Read Me File

This file describes the data and cleaning procedures to replicate AKK (2008). Two versions of the CPS data are used in the analysis – the March Annual Social and Economic Supplement (ASEC) and May Outgoing Rotation Group (MORG). The ASEC CPS data are downloaded from IPUMS for all years in question. Meanwhile the MORG CPS data are downloaded from two different NBER sources for the period 1973-1978 and 1979-2020, respectively.

Data Sources

CPS ASEC 1963-2022

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Megan Schouweiler, and Michael Westberry. IPUMS CPS: Version 11.0. Minneapolis, MN: IPUMS, 2023. https://doi.org/10.18128/D030.V11.0

CPS May Extracts 1973-1978

NBER, https://www.nber.org/research/data/current-population-survey-cps-may-extracts-1969-1987.

CPS MORG 1979-2020

NBER, https://www.nber.org/research/data/current-population-survey-cps-merged-outgoing-rotation-group-earnings-data.

GDP Personal Consumption Expenditure (PCE) Deflator

U.S. Bureau of Economic Analysis, "Table 1.1.4. Price Indexes for Gross Domestic Product".

Federal Minimum Wage

U.S. Department of Labor, Federal Minimum Wage Rate under the Federal Fair Labor Standards Act [STTMINWGFG], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/STTMINWGFG.

Unemployment Rate

U.S. Bureau of Labor Statistics, Unemployment Rate [UNRATE], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/UNRATE.

Unemployment Rate – married men

U.S. Bureau of Labor Statistics, Unemployment Rate - Married Men [LNS14000150], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/LNS14000150.

Unemployment Rate – married men

U.S. Bureau of Labor Statistics, Unemployment Rate - 25-54 Yrs., Men [LNS14000061], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/LNS14000061.

Do File Description: CPS ASEC

These do files synthesize publicly available cleaning codes from David from Autor's website for AKK (2008). Variable nomenclature is streamlined to be consistent with variable names from IPUMS CPS.

CPS-ASEC data can be downloaded from IPUMS: https://cps.ipums.org/cps/. IPUMS variables used in the analysis are: year, statefip, popstat, actnlfly, ahrsworkt, uhrsworkly, fullpart, empstat, educ, higrade, age, sex, race, asecwt, wkswork1, wkswork2, indly, classwly, inctot, incwage, incbus, incfarm, inclongj, oincwage, qinclong, qincwage, qoincwage, srcearn. After downloading, save the data as "cps_asec" and place it in the Data_Raw subfolder.

0_Paths

This file lays out initial housekeeping, defines directories used in the project, and describes each do file. The key path that needs to be changed is the global macro "main." The following folders should be located within this directory: Scripts, Data_Raw, Data_Output, Figures. These folders are then saved as their own global macros. The main directory can have other folders as well, but these four are the ones used in the do files for replication.

1_Run_DoFiles

This file executes the each do file in turn. The do files are listed in order in which they need to be executed, beginning with data cleaning before moving to figures and statistical analysis.

2_CPS_1976_1978

This do file cleans three years of the CPS ASEC: 1976–1978. This is the first step in the cleaning process because the years 1962–1975 are missing data on the number of weeks worked last year ("wkslyr" in AKK and "uhrsworkly" in IPUMS) and usual hours worked per week last year ("hrslyr" in AKK and "uhrsworkly" in IPUMS). There is a variable for weeks worked last year ("wkswork2" in IPUMS), but it is intervalled. Because of this, part of the function of this do file is to find the mean of weeks worked last year by sex and race, which is then imputed into variable "wkswork1" for the years 1962–1975. In addition to imputing weeks worked last year, this do file also runs a regression and saves the coefficients to imput usual hours worked per week last year ("hrslyr") for the years 1962–1975. The regression is:

$$\text{hours lyr}_i = \alpha + \beta_1 \text{hours lwk}_i + \beta_2 \text{fulltime}_i + \beta_3 \text{nilf}_i + \beta_4 \text{fulltime}_i \times \text{hours lwk}_i + \beta_5 \text{fulltime}_i \times \text{nilf}_i + \varepsilon_i$$

where hours lyr is the hours an individual worked per week last year (variable "uhrsworkly" in IPUMS), hours lwk is the hours worked last week (variable "ahrsworkt" in IPUMS), fulltime is an indicator if the worker is fulltime (worked 35+ hrs/wk last yr), nilf is an indicator if the worker's employment status is not in the labor force, interactions of being a fulltime worker and hours last week and not in the labor force, ε is an idiosyncratic error term. The regression is run for each gender and race group. Data are pooled for all three years. The regression is weighted using ASEC individual weights. The estimated coefficients are then saved to impute "hrslyr" for years 1962–1975.

