

ECON 142

APPLIED EXERCISE #1

Due September 27, In Class

This exercise examines 1) statistics that summarize the relationship between two variables; 2) the bivariate linear regression model; and 3) the multivariate linear regression model. The focus is on the relationship between individual earnings and educational attainment. Feel free to work cooperatively and in groups. However, each person must hand in his/her own problem set using his/her own words and interpretation of the results. **Summarize the results/answers to each question concisely and not by referencing a morass of STATA output.** STATA output may be attached to the end of the problem set for perusal by Justin for partial credit.

To examine these questions, we will analyze the following STATA data set:

Data Source: restricted92.dta

This data extract is from a 1992 survey of workers in Germany and comes from the DiNardo and Pischke paper, "The Returns to Computer Use Revisited: Have Pencils Changed the Wage Structure Too?" (1997). Reference the paper in the reader for more details. The observational unit of the data is the individual.

Data Notes:

1. Link to data via command: `ln -s /user8/e142/data/restricted92.dta .`
2. The data set has 20,042 observations on 19 variables.
3. For this problem set, the key variables are:
 - lnw = the natural log (ln) of the hourly wage reported by the individual
 - ed = the educational attainment of the individual
 - exp and exp2 = the work experience of a person and its square (experience²)
 - female = an indicator variable equal to one if the person is female, zero otherwise
 - mar = an indicator variable equal to one if the person is married, zero otherwise
 - computer = an indicator equal to one if the person uses a computer at work
 - pencil = an indicator equal to one if the person uses a pencil at work
 - telefon = indicator equal to one if person uses a telephone at work
 - calc = indicator equal to one if person uses a calculator at work
 - hammer = indicator equal to one if person uses a hammer at work
 - occ = 4-digit occupation codes

Question 1: Summarizing 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? Now comment on the linearity/nonlinearity of the relation between education and log-wages.
- b. What are the formulae for the sample covariance, variance, and correlation coefficient of two random variables, X and Y? Calculate the sample covariance of log-wages with education. Now calculate the

sample correlation coefficient of log-wages with education. Give a simple example of how the correlation between education and earnings may not be causal.

- c. Is the bivariate normal distribution a sensible functional form for the relation 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 their joint distribution? What is the formula for the population mean of log-wages conditional on education under the assumption of joint normality and in terms of the population parameters? Now substitute in the sample “analogs” of these parameters to calculate the constant and slope coefficient of the conditional mean of log-wages. Under the assumption of bivariate normality, how does the conditional variance of log-wages vary with education? What is the “technical” term for this?

Question 2: The bivariate linear regression model of earnings and education.

- a. Describe the 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 earnings? Under what assumptions will the least squares estimator be the best linear unbiased estimator? Describe briefly how the linear regression model presumes that the conditional expectation of log-wages (a.k.a., the regression function) is linear in education.
- b. Derive the “first-order conditions” for the least squares formula for the estimators of the constant and slope coefficient for a regression of log-wages on education. Briefly describe their intuition. Describe three properties of the least squares line.
- c. Now run the regression of log-wages on education and a constant. How do these estimates of the constant and slope coefficients compare to those in 1c? From the total sum of squares (TSS), explained sum of squares (ESS), and residual sum of squares (RSS), derive the R-squared of the regression. From the R-squared, derive the correlation coefficient of log-wages with education.
- d. How can the slope coefficient on education be interpreted as the percentage return to an additional year of schooling? From the root mean squared error (MSE) of the regression, derive an unbiased estimate of the variance of the residuals of the regression. Based on this estimate and the estimated standard error of the slope coefficient on education, what is the sample variance of education?
- e. Derive the 95% confidence interval for the return to education. Using a t-test, test the null hypotheses that the return to education is zero and five-percent, respectively. What is the p-value for the significance test on education (i.e., $H_0: \beta = 0$)? Use the ESS and RSS to derive the F-statistic for testing the significance of education. How is it related to the t-statistic/t-ratio? Now derive the t-ratio and F-statistic for the education coefficient using the R-squared of the log-wage regression.

Question 3: The multivariate earnings regression model.

- a. Now regress log-wages on a constant, education, experience, experience-squared, the gender and marital status indicators, and the computer indicator. 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 experience³ and experience⁴ to the regression. Does this substantially improve the fit of the regression model? Derive the F-test for whether experience³ and experience⁴ are important determinants of log-wages.
- b. Now create dummy variables for each of the ten levels of schooling (9-18) – for fractions of education, round to the nearest integer. 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 an F-test for the hypothesis that the returns to education are zero and use the R-squared from the regression to “recalculate” this same test. Now run the dummy variable regression for log-wages including experience, experience-squared, the gender and marital status indicators, and the computer indicator. 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, 10 years of experience, and who uses a computer at work?
- d. Now add the indicators for pencil, telephone, calculator, and hammer use to the regression you ran in 3a. Compare the estimated return to education and computer use to the ones from the 3a regression model. Are they different? Now run a regression that also controls for the individual’s occupation category as “fixed effects” – e.g., `areg y x, absorb(occ)`. Interpret the implications of your findings for the role of potential omitted variables bias in the OLS estimate of the effect of computer use on log-wages (see DiNardo and Pischke for their interpretation).