Create a publishable scatter plot of average partisan identification by average education *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 education on the x-axis. On the scatter plot, label the data points with state abbreviations. (Hint: Use the collapse command to average partisanship and education by state.)
5. Add a best-fit linear regression line to this graph.
6. Conduct the regression implied by # 5 and interpret the coefficients.
7. Does the relationship between partisan identification and education differ when analyzed at the state level vs the individual level? Briefly suggest an explanation for any difference you might see.

Part III: 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 IV: Functional forms (9 points, 3 points each)

There is a dataset in the Examples folder named gdp_co2. The variables, in addition to the identifiers, are GDP in U.S. dollars in 2010 and kilotons of CO_2 emissions in 2010. The unit of analysis is the country.

1. Explore the relationship between CO_2 emissions and GDP, with CO_2 emissions the dependent variable.
2. Run a regression in which the dependent variable is emissions and the independent variable is GDP. Write a couple of sentences that interpret the regression slope coefficient and the intercept. (To help with interpretation, you might want to transform the dependent variable so that it is expressed in terms of billions of dollars.)
3. Interpret the slope coefficient from this regression.
4. Because of the right-skew in both variables, this is a situation in which you may want to transform them using logarithms. Perform this transformation, run the regression on the transformed variables, and interpret the slope coefficient from this regression.
5. Questions (2) and (4) of this part asked you to describe the relationship between emissions (dependent variable) and GDP (independent variable) using two different

functional forms. Create a single graph in which you compare the different predictions about CO_2 emissions, given levels of GDP. (Hint: For Question 4, if you use the "predict" command after the regression to create predicted values, you would be predicting values of log(CO_2), not values of CO_2. What would you need to do to convert these values back into the original units?)

Part V: Calculation of regression coefficient (9 points, 3 points each)

This problem requires you to use the gdp_co2_pc_2010.dta dataset that is in the Examples folder.

You are interested in the relationship between per capita CO_2 emissions and per capita GDP. Unfortunately, the "regress" command in Stata is not working, so you have to calculate the regression coefficient by hand.

1. Calculate the slope coefficient for the regression of per capita CO_2 emissions on per capita GDP, using the formula for a slope coefficient that uses variances and covariances. (Hint: you can add a suboption to the correlate command in Stata to produce a variance-covariance matrix.)
2. Interpret the coefficient in substantive terms. (Hint: It might help to interpret your coefficient if you express per capita GDP in terms of thousands of dollars.)