JPSM The Joint Program in Survey Methodology

University of Maryland at College Park • University of Michigan • Westat

Synthetic Data: Balancing Data Confidentiality & Quality in Public Use Files

A two-day short course sponsored by the Joint Program in Survey Methodology

Presented by:

Joerg Drechsler, Ph.D.

Senior Researcher, Institute for Employment Research, Germany.

Jerry Reiter, Ph.D.

Professor of Statistical Science, Duke University.

December 3-4, 2019

Presented at

RTI, Washington DC.

JOINT PROGRAM
IN SURVEY
METHODOLOGY

A short course sponsored by the Joint Program in Survey Methodology

Synthetic Data: Balancing Confidentiality and Quality in Public Use Files

December 3-4, 2019
Presented at RTI, Washington DC.

JÖRG DRECHSLER
Senior Researcher, Institute for Employment Research, Germany

JERRY REITER
Professor of Statistical Science, Duke University

COURSE OBJECTIVES

This short course will provide a detailed overview of the topic, covering all important aspects relevant for the synthetic data approach. Starting with a short introduction to data confidentiality in general and synthetic data in particular, the workshop will discuss the different approaches to generating synthetic datasets in detail. Possible modeling strategies and analytical validity evaluations will be assessed and potential approaches to quantify the remaining risk of disclosure will be presented. The course will also briefly describe the how synthetic data could be used with differential privacy. To provide the participants with hands on experience, most of the second day will be devoted to practical sessions using R in which the students generate and evaluate synthetic data for various datasets.

WHO SHOULD ATTEND

The course intends to summarize the state of the art in synthetic data. The main focus will be on practical implementation and not so much on the motivation of the underlying statistical theory. Participants may be academic researchers or practitioners from statistical agencies working in the area of data confidentiality and data access. Some background in Bayesian statistics and R is helpful but not obligatory.

INSTRUCTORS

JÖRG DRECHSLER Jörg is distinguished researcher at the Department for Statistical Methods at the Institute for Employment Research in Nürnberg, Germany. He received his PhD in Social Science from the University in Bamberg in 2009 and his Habilitation in Statistics from the Ludwig-MaximiliansUniversität in Munich in 2015. He is also an adjunct associate professor in the Joint Program in Survey Methodology at the University of Maryland and honorary professor at the University of Mannheim, Germany. His main research interests are data confidentiality and nonresponse in surveys.

JERRY REITER is Professor of Statistical Science at Duke University in Durham, NC. He received his PhD in statistics from Harvard University in 1999. He has developed much of the theory and methodology for synthetic data, as well as supervised the creation of the Synthetic Longitudinal Database. He is the recipient of the 2014 Gertrude M. Cox Award.

COMPUTER

Students should bring their own laptop with R installed.
Prior to the course, students should install the latest version of R, which is available for free at http:// www.r-project.org/. Registrants should install the R package synthpop , which is available for free from CRAN at cran.r-project.org.

JOINT PROGRAM
IN SURVEY
METHODOLOGY

TENTATIVE SCHEDULE

Tuesday: December 3, 2019

08:00 – 09:00 Registrant Check-in and Continental Breakfast

09:00 – 09:30	Overview of data confidentiality
09:30 – 10:30	Introduction to synthetic data
10:30 – 10:45	**Coffee break**
10:45 – 12:15	Synthetic data models 1
12:15 – 01:45	**Lunch**
01:45 – 02:45	Synthetic data models 2
02:45 – 03:15	Utility checks
03:15 – 03:30	**Coffee break**
03:30 – 04:00	Disclosure risk
04:00 – 04:30	Synthetic Data and Differential Privacy
04:30	**Adjourn**

