Political Science 100a/200a
Fall 2001
Problem Set 4

1. This question asks you to test Stata's random number generator.

   (a) First draw 100 observations from a uniform distribution on [0, 1], and store the results in a variable called x. Then create a new variable y that equals 1 when $x < .2$, 2 when $x \in [.2, .4)$, 3 when $x \in [.4, .6)$, 4 when $x \in [.6, .8)$ and 5 when $x \in [.8, 1]$. Now tabulate y, and figure out how to use a χ^2 test to see if you can reject the null hypothesis that your sample comes from a uniform distribution. (Note that it will be much easier to figure out how to do the χ^2 test by hand here than to figure how to get Stata to do it for you.)

   (b) Suppose you had drawn 1000 observations instead of 100. Would you expect to reach a different conclusion?

2. *School spirit.* You want to test the hypothesis that undergrads at Stanford have more school spirit than do grad students. You propose to measure school spirit by asking students to estimate the likely difference between Stanford's and Cal's scores for the Big (football) Game, on the conjecture that greater school spirit will lead students to increase their estimate of Stanford's performance versus Cal's. You collect data on the 19 students in an intermediate social science statistics class who respond to a request for an estimate prior to the game. The data is summarized below.

Student #	Estimate	Grad Student (= 1)
1	17	0
2	22	0
3	17	0
4	21	0
5	10	0
6	24	0
7	30	0
8	28	0
9	46	0
10	35	1
11	-6	1
12	24	1
13	20	1
14	14	1
15	7	1
16	7	1
17	20	1
18	35	1
19	28	1

   (a) Conduct a difference of means test to see if you can reject the null hypothesis that there is no difference in school spirit across students types, by this measure, against the alternative hypothesis that undergrads have more school spirit. (You may use Stata to help with computing means and standard deviations, but show the algebra involved in computing and evaluating the relevant test statistic.)

(b) Now approach the same problem using bivariate regression, computing the least squares estimates for α and β in the model $ScoreDiff_i = \alpha + \beta * Grad + \epsilon_i$. Also compute the standard error estimated for β, the relevant test statistic and its p value. Interpret the results, and the substantive meaning of your estimates a and b.

(c) A friend tells you that even if your estimate b is not "statistically significant," it is still "unbiased." Explain what this means, and whether and how the observation is relevant in this context.

(d) Very briefly discuss all the reasons you can think of that your estimate b might differ from the true β, whether due to the research design, the statistical analysis, or specific features of the data (e.g., check for outliers).

3. Suppose you are doing research on deadly communal riots in India. You develop the hypothesis that the competence and integrity of city and local police forces varies across Indian states (India is a federation of states). Your theoretical argument suggests that larger states (by population) will on average have more trouble with highly politicized police forces, which you expect to be associated with more deadly episodes of rioting. You collect some preliminary data on deaths per capita by Indian state, and also on their population sizes, along with a few other variables. Download the data set **riots.dta** to see what "you" got. (In fact, this is data taken from a paper by Steve Wilkinson, an assistant professor in political science at Duke University. The "hypothesis" proposed here is my invention; Wilkinson's paper has a more sophisticated an plausible theory about party competition and communal riots.)

(a) Produce an informative looking scatterplot of the dependent and independent variables (**dthcap** is deaths from riots per million for 1955-1995, I believe; **pop** is state population in millions). What do you notice and what does your visual inspection suggest to you?

(b) Regress **dthcap** on **pop**. Interpret the Stata output (estimated coefficients, s.e.s, t stats, p values, confidence intervals, and the RMSE). Give substantive interpretations where possible (i.e., don't just say "the coefficient is x"; explain what this means, and explain so that the substantive size of the estimated effect of a change in the independent variable is clear).

(c) Use Stata to generate the predicted values, the $\hat{y}_i$'s, and produce a graph that plots the data points and the regression line. Use State to produce the estimated residuals, and do an analysis of the residuals.

(d) You will most likely find that there is an influential outlier. Redo the whole analysis above after dropping the outlier from the data (either by **drop in [observation #]** or by adding **... if caseno = ...** statements as appropriate).

(e) Use the **corr dthcap pop,covariance** command to get the information you need to check to see that Stata is properly estimating b and $se(b)$. Do this check.

(f) It occurs to you that a possible problem with your hypothesis is that the percentage of Muslims by state may be systematically related to state size, and that more riots and more deaths are more likely where the Muslim percentage is higher. Assess whether this is even potentially a problem by looking at the relationship between **muslim** and **pop**. Then do a regression analysis of the relationship between **dthcap** and **muslim**. Finally, use multiple regression to evaluate if omitting **muslim** is giving you a biased estimate of the effect of state size. (Throughout this part, drop the outlier.) What does this suggest, substantively?

(g) If you were actually doing this research, what do you think your next move would be? (This question doesn't have any one right answer.)