Match Bias or Nonignorable Nonresponse? Improved Imputation and Administrative Data in the CPS ASEC

Charles Hokayem
U.S. Census Bureau

Trivellore Raghunathan
University of Michigan

Jonathan Rothbaum
U.S. Census Bureau

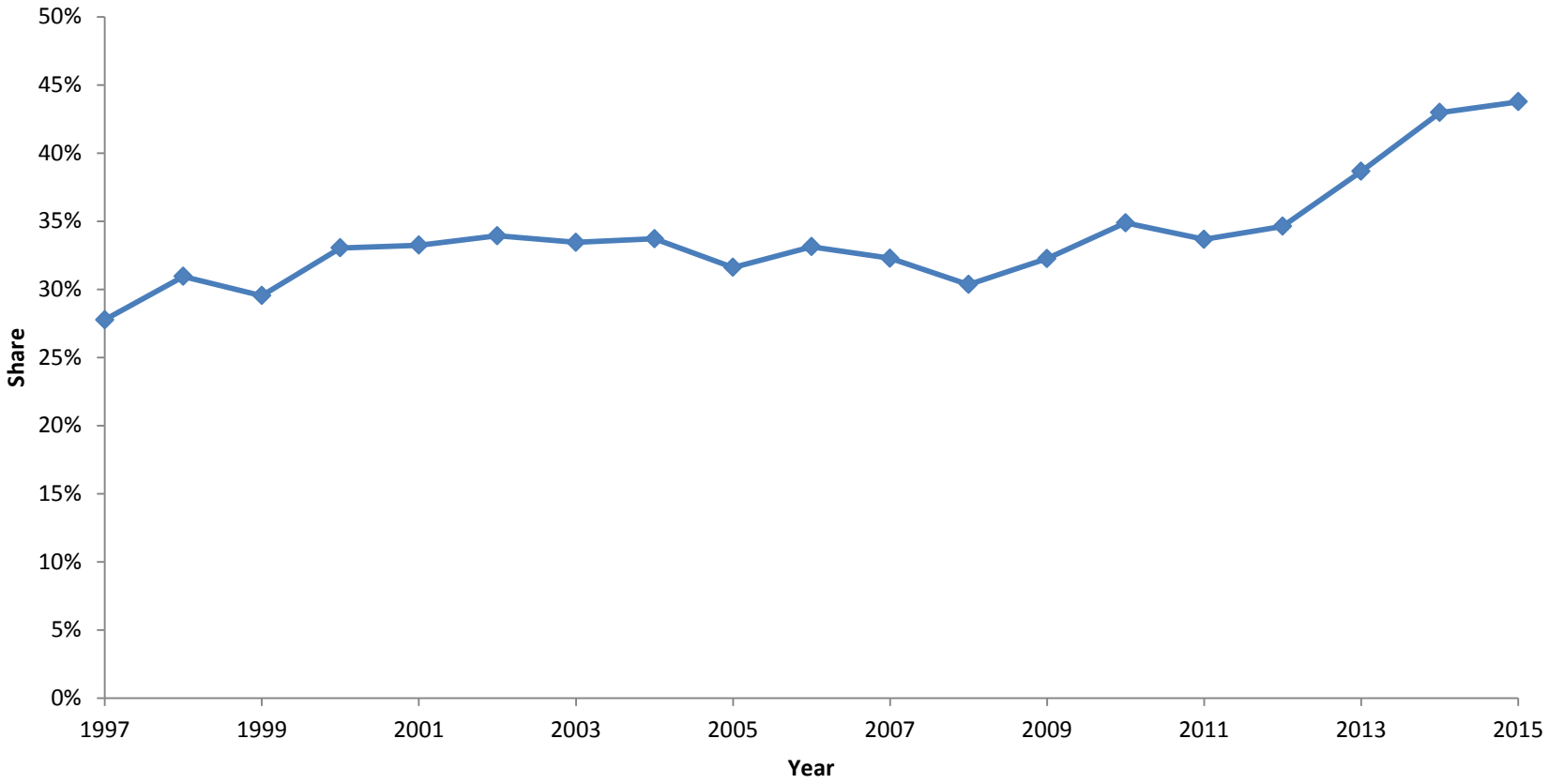
APPAM Fall Research Conference
November 4, 2017

