ECON7310: Elements of Econometrics

Research Project 1

Fu Ouyang

September 6, 2022

Instruction

Answer all questions following a similar format of the answers to your tutorial questions. When you use R to conduct empirical analysis, you should show your R script(s) and outputs (e.g., screenshots for commands, tables, and figures, etc.). You will lose *2 points* whenever you fail to provide R commands and outputs. When you are asked to explain or discuss something, your response should be brief and compact. To facilitate tutors' grading work, please clearly label all your answers. You should upload your research report (in PDF or Word format) via the "Turnitin" submission link (in the "Research Project 1" folder under "Assessment") by 11:59 AM on the due date September 13, 2022. Do not hand in a hard copy. You are allowed to work on this assignment in groups; that is, you can discuss how to answer these questions with your group members. However, this is *not* a group assignment, which means that you must answer all the questions in your own words and submit your report separately. The marking system will check the similarity, and UQ's student integrity and misconduct policies on plagiarism apply.

Background

You want to estimate the effect of education on earnings. The data file `cps4_small.csv` contains 1,000 observations on hourly wage rates, education, and other variables from the 2008 Current Population Survey (CPS):

- `wage`: earnings per hour
- `educ`: years of education
- `exper`: post education years experience
- `hrswk`: working hours per week
- `married`: dummy for married
- `female`: dummy for female
- `metro, midwest, south, west`: location dummies
- `black`: dummy for black
- `asian`: dummy for Asian

Research Questions

1. **(20 points)** Load this dataset in R (2 points). Obtain summary statistics (mean, standard deviation, 25, 50 (median), and 75 percentiles) for the variables `wage` and `educ` (5 points). Plot histograms for these two variables to explore their distributions. Make your histograms reader-friendly; that is, give informative titles and variable names instead of just using the default titles and variable names (6 points). For example, you could use `Years of Education` in place of `educ`. Create a new variable ln(`wage`) (2 points)[1] and draw a scatter plot of ln(`wage`) versus `educ` (3 points). Comment on the correlation between these two variables (2 points).

2. **(25 points)** Estimate the simple linear regression model:

$$\ln(\texttt{wage}_i) = \beta_0 + \beta_1 \texttt{educ}_i + e_i.$$

   where e_i is the error and β_0 and β_1 are the unknown population coefficients.

   (a) **(3 points)** Report the estimation results in a standard form as introduced in Lecture 5. For example, see page 5, where the estimates are presented in an equation form, along with standard errors (SE) and some measure for goodness of fit.

   (b) **(3 points)** Plot the estimated regression line you obtained in (a) on the scatter plot you constructed in Question 1.

   (c) **(6 points)** Interpret the estimated coefficient on `educ` (3 points) and test whether or not the population coefficient β_1 is zero at the 1% significance level (3 points).

   (d) **(6 points)** You suspect that the hourly wage could depend on working hours per week. Under what condition(s) would the estimates in (a) be biased and inconsistent due to the omission of the weekly working hours (2 points)? Give a reasonable and intuitive story on why omission of the weekly working hours would cause omitted variable bias in the regression in (a) (2 points). Based on your story, explain whether the coefficient on `educ` in (a) would be overestimated or underestimated (2 points). Hint: Review pages 4 and 5 of Lecture 4.

   (e) **(7 points)** The variable `hrswk` is the average weekly working hours for each individual in the data. Regress ln(`wage`) on `educ` and `hrswk` and report the estimation results in a standard form (3 points). Discuss the estimation results. In particular, how would you revise your answer in (c) (2 points)? Are the estimates statistically significant (2 points)?

3. **(40 points)** You are still concerned about omitted variable bias (OVB) in the regressions of Question 2. For that reason, you decide to regress ln(`wage`) on all other variables in the dataset and use this model as a benchmark.

   (a) **(11 points)** Report a 95% confidence interval for the slope coefficient on `educ` (3 points), explain the relationship between the confidence interval and hypothesis testing (4 points), and test the hypothesis that one year of additional education would increase hourly wage by 12% (4 points).

   (b) **(7 points)** Assuming there is no OVB, discuss the estimated coefficient on `female` in the benchmark model. In particular, explain what the estimated coefficient on `female` means on hourly wage (3 points), compare the effect of being female and the effect of one year of additional education (2 points), and discuss whether being female has a statistically significant effect on hourly wage (2 points).

[1] In R, the function log() computes logarithms, by default natural logarithms.

(c) **(5 points)** Using the estimation results of the benchmark model, test the hypothesis that the hourly wage is not affected by the geographic location (3 points). Explain how you reach your conclusion (2 points).

(d) **(5 points)** Using the estimation results of the benchmark model, test the hypothesis that the wage differential associated with African American is equal to the wage differential associated with Asian American (3 points). Explain how you reach your conclusion (2 points).

(e) **(7 points)** How would you modify the benchmark model to estimate the effects on hourly wage of one additional year of education separately for each gender (4 points). How do the effects of education differ between genders and is the difference statistically significant (3 points)? Hint: See pages 27–39 of Lecture 6.

(f) **(5 point)** Keoka is an African American woman, working in a metropolitan area. After she obtained her high school diploma, she got a job and started working instead of getting a higher education. She has never been married. Now she has a five-year of experience in the industry and is working full time (40 hours per week).[2] Using the benchmark model, predict her hourly wage.

4. **(15 points)** It may be more useful to estimate the effect on earnings of education by using the highest diploma/degree rather than years of schooling. Define four dummy variables to indicate educational achievements:

   - `lt_hs` $= 1$ if `educ` < 12
   - `hs` $= 1$ if `educ` $= 12$
   - `col` $= 1$ if `educ` ≥ 16
   - `some_col` $= 1$ for all other values of `educ`.

   (a) **(6 points)** Create the dummy variables `lt_hs`, `hs`, `col`, and `some_col` as defined above (4 points) and compute the sample means of hourly wage for each of the four education categories (2 points).

   (b) **(9 points)** Regress `wage` on the four dummies `lt_hs`, `hs`, `col`, and `some_col`. Can you obtain the OLS estimates? What is the problem here? Under what circumstances would you face this problem (4 points)? To avoid this problem, you now regress `wage` on three dummies (`lt_hs`, `col`, `some_col`) excluding `hs`. Interpret the estimated intercept (2 points) and compare the estimation results with the sample means calculated in (a) (3 points).

[2] Be careful! the left-hand side variable is ln(`wage`), but you are asked to predict Keoka's `wage`.