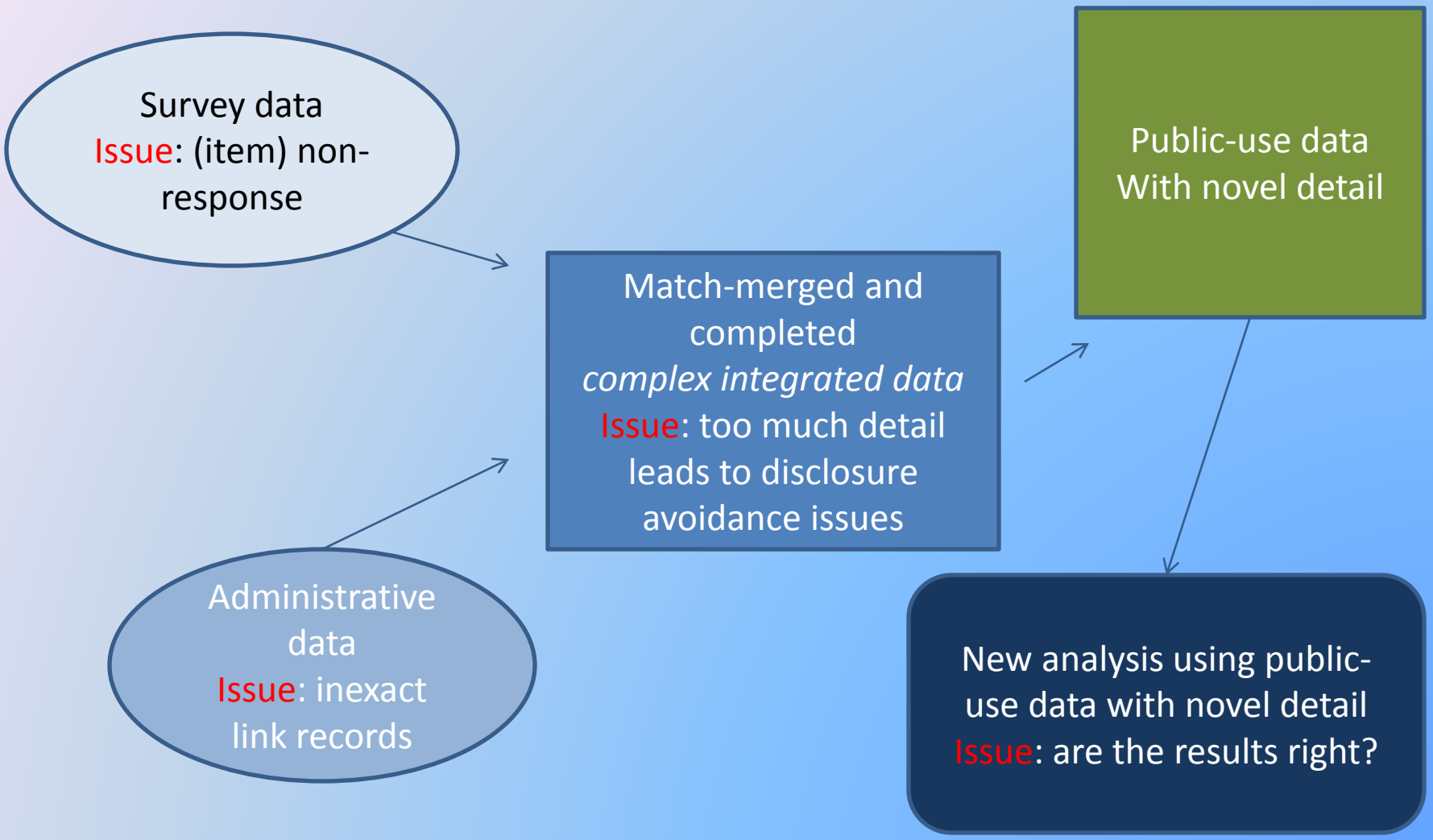
INFO 7470/ILRLE 7400 Survey of Income and Program Participation (SIPP) Synthetic Beta File

John M. Abowd and Lars Vilhuber

April 26, 2011

Elements



Survey of Income and Program Participation (SIPP)

