ECON 272 - 01, 02

Homework 2

Due Date: Februrary 13

Please submit a do files for all the Stata questions in Q.3. A submission without a "do file" will receive an immediately penalty of one "check point".

This homework involves data analysis using data from the Annual Survey of Industries. This is a nation-wide survey conducted in India for manufacturing establishments. The unit of observation is the manufacturing establishment. The data pertains to the year 2014, and all values are in U.S. dollars. The hypothesis of interest is workers' contribution to manufacturing output. The last page of the homework lists the definition variables in the dataset. You might want to refer to this first and familiarize yourself with the data.

Q.1a) Consider the following univariate population regression

$$Output_i = \beta_0 + \beta_1 Worker_i + \epsilon_i$$

i denotes the manufacturing establishment or firm. *A priori*, would you hypothesize β_1 to be positive or negative? Provide a 1/2 sentence justification.

Q.1b) Estimate this regression in the data and report the additional output (in USD) a firm can expect if it hires 10 additional workers? What information is conveyed by the R^2 of this regression?

Q.1c) In the present dataset, $Corr(Output_i, Worker_i) = .4388$, while $sd(Worker_i) = 805$, where $Corr$ refers to the sample correlation, and sd the standard deviation. Suppose you test this relationship in another dataset where the correlation between the independent and dependent variable is unchanged, but the estimated β_1 now equals 6,000. What can you infer from this about the variance of the independent variable?

Q.1d) Now augment the population regression with the following "controls":

$$Output_i = \beta_0 + \beta_1 Worker_i + \beta_2 Capital_i + \beta_3 Materials_i + \beta_4 Importer_i + \beta_5 Rural_i + \beta_6 Listed_i + \epsilon_i$$

Conditional on these controls, report the additional output a firm can expect to receive if it hires 10 additional workers? Is this effect larger or smaller than what you found in Q.1b)?

Q.1e) Using the sample mean of the dependent and independent variables, reason whether workers have a "large" or "small" impact on manufacturing output?

Q.1f) Interpret the estimated coefficient corresponding to β_5.

Q.1g) How has the R^2 changed in the regression estimated in Q.1d), compared to the one estimated in Q.1b)? Based on how the β_1 coefficient changed between Q.1b) and Q.1d), does the R^2 provide you any information about your hypothesis of interest?

Q.2a) Create a new variable after computing the natural log of manufacturing output, and estimate the equation:

$$\ln(Output_i) = \beta_0 + \beta_1 Worker_i + \beta_2 Capital_i + \beta_3 Materials_i + \beta_4 Importer_i + \beta_5 Rural_i + \beta_6 Listed_i + \epsilon_i$$

Interpret the β_1 coefficient. How much is the additional output a firm can expect if it hires 10 additional workers?

Q.2b) Create new variables after computing the natural log of total workers, capital stock, and raw materials, and estimate the following equation:

$$\ln(Output_i) = \beta_0 + \beta_1 \ln(Worker_i) + \beta_2 \ln(Capital_i) + \beta_3 \ln(Materials_i) + \beta_4 Importer_i + \beta_5 Rural_i + \beta_6 Listed_i + \epsilon_i$$

Interpret the estimated β_1 and β_2 coefficients. Based on these coefficients, what do you think is more impactful for a firm planning to increase output – hiring more workers, or investing in additional capital?

Q.2c) Based on the estimated coefficient corresponding to β_4, would you say firms importing inputs have higher or lower output?

Q.2d) Interpret the coefficient corresponding to β_5. How has your interpretation changed, compared to your answer in Q.1f)?

Q.2e) The age of a firm has been shown to affect its performance. Augment the population

regression function with age variable and estimate

$$\ln(Output_i) = \beta_0 + \beta_1 \ln(Worker_i) + \beta_2 \ln(Capital_i) + \beta_3 \ln(Materials_i) + \beta_4 Importer_i + \beta_5 Rural_i + \beta_6 Listed_i + \beta_7 Age_i + \epsilon_i$$

Based on your estimate for β_7, how much more output is produced by a firm which is 5 years older?

Q.2f) Test for non-linearities in the impact of age on output by creating a new variable Age_i^2 and including it in the regression estimated in Q.2e). Based on the coefficients corresponding to Age_i and Age_i^2, what would you infer about the relationship between a firm's age and its output?

Q.2g) Create 4 binary variables (you can call them *age1*, *age2*, *age3* and *age4*). Define *age1* so that it equals 1 for firms aged less than 10 years; *age2* for firms aged between 10 and 25 years; *age3* for firms aged between 25 and 50 years; *age4* for firms aged over 50 years. Drop age and its squared from the regression, but include the *age* binaries and estimate the regression from Q.2f). For what age group do firms have the highest output? Interpret the coefficient corresponding to the *age2* dummy.

Q.3a) Consider the population regression function:

$$\ln(Output_i) = \beta_0 + \beta_1 \ln(Worker_i) + \beta_2 \ln(Capital_i) + \beta_3 \ln(Materials_i) + \beta_4 Importer_i + \beta_5 Rural_i + \beta_6 Listed_i + \epsilon_i$$

Estimate this regression, but don't include the $\ln(Worker_i)$ variable. Use the *predict* command in *Stata* to compute the residuals, and store the residuals as a separate variable. Report the sum of squared residuals from this regression.

Q.3b) Now, estimate the following regression:

$$\ln(Worker_i) = \gamma_0 + \gamma_1 \ln(Capital_i) + \gamma_2 \ln(Materials_i) + \gamma_3 Importer_i + \gamma_4 Rural_i + \gamma_5 Listed_i + \mu_i$$

Store the residuals from this regression as a separate variable.

Q.3c) Now, regress the residuals obtained in Q.3a) on the residuals obtained in Q.3b) and report the estimated β_1 coefficient. How does this coefficient compare to the β_1 coefficient

estimated in Q.2b? What information does your answer convey to you about the functioning of the OLS estimator and what it means to "control" for variables other than your main hypothesis of interest?
(Hint: think of what information is contained in the residuals)

Variable Definitions

Variable names in the Stata dataset are in parentheses

State (state): state in which the firm is located.
District (dist01): district in which the firm is located.
Rural (rural): binary variable equaling 1 if the firm is located in a rural area.
Total workers (nototalworker): number of workers hired by the firm.
Total output (total_output): total output produced by the firm in the year (in USD).
Total wages (tot_wage_final): total wages paid by firm to workers (in USD).
Capital (avg_nfa): total capital stock of the firm (in USD).
Importer (importer): binary variable equaling 1 if the firm imported any input during the year.
Raw Materials (avg_raw_mat): value of raw materials used by the firm during the year (in USD).
Listing status (listed): binary variable equaling 1 if the firm was publicly listed in the stock market.
State-owned firm (psu): binary variable equaling 1 if the firm was owned by the government.
Age (age): age of the firm in years.