Research on Imputation Methods for the Survey of Income and Program Participation

Martha Stinson, U.S. Census Bureau

Analyzing Complex Survey Data with Missing Item Values

National Institute of Statistical Sciences Workshop

October 17, 2014

SIPP Background

- Few changes made to actual production imputation methods in many years
- Census has done a major re-design of the SIPP from 2006 - 2013
- Opportunity to consider how we might change and update imputation for item non-response cases

Research on Data Quality

- Abowd and Stinson (REStat Dec 2013) examined measurement error in SIPP earnings in the 1990-1996 SIPP panels.
  - Compare job-level annual earnings between the SIPP and W-2 tax data with at least one month of SIPP data imputed
    - white males with graduate degrees: SIPP 6% lower than W-2s
    - White females with graduate degrees: SIPP 8.6% higher than W-2s
  - Concern: Imputed values push everyone to the overall mean
  - In a mixed-effects model with person, employer, and “source” random effects, found that the variance of the DER effect is higher than the SIPP for imputed cases

Research on Methods

- Benedetto and Stinson (2009)
  - Modifications to SIPP imputation methods:
    - Model-based approach – helps handle small stratifying cell size problem
    - Use administrative data to mitigate problems caused when survey data are not “missing at random”
    - Multiple imputation – take account of the variance introduced by imputation
  - Tested on SIPP monthly earnings in first year of 2004 panel

Imputation Results

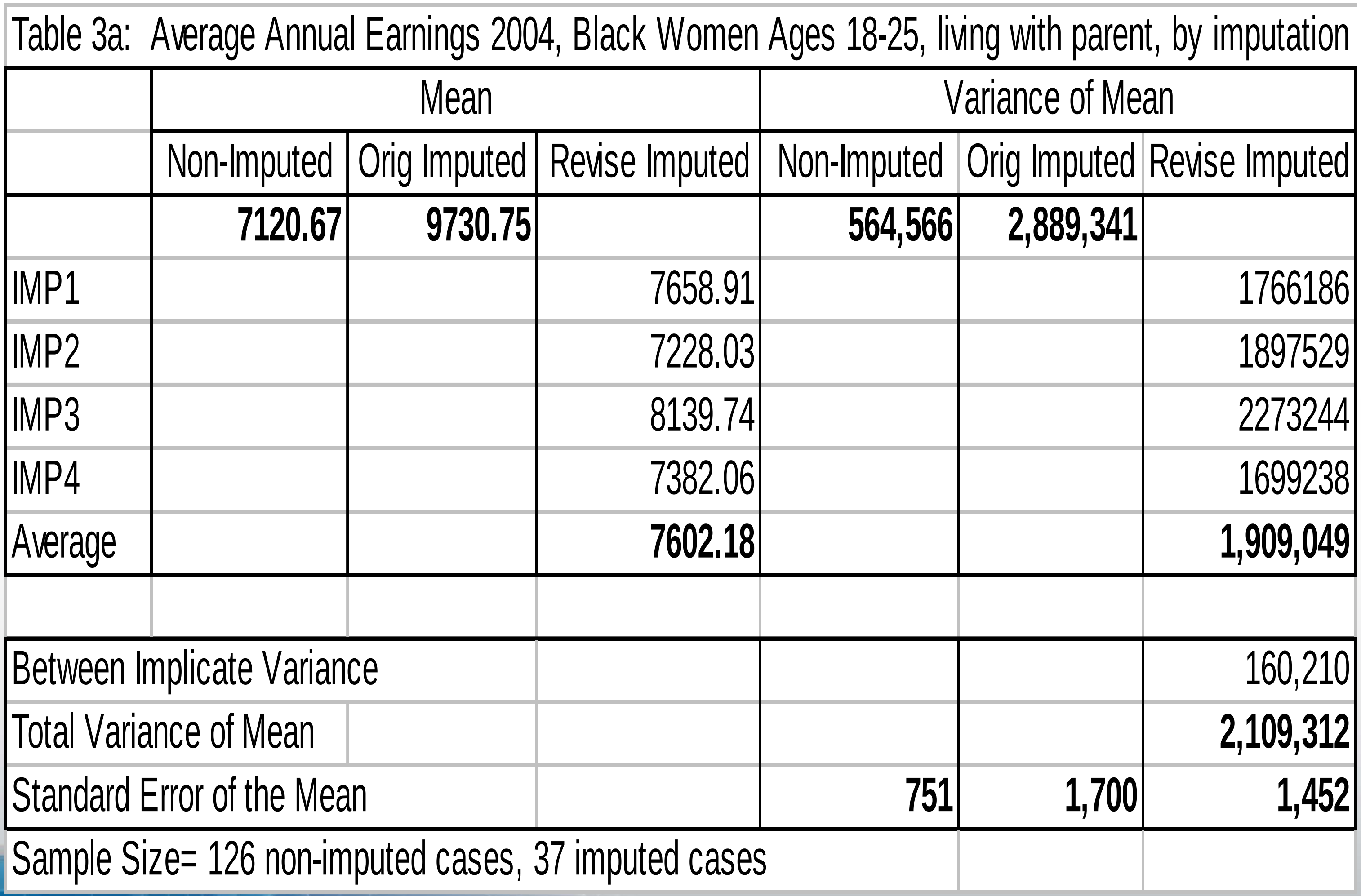

Table 3a: Average Annual Earnings 2004, Black Women Ages 18-25, living with parent, by imputation