- Input data: cps_asec
- Output data:
  - 1) ASEC_1976_1978_hrswks. This is an intermediate dataset, deleted at the end of the do file.
  - 2) ASEC_1976_1978_wkslyr. This is the dataset for weeks worked last year, which is imputed for years 1962–1975.
  - 3) ASEC_1976_1978_hrslyr. This is the dataset for usual hours worked per week last year, which is imputed for years 1962–1975.
  - 4) ASEC_1976_1978_cleaned. This is the cleaned dataset for these years, used in subsequent analysis.

3_CPS_1962_1975

This do file cleans the years 1962–1975. Over these years, the variable for weeks worked last year ("wkswork2") is intervalled. Hence the mean weeks worked last year by race and sex are imputed from the years 1976–78. Similarly, in the years 1962–1975, the variable for usual hours worked per week last year ("uhrsworkly") is not available. Hence there is a second imputation for "uhrsworkly" based on regressions run from the years 1976-78. The 1963 March CPS does not contain education variables, so this year is not included in the analysis. In AKK's code, 1966 is double-sampled, so they adjust the weights by dividing by two. However since then, IPUMS appears to have fixed this, so no longer have to do so. More details on this can be found at: https://cps.ipums.org/cps/asecsamplenotes.shtml.

- Input data: cps_asec
- Output data: ASEC_1962_1975_cleaned

4_CPS_1979_1987

This do file cleans the years 1979–1987.

- Input data: cps_asec
- Output data: ASEC_1979_1987_cleaned

5_CPS_1988_1991

This do file cleans the years 1988–1991. Starting in 1988, top-coded earnings and wages are multiplied by 1.5 times.

- Input data: cps_asec
- Output data:
  - 1) ASEC_1988_1991_cleaned_notop.
  - 2) ASEC_1988_1991_cleaned_top. Top-coded earnings and wages are multiplied by 1.5.

6_CPS_1992_2023

This do file cleans the years 1992–2023.

- Input data: cps_asec
- Output data:
  - 1) ASEC_1992_2023_cleaned_notop
  - 2) ASEC_1992_2023_cleaned_top. Top-coded earnings and wages are multiplied by 1.5.

7_AppendingYears

This do file appends the cleaned datasets from steps 2–6, to be used in subsequent analysis.

- Input data: ASEC_1962_1975_cleaned; ASEC_1976_1978_cleaned; ASEC_1979_1987_cleaned; ASEC_1988_1991_cl
  ASEC_1988_1991_cleaned_top; ASEC_1992_2023_cleaned_notop; ASEC_1992_2023_cleaned_top
- Output data:
  - 1) ASEC_all_cleaned_NOTOP. Dataset used for Figure 1.
  - 2) ASEC_all_cleaned_TOP. Dataset used for everything apart than Figure 1.

8_MarchCells

This do file creates earnings variables. It collapses the cleaned CPS ASEC dataset into year-education-experience-gender cells of weekly and hourly earnings, with different weighting definitions. Resulting dataset: MarchCells_1963_2023

- Input data: ASEC_all_cleaned_TOP
- Output data:
  - 1) precollapsemarch. This is an intermediate dataset, which is then merged with "marchcells1" to yield "marchcells_finalized."
  - 2) marchcells1. This dataset includes the count variables (with and without allocators), collapsed by year, education, experience, and gender. The count variables include different types of weighting (e.g. raw count of observations, ASEC weight, ASEC weight multiplied by weeks worked last year.
  - 3) MarchCells_1963_2023. This data set is the various earnings variables merged together with the count variables, collapsed by year, education, experience, and gender. The two main earnings/wage variables of interest are: 1) "lnrwinc", Log real weekly FT earnings, 2012$ and 2) "lnrhinc", Log real hourly FT wage, 2012$.

9_EfficiencyUnits

The do file calculates an average relative wage by year-education-experience-gender cell over the entire time period. It then calculates efficiency units by education, first not taking into account experience levels and then taking it into account. It calculates efficiency units for all individuals, and then broken down by gender.

- Input data: MarchCells_1963_2023
- Output data: Efficiency_units_1963_2023

10_PredictWages

This do file predicts weekly and hourly wages by gender for each year. It does so from a regression of real wages regressed on four education categories, three region dummies, race dummies for black and other, a quartic in experience, and interactions of education (3 broad groupings) with the experience quartic.

- Input data: ASEC_all_cleaned_NOTOP
- Output data: Predicted_wages

11_LaborSupplyWeights

This do file creates labor supply weights for year-school-experience-gender cells. The weights are equal to the sum of ASEC weights multiplied by weeks worked last year multiplied by usual hours worked in a week.

