Using the SIPP Synthetic Beta for Analysis

John M. Abowd, Gary Benedetto, Martha Stinson
Cornell University and U.S. Census Bureau

Census Training Meeting, Oct. 26, 2007

Outline

- Background
- Framework for Public Use Product
- Sources for the Gold Standard data
- Current System for data access
- Assessing analytical validity
- Creation of Synthetic Data

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

Framework for SIPP/SSA/IRS Synthetic Beta

- Link 5 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: sex, marital status, benefit status (initial and 2000); also link to spouse is not perturbed
- 616 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 problems

Create Gold Standard Data

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

SIPP Variables: Demographic

- Variables that will not be synthesized
  - Gender, marital status, link to spouse, and the same variables for the spouse, if present.
- Variables to be synthesized
  - Five category education, black, Hispanic, birth month/year, death month/year, disability limits work, disability prevents work, total number of children in family, history of marital events up to 4 marriages, age at which marital events occur, foreign born, decade arrive in US

SIPP Variables: Economic

- Labor Force Participation
  - Annual time series (1990-1999): Weeks worked with pay, weeks worked part-time, total hours worked, total earnings
  - Four category industry and occupation
- Income
  - Annual time series (1990-1999) Family poverty threshold, total family income, total personal income, family welfare participation and income, disability participation and income (non-SSA)

SIPP Variables: Wealth and Benefits

- Wealth
  - Total net worth, home ownership indicator, home equity, non-housing wealth
- Pension
  - Defined benefit plan, defined contribution plan
- Health insurance
  - Annual time series (1990-1999): health insurance indicator, health insurance from employer indicator

SSA/IRS Master Earnings File Variables

- Summary Earnings Record Extract 1951-2003
  - Annual total FICA covered earnings (capped)
  - Annual pattern of quarters worked 1951 – 1977
  - Total earnings from 1937 to 1950
- Detailed Earnings Record Extract 1978-2003
  - Wages, Tips, and other compensation (Box 1) (uncapped)
  - Deferred Wages (Box 13) – for example 401(k) contributions
  - FICA covered earnings (Box 3)
  - Summed across all employers

SSA Benefit Variables

- Master Beneficiary Record
  - Initial Benefit
    - Year first received benefits
    - type of benefit
    - amount of benefit
  - Current Benefit
    - year first began receiving current benefit
    - type of benefit
    - amount of benefit

Decennial Weight

- Purpose
  - Make the combined SIPP panels representative of U.S. population on Apr. 1, 2000
- Method
  - Divide Decennial 2000 population into same groups from which SIPP sample was drawn
  - Locate each SIPP respondent in the Decennial
  - Weight = #Decennial persons in group / #SIPP persons in group
  - Adjust the weight to match official U.S. population totals based on 1996 SIPP demographic subgroups
  - Create a synthetic weight for each synthetic implicate

Locating SIPP respondents in Decennial

- Match SIPP to Decennial by PIK (SSN)
- Locate remaining SIPP persons using probabilistic record linking
  - Assign each SIPP person a set of Decennial candidates based on blocking variables (race, gender, etc)
  - Choose a match from this set based on matching variables (birth month, # children, etc)

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 disclosable 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)

Survey of Income and Program Participation (SIPP) - Mozilla Firefox

File Edit View History Bookmarks Tools Help

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

Cornell University We... 7-Day Forecast for L... Main Page - CISERWiki Main Page - ISS Netw... LEHD Staff Only www.cyclingnews.co...

7-Day Forecast ... Codian MCU - C... Videoconferenci... VirtualRDC Proj... City Creek (Salt... OpenTable Survey of I...

U.S. Census Bureau

Survey of Income and Program Participation

SIPP

- Introduction to SIPP
- SIPP Survey Content
- Technical Information
- Using & Linking Files
- SIPP Publications
- Access SIPP Data
- Access SIPP Synthetic Data

- User Notes/ ListServe/News
- SIPP Users' Guide
- SIPP Tutorial
- Technical Documentation
- SIPP Help

Dynamics of Economic Well-being System

Contact DEWS

URL: http://www.sipp.census.gov/sipp/

Source: U.S. Census Bureau, Demographics Survey Division, Survey of Income and Program Participation branch
Created: February 14, 2002
Last revised: July 16, 2007

Census Bureau Links: Home • Search • Subjects A-Z • FAQs • Data Tools • Catalog • Census 2000 • Quality • Privacy Policy • Contact Us

USCENSUSBUREAU

Done

Synthetic Data

Some of the following documents are in the Portable Document Format (PDF). In order to view these files, you will need the Adobe(R) Acrobat(R) Reader which is available for free from the Adobe web site.

The SIPP Synthetic Beta (SSB) file was created by integrating data from the Survey of Program Participation (SIPP), Social Security Administration (SSA), and Internal Revenue Service (IRS) and then synthesizing these data. This work was performed as part of a joint project between the three data contributing agencies. The goal was to create a product that could be used by researchers outside of the regular Census restricted-access facilities. These synthetic data should reproduce the characteristics of the underlying confidential micro-data and, at the same time, assure the confidentiality of the actual data on the sampled individuals. The Census Disclosure Review Board, SSA, and IRS have cleared this file for use by individuals without Census Special Sworn Status. Researchers interested in using the file may submit questions to hhes.synthetic.data.use.list@census.gov. When researchers are ready to begin a project, they should submit the application posted here using the same email address.

The Census Bureau will not conduct a formal project review. Instead, applications will be judged solely on feasibility (i.e. the necessary variables are on the SSB). After projects are approved, researchers will be given accounts on the server housing these data. The document "Technical_Description_SIPP_Synthetic_Beta_July92007," also posted here, contains a codebook for this data set and further description of how th synthetic data were created.

- SIPP Synthetic Beta Application (in PDF format)
- Technical Description SIPP Synthetic Beta (7/9/2007) (MS WORD document)

Cornell University
Cornell Institute for Social and Economic Research

Search Cornell

VirtualRDC News @ CISER

2007 Joint Statistical Meetings

July 29th, 2007

July 29 – August 2, 2007
Salt Lake City, Utah.

2007 Joint Statistical Meetings to be held at the Salt Palace Convention Center.
JSM (the Joint Statistical Meetings) is the largest gathering of statisticians held in North America. It is held jointly with the American Statistical Association, the International Biometric Society (ENAR and WNAR), the Institute of Mathematical Statistics, and the Statistical Society of Canada. Attended by over 5000 people, activities of the meeting include oral presentations, panel sessions, poster presentations, continuing education courses, exhibit hall (with state-of-the-art statistical products and opportunities), career placement service, society and section business meetings, committee meetings, social activities, and networking opportunities. Salt Lake City is the host city for JSM 2007 and offers a wide range of possibilities for sharing time with friends and colleagues. For information, contact jsm@amstat.org or phone toll-free (866) 421-7169.

Posted in Events | No Comments »

Corrected OTM data for IL posted on July 13

July 24th, 2007

Site search

Search for this text:

Search

Site navigation

- Information about the VirtualRDC
- Available resources
- Data @ VirtualRDC
- Classes and Tutorials
- Help for RDC proposal writers

Recent articles

- Events
- General
- Grants
- Hardware
- Library
- Software

Related sites

- CISER
- NYCRDC

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 $p(Y,X,\theta)$ 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 $u_o \ldots u_i \ldots u_n$ 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 9 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