National Experimental Well-being Statistics (NEWS)

CNSTAT Meeting
Toward a Vision for a New Data Infrastructure for Federal Statistics and Social and Economic Research in the 21st Century

Jonathan Rothbaum
Research Economist
Economic Characteristics
Social, Economic, and Housing Statistics Division

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

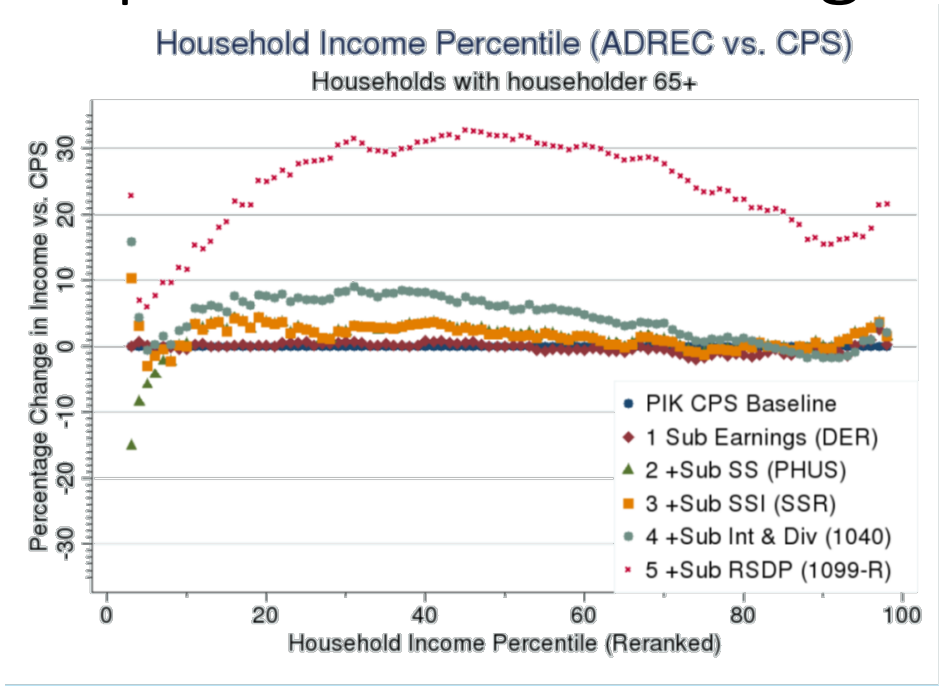
Using Linked Census/Survey and Administrative Data

- Linked microdata – NEWS project, Comprehensive Income Dataset (CID), Longitudinal Employer-Household Dynamics (LEHD), Opportunity Atlas, etc.
- Modeling – Small Area Income and Poverty Estimates (SAIPE), Small Area Health Insurance Estimates (SAIHE)

Goals for NEWS

- Rethink how we can produce income and resource statistics
  - What is the best possible estimate given all the data currently available at Census for a given income/resource statistic?
  - Expand the set of income and resource statistics we produce

Why Does This Matter?
Misreporting Example – Income for Age 65+ Households



Source: Bee and Mitchell, 2017, "Do Older Americans Have More Income Than We Think?" using 2013 CPS ASEC linked to W-2, 1040, and 1099-R forms for persons 65+.

Other Goals

- Experimental
  - Updated regularly with additional data and better methods
  - Expand income/resource concepts being measured
  - Longer term – move to regular production
- Transparent and replicable
  - Decisions about how to use survey and administrative income are well-documented, supported, and apolitical
  - Create linked microdata and code database that is accessible through the RDC system
  - Long term – create a set of synthetic data sets (akin to the SIPP Synthetic Beta) for public release?
- Timeline
  - 2022 – 1st set of statistics for a year or small number of years
  - 2023- Additional statistics, additional years, improved methods,…

Which Statistics?

1. Annual Income, Resource, and Poverty Statistics
   - Same general statistics we produce in existing official reports (simple moments, distributional/inequality statistics, poverty, etc.)

Which Statistics?

