Econ 9220, Advanced Labor Economics
CPS data assignment, Fall 2015
Barry Hirsch

The take-home assignment involves estimation of wage equations and wage differentials using data on wage and salary workers ages 18 to 65 from the 2014 monthly Current Population Survey Outgoing Rotation Group (CPS-ORG) earnings files. The data extract and documentation are posted on the course homepage (in Stata and SAS formats). The assignment is due by midnight, Friday, October 16. A copy of your computer program and output must be submitted with your paper.

Ground rules: This is an individual student assignment and should reflect only your individual work. If you have questions regarding the assignment, ask the instructor. Permitted help from others is limited to assistance in reading in the initial data set and help with generic Stata or SAS language.

A. Mincer wage equation, earnings nonresponse and imputation, wage-experience profiles, and returns to schooling.

Estimate a wage equation in which the natural log of hourly earnings (calculated as usual weekly earnings divided by usual hours worked per week) is a function of years of schooling; years of potential experience and experience squared, with experience measured by [age-schooling-6] or [age-18], whichever is smaller (if any experience values are negative, set them to zero); and dummies for female, black, and union membership. Show results. Now estimate the exact same regression, but omit all imputed earners. In all analysis that follows, continue to omit imputed earners.

Based on the second regression (the sample that omits imputed earners), what is the rate of return to schooling, the percentage union wage premium, the percentage difference between women and men, and the percentage wage difference of black relative to non-black workers?

Comparing coefficients in the regressions with and without imputed earners, which coefficient estimate is most affected? Which of the coefficient estimates is more reliable and which is highly biased? Provide an explanation why this coefficient estimate differs between the full sample and the sample that omits imputed earners. (See Hirsch and Schumacher, “Match Bias in Wage Gap Estimates Due to Earnings Imputation,” Journal of Labor Economics, July 2004). How do the R-square values compare between the regression with and without imputed earners? Explain why one is higher than the other.

In the regression without imputed earners, at how many years of experience do wages peak? What is the average age at which wages peak (mean schooling is 14.1 years). Show your calculations.

Starting with the regression just estimated (without imputed earners), add the six metropolitan area (city) size dummy variables, with the omitted reference group being rural and non-designated MSAs (coded missing '.'). Interpret the coefficients and provide the percentage wage differences between workers in MSAs of 5+ million versus rural areas? What is the percentage differential between workers in MSAs 5+million and workers in MSAs between 100 to 500 thousand residents?

How did adding the MSA size dummies affect your coefficient estimate on black? Provide a likely explanation for why the coefficient changed in this direction. Is the coefficient estimate on race more reliable with or without the MSA size dummies?

Next examine how schooling rate of return estimates differ with age, letting your age groups be ages 18-34, 35-49, and 50-65. Estimate two different equations. In the first, include schooling, schooling times Dummy35-49, and schooling times Dummy50-65 (plus experience and its square, female, black, and union). In the second, include schooling times Dummy18-34, schooling times Dummy35-49, and schooling times Dummy50-65 (plus the same non-schooling controls). What are the r.o.r. estimates for those ages 18-34, 35-49, and 50-65? Do you get equivalent or different estimates from the two equations? Explain. What is the difference in the r.o.r. for those 50-65 versus 18-34? Is this difference statistically significant?
B. Division bias and labor supply elasticities

Estimate a labor supply equation in which the dependent variable is the log of hours worked last week ($\text{hrs}_{\text{lstwk}}$) and the regressors are the log of the hourly wage (measured here by usual weekly earnings divided by $\text{hrs}_{\text{lstwk}}$), plus years of schooling, experience and experience squared, female, and black. At what age (not experience) do weekly work hours peak? What is the coefficient on the log wage? Interpret this coefficient, noting whether the income or substitution effect is dominant.

Similar to the approach in Borjas, now provide an alternative measure of the wage on the right-side of the equation, this time measured by the log of weekly earnings divided by usual weekly hours worked (as in parts A and C). Compare the two labor supply estimates. Explain how and why they differ? Which estimate do you regard as more trustworthy?

C. Blinder-Oaxaca Decomposition

What is the unadjusted mean logarithmic wage gap between Hispanic and non-Hispanic men (that’s right, you need not run a regression)? Now estimate separate wage equations for Hispanic and non-Hispanic men, including only the following eight variables (to keep things simple): schooling, experience (do not include quadratic), married, union, part-time, large metro (2.5 million and above), South (the combination of South Atlantic, East South Central, and West South Central states), and foreign born. Using a “Blinder-Oaxaca” decomposition, break-up the total wage gap into "explained" and "unexplained" portions, using the non-Hispanic male coefficients as the "nondiscriminatory" betas. How much of the total wage gap is explained by differences in measured characteristics? Show the contribution that each of the explanatory variables provides to the explained portion of the wage gap. Which characteristics are most important? How much of the wage gap is unexplained by differences in characteristics? Now re-estimate the Hispanic wage differential using a single equation (with Hispanic and non-Hispanic men pooled) that includes the same eight variables, plus a Hispanic dummy. How do estimates of the explained and unexplained portions using the single equation dummy variable approach differ from the B-O decomposition results?

[Note: You should calculate the B-O decomposition by hand (i.e., using a spreadsheet) or programmed yourself in Stata or SAS. Do not provide results using an available B-O Stata .do file.]
The data set cps2014class (available in SAS and Stata formats) includes 137,374 non-student wage and salary workers, ages 18 to 65, in 2014, extracted from the Census Bureau's Current Population Survey Outgoing Rotation Group files (i.e., the ¼ sample asked earnings and union questions each month). Below is a brief description of the variables you will need to work with in your take-home exercise (a small number of other variables are included, but not defined below). Variable names in Stata dataset are in lower case.

SCHOOL Years of schooling completed (0-20)
AGE 18 to 65
FEMALE 1=females, 0=male
BLACK 1=black, 0=other race
HISPANIC 1=Hispanic; 0=other ethnicity
FORBORN 1=foreign born (includes citizens and non-citizens); 0=native born
MARRIED 1=married spouse present; 0=ever married, spouse not present; never married
UNION 1=union member; 0=not a member
WKEARN Usual weekly earnings ($1 to estimated mean above the topcoode)
IMPUTE 1=earnings non-respondent, earnings imputed; 0=earnings respondent, not imputed
WKHOURS Usual hours worked per week (1-99)
HRSLSTWK Hours worked last per week (1-99)
PARTTIME 1=less than 35 usual hours worked per week; 0=35+ usual hours
MSASIZE . =not in MSA or small MSA not identified; 2=100t-250t, 3=250t-500t, 4=500t-1m, 5=1m-2.5m, 6=2.5m-5m, 7=5m+
STATE State codes (11-95); regional breakdown shown below
Northeast 10<state<20
Middle Atlantic 20<state<30
East North Central 30<state<40
West North Central 40<state<50
South Atlantic 50<state<60
East South Central 60<state<70
West South Central 70<state<80
Mountain 80<state<90
Pacific 90<state<100

Other variables included in data set (not needed in assignment and not defined here): race, class, month, fips, marital, degree, ind3, occ3, citizen, weight.