

### **Applied Exercise #1**

– Due Thursday 02/21/08 –

The focus of this exercise is on the relationship between educational attainment and earnings. It examines statistics that summarize the relationship between two random variables, the bivariate linear regression model, and the multivariate linear regression model.

Feel free to work cooperatively and in groups. However, each student must hand in their own problem set using own words to explain the results. Try to summarize the answers/results concisely. STATA output may be attached to the end of the problem set for partial credit.

Data: The data is an extract of a 1992 survey of German workers. It comes from the paper by John DiNardo and Steven Pischke “The returns to computer use revisited: Have pencils changed the wage structure too?” (1997). For more information see the reference to the paper on the reader.

Access the data on bspace site for Econ C142: problem set 1/ restricted92.dta (STATA 9 format).

The data set has 20,042 observations and 19 variables.

The unit of observation is the individual.

The key variables for this exercise are:

- ed completed years of education
- exp and exp2 years of labor market experience and its square (experience<sup>2</sup>)
- female indicator variable equal to 1 if the person is female
- mar indicator variable equal to 1 if the person is married
- computer indicator variable equal to 1 if the individual uses a computer at work
- pencil indicator variable equal to 1 if the individual uses a pencil at work
- telefon indicator variable equal to 1 if the individual uses a telefon at work
- calc indicator variable equal to 1 if the individual uses a calculator at work
- hammer indicator variable equal to 1 if the individual uses a hammer at work
- occ 4-digit occupational codes
- lnw log of the hourly wage reported by the individual

Question 1: Summarize the earnings education relationship.

- a) Graph a scatter plot of log-hourly wages on the y-axis and education on the x-axis. From the plot, what is the likely sign of the covariance and correlation coefficient of log-wages and education? Comment on the linearity/non-linearity of the relationship.
- b) What are the formulae for the sample covariance, variance and correlation coefficient of two random variables X and Y? Calculate the sample covariance and the correlation coefficient for log-wages and education. Give an example of how the correlation between education and earnings may not be causal.
- c) Is the bivariate distribution a sensible functional form for the relationship between log-wages and education? Why or why not? Suppose that log-wages and education had a bivariate normal distribution, what five population parameters would fully describe the joint distribution? Under the assumption of joint normality what is the formula of the population mean of log-wages conditional on education in terms of the population parameters? Now substitute the sample “analogs” of the population parameters to calculate the slope coefficient of the conditional mean of log-wages. Under the assumption of joint normality, how does the conditional variance of log-wages vary with education? What is the “technical” term for this?

Question 2: Bivariate linear regression model for log-wages and education.

- a) Describe conditions under which a linear regression model of log-wages as the dependent variable and education as the independent variable will result in an unbiased estimate of the causal effect of education on log-wages. Under what assumptions will the OLS estimator be the best linear unbiased estimator? Describe briefly how the linear regression model presumes that the conditional expectation of log-wages (or the regression model) is linear in education.
- b) Derive the “first-order” conditions from the least squares minimization procedure for the estimators of the constant and slope parameters. Briefly describe their intuition. Describe three properties of the regression line.
- c) Now run the regression of log-wages on education and a constant. How do these estimates of the constant and slope compare to the result in 1c)? From the total sum of squares (SST), explained sum of squares (SSE), and residual sum of squares (SSR), derive the R-squared of the regression.

- d) How can the slope coefficient on education be interpreted as the percentage effect of an additional year of schooling on the wage? From the mean root squared error (MSE) derive an unbiased estimate of the variance of the residuals in the regression.
- e) Derive the 95% confidence interval for the return to education. Using a t-test, test the null-hypothesis that the return to education is zero and 10%, respectively. What is the p-value for the significance test on education (i.e. for  $H_0 : \beta = 0$ )? Use SSE and SSR to derive the F-statistic for testing the significance of education. How is this related to the t-statistic? Now derive the t-statistic and F-statistic for the education coefficient using the R-squared of the log-wage regression.

Question 3: Multivariate earnings regression model.

- a) Regress log-wages on a constant, education, experience, experience-squared, the gender, marital status, and computer indicators. Briefly interpret the “economic meaning” of each slope coefficient. What do the coefficients on experience and experience-squared imply about the life-cycle profile of earnings? Would including just a linear term for experience lead to a more appropriate regression model? Explain. Now add experience<sup>3</sup> and experience<sup>4</sup> to the regression. Does this substantially improve the fit of the regression model? Derive the F-test for whether experience<sup>3</sup> and experience<sup>4</sup> are important determinants of log-wages.
- b) Now create dummy variables for each of the ten levels of schooling (note: round the education variable to the next integer value to get levels of education 9-18). Regress log-wages on just the ten dummy variables. Why does STATA drop one of the variables from the regression? Is the effect of education on log-wages linear in education? Describe where the “nonlinearities” are, if any. Calculate the F-test for the hypothesis that the returns to education are zero and use the R-squared from the regression to “recalculate” the same test. Now run the dummy variable regression for log-wages including experience, experience-squared, gender, marital status, and computer indicators. Does allowing for nonlinearities in the return to education improve the fit of the regression model? Explain.
- c) Using the estimated log-wage model in 3a) what is the predicted level of wages for a single woman with 12 years of education, and 9 years of experience?
- d) Now add the indicators for pencil, telefon, calculator, and hammer use to the regression you estimated in 3.a). Compare the estimated returns to education and computer use to the ones obtained in 3.a). Are they different?

Finally, estimate a model that controls for the individual's occupation as a "fixed effect" – `areg y x, absorb(occ)`. Interpret the implications of your findings for the role of potential omitted variables bias in the OLS estimates of the effect of computer use on log-wages. (See DiNardo and Pischke for their interpretation.)