Topic Flag Imputation in 2014 SIPP

- What are topic flags?
- What is their purpose?
- How will they help improve the SIPP?

Description of topic flags

- Instrument is divided into subject areas
  - "lines" in the EHC
  - question blocks after EHC
- 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
  - For any respondent who does not answer an entire topic, we will use a model to impute the answer
    - “yes – I had topic X”
    - “no – I did not have topic X”
- 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

Why use topic flags?

- Handle individuals with large amounts of missing data differently than in past
  - Allow whatever data is reported to be used
  - Handle missing topics consistently for everyone whether missing one or all topics
- Allow introduction of model-based imputation
  - Can include many more RHS control variables
    - Administrative data
    - Reported information from other family members
  - Use SRMI to model all topic flags jointly
  - Manageable number of variables to impute using new method

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

SIPP Synthetic Beta

- Purpose: Allow users to access administrative data linked to SIPP outside a secure Census worksite
- Source data: multiple SIPP panels linked to lifetime earnings and benefit histories
  - Master Earnings File: record of earnings from 1951-2011
  - Master Beneficiary Record: record of benefits from 1962-2012
- Creation:
  - Gold Standard File (GSF)
    - Standardize SIPP variables across panels
    - Create useful research variables from the administrative data
  - Employ a disclosure protection technique that does not allow users to re-identify source records in existing SIPP public use file

Confidentiality Protection

- Keep a few key variables unchanged
  - Gender and link to spouse

- Choose list of variables that is:
  - long enough to be useful (emphasis on retirement and disability research)
  - short enough to be protected
  - short enough to be created and protected in a “reasonable” amount of time

- Goal: to preserve as many multivariate relationships among variables as possible

- Chosen Method: Partially Synthetic Data

Synthesis Process

- Treat GSF as draws from large multivariate distribution
- Estimate the parameters of this distribution using GSF data
  - Measured with uncertainty
- Estimate distributions for the parameters
  - accounts for the variance introduced by the GSF sample
- Use a Posterior Predictive Distribution to create synthetic data
  - Prob(Y given X, β, and σ^2)Prob(β, σ^2 given X)
  - We take multiple draws and create 16 separate data sets.
- Use SRMI process to approximate PPD
  - Details in our technical paper “The Creation and Use of the SIPP Synthetic Beta” posted on the SSB website

User Access

- Researchers submit applications to use the SSB
  - Application is only rejected if request variables that are not on the SSB
- Account on SDS server at Cornell is created
  - Approval/account creation done within one week
  - Server has both Stata and SAS, mimics Census internal computing environment
  - Funded by NSF grant
- No data downloads currently allowed
- Number of users:
  - 52 total since 2007
  - About 15 currently active
  - Add 2-3 new users per month

Validation of Results

- Users strongly encouraged to send programs to Census for validation
- Analysis programs transferred from SDS server to internal Census server.
- Census staff runs programs on internal, confidential data
- Review results for disclosure risk
- If approved, release to user
- Have done 8 validation analyses in the past 12 months

SSB Development: Short-term plans

- Version 5.1 currently available
- Plan to release version 6.0 in June 2014
  - Adds 2008, 1984 panels
  - Longer administrative data history (through 2011)
  - New set of SSDI variables created: applications, rejections, acceptances, diagnosis codes
- SIPP SSB workshop June 23-27 at University of Michigan

SSB Development: Long-term plans

- Add job component to SSB
  - History of jobs and employers
  - Employer characteristics from Census business data
- Add links between parents and children
  - Contribute to research on inter-generational correlations
  - Correlation of events from childhood with adult earnings and labor market outcomes
- Add more administrative data
  - Income history
  - Residence history
  - Education
  - Health insurance