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

Gary Benedetto, U.S. Census Bureau
Joanna Motro, U.S. Census Bureau
Martha Stinson, U.S. Census Bureau

FedCASIC

March 5, 2015

SIPP Background

- Few changes made to actual production imputation methods in many years
- All SIPP variables with missing values were imputed using hot decks
- 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

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
  - Implement new methods only for these 40+ variables

Modifications to SIPP imputation methods

- Replace item-level hot deck with parametric model-based approach
  - Helps handle small stratifying cell size problem
  - Allows inclusion of many more predictor variables
- Use administrative data to mitigate problems caused when survey data are not “missing at random”
- Use topic flags as alternative to whole-record donation for cases where respondent did not complete the majority of the survey

Description of topic flags

- Survey Instrument is divided into subject areas
- Each subject has 1 or 2 screener questions that determine if a respondent is asked the detailed 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 respondent held a job in 2013
  - = 0 if the respondent did not hold a job in 2013
  - = missing if the respondent skipped the topic completely

Purpose of topic flags

- Measure number of missing topics
- Facilitate imputation of missing data
  - Stop whole-person substitution
  - Preserve correlation across topics by estimating a joint distribution for imputation
  - Allow any reported data to be used, including from other family members
  - Use administrative data as additional predictors
- Use in downstream edits:
  - Topic flag sets the universe for follow-up questions
  - Flags from other topics can be used in edits and hot decks

List of Topic Flags in 2014 SIPP

- **Education Enrollment**
- **Employment (job lines 1-7)**
- **Program Participation**
  - General Assistance
  - SNAP
  - SSI
  - TANF
  - WIC
- **Health Insurance**
  - Private
  - Medicaid
  - Medicare
  - Military
  - Other
- **Biological Parent (fertility)**
- **Disability**
  - functional limitations
  - difficulty finding/keeping job

Other Sources of Income

- Disability Payments
- Energy Assistance
- Lump Sum Payments
- Retirement/Retirement Payments
- Life Insurance
- School Breakfast and Lunch
- Social Security- Adults
- Socials Security- Kids
- Survivor Payments
- Unemployment Compensation
- Veterans Affairs Benefits
- Worker's Compensation
- Payments to cover costs of Dependent Care

Imputation Methodology

- Sequential Regression Multivariate 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, Θ)Prob(Θ given X)

Imputation Methodology (cont.)

- Why SRMI?
  - Predictor variables, including admin. data, have missing values with non-monotone missing patterns
  - Easy to implement and interpret
- Complete all variables using iterative process
- Send only completed topic flags to next stage of the edit process

Topic Flag Imputation Specifics

- Stratify sample by short list of demographic and administrative variables to create homogeneous sub-samples
- Topic flags imputed using separate logistic regression models for each sub-sample
- After each SRMI iteration, merge latest values of parent and spouse variables onto person's record
- Some topics were modeled at the family level to take account of complex survey design

Topics with special models

- General Assistance, TANF, and SNAP
  - answered by one respondent who represents a “clump” of people that would be expected to receive these benefits as a group
  - Only impute for the “clump” respondent
- Health Insurance
  - Jointly model health insurance receipt for all family members (adult 1, adult 2, child 1, child 2, all other children)
  - Imputation done at the individual level, conditional on values for other family members
- SSI
  - Also model year began receiving SSI because of presence of good administrative data
  - Use administrative data to evaluate “program confusion” between SSI and OASDI

Imputation of Administrative Data

- Use Logistic Regression to model two high level indicator variables
  - Did respondent receive OASDI benefits during the reference year?
  - Did respondent receive SSI benefits during the reference year?
- Use Bayes Bootstrap to find donors for all other administrative indicator variables
  - Did respondent have positive W-2 earnings?
  - What kind of OASDI benefits did the respondent receive?
- Use Linear Regression to model continuous variables
  - W-2 earnings amount
  - OASDI and SSI benefit amounts
  - Age began receiving OASDI and SSI benefits
- 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)

Results

Overall Percentages for cases where SIPP respondent answered the first question about jobs held (94.5% of in-universe respondents)			
Worked for pay in 2013?		W-2/Schedule C positive earnings in 2012?	
Yes	58.2	Yes	58.1
No	41.8	No	41.9

Overall Percentages for cases where SIPP respondent DID NOT answer the first question about jobs held and TF was imputed (5.5% of in-universe respondents)			
Worked for pay in 2013?		W-2/Schedule C positive earnings in 2012?	
Yes	61.5	Yes	60.4
No	38.5	No	39.6

Next steps for future waves of 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

Conclusion

- Model-based imputation is feasible in a production environment for a large-scale survey
- Outside data sources (such as administrative data) are valuable
  - Additional predictor variables in a model
  - Independent of survey non-response mechanism