2. Longitudinal Income, Resource, and Poverty Statistics – MOVS project (Mobility, Opportunity, and Volatility Statistics)
   - Income and earnings dynamics

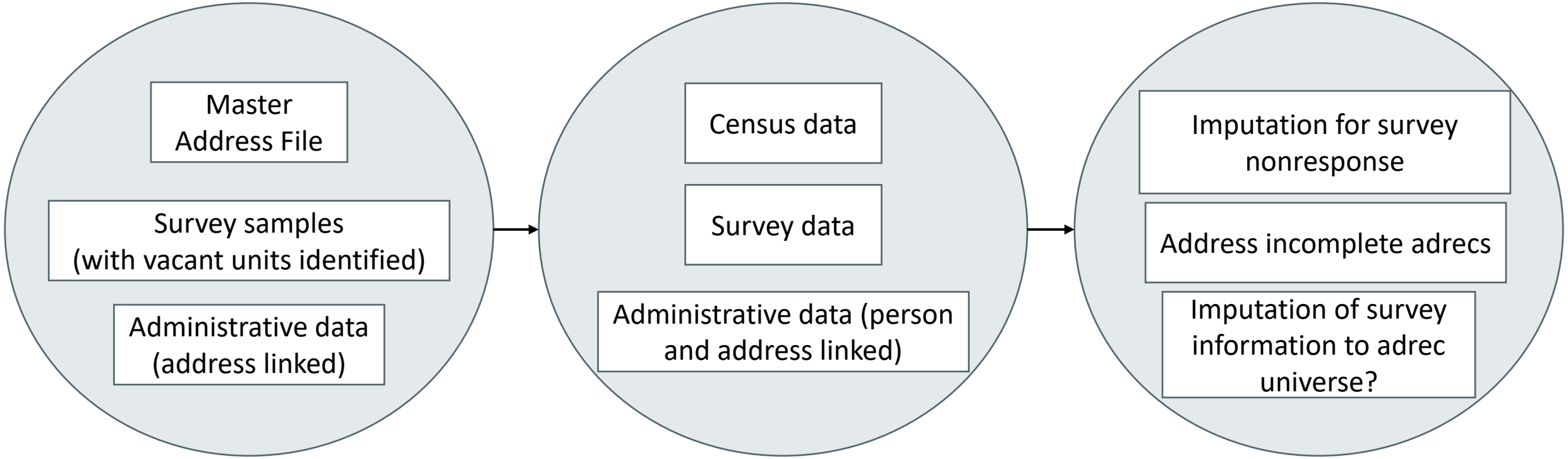
Source Data – Survey and Census Data

- Information not available in administrative data
  - Demographics and socioeconomic characteristics (Race, education, etc.)
  - Income and benefits – address linkage and income coverage issues
  - Survey frames potentially provide sampling information needed for estimates (random sampling + vacancy assessments)
- Including:
  - CPS ASEC
  - ACS
  - Decennial Census

Source Data – Administrative Data

- IRS and SSA income data
  - 1040, W-2, 1099-R, 1099-IRMF, DER, social security and SSI payment data, etc.
- LEHD
- Numident
- Master Address File
- State and federal program data
  - SNAP, TANF, WIC, HUD, VA, Medicare/Medicaid data, etc.
- Firm Data
  - Business Register, Longitudinal Business Database, Form 5500 filings
- Third-Party Data
  - Black Knight data on home values

General Approach



Frame
(Housing-unit level)

Extract
(Individual/Household Level)

Processing and Estimation

Challenges

- Measurement error in administrative data – earnings in particular
- Linkage challenges – incomplete linkage and errors in linkage
- Coverage and representativeness
- Incomplete geographic coverage of administrative data
- Conceptual misalignment or incomplete income coverage in administrative data
- Timeliness/availability of administrative and survey data
- Changes in administrative data that may be unrelated to changes in the underlying income or resources

Described in “The Administrative Income Statistics (AIS) Project: Research on the Use of Administrative Records to Improve Income and Resource Estimates” (Bee and Rothbaum, 2019)

Challenge
Measurement Error in Administrative Data

- Earnings - 80% of income
  - Wage and salary earnings is probably the best reported of any income category in surveys (70% of income)
    - Particularly for aggregates and extensive margin agreement
    - Still, error in earnings matters more than in any other income type

Challenge
Measurement Error in Administrative Data

- Wage and salary – under-the-table earnings
  - Detailed occupation level differences in administrative and survey earnings largely match expectations about workers that are likely to be paid under the table (construction, food service/bartending, etc., from Bollinger et al., 2015 and our work with linked ACS data)
- Self-employment – tax avoidance
  - Confirmed by audit studies and consumption/income relationship for the self-employed
  - Nearly ½ of self-employment income in the National Income and Product Accounts is imputed due to under-reporting to the IRS