- Input data: MarchCells_1963_2023
- Output data: March_labor_supply_weights

12_AssembleWages

This do file merges the predicted wages with the labor supply dataset. It then creates a variable for the labor supply share in each cell-year. It also finds the average labor supply share in each cell between 1963-2019.

- Input data: Predicted_wages (from do file 10_PredictWages)
- Output data: Predicted_wages_1964_2023

13_WageGaps

This do file uses the predicted wages from the previous do file to calculate the wage gaps by education, and education-experience. The wage differential is estimated by year, age, gender, and experience. This is weighted using fixed share weights averaged over all years 1963–2023.

- Input data: Predicted_wages_1964_2023
- Output data: College_HS_wage_premium_exp

CPS_education_post92

This do file is executed within the "6_CPS_1992_2023" file. It replaces values for "educomp" from those by AKK for 1992 on.

Deflator_gdp_pce

This do file imports an anuual series of personal consumption expenditures from the BEA to construct a deflator and put earnings variables in real terms.

Variable nomenclature

The table below shows a concordance between variable names used in the original AKK do files and the corresponding variables in IPUMS.

AKK (2008)	**IPUMS CPS**	**Description**
age	age	Age
sex	sex	Gender
race	race	Race
_popstat	popstat	Adult civilian, armed forces, or child
_grdhi	higrade	Highest grade of school
grdcom	NA	No analogous variable in IPUMS. educ combines "grdhi" and "grdcom".
grdatn	educ	Educational attainment recode
_state	statefip	State (FIPS code)
empstat	empstat	Employment status
esr	empstat	Employment status
wgt	asecwt	Annual social and economic supplement (ASEC) weight
ftpt	fullpart	Worked full or part time last year
pyrsn	actnlfly	Activity when not in labor force last year (part year workers)
hrslyr	uhrsworkly	Usual hours worked per week (last yr). Use imputed data from 1976-1978 for years 1962-1975.
hours	ahrsworkt	Hours worked last week
ftpt	fullpart	Worked full or part time last year
_wkslyr	wkswork1	Weeks worked last year (begins in year 1976)
_wkslyr	wkswork2	Weeks worked last year, intervalled (used for 1962-1975)
indly	indly	Industry last year
clslyr	classwly	Class of worker, last year
_incwag	incwage	Wage and salary income
incer1	inclongj	Earnings from longest job
incwg1	oincwage	Earnings from other work included wage and salary earnings
incern	inctot	Total personal income
_incse	incbus	Non farm business income
_incfrm	incfarm	Farm income
aincwag	qincwage	Data quality flag for incwage
aincwg1	qoincwage	Data quality flag for oincwage
aincer1	qinclong	Data quality flag for inclongj
ernsrc	srcearn	Source of earnings from longest job

Do File Description: CPS MORG

0_Paths

This file lays out initial housekeeping, defines directories used in the project, and describes each do file.

1_Run_DoFiles

This file executes the each do file in turn.

2_CPS_1973_1978

This do file cleans the May Outgoing Rotation Group for years 1973-1978.

- Input data: annual data from NBER
- Output data: MayCPS_1973_1978

3_MORG_1979_2020

This do file cleans the MORG data from 1979 to 2020.

- Input data: annual data from NBER
- Output data: MORG_1979_2020

4_AppendingYears

The do file appends the cleaned datasets together.

- Input data: MayCPS_1973_1978 and MORG_1979_2020
- Output data: MayCPS_MORG_cleaned

5_PredictWages

This dataset predicts hourly wages by gender for each year. It does so from a regression of real wages regressed on four education categories, race dummies for black and other, a quartic in experience, and interactions of education (3 broad groupings) with the experience quartic.

- Input data: MayCPS_MORG_cleaned
- Output data: Predicted_wages_MORG

6_AssembleWages

This do file merges the predicted hourly wages with the labor supply dataset. It then creates a variable for the labor supply share in each cell-year. It also finds the average labor supply share in each cell between 1963-2020.

- Input data: Predicted_wages_MORG
- Output data:Predicted_wages_1979_2020_MORG

7_WageGaps

This do file uses the predicted wages from the previous do file to calculate the wage gaps by education, and education-experience.

- Input data: Predicted_wages_1979_2020_MORG
- Output data:College_HS_wage_premium_exp_MORG

CPS_education_post92

This do file is executed within the "3_MORG_1979_2020" file. It replaces values for "educomp" from those by AKK for 1992 on.

Labeling_variables_May_CPS

This do file labels the variables for the MORG dataset from the NBER.