GENERAL COMMENTS ON ANALYSIS & REGRESSIONS
(YES, this information pertains to YOU!)

1. Dummy variable issues
   (a) If you have N categories, you can only include N-1 dummy variables (e.g., if race/ethnicity categories are white, black, Hispanic, Asian, and other, only 4 of these 5 dummy variables get included in the regression).
      i. This is true for racial/ethnic proportions with state-level data, too! You can't include % white, % black, % Hispanic, % Asian, and % other… one has to be left out.
   (b) If white is the excluded dummy variable and the coefficient estimate on the Asian dummy variable is 0.14, then the correct interpretation is: being Asian is associated with a 0.14 unit higher value of Y than being white on average and holding everything else constant.
      i. With state-level proportions data, if the % white is left out and the coefficient estimate on % Asian is 0.14, then the correct interpretation is: a 1 percentage point increase in the proportion Asian relative to the proportion white is associated with a 0.14 unit higher value of Y.

2. Scatter plots are descriptive, while regression results are actually used to analyze the data and test your hypotheses. Thus, scatter plots should always be presented and discussed before regression results.
   (a) The only exception to this rule is if you get an unanticipated result in your regression and then want to take a closer look at one particular variable.

3. You need to be able to interpret both the SIGN AND MAGNITUDE of each coefficient estimate in your regression results table. Most of you wimped out.
   (a) If the t-stat and/or p-value indicates that a coefficient estimate is not statistically different from zero, then you CANNOT interpret it as being positive or negative… IT IS ZERO!
   (b) If you do not have information about the units for each explanatory variable, it will be impossible to interpret your coefficient estimates correctly. For example, if a variable in your table is labeled "INCOME", it is necessary to clarify whether this is in dollars, thousands of dollars, real or nominal, per capita or not. These details should, at a minimum, be clearly provided in the table where you first define your variables and provide summary statistics.

4. The interpretation of coefficient estimates is different if you have state-level data than if you have individual-level data.
   (a) Coefficient estimate on male dummy variable is -0.56 and you have individual-level data: males have a 0.56 unit lower value of Y than females on average and holding everything else constant.
   (b) Coefficient estimate on the proportion male in a state is -0.56 and you have state-level data: a 1 percentage point increase in the male population proportion in a state is associated with a 0.56 unit decrease in Y, on average and all else constant.

5. If you have a coefficient estimate that is very small, you should rescale that variable before including it in the regression.
   (a) If income is one of your explanatory variables and the coefficient estimate is 0.00004, this is the effect on Y of a $1 increase in income. If you divide your income variable by 1000 and run the regression with this transformed income measure, the coefficient estimate will now be 0.04 and it is the effect on Y of a $1000 increase in income.

6. Tables are numbered (Table 1, Table 2, Table 3, …), Graphs are also numbered (Figure 1, Figure 2, Figure 3, …) and both have titles that convey clear information about what they contain (for example, it's nice to make sure the reader knows what the dependent variable is in a table of regression results).
   (a) Information about the sample size and the adjusted R-squared belong at the bottom of the table.

And now, here is some information that you haven't previously heard from me regarding problems that your regressions likely have:

1. **Omitted Variables Bias**
   If many of your coefficient estimates are statistically significant, but your adjusted R-squared is pretty low, then you are missing important variables that should be included to help explain variation in Y.

   If an omitted variable is correlated with one of the explanatory variables you *do* include, then the coefficient estimate on the included variable will be biased.

   For example, if Y=college GPA and you do not have some measure of pre-collegiate academic ability (like high school GPA or SAT score) included as an explanatory variable, the effect of pre-collegiate academic ability will bias other included variables that you *do* include (like race). Underrepresented minorities have lower measured academic ability, so if you don't include a variable like high school GPA in your regression, you will get larger negative coefficient estimates on black and Hispanic dummy variables because of the negative correlation between belonging to these race/ethnicity categories and measured academic ability.

   The solution: Include as many explanatory variables as you think belong in your regression. If important variables are unavailable, you have to at least think about and acknowledge the way in which their omission affects your regression results.

2. **Multicollinearity**
   If your adjusted R-squared is quite high but many of your coefficient estimates are statistically insignificant, you are probably including some variables that are too highly correlated with one another.

   Check the correlation between variables that you think may measure roughly the same thing. If the correlation is high, drop one variable. (The command in Stata is "corr"; in Excel, there is a Correlation option under Data Analysis.)

   As an extreme example, you cannot include *BOTH* GDP growth rate and GDP per capita as explanatory variables.

   The solution: Drop variables that are redundant.

3. **Heteroskedasticity and/or Serial Autocorrelation**
   These are problems with the residuals (the difference between actual and predicted Y values). The Ordinary Least Squares regression model that all of you are using assumes that errors are independent of one another across observations and over time (e.g. large values of Y are no more likely to have large errors than small values of Y → homoskedasticity; an error in one month is no more likely to be followed by a positive error than a negative error → serially uncorrelated).

   Heteroskedasticity: Assume that it is there and correct for it. This is only done easily in Stata… simple add ", robust" to the end of your "regress" command. Another option, available in Excel, is to transform the dependent variable by taking the natural log [new Y is defined as ln(Y)]. The ln(Y) is a non-linear function and so it may fit the data better.

   Autocorrelation: Test for it using a Durbin-Watson statistic in Stata and then transform the dependent variable by taking natural logs and then first differences [new Y is defined as $\ln(Y_t) - \ln(Y_{t-1})$].