Replication Data for: "How Does the Earned Income Tax Credit Work? Exploring the Role of Commuting and Personal Transportation"

Owen Davis

December 3, 2025

1 Overview

This replication package reproduces the data setup, regression models, tables, and figures for "How Does the Earned Income Tax Credit Work? Exploring the Role of Commuting and Personal Transportation," published in the *Journal of Labor Economics*. All code is written for Stata.

Both the raw datasets and final processed data files required to replicate the paper are included in the replication package. In brief, a full replication requires the following:

1. Download and unzip the full replication package to a directory.
2. Open Do/_master.do and change the global macro `homedir` in line 2 to the local directory where you have unzipped the full contents of the replication package.
3. Run _master.do and allow several hours (likely 10+) for it to complete. The intermediate and final data files as well as all graphs and tables will be replicated.

2 Directory structure

```
eitc_jole/
├── Data/
│   ├── Raw/
│   ├── Processed/
│   └── ...
├── Do/
│   └── ...
└── Output/
    ├── Estimates/
    │   ├── ASEC/
    │   ├── Census/
    │   ├── CPS/
    │   └── SIPP/
    ├── Graphs/
    └── Tables/
```

As noted above, it will be necessary for the replicator to replace the directory specified at the top of the _master.do file with their local path. This is done on line 2 with: `global homedir "[your dir here]"`.

3 Datasets

The project makes use of four primary datasets and numerous secondary data sources.

- **Primary datasets**
  - Current Population Survey, Annual Social and Economic Supplement (ASEC), 1986–2019
  - Current Population Survey, Basic Monthly Survey (CPS), 1986m1–2019m12
  - Census 5% Sample, 1990 and 2000
  - Survey of Income and Program Participation (SIPP), 1990–2001 panels
- **Secondary datasets**
  - Federal EITC parameters from Tax Policy Center (2025): "EITC Parameters, 1975 to 2024"
  - State EITC parameters from Komro, et al. (2020): "Anti-Poverty Policy And Health: Attributes And Diffusion Of State Earned Income Tax Credits Across U.S. States From 1980 To 2020," PLOS ONE 15(11)
  - State welfare waivers from Table B in: Department of Health and Human Services (1999): "State Implementation of Major Changes to Welfare Policies, 1992-1998"
  - Urban Institute Welfare Rules Database
  - Bureau of Economic Analysis, GDP by state
  - State minimum wages from Vaghul and Zipperer (2016): "State and Sub-State Historical Minimum Wage Data"
  - County population totals from U.S. Census Bureau, Population Estimates program
  - Crosswalks between geographies and for metropolitan areas over time; some of these (e.g., Census_metareas.csv) are made by hand
  - Department of Housing and Urban Development Picture of Subsidized Housing, 1996
  - Quarterly Census of Employment and Wages

Where possible, the first time an externally produced dataset is referenced, the script first checks if that dataset already exists in the expected directory and, if not, attempts to download that file (or files) from the web. This should be unnecessary, however, since the replication package includes all datasets used in the project.

The exception to the above is the raw QCEW data, which is excluded from the replication package to preserve space. The smaller processed file derived from the QCEW data is included, however (found in Data/Processed/qcew.dta). By default the code skips the script that builds qcew.dta. A user who would like to run this script and build the qcew.dta file from scratch can do so by setting the value of the global macro $download at the top of qcew.do to 1.

The project initially used a Python script to call the IPUMS API to download CPS and Census data. Alternatively, a researcher wishing to replicate the data construction from scratch can use the IPUMS website to download the necessary extracts. The IPUMS variables and date ranges needed are listed below.

- **CPS ASEC**
  - Samples: 1986–2019
  - Variables: `YEAR, SERIAL, CPSID, PERNUM, CPSIDP, ASECWT, ASECFLAG, STATEFIP, METAREA, METRO, REGION, RELATE, MOMLOC, POPLOC, SPLOC, AGE, SEX, RACE, MARST, NCHILD, NCHLT5, ELDCH, YNGCH, HISPAN, EMPSTAT, LABFORCE, UHRSWORKT, AHRSWORKT, WKSTAT, EDUC, WORKLY, WKSWORK1, FULLPART, NUMEMPS, INCTOT, INCWAGE, INCINT, INCDIVID, INCRENT, SRCEARN, OINCBUS, INCBUS, PUBHOUS, RENTSUB`
- **CPS monthly**
  - Samples: 1986m1–2019m12
  - Variables: `YEAR, SERIAL, MONTH, CPSIDP, PERNUM, WTFINL, STATEFIP, METAREA, METFIPS, RELATE, MOMLOC, MOMLOC2, AGE, SEX, RACE, MARST, NCHILD, NCHLT5, ELDCH, YNGCH, HISPAN, EMPSTAT, LABFORCE, UHRSWORKT, AHRSWORKT, EDUC, EDATT, EDGRADE, EDPUPR`
- **Census 5%**
  - Samples: 1990 and 2000
  - Variables: `YEAR, CLUSTER, STRATA, FAMUNIT, POPLOC, NCHILD, SEX, RACED, EDUCD, INCBUS00, SAMPLE, STATEFIP, GQ, FAMSIZE, SPLOC, NCHLT5, AGE, HISPAN, WORKEDYR, INCINVST, SERIAL, METAREA, PERNUM, SUBFAM, MOMLOC2, ELDCH, MARST, HISPAND, INCWAGE, TRANWORK, HHWT, METAREAD, PERWT, MOMLOC, POPLOC2, YNGCH, RACE, EDUC, INCBUS, TRANTIME`

The replication package includes a script that (optionally) downloads the raw SIPP files from the NBER collection of SIPP data and keeps only the necessary variables. The resulting "raw" SIPP files are also included in the replication package, so this step is not required and will not download any files if the replication package is run as-is.