Disclaimer: This presentation is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Issues Facing Income Surveys

- Increasing nonresponse
  - Unit (no information at all)
  - **Item (no information for a particular question)**
- Measurement Error/Misreporting

Share of All Income Imputed



Source: Author's calculation from the CPS ASEC

Motivation

- Non-response is a growing problem in surveys, including the CPS ASEC
- Hot deck procedure for imputing non-response in CPS ASEC has been in place with few changes since 1989
- Explore two possible biases in current imputation
  1. Match Bias – compare hot deck to model-based method that permits more covariates
  2. Nonignorable nonresponse – add administrative data to model to evaluate impact of nonignorable nonresponse on data

Missing at Random and Match Bias

- Missing at Random – assumed by nearly all imputation models
  - Given Observables O, Unobservables U, and R as response indicator

$$p(R = 1|O, U) = p(R = 1|O)$$

  - For a given statistic Q:

$$E(\hat{Q}|O, U) = E(\hat{Q}|O) = Q$$

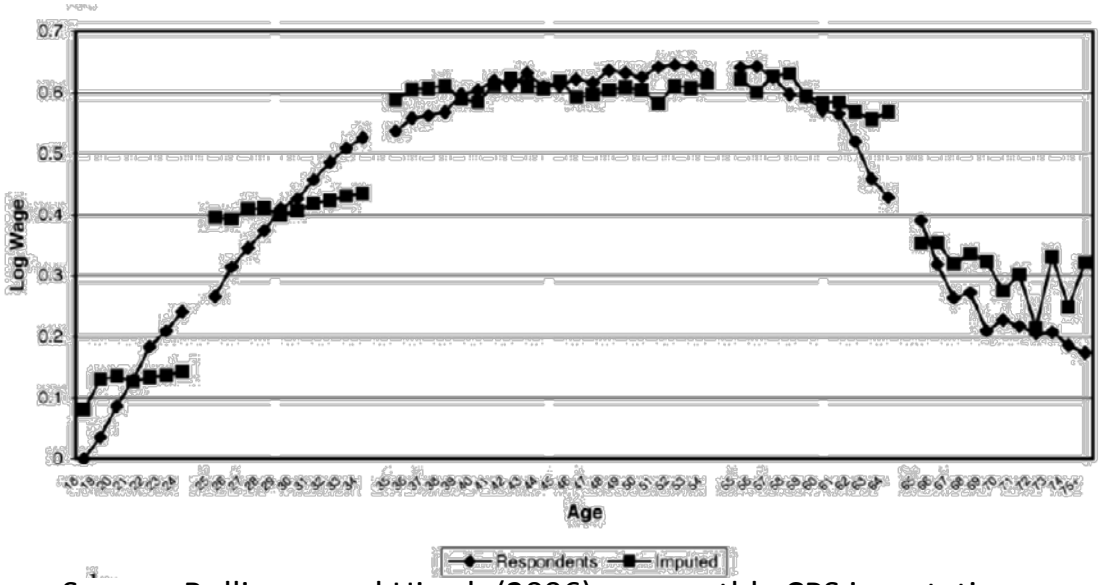
- Match bias – only a subset of variables are in the model (M) and:

$$E(\hat{Q}|O) \neq E(\hat{Q}|M)$$

  - Exclusion of $O_{\backslash \mathrm{M}}$ biases results ($O = \{M, O_{\backslash M}\}$

Match Bias - Examples

- Union status (Hirsch and Schumaker, 2004) – not in CPS imputation model
  - Estimates of wage differences between union/non-union worker attenuated by imputation model's assumption that there is no relationship conditional on M
- Earnings and Experience (Bollinger and Hirsch, 2006) – in CPS model, but grouped
  - Attenuates estimates of returns to experience