Wednesday: December 4, 2019

08:00 – 09:00 Registrant check-in and Continental Breakfast

09:00 – 10:00 Exemplary applications

10:00 – 10:15 Coffee break

10:15 – 11:00 Introduction to synthpop package in R

11:00 – 11:45 Students generate synthetic data in small groups

11:45 – 01:15 Lunch

01:15 – 02:00 Utility checks

02:00 – 03:00 Disclosure checks

03:00 – 03:15 Coffee break

03:15 – 04:00 Discussion among class

04:00 – 04:30 Wrap up

4:30 Adjourn

Overview of Data Confidentiality

Synthetic Data

Balancing Confidentiality and Quality in Public Use Files

**Short Course sponsored by the
Joint Program in Survey Methodology**

Jörg Drechsler
&
Jerry Reiter

Outline First Day – Theory

- Overview of Data Confidentiality
- Introduction to Synthetic Data
- Synthetic Data Models
- Utility Checks
- Disclosure Risk Assessment

Outline Second Day – Practical Applications

- Exemplary Applications
- Students Generate Synthetic Data
- Utility Checks in Practice
- Disclosure Risk Assessment in Practice

History of Data Confidentiality

- Data confidentiality is a hot topic
- But only since the last 2-3 decades
- Personal information has been collected for thousands of years
- In the early days most data collected by statistical agencies
- Only confidentiality breaches: sharing data with other government agencies
- Otherwise all information was published only in tables
- Access to the microdata for external researchers was unthinkable and nobody else stored any data

History of Data Confidentiality

- Research on data confidentiality mainly focused on tabular data
- Confidentiality for tabular data still a very important topic for statistical agencies
- Nowadays massive amounts of data are collected (and analyzed) daily
- Most data no longer collected by the government (internet search logs, Twitter, supermarket scanners…)
- Question how to share collected information without violating privacy guarantees becomes more relevant

History of Data Confidentiality

- First papers on microdata confidentiality in the early eighties (Data swapping, Dalenius and Reiss (1982))
- Three famous privacy breaches stimulate the discussions on data confidentiality
  - Identification of a city mayor in “anonymised” medical records in Massachusetts
  - A Face Is Exposed for AOL Searcher No. 4417749
  - Netflix Spilled Your Brokeback Mountain Secret
- Data confidentiality for microdata can be achieved in two ways
  - Information reduction
  - Data perturbation

Information Reduction

- Information that poses a possible risk of re-identification is suppressed
- Possible methods:

  - top coding - local suppression - rounding
  - global recoding - dropping variables - sampling
  - …

- Advantage
  - All released information is unaltered
- Disadvantage
  - Important information is lost
  - Information reduction might be so severe for sensitive data that the dataset will become useless

Data Perturbation

- All variables remain in the dataset but individual records are altered to guarantee data confidentiality
- Possible methods:

  - swapping - microaggregation - PRAM
  - noise infusion - …

- Advantage
  - All information is still available in the released data
- Disadvantage
  - Data have been altered
  - Important relationships found in the original data might be distorted

Problems with traditional SDC methods

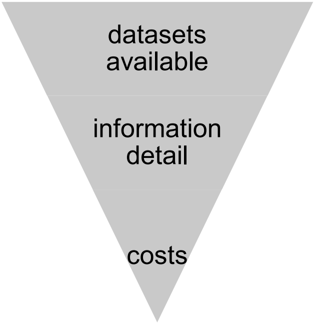
- Recoding
  - Loses information in tails
  - Disables fine spatial analysis
  - Creates ecological fallacies
- Suppression
  - Creates nonignorable missing data
  - May not be fully protective
- Swapping
  - Attenuates correlations
  - Protection based on perception
- Noise Infusion
  - Inflates variances
  - Distorts distributions
  - Attenuates correlations
  - May need large noise variances

Recent Developments

- Two alternatives to data dissemination
- Research data centers
  - Advantages: - more datasets available
- more detailed information available
  - Disadvantages: - burdensome for the researcher
- cost intensive for the agency
- Remote analysis servers/remote access
  - Advantages: - more convenient for researchers
- less costs for agency
  - Disadvantages: - only limited analyses possible for remote servers
- disclosure risk not fully evaluated for remote access

Current Data Dissemination Practice

- Three access channels
  - Onsite Access
  - Remote Execution
  - Public-Use-Files



Introduction to Synthetic Data

Outline First Day – Theory

- Overview of Data Confidentiality
- Introduction to Synthetic Data
  - Synthetic Data Approaches
  - Analyzing Synthetic Datasets
- Synthetic Data Models
- Utility Checks
- Disclosure Risk Assessment

Where Do We Start From?

