Constructing Age-earnings Profiles From CPS Data

Lutz Hendricks*

November 20, 2015

Abstract

This document describes the construction of age-earnings profiles from CPS data. It serves as documentation for the data used in a number of my papers.

*University of North Carolina, Chapel Hill; lutz@lhendricks.org

1 Introduction

This document describes the construction of age-earnings or age-wage profiles from March CPS data. The same procedures are used in a number of my papers. Documenting them in one place avoids duplication.

Details, such as the age ranges or cohorts included in the sample, differ across papers. This document therefore replaces specific values of the parameters governing these choices with variables. The values of these variables are specified in the papers that use the data.

2 Sample

The data are taken from the March Current Population Survey (King et al., 2010). The data waves used and the sampling criterisa (age and birth year ranges, etc.) vary across applications and are documented in the papers that use the data. See `import_cpsearn.filter` and `import_cpsearn.import` (this is a pointer to the program file that executes this step).

3 Individual Variables

The construction of individual variables is based on Bowlus and Robinson (2012).

Dollar values are deflated using the Consumer Price Index (all items, U.S. city average, series Id: CUUR0000SA0; see bls.gov).

Schooling: Individuals are assigned one of 4 schooling levels: high school dropout (HSD), high school graduate (HSG), college dropout (CD), college graduate (CG). See `var_cpsearn.school_create`.

As discussed in Jaeger (1997), the coding of schooling changes in 1991. I use the coding scheme proposed in his tables 2 and 7 to recode HIGRADE and EDUC99 into the highest degree completed and the highest grade completed.

Prior to 1992, we have information about completed years of schooling (variable higrade). During this period, we define high school graduates as those

completing 12 years of schooling (higrade=150), college dropouts as those with less than four years of college (151,...,181), and college graduates as those with 16+ years of schooling (190 and above). Beginning in 1992, the CPS reports education according to the highest degree attained (educ99). For this period, we define high school graduates as those with a high school diploma or GED (educ99=10), college dropouts as those with "some college no degree," "associate degree/occupational program," "associate degree/academic program" (11,12,13,14). College graduates are those with a bachelors, masters, professional, or doctorate degree (15,16,17,18).

Hours worked: Hours worked per year are defined as the product of hours worked last week and weeks worked last year.

Weeks worked per year are intervalled until 1975 (WKSWORK2). For consistency, I use the intervalled variable for all years. Each interval is assigned the interval midpoint.

Until 1975, hours worked per week are only available for the previous week (HRSWORK). For consistency, I use this variable for all years.

Income variables: Labor **earnings** are defined as the sum of wage and salary incomes (INCWAGE). In some cases, I consider a broad measure of labor income that includes a fraction of self-employment income (INCBUS).

Wages are defined as labor earnings divided by weeks worked.

I construct 2 versions of earnings and wages. For the **full sample**, I retain all observations with non-negative earnings or wages. For the **wage sample**, I drop individuals who work fewer than a minimum number of weeks or whose wages are outside a fixed range around the median (presumed to be outliers or persons not truly working for pay).

Income variables are top-coded. As discussed in Bowlus and Robinson (2012), the frequency of top-coding and the top-coded amounts vary substantially over time. In addition, top-coding flags contain obvious errors. In most years, fewer than 2% of labor earnings observations appear to be top-coded. Following Bowlus and Robinson (2012), I prefer to use median rather than mean log wages to avoid this problem. When using mean log wages, I follow Autor et al. (2008) and replace top-coded incomes with 1.5 times their top-coded values.

See `var_cpsearn.wage_create`.

Weights: Observations are weighted by WTSUPP which excludes the armed forces and the Hispanic oversample.

Work status: Individuals who report working for wages (CLASSWKR) are classified as wage earners.

4 Aggregate Statistics

All aggregate statistics are computed for the full sample and the wage sample.
For each [age, school, year, sample] cell, I compute (`aggr_cpsearn.age_school_year_stats`):

1. number of observations
2. mass
3. for earnings and wages: median, mean log
4. mean weeks worked
5. mean years of schooling.

I compute similar statistics (`aggr_cpsearn.aggr_stats`)

1. by [schooling, year], pooling a fixed age range
2. by [age range, schooling, year]. This is mainly for computing the college wage premium for young and old workers as in Card and Lemieux (2001).

5 Cohort Lifetime Earnings

For each [schooling, cohort] cell, I compute the present value of lifetime earnings over a fixed age range (that depends on schooling) (`profiles_cpsearn.cohort_earn_profiles`

Since full age profiles are not observed for most cohorts, some imputation is needed. This involves the following steps:

1. Regress log median wages on time dummies and age dummies.
2. Smooth each cohort's observed age-wage profiles using a Lowess filter.
3. Linearly interpolate interior missing values (these are rare and typically due to small numbers of observations in some cells).
4. Impute missing observations at the start or at the end of the cohort's career using the estimated age-wage profile. This is scaled to match the mean of 5 overlapping periods.

6 Code Documentation

Much of the code documention exists as program comments.

`run_all_cpsearn` runs all routines in sequence. As inputs, this expects:

1. CPS data files and code to load / recode them.
2. A `filterCpsearn` object that specifies sample selection rules.
3. A `profile` object that specifies how cohort wage profiles are to be constructed.
4. The program directory and some general purpose code need to be on the Matlab path.

References

Autor, D. H., L. F. Katz, and M. S. Kearney (2008): "Trends in U.S. Wage Inequality: Revising the Revisionists," *Review of Economics and Statistics*, 90, 300–323.

Bowlus, A. J. and C. Robinson (2012): "Human Capital Prices, Productivity, and Growth," *The American Economic Review*, 102, 3483–3515.

Card, D. and T. Lemieux (2001): "Can Falling Supply Explain the Rising Return to College for Younger Men? A Cohort-Based Analysis," *The Quarterly Journal of Economics*, 116, pp. 705–746.

Jaeger, D. A. (1997): "Reconciling the Old and New Census Bureau Education Questions: Recommendations for Researchers," *Journal of Business and Economic Statistics*, 15, pp. 300–309.

King, M., S. Ruggles, J. T. Alexander, S. Flood, K. Genadek, M. B. Schroeder, B. Trampe, and R. Vick (2010): "Integrated Public Use Microdata Series, Current Population Survey: Version 3.0. [Machine-readable database]," Minneapolis: University of Minnesota.