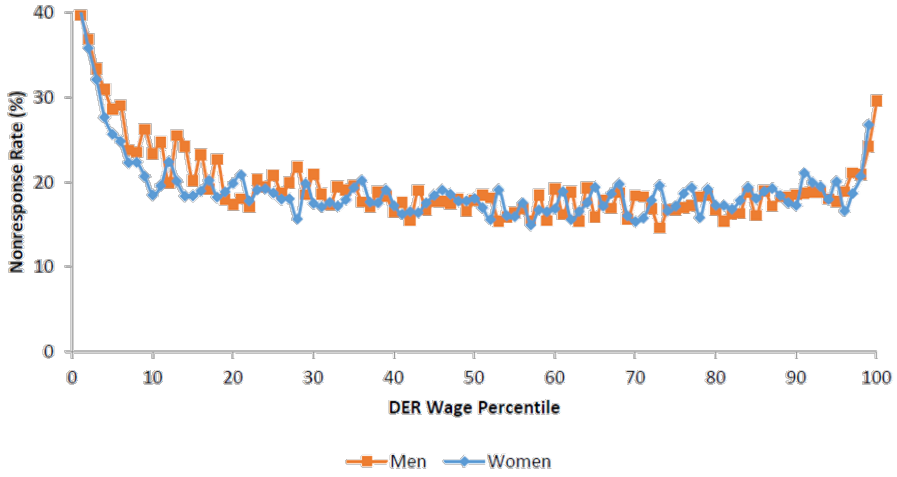
Source: Bollinger and Hirsch (2006) on monthly CPS imputation.

Nonignorable Nonresponse

- Data not missing at random

$$E\left(\hat{Q}\middle|O,U\right) \neq E\left(\hat{Q}\middle|O\right)$$

  - Exclusion of U biases results
- Example - Trouble in the Tails (Bollinger et al., 2015)
  - Nonresponse is a function of the missing variable, earnings



Source: Bollinger et al. (2015) from CPS ASEC linked to W2 records

Existing Approach: Hot Deck

- Match non-respondents to "similar" respondents along a set of characteristics in the model
- Donate response as imputation from respondent to non-respondent

- Example: 2 variables, 2 categories each – 4 cells
  1. Race: White/non-White
  2. Gender: male/female

  Two non-respondents (A and B)

  Person A: white, female – randomly select a white, female respondent and use her response as the imputed value

  Person B: non-white, male – randomly select a non-White, male respondent and use his response as the imputed value

Hot Deck Limitations

- Dimensionality
  - Limited number of variables can be included
    - Suppose there are 20 variables you believe are correlated with your outcome of interest
    - Divide each into only 3 categories
    - $3^{20} \approx 3.5$ billion possible cells for each individual
  - Must exclude predictors from the model
- Implied model places emphasizes all possible interaction terms of a small set of variables over the inclusion of more predictors
  - Equivalent to imputation by a regression model with dummies for each variable/category + all possible interactions with random draws from errors (within variable/category strata)

Hot Deck Limitations – Playing out in the CPS ASEC

Variables and Categories for "Earnings from the Longest Job Only" Hot Deck Match

	Match Level					
Match Variable	1	2	3	4	5	6
Sex	2	2	2	2	2	2
Race	3	2	2			
Age	9	6	3	3		
Relationship	7	7	4	4	4	
Years of School Completed	6	5	5	4	4	4
Marital Status	4	4				
Presence of Children	3					
Labor Force Status of Spouse	3					
Weeks Worked	5	5	4	4	4	4
Hours Worked	3	3	3	3	2	
Occupation	528	528	66	66	66	
Class of Worker	5	5	5	3	3	3
Other Earnings	8	8				
Type of Residence	3	2	2			
Region	4	4				
Transfers payments receipt	2	2	2	2		
Number of Donor-Recipient Cells	**620,786,073,600**	**17,031,168,000**	**3,801,600**	**456,192**	**50,688**	**96**
Percent of Missing Matched (Weighted)	**6.8**	**14.6**	**52.7**	**12.7**	**8.2**	**5.0**

- SRMI
  - Flexible imputation technique
  - Fixes issue of sequential imputation
    - Another source of match bias – cannot condition on Y_2 in model for Y_1 with current approach
- Regression Models
  - Allow inclusion of additional variables in model

Data

- 2011 Current Population Survey Annual Social and Economic Supplement (CPS ASEC)
  - Survey of ~100,000 addresses
    - About 200,000 individuals
  - Official source of US poverty estimates
  - Income from 2010 calendar year
- Social Security Administration Detailed Earnings Records (DER)
  - W-2 data linked to CPS ASEC using Protected Identification Key (PIK)
  - Includes W-2 earnings, deferred contributions (i.e. 401k), and reported SSA covered self-employment earnings

Nonresponse by Income Type

Variable	Non-response rate (%)	Share of Income Imputed (%)
Earnings Recipiency	0.1	
Wage Earnings (Primary Job)	12.7	20.7
Social Security	4.4	23.9
Interest Income	16.5	59.7
Supplement Non-response	12.9	12.9
Total Non-response		
Any Recipiency	22.7	
Any Value	44.2	34.7

Note: Share of income imputed is for income in the given category. For Supplement non-response and total non-response, the share is of all income in the CPS ASEC.
Source: Authors' calculations from the 2011 CPS ASEC.