- Easy to implement SDC methods either fail to protect the data or drastically reduce the analytical validity
- Other methods only preserve pre-specified statistics like the mean and the variance
- Remote analysis servers helpful tool for the public but not so much for the scientific researcher
- Remote access promising tool with a number of open questions regarding the level of confidentiality that can be guaranteed
- Releasing synthetic data can be a viable alternative

The Basic Concept

- Idea is closely related to multiple imputation for nonresponse
- Generate synthetic datasets by drawing from a model fitted to the original data
- Not the missing values but the sensitive values are replaced with a set of plausible values given the original data
- Generate multiple draws to be able to obtain valid variance estimates from the synthetic data

The Basic Concept

- Three steps necessary for data release:
  - Fit model to the original data
  - Repeatedly draw from that model to generate multiple synthetic datasets
  - Release these datasets to the public
- Over the years different designs for generating synthetic data evolved
- Two main approaches: fully synthetic datasets and partially synthetic datasets

Fully Synthetic Datasets

- Goes back to Rubin (1993)
- A useful SDC method should fulfil three goals
  - Preserve confidentiality
  - Maintain valid inferences
  - Allow the user to rely on standard statistical software
- Masking techniques very popular at that time
- Can fulfill the first two goals in certain settings
- Rubin criticizes masking as an approach to protect confidentiality

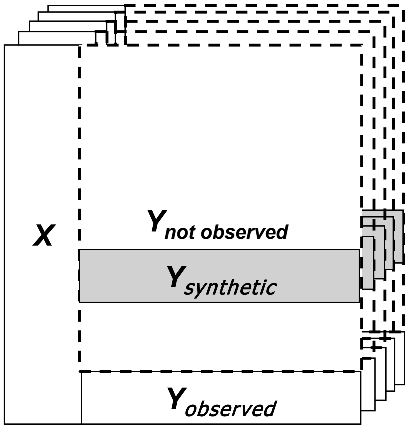
Masking Techniques

- Requires special software to obtain valid inferences
- Requires complicated error-in-variables models
- No special software will be developed for each *analysis method x masking method x database type*
- Users have their own science to worry about
- Shouldn’t be expected to become experts in demasking programs

Fully Synthetic Datasets

- Rubin suggests an alternative approach for releasing confidential microdata
- Instead of applying masking procedures, completely synthetic data should be released
- Approach is based on the ideas of multiple imputation
- All units that did not participate in the survey are treated as missing data
- Missing data are multiply imputed
- Samples from the generated synthetic populations are released to the public

Fully Synthetic Datasets



Pros and Cons of the Approach

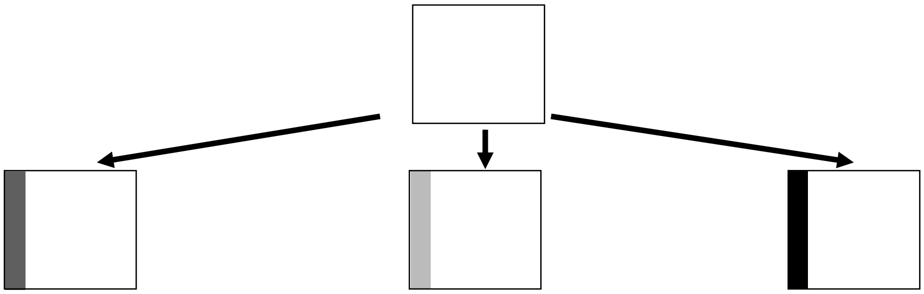
- Advantages of the approach
  - Data are fully synthetic
  - Re-identification of single units almost impossible
  - No need to decide which values to alter nor which variables are quasi-identifiers
  - Protection does not depend on hiding nature of SDL to public
  - All variables are still fully available
  - Valid inferences can be obtained using simple combining rules
- Disadvantages of the approach
  - Strong dependence on the imputation model
  - Setting up a model might be difficult/impossible
- Not always necessary to synthesize all variables
- Alternative: partially synthetic data