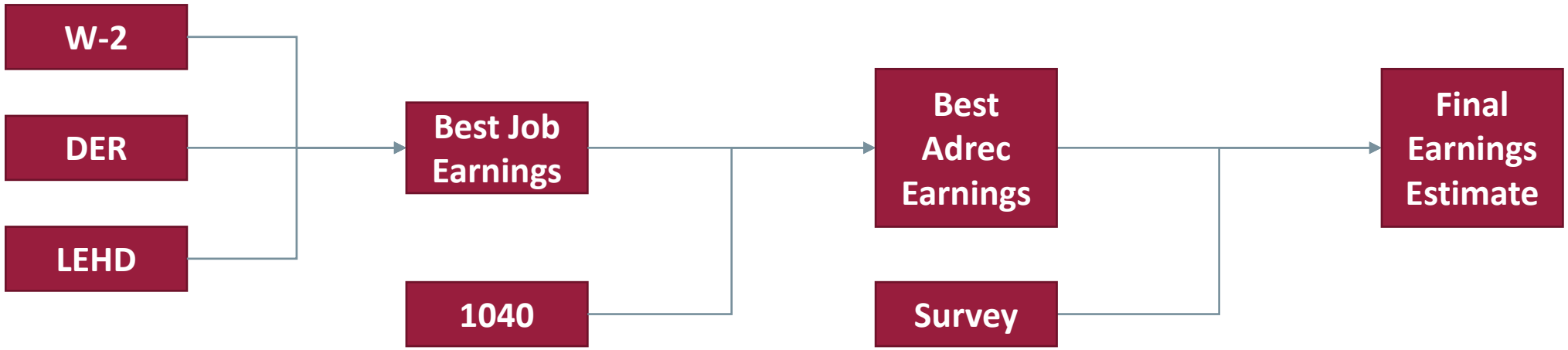
Combining Survey and Adrec Earnings
(Bee, Mitchell, and Rothbaum 2020)

1. Use job-level Information to get "best possible" administrative job-level earnings

2. Compare to 1040 to check for missing earnings (at tax-unit level)

3. Compare to survey and decide for which individuals to use adrec or survey earnings

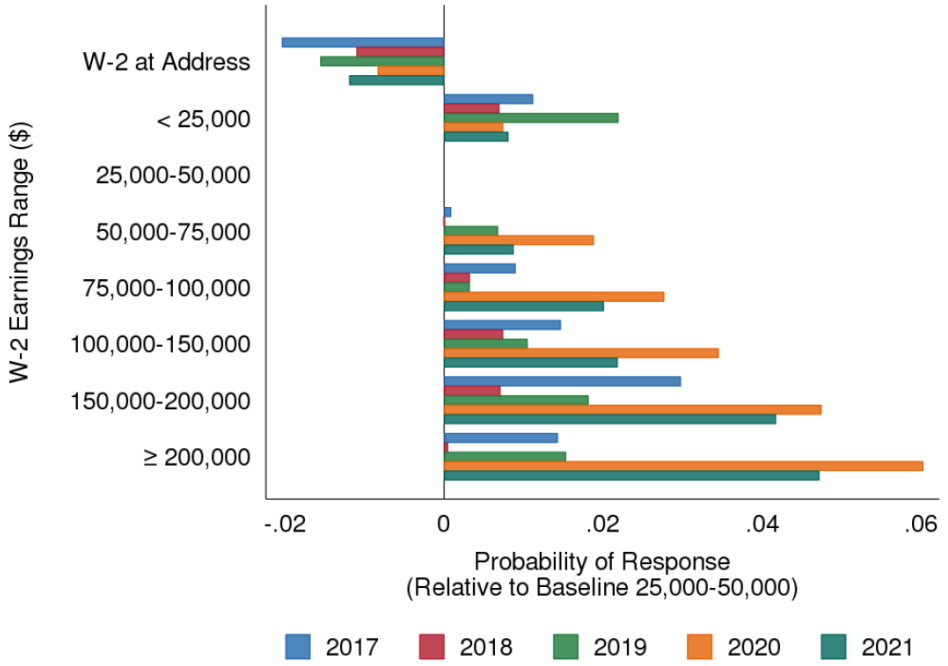
4. Final "best" estimate of earnings for each individual/household