Modeling – Throw in the Kitchen Sink!

- Any imputation model assumes some $f(Y|O, U, \theta)$
- Regression models (f)
  - OLS for continuous variables
  - Logit for binary and categorical (separated into binary trees)
- Variables imputed (Y)
  - Recipiency and value for all income types (45 variables), weeks worked in previous year, hours worked per week, occupation (11 separate categories)
- Explanatory variables
  - Observables (O): Among others, includes gender, relationship to householder, education, marital/cohabiting status, spouse/partner earnings, number of children, urban/rural status, small or large metropolitan area, Census region, means-tested benefits, health insurance status and type, renter/homeowner, unemployment status, school enrollment, citizenship, race, age
  - Unobservables (U): DER – number of separate W-2 jobs, total wages, total self-employment earnings
  - Interaction terms for all possible combinations of a subset of Y and O variables
  - Over 3,000 potential predictors in DER SRMI (given recoding of categorical variables as sets of dummies)

Modeling Challenges

1. Handling non-normal distributions
   - Highly skewed
   - Bunching

2. Selecting variables to include in the regression models
   - Too many possible variables and interactions to pick from
   - Want to avoid imposing too many modeling assumptions

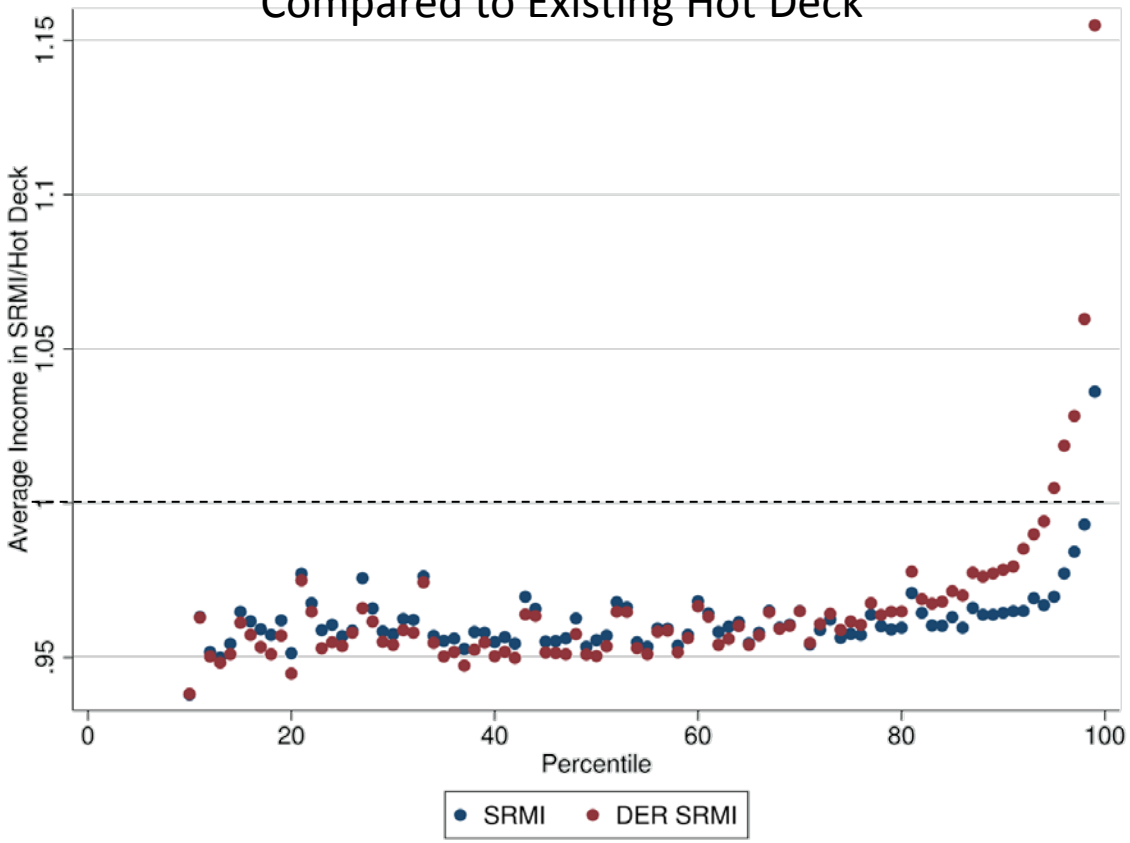
3. Accounting for model uncertainty

SRMI Steps

1. Empirical normal transformation to all continuous variables in y and U (Non-Normality)
2. Create all interaction terms
3. First model-selection stage for each Y_i (Too many variables)
4. Reverse empirical normal transformation (Non-Normality)
5. SRMI steps at each iteration
   a. Normal transformation again (Non-Normality)