Partially Synthetic Datasets

- Originally proposed by Little (1993)
- Not all information in a dataset is sensitive
- Replace only those variables/records that lead to an unacceptable risk of disclosure
- Replaced variables could be sensitive variables or key variables that could be used for re-identification purposes
- Not necessary to replace all records of one variable
- Only the records at risk need to be replaced
- Every unchanged record will increase the analytical validity

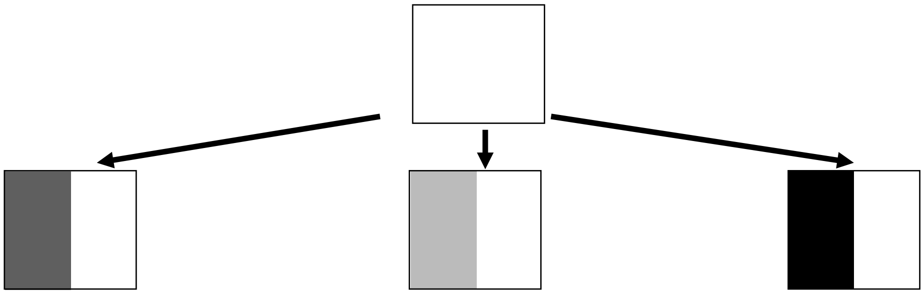
Partially Synthetic Datasets

- Only potentially identifying or sensitive variables are replaced



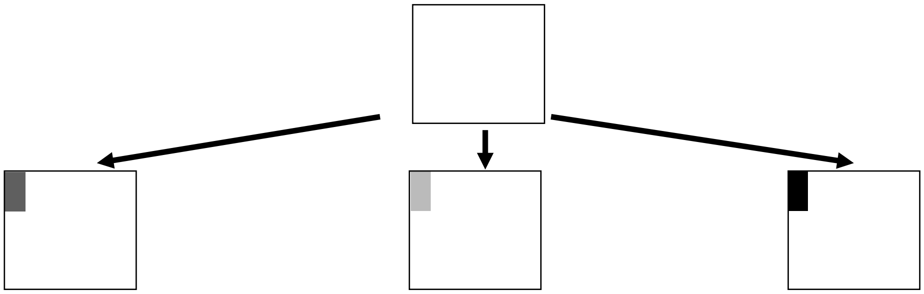
Partially Synthetic Datasets

- Only potentially identifying or sensitive variables are replaced



Partially Synthetic Datasets

- Only potentially identifying or sensitive variables are replaced



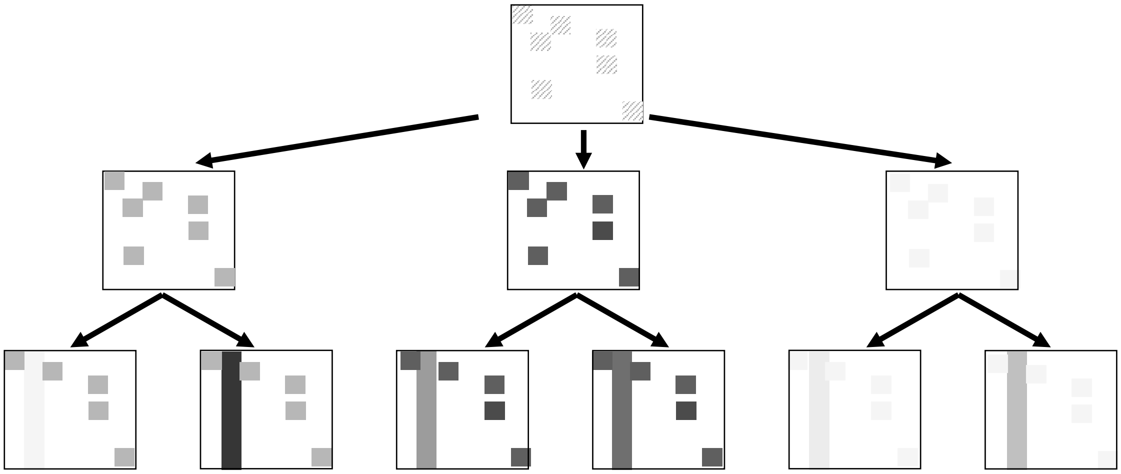
Pros and Cons of the Approach

- Advantages of the approach
  - Model dependence decreases
  - Models are easier to set up
- Disadvantages of the approach
  - True values remain in the dataset
  - Disclosure might still be possible
- Careful disclosure risk evaluation necessary

Dealing with Missing Data

- Missing data are a common problem in surveys
- Most SDC techniques cannot deal with missing values
- Straightforward to address the problem with synthetic data
- Imputation in two stages:
  - Multiply impute missing values on stage one *r* times
  - Generate synthetic datasets for each one stage nest on stage two *m* times
- Possible (and likely) to use different models for imputation and synthesis
- Incorporates the estimation uncertainty on both levels
- New combining rules necessary

Multiple Imputation for Nonresponse and Confidentiality



Synthetic Data Compared to Other SDC Techniques

Advantages

- Tries to preserve the multivariate relationship between the variables and not only specific statistics
- Suitable for any variable type
- Most SDC methods cannot address some of the problems typically encountered in practice
  - Item nonresponse
  - Skip patterns
  - Logical constraints

Disadvantages

- Lot of work
- Depends heavily on the quality of the imputation models

Synthetic Datasets in Practice

- Original proposal confronted with disbelief
