

**University of California, Santa Barbara**  
**Economics 594SE: Advanced Econometrics**  
**Assignment 1: OLS Assumptions; Introduction to Stata**  
*Due Monday, June 28, in class.*

1. You have data on output,  $Q_i$ , capital input,  $K_i$ , and labor input,  $L_i$ , for a sample of firms in an industry. You believe the production function of a firm in this industry takes the Cobb Douglas form,  $Q_i = A L_i^\alpha K_i^\beta \varepsilon_i$ , where  $i$  indexes firms and  $\varepsilon_i$  is an error term. Can this production function be estimated using the Classical Linear Regression Model? Why or why not? If yes, under what conditions, and are these conditions likely to be realistic, in your opinion?

2. You are interested in estimating the effect of years of completed schooling,  $S_i$ , on annual earnings,  $Y_i$ , in a sample of individuals,  $i$ . Suppose that the only source of error in the relationship between  $S_i$  and  $Y_i$  is due to reporting error (people's reported income deviates randomly from their true income). Suppose also that the standard deviation of the reporting error is strictly proportional to an individual's income (in other words, on average, high-earnings peoples' incomes are misreported by the same *percentage* as those of low-earnings people). Can the Classical Linear Regression Model be used to estimate the relationship between  $S_i$  and  $Y_i$  in this context? Why or why not? If yes, under what conditions, and are these conditions likely to be realistic, in your opinion?

3. Download the CPS Merged Outgoing Rotation Group (MORG) *Stata* data sets for 1998 and 2008 --morg98.dta and morg08.dta--, plus the data documentation, cps79\_06.doc, from the course [web site](#). The MORG provide earnings and work activity information for a representative sample of American households.

(a) Data Preparation. For each of these two years, create a data set with the following variables: earnings per week (in dollars; use *earnwke*); age (in years); years of education completed (use *ihigrdc*), and 0-1 indicator variables for being *female*, *black* and *hispanic*. (Note that the coding for race and hispanic (*ethnic*) changed between these years; you will need to choose a way to reconcile these so they match across surveys; it may not be perfect.) Restrict your data set to persons between the ages of 20 and 64 inclusive, and with weekly earnings between \$50 and \$2884 inclusive (why?). Retaining the variable *year* (already supplied in both data sets) to distinguish the two data sets, append the two data sets together, keeping only the observations with nonmissing weekly earnings. Convert the 1998 weekly earnings in your data to 2008 dollars (look up the CPI on the web), then take the log of real weekly earnings.

Now, create a variable for each person in your data equal to that person's year of birth (*yob*). Social scientists often refer to people born in the same year (or group of years) as a cohort, and argue that some things that cohorts share in common, such as the education systems they went through and the size of their cohort –e.g. the baby boom— could have important effects on their earnings. For now, we'll use the *yob* variable to capture cohort effects.

In addition to the effects of education, race, ethnicity and cohort, we are also interested in the effects of a person's age on his/her weekly earnings: peoples' earnings tend to rise with age, as they accumulate experience and skills. Finally, we note that the earnings of all workers tend to vary from year to year, due for example to business cycles such as the current recession or longer-term economic growth/decline. To capture these kinds of effects, create a dummy variable, *y2008*, that equals one for the observations from the 2008 data and 0 for 1998.

(b) To estimate all the above effects, run the following OLS regression:

$$\text{Log}(\textit{weekly earnings}) = a + b(\textit{age}) + c(\textit{education}) + d(\textit{female}) + f(\textit{black}) + g(\textit{hispanic}) + h(\textit{yob}) + k(\textit{y2008}) + e.$$

What happens when you ask Stata to run this regression? Explain in terms of the assumptions of the Classical Linear Regression Model.

(c) Now try this regression slightly differently. Instead of forcing cohort effects to be linear in the year of birth, assume it's a step function. Specifically, create an indicator variable, call it  $b_{30}$  (for the 1930s birth cohort) for whether the individual was born in the 1930s (equal to 1 if he/she was born in that decade, 0 otherwise); do the same for  $b_{40}$  through  $b_{80}$ . Now estimate:

$$\text{Log}(\textit{weekly earnings}) = a + b(\textit{age}) + c(\textit{education}) + d(\textit{female}) + f(\textit{black}) + g(\textit{hispanic}) + h^{40}b_{40} + h^{50}b_{50} + \dots + h^{80}b_{80} + k(\textit{y2008}) + e.$$

(Note that  $b_{30}$  is not included—why do we need to do this, and what does it imply?) What happens? Why?

(d) Next, instead of forcing the effect of age on earnings to be linear, let's represent age by a step function too. Specifically, generate a set of dummies for the following age intervals: 20-28, 29-38, 39-48, 49-58, and 59-64. Run the above regression replacing the linear age variable by the dummies for age groups 29-38 through 59-64. What happens and why?

(e) Finally, let's tweak the above regression just a little. Specifically, define the age categories as 20-29, 30-39, 40-49, 50-59, and 60-64 instead. Comment on how things change and why. Submit your Stata do- and log files with your assignment.

4. Download the Stata data set CharnessKuhn4AEASTP.dta from the course web site and skim the article from which it was drawn (also available on the web site). This data contains information from the *Public Wage Regime* in the experiment only, and on three variables: Type-1 (low ability) workers' effort ( $e1$ ), and the wages paid to both workers in the 'firm':  $w1$  and  $w2$ . Run an OLS regression of  $e1$  on  $w1$  and  $w2$  and interpret the result. Then, using the *outsheet* command, export the data as a simple, comma-delimited text file with three columns, one for each variable. Finally, write a .m file in MATLAB that represents this data as the following matrices:

$\mathbf{y}$  is the ( $n \times 1$ ) column vector of worker 1 effort ( $e1$ ); in this case you will have  $n = 555$ .

$\mathbf{X}$  is the ( $n \times K$ ) data matrix, including a constant term. In this case,  $K=3$  ( $w1$ ,  $w2$ , and a vector of ones are the three variables).

We will work with these matrices in MATLAB in the next Assignment. Submit your do files, Stata Output, and .m file.