- Goal of SIPP: accurate info about income and program participation of individuals and households and its principal determinants
- Information:
  - cash and noncash income on a sub-annual basis.
  - taxes, assets, liabilities
  - participation in government transfer programs

  http://www.census.gov/sipp/intro.html

Background

- In 2001, a new regulation authorized the Census Bureau and SSA to link SIPP and CPS data to SSA and IRS administrative data for research purposes
- Idea for a public use file was motivated by a desire to allow outside access to long administrative record histories of earnings and benefits linked to household demographic data
- These data allow detailed statistical and simulation study of retirement and disability programs
- Census Bureau, Social Security Administration, Internal Revenue Service, and Congressional Budget Office all participated in development

Genesis of SSB

- A portion of the SIPP user community was primarily interested in national retirement and disability programs
- SIPP augmented with
  - earnings histories from the IRS data maintained at SSA (W-2)
  - benefit data from SSA's master beneficiary records.
- Feasibility assessment (confidentiality!) of adding SIPP variables to earnings/benefit data in a public-use file (PUF)
  - set of variables that could be added without compromising the confidentiality protection of the existing SIPP public use files was VERY limited
- alternative methods explored

SSB basic methodology

- Experiment using "synthetic data"
- In fact: *partially synthetic data* with *multiple imputation* of missing items
- "partially synthetic data":
  - some (at least one!) variables are actual responses
  - other variables are replaced by values sampled from the posterior predictive distribution for that record, conditional on all of the confidential data.

History of SSB

- 2003-2005: Creation, but not release, of three versions of the "SIPP/SSA/IRS-PUF" (SSB)
- 2006: Release to limited public access of SSB V4.2
  - Access to general public only at Cornell-hosted Virtual RDC (SSB server: restricted-access setup)
    - With promise of evaluation of Virtual RDC-run programs on internal Gold Standard
  - Ongoing SSA evaluation
  - Ongoing evaluation at Census (in RDC)
- 2010: Release of SSB V5 at Census and on the Virtual RDC (codebook: http://www.census.gov/sipp/SSB_Codebook.pdf)

Basic structure of SSB V4

- SIPP
  - Core set of 125 SIPP variables in a standardized extract of SIPP panels 1990-1993 and 1996
  - All missing data items (except for structurally missing) are marked for imputation
- IRS
  - maintained at SSA, but derived from IRS records
  - Master summary earnings records (SER)
  - Master detailed earnings records (DER)

Basic structure SSB v4 (2)

- SSA
  - Master Beneficiary Record (MBR)
- Census
  - Numident: administrative birth and death dates

All files combined using (verified) SSNs

=> “Gold Standard”

Basic Structure of SSB V5

- Panels: 1990, 1991, 1992, 1993, 1996, 2001, and 2004 (this variable is now in the SSB)
- Couple-level linkage: the first person to whom the SIPP respondent was married during the time period covered by the SIPP panel
- SIPP variables only appear in years appropriate for the panel indicated by the PANEL variable (biggest change from V4.2)

Missing values in Gold Standard

- Values may be missing due to
  - [survey] Non-response
  - [survey] Question not being asked in a particular panel
  - [admin] Failure to link to administrative record (non-validated SSN)
  - [both] structural missing (e.g., income of spouse if not married)

Scope of Synthesis

- Never missing and not synthesized
  - gender,
  - marital status
  - spouse's gender
  - initial type of Social Security benefits
  - type of Social Security benefits in 2000
  - spouse's benefits type variables
- All other variables in the public use file were synthesized.

Common Structure to Multiple Imputation and Synthesis

- Hierarchical tree of variable relationships (parent-child relationship, accounting for structure)
- At each node, independent SRMI is used
  - a statistical model is estimated for each of the variables at the same level
    - Bayesian bootstrap,
    - logistic regression, or
    - linear regression
  - Statistical models are estimated separately for groups of individuals
  - Then, a proper posterior predictive distribution is estimated
  - Given a PPD, each variable is imputed /synthesized, conditional on all values of all other variables for that record
- The next node is processed

MI and Synth

- Initial iterations for missing data imputation, keeping all observed values where available
- Final iteration is for data synthesis (replacing all observed values, see exceptions)

Latest Release of SSB

- 2010: Release of limited public access of SSB v5

Framework for SIPP/SSA/IRS Synthetic Beta

- Link 1990-2004 SIPP panels with lifetime earnings and benefit histories from IRS and SSA
- Keep a few key variables unchanged
- Choose and synthesize other SIPP, IRS, and SSA variables subject to the following requirements:
  - List of variables must be long enough to be useful to some group of researchers
  - Multivariate relationships across synthesized variables must be analytically valid: shorter lists, less distortion
  - disclosure avoidance challenge: users cannot re-identify source records in existing SIPP public use file: shorter lists, more distortion