4 Scripts

The package contains eight sets of do files detailed below. The file _master.do calls all the scripts below in order (and also contains much of the information contained here).

A: Extract IPUMS files

The script in this section builds raw CPS, ASEC and Census 5% files from IPUMS .dat/.dct/.do files downloaded via the IPUMS API (a Python script to call the IPUMS API and download the extracts is not included in the replication package). This file only does anything if CPS/ASEC/Census .dta files are not yet extracted from IPUMS raw files (.dat and .dct). This step is not necessary for replication and is kept here only for completeness.

- A01_ipums_build.do

B: Secondary datasets

These files set up all the necessary secondary data, including tax rates, welfare waiver indicators, EITC parameters, consistent metro area identifiers, and several covariates used in regression modeling.

- B01_taxrates.do
- B02_exempt.do
- B03_states.do
- B04_eitc.do
- B05_metarea_xw.do
- B06_county_metarea.do
- B07_qcew.do
- B08_county_pops.do
- B09_hud_data.do

C: CPS/ASEC initial data setup

These files perform the initial data setup steps for the CPS ASEC and monthly CPS datasets as well as for the Census 5% data. They produce intermediate files with the suffix "_pre_final" as well as a few supplemental data files for variables used as covariates in some analyses (e.g., metro-level public housing indicators in pubhous.dta). The Census file creates some variables necessary for the simulated EITC creation in D01.

- C01_asec_setup.do
- C02_cps_setup.do
- C03_census_setup.do

D: Simulated EITC computation

This file creates the simulated instrument by: (1) building a dataset that projects individual-level economic characteristics forward in time from the 1990 5% Census base, (2) using NBER taxsim to calculate EITCs for these simulated recipients, and (3) collapsing these by year x metro x marital x family-size cells. This is repeated for the 2000 base year. This file also creates a Stata log file eitc_fedtax_stats.smcl that reports two statistics about simulated EITC values and tax refunds reported in the paper.

- D01_taxsim_eitc.do

E: Finalize CPS/ASEC datasets

These files finalize the ASEC, monthly CPS, and Census datasets. The initial file (E00) combines the various metro-level covariates. The next three files (E01-E03) join this dataset as well as state-level data and the simulated instruments to the primary datasets.

- E00_metarea_controls.do
- E01_asec_finalize.do
- E02_cps_finalize.do
- E03_census_finalize.do

F: CPS/ASEC regression models

These files run the regression models and save the results.

- F01_asec_regs.do
- F02_cps_regs.do
- F03_census_regs.do

G: CPS/ASEC output

These files create the tables and figures derived from CPS, ASEC, and Census 5% data and analyses. The summary_stats file also includes a few one-off calculations and estimates for a few numbers cited in the study.

- G01_output.do
- G02_summary_stats.do

S: SIPP download, setup, regressions, output

These files mirror the data setup and analysis above, but for SIPP as the primary data source. This section is not a standalone one, however, as it relies on some data files created in previous steps. The first file downloads raw SIPP files directly from the NBER website and keeps only the necessary variables. Since the raw SIPP files that result from this script are already included in the replication package, this step will be unnecessary; it is kept here only for completeness.

- S00_download.do
- S01_metarea_xw.do
- S02_topicsetup.do
- S03_datasetup.do
- S04_census_setup_sipp.do
- S05_taxsim_eitc_sipp.do
- S06_sipp_finalize.do
- S07_sipp_regs.do
- S08_sipp_output.do

5 Other considerations

- As noted above, the replicator will need to manually replace the directory at the top of the _master.do file with their local path.
- Because the construction of the simulated instrument requires some random sampling (as detailed in the main text), final results are likely to differ slightly from those the published paper. These differences should be quite minor.

- I use a custom Stata command called `frame2` to simplify the creation and use of frames in Stata. The replicator will need to copy frame2.ado into one of their ado-path directories, which can be identified using the command `adopath` in Stata. Calling `frame2 <name>` is the same as successively calling `frame change default`, `cap frame drop <name>`, `frame create <name>`, and `frame change <name>`.
- The following community-contributed programs will also need to be installed in Stata: reghdfe, taxsimlocal35, grc1leg2, and coefplot.
- The command `import fred` is used, which requires the user to obtain a FRED API and set this value in Stata (see here).
- The entire replication took about 14 hours to run on a 2019 MacBook Pro with a 2.6 GHz processor and 32 GB of RAM, Stata 17/MP.