Challenge
Coverage and Representativeness

Response Propensity by W-2 Earnings at Address

(Controlling for linked demographics, such as race and age)



Source: Rothbaum and Bee, 2020. "Coronavirus Infects Surveys, Too: Nonresponse Bias During the Pandemic in the CPS ASEC"

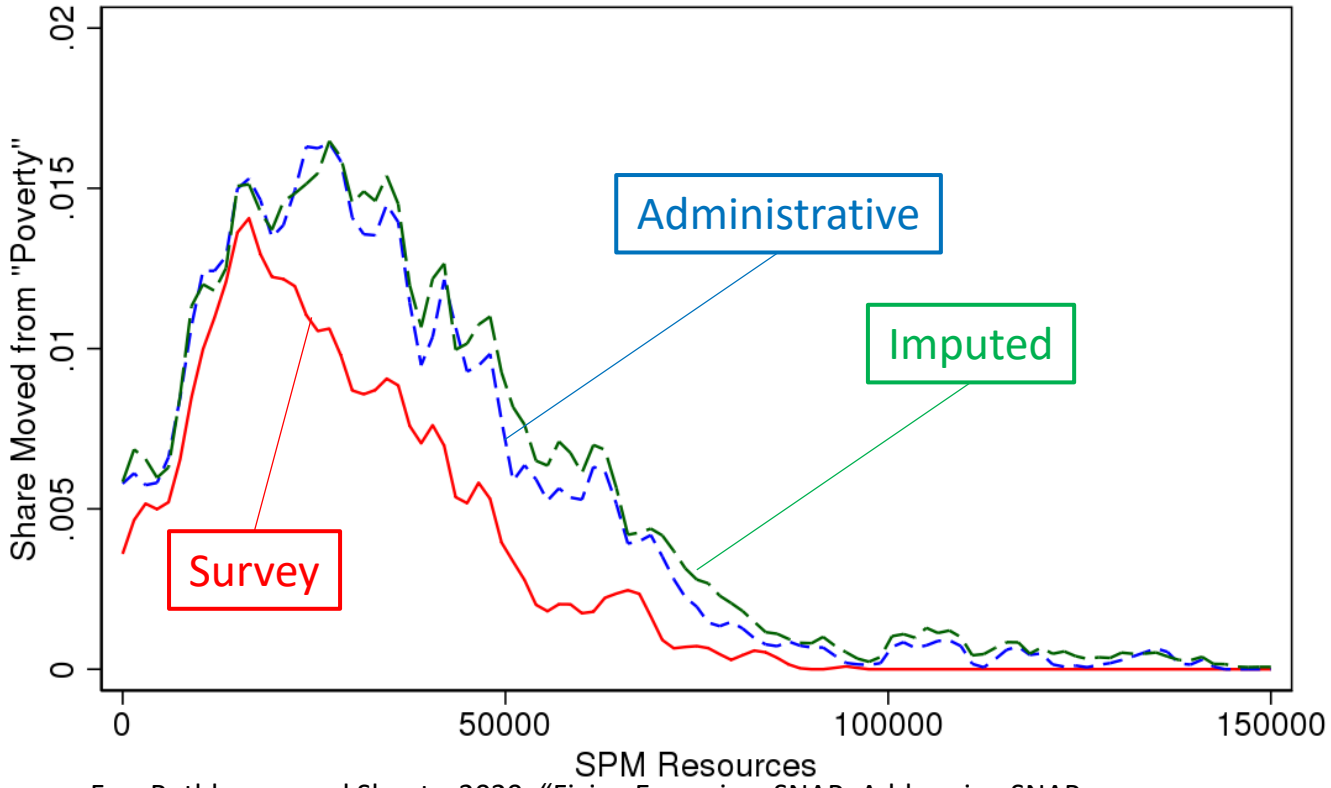
Challenge
Linkage Issues

- Addressing misreporting
  - ~10% of individuals in a survey cannot be linked to their SSN
- Representativeness/Weighting
  - Administrative records may come from nonrepresentative samples
  - Surveys have random samples but nonrandom selection into response
- Linkage error – understudied