Important Design Decisions

- Four variables remain unsynthesized: gender, marital status, benefit status (initial and 2000); also link to spouse is not perturbed
- Hundreds variables would be synthesized
- Variables chosen for target users: disability and retirement research communities
- Use SRMI and Bayesian Bootstrap methods for data synthesis
- Create “gold standard” data and compare to synthetic data to assess analytic validity
- Try to match back to public use SIPP to test disclosure avoidance problems

Create Gold Standard Data

- Create a data extract from the SIPP panels conducted in the 1990s
  - Seven panels: 1990, 1991, 1992, 1993, 1996, 2001, 2004
  - Data from core and topical module survey questions
- Standardize variables across panels
- Link to Summary/Detailed Earnings Records and SSA benefits data
- These data are the “truth.” Any synthetic data should preserve the characteristics of and relationships among the variables on this file.

SIPP Variables

- Codebook

Synthetic Data Creation

- Purpose of synthetic data is to create micro data that can be used by researchers in the same manner as the original data while preserving the confidentiality of respondents' identities
- Fundamental trade-off: usefulness and analytic validity of data versus protection from disclosure
- Our goal: not be able to re-identify anyone in the already released SIPP public use files while still preserving regression results

Multiple Imputation Confidentiality Protection

- Denote confidential data by Y and non-confidential data by X.
- Y contains missing data so that $Y=(Y_{obs}, Y_{mis})$ and X has no missing data.
- Use the posterior predictive distribution(PPD) $p(Y_{mis} | Y_{obs}, X)$ to complete missing data and $p(Y | Y_m, X)$ to create synthetic data
- Data synthesis is same procedure as missing data imputation, just done for all observations
- Major emphasis is to find a good estimate of the PPD

Testing Analytical Validity

- Run regressions on each synthetic implicate
  - Average coefficients
  - Combine standard errors using formulae that take account of average variance of estimates (within implicate variance) and differences in variance across estimates (between implicate variance).
- Run regressions on gold standard data
- Compare average synthetic coefficient and standard error to g.s. coefficient and s.e.
- Data are analytically valid if coefficient is unbiased and the same inferences are drawn

Formulae: Completed Data only

- Notation
  - script ℓ is index for missing data implicate
  - m is total number of missing data implicates
- Estimate from one completed implicate

$$q^{\ell} = q(D^{\ell}).$$

- Average of statistic across implicates

$$\bar{q}_m = \sum_{\ell=1}^{m} \frac{q^{(\ell)}}{m}.$$

Formulae: Total Variance
Between Variance – variation due to differences between implicates

- Total variance of average statistic

$$T_m = \bar{u}_m + \left(1 + \frac{1}{m}\right) b_m$$

- Variance of the statistic across implicates: between variance

$$b_m = \sum_{\ell=1}^{m} \frac{(q^{(\ell)} - \bar{q}_m)(q^{(\ell)} - \bar{q}_m)'}{m-1}$$

Formulae: Within Variance
Variation due to differences within each implicate

- Variance of the statistic from each completed implicate

$$u^{(\ell)} = u(D^{(\ell)})$$

- Average variance of statistic: within variance

$$\bar{u}_m = \sum_{\ell=1}^{m} \frac{u^{(\ell)}}{m}.$$

Formulae: Synthetic and Completed Implicates

- Notation
  - script ℓ is index for missing data implicate
  - script k is index for synthetic data implicate
  - m is total number of missing data implicates
  - r is total number of synthetic implicates per missing data implicate
- Estimate from one synthetic implicate

$$q^{(\ell,k)} = q\left(D^{(\ell,k)}\right).$$

- Average of statistic across synthetic implicates

$$\bar{q}^{(\ell)} = \sum_{k=1}^{r} \frac{q^{(\ell,k)}}{r}$$

Formulae: Grand Mean and Total Variance

- Average of statistic across all implicates

$$\bar{q}_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{q^{(\ell,k)}}{mr} = \sum_{\ell=1}^{m} \frac{\bar{q}^{(\ell)}}{m}.$$

- Total variance of average statistic

$$T_M = \left(1 + \frac{1}{m}\right) B_M - \frac{b_M}{r} + \bar{u}_M.$$

Formulae: Between Variance

Variation due to differences between implicates

- Variance of the statistic across missing data implicates: between m implicate variance

$$B_M = \sum_{\ell=1}^{m} \frac{(\bar{q}^{(\ell)} - \bar{q}_M)(\bar{q}^{(\ell)} - \bar{q}_M)'}{m-1}.$$