   b. Calculate derived variables used as predictors (spouse, HH variables for example) and interaction terms
   c. Stratify sample by race and gender
   d. Impute each Y_i sequentially, where for each Y_i
      i. Select regression sample by Bayes' Bootstrap for each race-gender stratum (Model Uncertainty)
      ii. Within each stratum, run second stage model selection to select predictors (Too many variables)
      iii. By stratum, impute the missing value using logistic or OLS regression and sampling from error distribution
   e. Reverse transformation (Non-Normality)

Household Income by Percentile Relative to Official Estimates

Model-Based Imputation (with and without Tax Data)
Compared to Existing Hot Deck



Note: Figure truncated at 99th percentile for scale
SRMI: addresses match bias
DER SRMI: addresses nonignorable nonresponse for earnings

Poverty by Selected Characteristics

Characteristic	Poverty Rate		
	Hot Deck	SRMI (Correction for Match Bias)	DER SRMI (Correction for Nonignorable Nonresponse)
Total	15.1	15.9***	16.0***
Race and Hispanic Origin			
White alone, Non-Hispanic	9.9	10.5***	10.5***
Black alone	27.4	29.8***	30.4
Hispanic (of any race)	26.5	26.9	27.2
Children (< 18)	22.1	21.0***	21.1***
Aged 65+	9.0	9.3	10.0

Match Bias – children dropped from imputation for 93% of earners and marital status for 80%

Asterisks are for statistical significa
and * at 0.1 level). No differences
and DER SRMI standard errors incorporate multiple imputation uncertainty. However, hot deck standard errors do not.

Median Household Income by Selected Characteristics

Characteristic	Hot Deck	SRMI	DER SRMI
All Households	49,445	47,144***	46,981***
Family Households	61,544	59,240***	59,153**
Race and Hispanic Origin			
White alone, Non-Hispanic	54,620	51,854***	51,875***
Black alone	32,068	30,292*	29,898*
Hispanic (of any race)	37,759	37,485	36,864

