

636, Fall 2010

Homework 6 and 7

Data Generation, Binary Terms, Multicollinearity, Chow Tests, etc.

Due **Thursday: November 4** before class begins. Note, put all regressions sequentially in an appendix and make sure they are carefully labeled. While you are free to copy and paste any parts of your STATA results, all answers should be on separate sheets of paper: I should be able to quickly score homeworks with the correct regressions. Unless your regressions are wrong, I should not have to sift and winnow through mountains of output to find out what you want me to look at. I want you to check with at least one other person to verify your regressions.

Question 1

The model you are about to run is an example of how data could (in theory) be *generated*. You should see that the data you collect, whether it be from a survey, a retail scanner device, a government web site, etc., is indeed generated by the underlying beliefs and decisions of the economic agents as well as the reactions to events that frame and perhaps alter behavior. In this example, we will generate 1000 observations representing 1000 weeks of actual production data aggregated for a small industry.

Here are the steps:

1. Open Stata
2. Type in

```
. set obs 1000

. set seed 77
. gen RN1= -3+(3+3)*runiform()
. gen RN2= -3+(3+3)*runiform()
. gen RNY= 0.9+(1.1-0.9)*runiform()
```
3. Calculate

```
gen X2=10+RN1
gen X3=30+RN2
gen Y=1.5*(X2^0.6)*(X3^0.4)+RNY
```

Note: you just generate data based on a Cobb-Douglass production function given by $Y_i = AX_{2i}^{\beta_2} X_{3i}^{\beta_3} + RN_i$, where $A=1.5$ is the technology scale value, RN_1 , RN_2 , and RNY add randomness to the data and $\beta_2 = .6$ and $\beta_3 = .4$ are the supply elasticities.

4. Calculate

```
. clear
gen lnY=log(Y)
gen lnX2=log(X2)
gen lnX3=log(X3)
```

reg lnY lnX2 lnX3 in 351/1000

Now you are ready to conduct some analysis.

- A. Estimate the model $Y_i = AX_{2i}^{\beta_2} X_{3i}^{\beta_3} \exp^{e_i}$ using all your data.
- B. Do a chow test using n1=350 and n2=650.
- C. Suppose now that in the first 350 weeks, firms operated under the A=1.5 technology while in weeks 351-1000, firms used a technology that led to 20% more output when using the same inputs (i.e. A=1.8 for i>350). Change your output vector accordingly and rerun the three regressions that will give you an appropriate chow test. Discuss your results.
- D. Redo part C using a dummy variable structure that tests for a structural change. Discuss your result.
- E. Explain using Figure 1.1 from page 2 of our notes what it means to think about the “data generation process.” (yes, I know, this is the Stiggum’s figure!)

Question 2

(Data from CD accompanying Schmidt, *Econometrics*, 2005, McGraw-Hill) “In the U.S. men earn substantially more money in labor markets than do women. This may be evidence of discrimination, or it may have other causes, such as differences in educational levels between men and women. In this problem, we’ll look at data on labor market earnings of men and women, taken from the U.S. Census Bureau’s Current Population Survey...” pg 214.

Load the ds67 file from the AAE636 web page, which contains two sheets of data. The data from the first sheet are:

- Earnings: total earned income
 - WageSalary: income from work
 - Female (F), Africamer (A), and Hispanic (H) are binary variables (=1 if yes, =0 if not). Caucasian/Asian (CA) and Male (M) is the category if all binaries are zero.
 - Age and Hours are the workers age and the number of hours worked.
 - Education is the number of years of education.
 - (data implies that earnings (i.e. wagesalary) is a proxy for all income. We will use **wagesalary** as our dependent variable. Using the “drop” command or the “if” command, you will always run regressions with only positive wagesalary values.
1. Run an ANOVA regression (call it model X) just to test a null hypothesis about the first sentence in the above quote. Do not control for race in this model: $\text{wagesalary} = b_1 + b_2 * F$. Interpret the regression
 2. Run a regression adding education and age to model X. This is called Model Y. Are these two variables jointly different from zero?

3. Add an interaction term to Model Y. The term is $F \cdot \text{education}$. At the mean education level, run a test if gender affects wages (hint: what is the marginal effect of 'female' on wagesalary?).
4. Add to model Y a dummy variable structure much like equation 7.12 in our notes to capture all gender and race effects. Note that you have $j=6$ classifications (F-C/A, M-C/A, F-H, M-H, F-A, M-A), thus you will need $j-1=5$ dummy variables. Determine if gender matters within each race category. Determine if race matters in each gender category.

Question 3

The second sheet contains an exercise evaluating a model for multicollinearity in the X matrix and begins the step of developing a model along theoretical lines.

- A. Estimate the log-log model: $\ln Y = f(\ln X_2, \dots, \ln X_6)$. Is this a good model? Explain Briefly.
- B. Run the auxiliary regressions to get the VIF for the coefficient on $\ln X_2$. Determine if multicollinearity is severe. Interpret.
- C. Run . estate VIF, and check all the VIF's. Interpret.
- D. Run the Haitovsky test using the critical Chi-square level for $\alpha=.10$. Interpret.

Question 4 (not Graded but used in future classes). No need to turn in.

Lets try and do a minimal but full battery of tests that one could envision about a regression. These relate to normality of the error, multicollinearity, heteroskedasticity, autocorrelation, and model specification. These are all pretests in the sense that we are seeing if the selected OLS model is reasonable.

bring in some data off the web by typing:

. use <http://fmwww.bc.edu/ec-p/data/wooldridge/CEOSAL1.dta>

Salary=CEO Salaries

Sales=Company sales

ROE=return on equity

ROS=return on sales

Note : prefix (pc) implies percent change; prefix (l) implies log transformation

. regress salary sales roe pcroe indus finance consprod utility

1. Why do you suppose Stata dropped one of the regressors? Show how you might verify your answer.

2. let run some tests for normality of the errors: `. swilks` and `. sfrancia`
3. get a graph of the **residuals versus fitted plot**: `rvfplot`
4. get your variance inflation factors: `. estat vif`
5. run three versions of a test for heteroskedasticity:
 - a) `. estat hett`
 - b) `estat hett, rhs`
 - c) `.estat hett, iid`
6. run a test for autocorrelation: `. estat dwatson` (any idea why this won't work?)
7. run two tests for model specificity: `. estat ic`
8. run the linktest to evaluate model specificity: `. linktest`
9. run the RESET test to evaluate model specificity: `estat ovtest`

In ½ page or less (double spaced) try and make some sense about the overall quality of this regression. Remember the assumptions by which the parameter have arrived.

Question 4 is not graded. We will go over this in class.