Challenge
Incomplete geographic coverage of administrative data

- Some data is only available for some locations (and in some years)
- Examples
  - **SNAP**
  - LEHD (in some years)
  - TANF
  - WIC
- Missing information problem - Impute

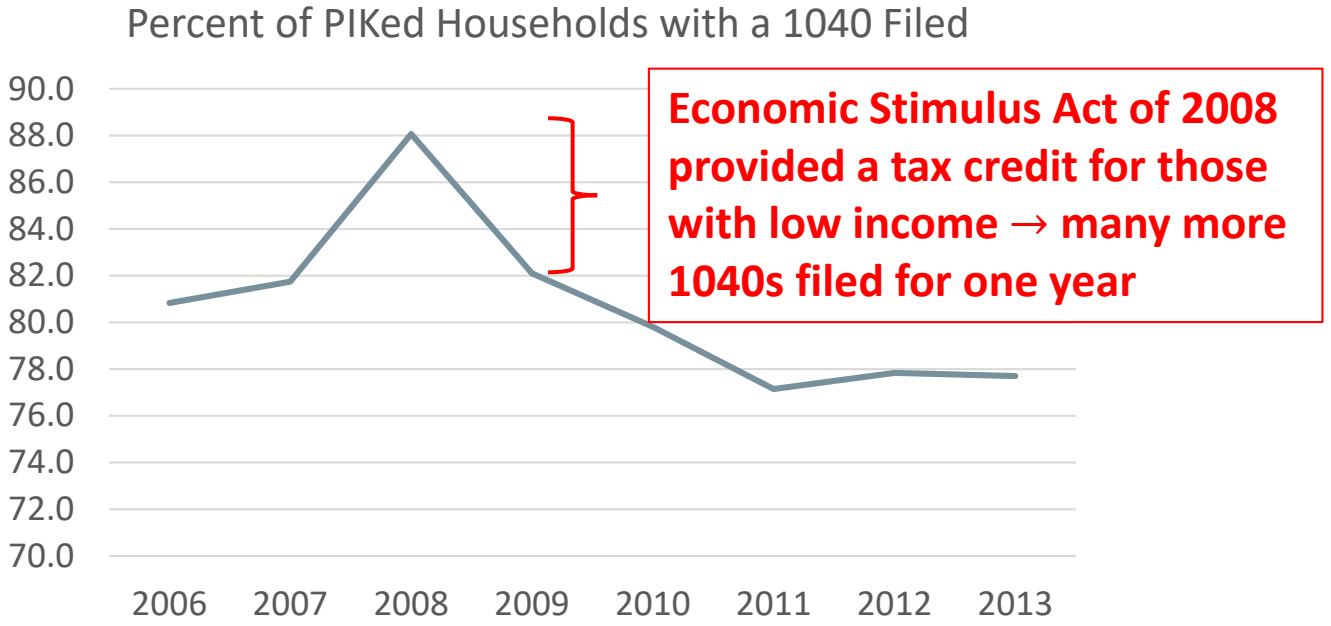
Addressing Incomplete Geographic Coverage Imputing to States without Adrecs



Source: Fox, Rothbaum, and Shantz, 2020. "Fixing Errors in a SNAP: Addressing SNAP Under-Reporting to Evaluate Poverty"

Changes in Administrative Data

- Can change over time due to statutory/regulatory changes that affect programs and agencies



Source: Rothbaum, 2018
2006-2013 ACS Linked to IRS records.

Changes in Administrative Data

- Can change over time due to statutory/regulatory changes that affect programs and agencies
  - Auten and Splinter (2018) argue that much of the inequality increase in tax data from 1960 to present in work by Piketty, Saez, and Zucman is due to changes in the tax code and the nature of tax reporting, not in actual underlying income changes

