GENERAL COMMENTS PERTAINING TO ASSIGNMENT #6

1. HOW TO WRITE YOUR REGRESSION EQUATION LIKE A PRO

First, use equation editor so that it looks professional. Second, decide whether to use generic Y and X or variable names that are more descriptive of your specific regression. Options:

(1) $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$

(2) $SAT_i = \beta_0 + \beta_1 GPA_i + \beta_2 FamIncome_i + \beta_3 PublicHS_i + \varepsilon_i$

With either equation (1) or (2) you must define both the unit of observation (what does *i* index?) and each of your variables (making it clear what units they are measured in). I find it easiest to do this in a table if there are more than 5 variables (see the SAT exercise for an example of this). If you have a lot of explanatory variables (say, 10+), it is probably easier to lump your variables into categories and write your regression equation slightly differently:

(3) $Y_i = \alpha X_i + \beta Z_i + \varepsilon_i$

In equation (3) there are two categories of variables, the X's and the Z's. Perhaps X contains 10 variables that measure the attributes of student *i* while Z contains 6 variables that measure the attributes of student *i*'s high school. Rather than list all of them, this specification can be combined with a table of the variables in X and Z, and readers will realize that each X variable will have a different α while each Z variable will have a different β parameter estimate.

Other things to note:

a. Subscripts indicate which variables change over your unit of analysis. If you have panel data (like states over time), you need two subscripts (i.e., *s* to signify state and *t* to signify time… Y_{st}).
b. A regression equation needs an error term otherwise it would be deterministic.
c. In equation (2), the names that I have chosen for the variables are shortened, but there is still some possibility that people will know what they are just by looking at them.
d. Number your equations and refer to them by those numbers in the text. Don't refer to "regression (1)", refer to "running a regression on equation (1)" or "estimating the parameters in equation (1) using regression."

2. TABLES & FIGURES

PLACEMENT: Please place the discussion/analysis of a table or chart ***before*** the table or chart appears in the paper. Doing the opposite means that your reader encounters a table or chart, feels confused, and then goes into your description feeling confused already. Additionally, scatter plots, histograms, and other charts are more simplistic than a regression and should come before your regressions in the paper. Think of a scatter plot as ***describing*** the data and the regression as actually ***analyzing*** the data.

REFERENCING & LABELING: Don't forget to number all tables and figures and then reference them by number in the text. All tables and figures should have a number and a descriptive title at the top. All table column headings must contain clear labels. Tables of regression results should indicate (how about in the title!?) what the dependent variable is. Use complete variable names in your tables and figures where possible (i.e., why use 'FamIncome' when 'Real Family Income ($000s)' fits?).

DO NOT CUT & PASTE REGRESSION OUTPUT FROM EXCEL: If you are only going to talk about the coefficient estimates and the t-statistics, why do you presume that your reader wants to see all the other crap? Additionally, Excel's output comes with your cryptic variable names, which breaks the rule

described above. Create a table in Word (Insert menu) and enter just the information that your reader needs. The only other things I asked for are the number of observations and the adjusted R^2. You might also consider putting the F-statistic significance level if you want to discuss the joint significance of your full set of explanatory variables (another measure of overall model fit).

3. BE CLEAR ABOUT HOW TO INTERPRET YOUR $\beta's$. DON'T WIMP OUT.

If you run a regression of $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$, then β_1 tells you something very specific about the relationship between X_1 and Y. Your interpretation depends on whether X_1 is a dummy variable or not.

a. If X_1 is a dummy variable, like $male_i$, then being a male is associated with a β_1 increase in Y relative to being female (note that the interpretation is also relative to the reference group, which is captured in β_0).
b. If X_1 is not a dummy variable, like $income_i$, then a 1-unit increase in X_1 ($1 or $1000 depending on your units) is associated with a β_1 increase in Y.

There is a big difference between 1 percentage point and 1 percent. The current unemployment rate is 8.5 percent. If it increases by "1 percentage point", it would go up to 9.6 percent. If it increases by "1 percent", it would go up to 8.585 (because 1% of 8.5 is .085). Nearly all of you confuse these two. If your variables are in percent form (like unemployment rates or percent changes that you calculated), then a 1-unit change is a "1 percentage *point*" change.

Note that any discussion of your parameter estimates should be accompanied by a discussion of statistical significance.

Also note that most of you are not uncovering ***causal*** relationships between Xs and Y, you are uncovering ***correlations***. This means that you should be very careful about your word choice in interpreting results. For example, causality is implied in a statement like, "A $1 increase in family income increases Y by 45.33." I like "is associated with" (as used above).

4. OTHER DETAILS:

a. I don't want to see more than three decimal places unless you have a very good reason. Who wants to look at a table full of numbers like 54.128546665889 when 54.13 would really do the job? Note that the decimal warning pertains to figures, too (why have 1.00, 2.00, 3.00 on your axes?). Also, if you clean up the number of decimal places and then right-align the contents of the cells, your tables will appear neat and tidy!
b. Spell out "percent" rather than utilize "%". Similarly, spell out any number ten or lower.
c. Many of you have a penchant for randomly capitalized letters, particularly when referring to variable names in your text. Just because one of your tables include "Family Income" doesn't mean that it is appropriate to talk about average **F**amily **I**ncome of $50,000. Basic sentence rules still apply.
d. Most of you are still slipping into past or future tense. It is ***not*** appropriate to say:
   - I ***ran*** a regression… (instead: A regression of Y on X reveals that…)
   - I ***looked at*** a scatter plot of… (instead: In Figure 1 I examine a scatter plot of…)
   - I ***will use*** my t-statistics to test the null hypothesis… (instead: I use t-statistics to test…)
e. **One final warning about spell-check and grammar-check before I go all postal worker on you. Use them.**