POLS/CSSS 503:
Advanced Quantitative Political Methodology

Problem Set 2

Professor: Chris Adolph, Political Science and CSSS

Spring Quarter 2014

Due in section, 25 April 2014

General instructions for homeworks: Homework can be handwritten or typed. For any exercises done with R or other statistical packages, you should attach all code you have written and all (interesting) output. Materials should be stapled together in order by problem. The most readable and elegant format for homework answers incorporates student comments, code, output, and graphics into a seamless narrative, as one would see in a textbook.

Problem 1: Running Regressions

This problem uses `sprinters.csv`, which contains the winning times from the 100 meter sprint in Olympic competitions going back to 1900. *Source:* A. J. Tatem, C. A. Guerra, P. M. Atkinson and S. I. Hay, *Nature* Vol. 431, p. 525 (2004).

Variable	Description
`finish`	best time in seconds in the 100 meter sprint
`year`	the year of the competition
`women`	1 = the women's best time time, 0 = the men's best time

a. In R, Create a matrix $\mathbf{X}$ comprised of three columns: a column of ones, a column made of the variable `year`, and a column made up of the variable `women`. Create a matrix $\mathbf{y}$ comprised of a single column, made up of the variable `finish`. Now compute the following using R's matrix commands (note that you will need to use the matrix multiplication operator `%*%`):

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Report the result of this calculation.

b. Using `lm()`, run a regression of `finish` on `year` and `women`. Compare your results to the calculation you did in part **a.**

c. Make a nice plot summarizing this regression. On a single graph, plot the data and the regression line.[1] Make sure the graph is labeled nicely, so that anyone who does not know your variable names could still read it.

d. Rerun the regression, adding an interaction between `women` and `year`.

e. Redo the plot with new fit, one for each level of `women`. Make sure that we can see which points are female finishing times, and which are male.

f. Suppose that an Olympics had been held in 2001. Use the `predict()` command to calculate the expected finishing time for men and for women. Be sure to calculate 95% confidence intervals.

g. The authors of the *Nature* article were interested in predicting the finishing times for the 2156 Olympics. Use `predict()` to do so, for both men and women, and report 95% confidence intervals for your results.

h. Do you trust the model's predictions? Is there reason to trust the 2001 prediction more than the 2156 prediction? Is any assumption of the model being abused or overworked to make this prediction? *Hint:* Try predicting the finishing times in the year 3000.

1 Your R code should run without "help"; e.g., you shouldn't need to copy and paste any results from `summary()` or elsewhere in order to plot the regression line.

Problem 2: Project Checkpoint

a. Identify the dependent variable for your final project, and your collaborators, if any.

b. Indicate how you plan to obtain the data.

c. Describe, in as precise terms as possible, the distribution of the data. If you have the data in hand, a histogram would be ideal; if you do not, give a verbal description of what you expect the distribution to look like. Be sure to indicate if the data are continuous or categorical.

d. What challenges would your data pose for analysis by least squares regression? Be sure to discuss any potential violations of the assumptions of the Gauss-Markov theorem, as well as any other complications or difficulties you see in modeling your data.