17.871, Political Science Lab
Spring 2009
Problem set # 2

Handed out: February 25
Due: March 4

Submit a paper copy of your answers to each question, including code for each command, your scatter plots, and calculations. Work in groups. Write up separately.

Part I: Golf putting data and regression (one point each)

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. As in the previous problem set, create a scatter plot of success rate (y-axis) by distance (x-axis), but this time ad the regression line (see lecture slides for code).

Part II: Scatter plots and ecological inference (three points each)

Using the **CCES.dta** file in the Examples folder of the course locker. The CCES is a survey about the 2006 election conducted in 2006 by professor Ansolabehere. 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.
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-access. 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 the within region level? (If it doesn't, you have made a mistake.) Briefly suggest an explanation for any difference.

Part III: Interpreting regression coefficients (two points each)

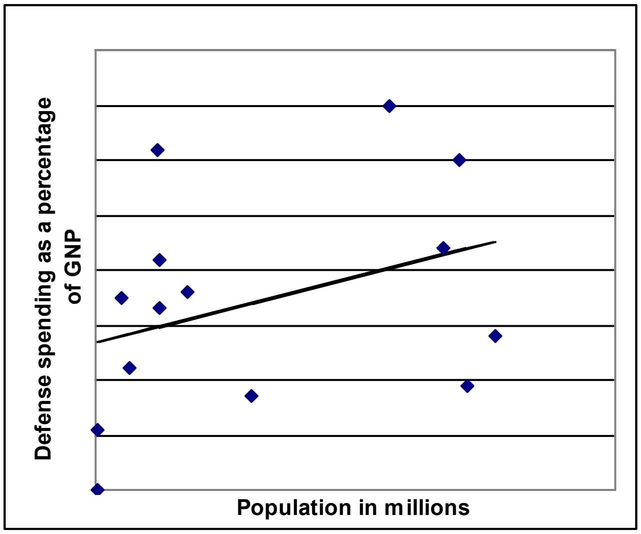
Using the data set **quartet.dta** in the Examples folder of the course locker:

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 table should contain the coefficients, the p values, and the standard error of regression (Root MSE).
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) from this example?

Part IV: Interpreting regression coefficients (two points each)

The following table and figure shows the population (in millions) and defense expenditures (as a percentage of GNP) for 14 NATO allies in 1981.

		Defense
	Population	**Expenditures**
Country	(millions)	(% of GNP)
Iceland	0.2	0
Italy	57.2	1.9
W. Ger	61.4	2.8
Canada	23.9	1.7
Luxembourg	0.4	1.1
Denmark	5.1	2.2
France	53.5	4.4
Belgium	9.9	3.3
Netherlands	14.1	3.6
Norway	4.1	3.5
Portugal	9.9	4.2
UK	56	6
Turkey	45.2	7
Greece	9.5	6.2
Means	25.03	3.42
Variance	570.18	4.08
Covariance	16.92	



1. Based on the scatter plot, which is the dependent variable and which is the explanatory variable?
2. Calculate the correlation coefficient by hand.
3. Calculate the least squares regression coefficient by hand. Interpret it.
4. In his theory of collective action, Mancur Olson hypothesized that bigger partners in alliances will tend to bear more than their proportional share of total costs. What light, if any, do these data shed on his hypothesis?
5. Which of the 14 countries included in the table had higher defense expenditures than would have been expected given their populations? What factors might account for these anomalies?
6. In 1981, United States had a population of 230.1 million. What is the expected level of defense expenditures corresponding to that population, given the slope and intercept of the least squares regression line? (To answer this, you need to know the constant: 2.7). Is this a reasonable use of the regression model? Why or why not?
7. The Standard Error of Regression (SER) is 1.97. Interpret.

Part V: Interpreting and comparing regression coefficients (two points each except the last)

Using the **CCES.dta** file in the Examples folder of the course locker:

1. Recode the housevote variable so that it equals 1 for the Republican candidate and 0 for the Democratic candidate.
2. Regress housevote on partisan identification. Before doing so, recode party identification to vary between 0 and 1. Interpret the effect of party identification and interpret its confidence interval.
3. Regress housevote on Bush approval (gwbapp). Before doing so, recode Bush approval to vary between 0 and 1. Interpret the effect of Bush approval and its confidence interval.
4. (four points) Does party identification or Bush approval have the largest effect on votes for US House of Representatives? How sure are you that these effects are causal? Finally, which is most vulnerable to an alternative explanation?