- Variance of the statistic across synthetic data implicates: between r implicate variance

$$b_M = \sum_{\ell=1}^{m} \sum_{k=1}^{r} \frac{\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)\left(q^{(\ell,k)} - \bar{q}^{(\ell)}\right)'}{m(r-1)} = \sum_{\ell=1}^{m} \frac{b^{(\ell)}}{m}.$$

Formulae: Within Variance
Variation due to differences within each implicate

- Variance of the statistic on each implicate

$$u^{(\ell,k)} = u\left(D^{(\ell,k)}\right)$$

- Average variance of statistic: within variance

$$\bar{u}_M = \sum_{\ell=1}^{m}\sum_{k=1}^{r}\frac{u^{(\ell,k)}}{mr} = \sum_{\ell=1}^{m}\frac{\bar{u}^{(\ell)}}{m}$$

- Source: Reiter, *Survey Methodology* (2004): 235-42.

Example:
Average AIME/AMW

- Estimate average on each of synthetic implicates
  - $AvgAIME^{(1,1)}$, $AvgAIME^{(1,2)}$, $AvgAIME^{(1,3)}$, $AvgAIME^{(1,4)}$,
  - $AvgAIME^{(2,1)}$, $AvgAIME^{(2,2)}$, $AvgAIME^{(2,3)}$, $AvgAIME^{(2,4)}$,
  - $AvgAIME^{(3,1)}$, $AvgAIME^{(3,2)}$, $AvgAIME^{(3,3)}$, $AvgAIME^{(3,4)}$,
  - $AvgAIME^{(4,1)}$, $AvgAIME^{(4,2)}$, $AvgAIME^{(4,3)}$, $AvgAIME^{(4,4)}$
- Estimate mean for each set of synthetic implicates that correspond to one completed implicate
  - $AvgAIMEAVG^{(1)}$, $AvgAIMEAVG^{(2)}$, $AvgAIMEAVG^{(3)}$, $AvgAIMEAVG^{(4)}$
- Estimate grand mean of all implicates
  - AvgAIMEGRANDAVG

Example (cont.)

- Between m implicate variance

$$B_M = \sum_{\ell=1}^{4} \frac{(avgAIMEavg^{(\ell)} - avgAIMEGrandavg)(avgAIMEavg^{(\ell)} - avgAIMEGrandavg)'}{3}.$$

- Between r implicate variance

$$b_M = \sum_{\ell=1}^{4}\sum_{k=1}^{4} \frac{(avgAIME^{(\ell,k)} - avgAIMEavg^{(\ell)})(avgAIME^{(\ell,k)} - avgAIMEavg^{(\ell)})'}{4(3)}.$$

Example (cont.)

- Variance of mean from each implicate
  - VAR[AvgAIME$^{(1,1)}$] , VAR[AvgAIME$^{(1,2)}$] , VAR[AvgAIME$^{(1,3)}$] , VAR[AvgAIME$^{(1,4)}$]
  - VAR[AvgAIME$^{(2,1)}$] , VAR[AvgAIME$^{(2,2)}$] , VAR[AvgAIME$^{(2,3)}$] , VAR[AvgAIME$^{(2,4)}$]
  - VAR[AvgAIME$^{(3,1)}$] , VAR[AvgAIME$^{(3,2)}$] , VAR[AvgAIME$^{(3,3)}$] , VAR[AvgAIME$^{(3,4)}$]
  - VAR[AvgAIME$^{(4,1)}$] , VAR[AvgAIME$^{(4,2)}$] , VAR[AvgAIME$^{(4,3)}$] , VAR[AvgAIME$^{(4,4)}$]
- Within variance

$$\bar{u}_M = \sum_{\ell=1}^{4} \sum_{k=1}^{4} \frac{Var[avgAIME^{(\ell,k)}]}{4(4)}$$

Example (cont.)