Contact Information

Jonathan Rothbaum

Research Economist

Social, Economic, and Housing Statistics Division

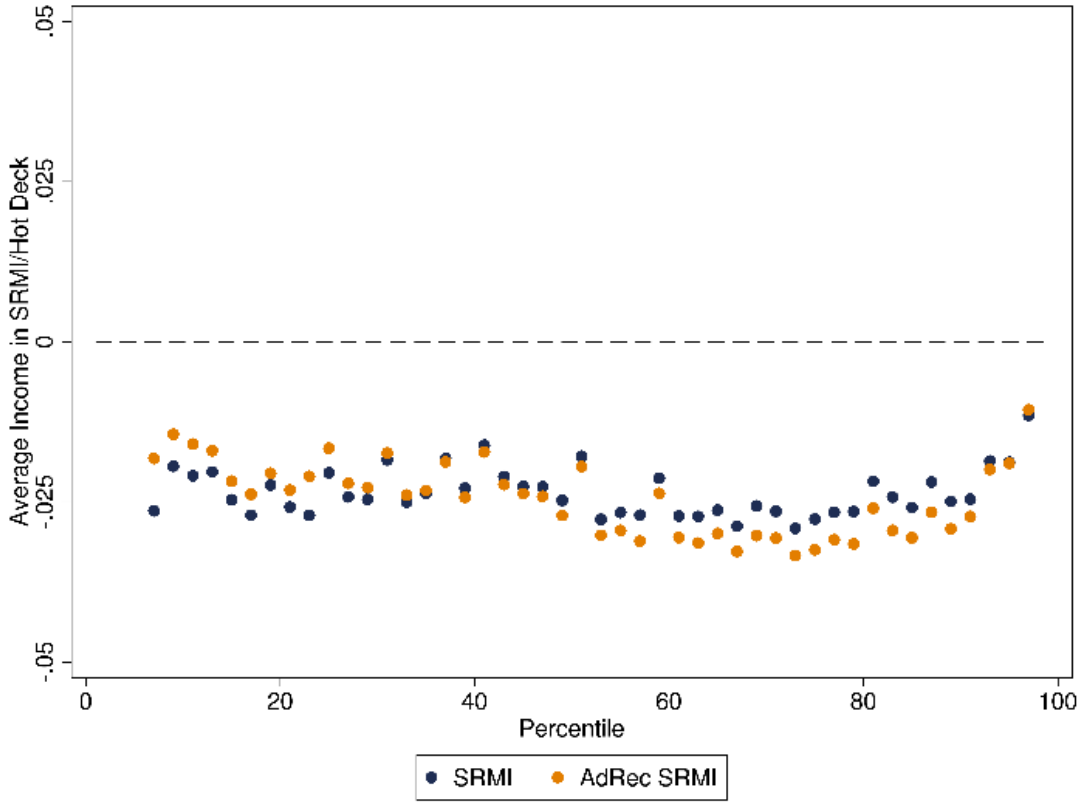
jonathan.l.rothbaum@census.gov

(301) 763-9681

Extra Material

Combining Survey and Adrec Earnings

Survey Nonresponse



- Improved imputation
  - Poverty ↑
  - Median household income ↓
  - Inequality ↑

Source: Hokayem, Raghunathan, and Rothbaum, 2022, "Match Bias or Nonignorable Nonresponse: Improved Imputation and Administrative Data in the CPS ASEC"

Coverage and Representativeness
Our Strategy

- Use random samples from surveys + vacancy classification (since adrecs exist at vacant addresses)
- Link all non-vacant addresses to as much available data as possible
  - 1099 IRMF – link addresses to people
  - Adrec universe income data – 1040, W-2s, LEHD, 1099-Rs
  - Survey and census data – short form census (race), ACS (education for linkable individuals)
- Estimate the distribution of characteristics in the linked sample, including respondents and nonrespondents
- Reweight the full sample of individuals in the data (whether from a given survey or administrative data linked to several surveys, …) to match that distribution of linked characteristics *and* known population aggregates

Rothbaum and Bee, 2020. “Coronavirus Infects Surveys, Too: Nonresponse Bias During the Pandemic in the CPS ASEC”

- We implement this approach with the CPS ASEC
- Use entropy balancing
  - Hainmueller, 2012. “Entropy balancing for causal effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies”
  - Estimates weights subject to set of constraints (set of target moments)
    - Minimizes “distance” between initial weights (such as probability of selection into a survey sample) and final weights given the constraints

Entropy Balancing

- Can be used to reweight any linked data set to any set of targets in the linked data (as would be the case in an inverse probability weight approach) and external aggregates (as wouldn't be)
- Efficiently reweight to very high-dimensional set of targets
  - External aggregates – existing CPS ASEC pop targets race x age x gender x state cells
  - Linked data targets – summary stats from linked 1040s, 1099s, W-2s, census, and ACS data