Homework #5
***** Due Tuesday, March 19, 2013 at the beginning of lecture (9:30 a.m.)*****

You will have lab problems and written problems for this assignment.

You will be using the data set `mlb1.dta` located at Beachboard to answer the Lab Problems. Please work on the STATA code during lab so you can ask questions of the lab instructor. For the lab problems, I want you to submit your log file and your **typed** answers in complete sentences. You do not need to submit your do file.

The answers to the written problems are to be neatly written out or typed. The answers to anything that asks you to explain should also be in complete sentences.

Please submit everything in one package in the order assigned in this homework.

Lab Problems

L1. The dataset `mlb1.dta` contains data on 353 Major League Baseball players. This question looks at the determinants of their salaries. Create a do file `mlblab.do` and a log file `mlblab.log` for this problem. Submit the log file with your answers.

- **a.** Run a regression where the dependent variable is `salary` and the explanatory variables are `teamsal pcinc yrsallst years hits runsyr.` In other words, the total team salary, the average income of the city where the team is located, the number of years on the All-Star team, the number of year the player has been in the League, the number of all-time hits (over all years), and runs per year.
- **b.** Look at the simple correlation between the variables in your model, using `correlate`.
- **c.** Test for multicollinearity by using the `vif` command.
- **d.** Look at your regression results, the test for multicollinearity, and the test for correlation. Is there a problem with multicollinearity in the model? Look at the variables in the model. If you believe there is a problem with multicollinearity, explain why you might have that problem.
- **e.** Now, use the `regression` command to estimate a linear regression model where the dependent variable is `salary` and the explanatory variables are `teamsal pcinc allstar slugavg fldperc frstbase scndbase shrtstop thrdbase outfield.` In this new model, the explanatory variables are team salary, per capita income of the city where the team is located, percentage of years as an All-Star, career slugging average, career fielding percentage, and dummy variables for the various positions on the team, with catcher excluded.
- **f.** Test for multicollinearity again for the model in part e.
- **g.** Is multicollinearity a problem for the model in part e. Why do you think this result is different from before?

h. Now, predict the residuals for each observation in the model from part e.

i. Create a new variable `resid_2` using the `generate` command.

j. Regress `resid_2` on the explanatory variables (not `salary`) from part f.

k. Run the regression from part e again.

l. Use the standard STATA Breusch-Pagan test for heteroskedasticity (`estat hettest`).

m. Run the standard STATA White's test for heteroskedasticity (`imtest, white`).

n. For now, assume there is a problem with heteroskedasticity in the model from part e, estimate the model again with robust standard errors.

o. Pick a factor that you believe is missing from the regression equation in part e. Add that variable to the regression equation for `salary` using STATA. Test for multicollinearity and heteroskedasticity (using the standard STATA Breusch-Pagan test) in your model. Rerun your model to account for heteroskedasticity, if appropriate.

p. Why did you decide to include the variable you chose for your model in part o?

q. Did your estimate and statistical significance for `allstar` (percent of years as an All-Star) change from the model in part e to the final estimation model in part o (controlling for heteroskedasticity, if appropriate)? Explain why you might have a difference.

Questions

Q1. Which of the following factors can cause OLS estimators to be biased? Explain.

a. Heteroskedasticity.
b. Omitting an important variable.
c. A sample correlation coefficient of 0.95 between two independent variables both included in the model.

Q2. Which of the following factors will lower the variance of an OLS estimator. Explain.

a. Increasing the number of observations.
b. Increasing the number of explanatory variables that reduce the error variance in the model.
c. Minimizing the multicollinearity in the model.

Q3. Use your results from L1 for this problem.

a. Use your results from part j of L1. Calculate the Chi-square statistic for the Breusch-Pagan test. Since this model has 10 explanatory variables, this statistic has 10 degrees of freedom and the critical Chi-Square value is 18.31. Write down the null and alternative hypothesis for the Breusch-Pagan test and whether or not you reject the null (or fail to reject the null) based on your results.

b. Is your conclusion about heteroskedasticity in part a of this problem the same as what you found using the standard STATA Breusch-Pagan test (L1, part l)? I know the values will be different, but did you come to the same conclusion regarding the errors?
c. Based on your answer to part a of this problem and the results of the STATA Breusch-Pagan and White's test in L1 (parts l and m), which model (L1 part e or L1 part n) is consistent with your results?
d. Using the best model (you identified in part c) interpret the coefficient on team salary (`teamsal`).
e. Are there any variables missing from this model that might be biasing the results? Which ones? Explain.