- Total Variance

$$T_M = \left(1 + \frac{1}{4}\right)B_M - \frac{b_M}{4} + \bar{u}_M.$$

- Use AvgAIMEGRANDAVG and Total Variance to calculate confidence intervals and compare to estimate from completed data

SAS Programs

- Sample programs to calculate total variance and confidence intervals

Results: Average AIME

Average of AIME (Average Indexed Monthly Earnings)/AMW(Average Monthly Wage)

	AVG STAT	Total VAR	Betw. M Var	Betw. R Var	Betw. Var	Within Var	confidence interval	
synthetic	1094.2	91.8	59.3	13.3		21.1	1074.5	1113.9
completed	1142.5	52.8			23.4	23.7	1129.3	1155.7

*All individuals with TOB_2000=1

Public Use of the SIPP Synthetic Beta

- Full version (16 implicates) released to the Cornell Virtual RDC
- Any researcher may use these data
- During the testing phase, all analyses must be performed on the Virtual RDC
- Census Bureau research team will run the same analysis on the completed confidential data
- Results of the comparison will be released to the researcher, Census Bureau, SSA, and IRS (after traditional disclosure avoidance analysis of the runs on the confidential data)

Methods for Estimating the PPD

- Sequential Regression Multivariate Imputation (SRMI) is a parametric method where PPD is defined as

$$p\left(\tilde{Y} \mid Y_{obs}, X_{obs}\right) = \int p\left(\tilde{Y} \mid Y_{obs}, X_{obs}, \theta\right) p\left(\theta \mid Y_{obs}, X_{obs}\right) d\theta$$

- The BB is a non-parametric method of taking draws from the posterior predictive distribution of a group of variables that allows for uncertainty in the sample CDF
- We use BB for a few groups of variables with particularly complex relationships and use SRMI for all other variables

SRMI Method Details

- Assume a joint density $p(Y,X,\vartheta)$ that defines parametric relationships between all observed variables.
- Approximate the joint density by a sequence of conditional densities defined by generalized linear models.
- Same process for completing and synthesizing data
- Synthetic values of some $y_k \in Y$ are draws from:

$$p_k\left(\tilde{y}_k \mid Y^m, X^m\right) = \int p_k\left(\tilde{y}_k \mid Y^m_{\sim k}, X^m, \theta\right) p_k\left(\theta \mid Y^m, X^m\right) d\theta$$

where Y^m, X^m are completed data, and densities p_k are defined by an appropriate generalized linear model and prior.

SRMI Details: KDE Transforms

- The SRMI models for continuous variables assume that they are conditionally normal
- This assumption is relaxed by performing a KDE-based transform of groups of related variables
- All variables in the group are transformed to normality, then the PPD is estimated
- The sampled values from PPD are inverse transformed back to the original distribution using the inverse cumulative distribution

SRMI Example: Synthesizing Date of Birth

- Divide individuals into homogeneous groups using stratification variables
  - example: male, black, age categories, education categories, marital status
  - example: decile of lifetime earnings distribution, decile of lifetime years worked distribution, worked previous year, worked current year
- For each group, estimate an independent linear regression of date of birth on other variables (not used for stratification) that are strongly related

SRMI Example: Synthesizing Date of Birth

- Synthetic date of birth is a random variable
- Before analysis, it is transformed to normal using the KDE-based procedure
- Distribution has two sources of variation:
  - variation in error term in regression model
  - variation in estimated parameters: β's and σ^2
- Synthetic values are draws from this distribution
- Synthetic values are inverse transformed back to the original distribution using the inverse cumulative distribution.

Bayesian Bootstrap Method Details

- Divide data into homogeneous groups using similar stratification variables as in SRMI
- Within groups do a Bayesian bootstrap of all variables to be synthesized at the same time.
  - n observations in a group, draw 1-n random variables from uniform (0,1) distribution
  - let u_o … u_i … u_n define the ordering of the observations in the group
  - $u_i - u_{i-1}$ is the probability of sampling observation i from the group to replace missing data or synthesize data in observation j
  - conventional bootstrap, probability of sampling is 1/n

Creating Synthetic Data

- Begin with base data set that contains only non-missing values
- Use BB to complete missing administrative data – i.e. find donor SSN based on non-missing SIPP variables
- Use SRMI to complete missing SIPP data
- Iterate multiple times – input for iteration 2 is completed data set from iteration 1
- On last iteration, run 4 separate processes to create 4 separate data sets or implicates

Creating Synthetic Data, Cont.

- Synthesis is like one more iteration of data completion, except all observations are treated as missing
- Each completed implicate serves as a separate input file
- Run 16 separate processes to create 16 different synthetic data sets or implicates
- The separate processes to create implicates have different stratification variables
- Need enough implicates to produce enough variation to ensure that averages across the implicates will be close to “truth”

Features of our Synthesizing Routines

- Parent-child relationships
  - foreign-born and decade arrive in US
  - welfare participation and welfare amount
  - presence of earnings, amount of earnings
- Restrictions on draws from PPD
  - Some draws must be within a pre-specified range from the original value: example MBA is +/- $50 of original value.
  - impose maximum and minimum values on some variables