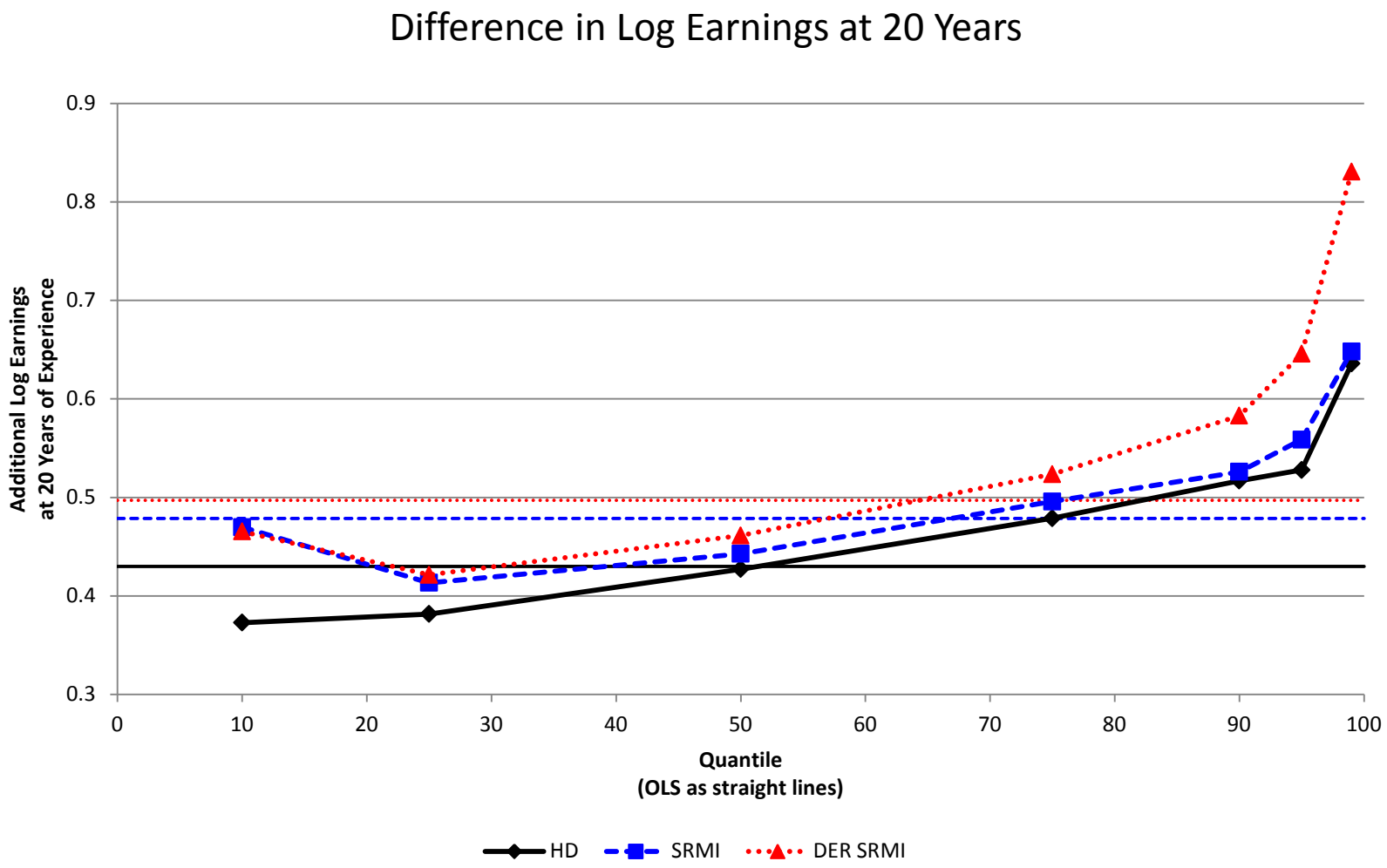
Asterisks are for statistical significance compared to the Hot Deck (*** at 0.01 level, ** at 0.05 level, and * at 0.1 level). No differences between SRMI and DER SRMI are statistically significant. SRMI and DER SRMI standard errors incorporate multiple imputation uncertainty. However, hot deck standard errors do not.

Inequality

	Hot Deck	SRMI	DER SRMI
Share of Income (%) in			
1st Quintile	3.3	3.0	2.9
2nd Quintile	8.5	8.0	7.6
3rd Quintile	14.6	13.7	13.1
4th Quintile	23.4	21.9	21.1
5th Quintile	50.3	53.4	55.2
Top 5 Percent	21.3	26.1	28.5
Top 1 Percent	7.8	12.9	14.9
GINI	0.470	0.503	0.521

All differences between SRMIs and Hot Deck are significant at the 1 percent level. All differences between SRMI and DER SRMI significant at the 5 percent level except 1st quintile (not significant) and the top 1 percent (10 percent level).

Returns to Experience (Mincer Earnings Regression)



Future Research

1. Add more sources of administrative records
   - 1040 information
   - 1099Rs
   - SSA Records – OASDI and SSI payments
   - State-provided means-tested program benefits

2. Add more years to understand if/how nonresponse bias has changed over time

3. Include more summary information by geography to better capture associations between state and local area characteristics

Contact Information

Jonathan Rothbaum

Chief, Income Statistics Branch

Social, Economic, and Housing Statistics Division

jonathan.l.rothbaum@census.gov

(301) 763-9681