

This assignment gives you a chance to demonstrate your mastery of multiple regression techniques. Choose a single dependent variable and some independent variables, presumably on the basis of some substantive theory or argument that you can sketch briefly. Estimate the relationships with OLS. The details are up to you--how many variables, what kind of model, etc., but you should try not to get in over your head. Don't throw every variable in "just for the heck of it." The emphasis is on specifying an interesting problem and correctly working through it. Make sure at least one of your independent variables is a dummy (dichotomous) variable. And choose a dependent variable that you can treat as if it were continuous.

Write 5-10 pages that do the same things as assignment 1. Use the same outline. Describe your research questions, the data, and your objectives. Present the estimates and interpret them to demonstrate your proficiency in multiple regression analysis. Translate your outputs into a professionally acceptable form. Explain what the estimated coefficients mean, the standard errors, the  $R^2$ , and whatever else you find to be important.

It is always great to try to make plots that illustrate the data and the results. With a little effort, you might be able to create interesting illustrations.

In addition to your presentation of the model itself, please report on the following diagnostic components.

1. Check for multicollinearity. Conduct whatever diagnostic tests you think are appropriate. Discuss your diagnostic efforts, report whether or not there is multicollinearity in your model and discuss it.
2. Test a hypothesis of the sort  $b_3=b_4$ , meaning that two parameters are equal. Write out the null hypothesis and the alternative. Explain why this could be an important question and outline the testing procedure. To do this, you MUST obtain an estimate of the variance-covariance matrix of the coefficient estimates. In R, it is as simple as fitting a model "mod1" and then running the command `vcov(mod1)`. (Note, by the way, that the square roots of the diagonal elements of the Var-Covar matrix are the standard errors that the computer prints out for the estimates of the  $b$ 's. If they aren't, you've calculated the Var-Covar matrix incorrectly.)
3. You MUST conduct some kind of Chow or "F-test" to find out if a subset of the model parameters is likely to be zero (simultaneously). The null is  $b_3=b_4=b_5=0$ , for example. It is not necessary to give a formal proof of the process that leads up to the F test, but you can (at least) give an explanation of the test **in your own words**. In R, you can estimate the restricted and unrestricted models and then use the `anova()` command to find out if the difference between them is statistically significant. The `anova.lm` help page will have information on how you can make sure it conducts an F test, as opposed to a likelihood ratio or some other kind of test.
4. Conduct any kind of check for heteroskedasticity. Explain the justification your test and its results.
5. Calculate the "heteroskedasticity-corrected standard errors" for your parameter estimates. Compare against the ordinary standard errors.

The exercise outlined is absolutely fundamental to your mastery of multiple regression. Hence you should emphasize these methodological aspects. Don't spread your effort too thin by reporting lots of estimated equations or trying to provide a comprehensive picture of all the relationships that exist in a dataset.