ECO 348K, Applied Econometrics and Data Analysis **Prof. Tom Vogl**

Spring 2019 **University of Texas at Austin**

PROBLEM SET 2: ESTIMATING THE RETURNS TO EDUCATION

DUE AT 4 PM ON FRIDAY 2/22

A surprisingly large share of the past half-century of research in labor economics has focused on the return to education: the added earnings power that an individual obtains by staying in school an extra year. In the 1970s, the late economist Jacob Mincer formulated what is now seen as the standard relation between human capital and wages:

$$ln(w_i) = \beta_0 + \beta_1 ed_i + \beta_3 exper_i + \beta_4 exper_i^2 + \varepsilon_i$$

where w_i is the hourly wage, ed_i is years of education, and $exper_i$ is years of labor market experience. This equation is known as the Mincerian Wage Equation. In this problem set, we will explore the difficulties that arise in estimating the returns to education using OLS.

The zip file contains two datasets, both containing data on labor earnings and education among US adults. One is a sample of working-age (25-64) adults in the Current Population Survey, a nationally-representative survey of the non-institutionalized population that takes place monthly. This dataset is from March 2018, with data on labor market outcomes in 2017. The data are in raw format, but the associated do file processes them into a Stata dataset. The other dataset comes from the National Longitudinal Survey of Youth, a study that first surveyed a sample of 14-21 year olds in 1979 and then re-surveyed them annually or bienially to the present. The labor market data are for 2007, when the cohort was aged 42-49.

1. Interpret the Mincerian Wage Equation conceptually. If one assumes that education and experience are exogenous, how should one interpret β_1? Why do you think the equation has a squared term in experience?
2. Start with the CPS data. Run the data processing do file. Generate a log hourly wage variable, where the hourly wage equals annual labor earnings divided by annual work hours. Generate race dummies for the categories "white," "black," and "other." Generate a new education variable to measure years of schooling (type `label list educ_lbl` to view the labels for education). For intervalled education categories, you may assign the midpoint of the category (e.g., "5th to 6th grade" becomes 5.5 years). Generate a "potential experience" variable as follows: $exper_i = age_i - ed_i - 5$. Also generate $exper_i^2$. Drop anyone who worked fewer than 50 weeks or fewer than 35 hours in a typical week. Summarize the data.

3. Estimate a univariate regression of the log hourly wage on education. What is the estimated return to education? Based on your regression coefficient and the summary statistics in your answer to question (2), calculate the correlation between education and the log hourly wage. Confirm that your calculation is correct using Stata's `corr` command or by adding the option `beta` to your regression command. How does the correlation compare to the R^2 of the regression? Why? Show mathematically how the correlation coefficient relates to the regression coefficient and the R^2.
4. Estimate the Mincerian Wage Equation. What is the estimated return to education?
   BONUS: Use a series of bivariate regressions to obtain the same coefficient on education you estimated through multiple regression. (This bonus problem involves an application of the Frisch-Waugh theorem. You will need to estimate six separate bivariate regressions.) Do you obtain exactly the same coefficient as you did in your answer to the main part of question (4)? Is the standard error the same? Why?
5. Estimate an "extended" Mincerian Wage Equation that controls for race and sex? Does the estimated return to education change after controlling for these covariates?
6. In the "extended" regression, is the black-white wage log gap statistically different from the female-male log wage gap?
7. Add a gender-education interaction term to the "extended regression." Is the estimated return to education significantly different for men and women? Estimate the ratio of the return for women to the return for men. Is it significantly different from 1?
8. Now move on to the NLSY data. Generate a log hourly wage variable and a "potential experience" variable as above. Drop anyone who worked less than 35 hrs/week for 50 weeks. Summarize the data.
9. Estimate an extended Mincerian Wage Equation with controls for race/ethnicity and sex. How do your estimates of the return to education and the return to experience compare to the estimates from the CPS? If there are differences, hypothesize why.
10. Do you think your estimate of β_1 represents the causal effect of education? Explain why or why not.
11. NLSY respondents took a cognitive test, the Armed Forces Qualifying Test (AFQT), in 1981. They also responded to several questions on their childhood environment. The dataset contains both the cognitive test scores and the measures of the childhood environment. Do you think any of these variables would be appropriate as control variables in the Mincerian Wage Equation? If so, re-estimate the equation, controlling for race/ethnicity, sex, and any other variables as you see appropriate. What happens to the estimated return to education? Interpret any changes you observe.
12. Based on the results from the NLSY, what do you conclude about the ability of OLS to deliver causal estimates of the Mincerian Wage Equation?