ECON 121, Applied Econometrics and Data Analysis **Prof. Tom Vogl**

Winter 2020 **University of California San Diego**

PROBLEM SET 1: ESTIMATING THE RETURNS TO EDUCATION

DUE BY 11:59 PM ON FRIDAY 1/31

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 the Mincerian Wage Equation. What is the estimated return to education?
4. Estimate an "extended" Mincerian Wage Equation that controls for race and sex? Does the estimated return to education change after controlling for these covariates?
5. In the "extended" regression, is the black-white wage log gap statistically different from the female-male log wage gap?
6. Run the "extended" regression separately for men and women. By how much do estimated returns differ by sex? Based on the two sets of regression results, assess whether the difference is statistically significant.
7. Estimate the male-female difference in returns by adding interaction terms to the "extended" regression in the full sample. Do you get the same answer? (You should.) Instead of the *difference* in returns between men and women, we might instead be interested in the *ratio* of returns. Use the delta method and the bootstrap to estimate the ratio of returns for women to the return for men. Do these two methods suggest the ratio is significantly different from 1?
8. Now move on to the NLSY data. The NLSY oversampled black and Hispanic/Latino respondents. The variable `perweight` is a sampling weight that can be used to obtain statistics that are representative of the population. Summarize the variables `black` and `hisp` with and without using sampling weights. Which summary statistics provide unbiased estimates of the racial/ethnic composition of US adults who were teenagers in 1979? Explain.
9. Generate a log hourly wage variable and a "potential experience" variable as above. Drop anyone who worked less than full time (e.g., 35 hrs/week for 50 weeks). Estimate an extended Mincerian Wage Equation (controlling for race/ethnicity and sex), with and without using sampling weights. How does the use of sampling weights change the results? Decide whether you want to use sampling weights for the rest of the analysis, and justify your choice. For the remainder of the analysis, only use your preferred method.
10. How do your preferred estimates of the return to education and the return to experience compare to the estimates from the CPS? If there are differences, hypothesize why.
11. Do you think your preferred estimate of β_1 represents the causal effect of education? Explain.
12. 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.

13. Based on the results from the NLSY, what do you conclude about the ability of OLS to deliver causal estimates of the return to education?