	Mean			Variance of Mean		
	Non-Imputed	Orig Imputed	Revise Imputed	Non-Imputed	Orig Imputed	Revise Imputed
	7120.67	**9730.75**		**564,566**	**2,889,341**	
IMP1			7658.91			1766186
IMP2			7228.03			1897529
IMP3			8139.74			2273244
IMP4			7382.06			1699238
Average			**7602.18**			**1,909,049**
Between Implicate Variance						160,210
Total Variance of Mean						**2,109,312**
Standard Error of the Mean				**751**	**1,700**	**1,452**
Sample Size= 126 non-imputed cases, 37 imputed cases						

2014 SIPP Production

- Question faced by SIPP Survey Director:
  - How to implement new imputation methods and still release data in a timely manner for a survey with 11,000 variables?
- Solution
  - Topic Flags: indicator variables for all the major topics covered by SIPP
  - Use model-based imputation and administrative data to impute these flags

Description of topic flags

- Survey Instrument is divided into subject areas
  - “lines” in the Event History Calendar
  - question blocks after the Event History Calendar
- Each subject has 1 or 2 screeners that determine if a respondent is asked the questions for that topic.
  - “Do you currently have a job or business or do any kind of work for pay?”
  - “Did you have a job or business or do any kind of work for pay at all since January 1, 2013?”
- Topic flags will summarize information contained in the screeners:
  - = 1 if the respondent answered “YES” to either screener
  - = 0 if the respondent answered “NO” to both screeners
  - = missing if the respondent skipped the topic completely

Purpose of topic flags

- Measure missing data
  - We will be able to quantify how many topics each respondent answered
- Facilitate imputation of missing data
  - Stop whole-person substitution when interview didn't reach point designated as "sufficient partial"
  - Allow whatever data is reported to be used
  - Handle missing topics consistently whether missing one or all topics
  - Many more RHS control variables
    - Administrative data
    - Reported information from other family members
- Use in downstream edits:
  - Each topic will use its flag to set the universe for who receives edited data for questions about that topic
  - Flags from other topics can be used in imputations

List of Topic Flags in 2014 SIPP

EHC topics:

- Education Enrollment
- Employment (job lines 1-7)
- General Assistance
- SNAP
- SSI
- TANF
- WIC
- Health insurance
  - Private
  - Medicaid
  - Medicare
  - Military
  - Other

Non-EHC topics:

Alimony received
Biological Parent (fertility)
Children living outside the home
Child support paid
Child support received
Dependent care
Disability (has a disability: seeing, hearing, etc.)
Disability (not being able to work because of disability)
Disability payments
Energy Assistance
Foster child support received
Lump Sum Payments
Retirement
Retirement payments
School lunch
School breakfast
Social Security- Adults
Socials Security- Kids
Survivor payments
Unemployment compensation
Veterans affairs benefits
Worker's compensation

Imputation Methodology

- Sequential Regression Multiple Imputation (SRMI)
  - Raghunathan, Lepkowski, van Hoewyk, Solenberger (2001) *Survey Methodology,* "A Multivariate Technique for Multiply Imputing Missing Values Using a Series of Regression Models"
  - Iterative Method of arriving at the Posterior Predictive Distribution (PPD)
    - Prob(Y given X, β, and σ^2)Prob(β, σ^2 given X)
  - Iterations used to handle our non-monotone missing data

Imputation of Administrative Data

- Use Bayes Bootstrap to find donors for administrative indicator variables
  - Did respondent have positive W-2 earnings?
  - Did respondent receive OASDI benefits?
- Use Linear Regression to model continuous variables
  - W-2 earnings amount
  - OASDI benefit amount
  - Age began receiving OASDI benefit
- Apply a KDE transform method to continuous variables before modeling to make distribution more approximately normal.
  - Benedetto and Woodcock (2009) *Computational Statistics and Data Analysis* 53 (12)

Imputation of Missing SIPP Data

- Some demographic characteristics already imputed by hot deck (age, gender, education, race, ethnicity, links to family members)
- Topic flags imputed using logistic regression models
- After first SRMI iteration, merge parent and spouse administrative and topic flag variables onto person’s record; re-merge after each subsequent iteration

More Research needed

- Recognize this is not frontier of statistics research
- Tension between research and implementation – engineering research needed
  - Models that can process a variable in 5 minutes or under
  - Systems that can manage large data sets

Next Steps for the SIPP

- Model respondent-reported earnings
- Model beginning and end of spells
  - Help mitigate seam bias
- Model more topics
  - Defined benefit pension contributions
- How to best take account of spouse/parent/sibling relationships in the data when modeling