- Some other theoretical papers followed (Fienberg, 1994; Fienberg et al., 1998)
- First application of partially synthetic data in practice: Survey of Consumer Finances (Kennickell, 1997)
- Other important early contributions: Abowd and Woodcock (2001,2004) evaluate the approach on a French longitudinal business dataset
- Raghunathan et al. (2003) and Reiter (2003, 2004) derive the correct combining rules for valid inferences from fully and partially synthetic datasets

Current Situation

- Main driving force: US Census Bureau
- List of products based on synthetic data released so far:
  - SIPP synthetic data (combination of the SIPP, selected variables from the Internal Revenue Service's (IRS) lifetime earnings data, and the individual benefit data from the Social Security Administration (SSA))
  - OntheMap
  - Parts of the American Community Survey
  - Longitudinal Business Database (LBD)
- Other products are in the development stage
- Outside the US, the approach is also investigated in Australia, Canada, Germany, Scotland, England, and New Zealand

Outline First Day – Theory

- Overview of Data Confidentiality
- Introduction to Synthetic Data
  - Synthetic Data Approaches
  - Analyzing Synthetic Datasets
- Synthetic Data Models
- Utility Checks
- Disclosure Risk Assessment

Analyzing Synthetic Datasets

- Analysis based on the synthetic data is straightforward for the user
  - Analyse each synthetic dataset separately using standard methods
  - Combine the results from the different datasets to obtain final estimates
- Comparable to combining procedures for multiple imputation for nonresponse
- Combining procedures for the estimated variance of the parameter estimates differs between the different settings

Synthetic Data Analysis

- Let Q be the parameter of interest in the population
- Let q be the point estimate for Q that would have been used if the original data were available
- Let u be the variance estimate for the point estimate
- Let q_i and u_i be the obtained estimates from synthetic dataset D_i, with $i=1,\dots,m$

Synthetic Data Analysis

- The following quantities are needed for inferences

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m$$

$$b_m = \sum (q_i - \bar{q}_m)^2 / (m-1)$$

$$\bar{u}_m = \sum_{i=1}^{m} u_i / m$$

Analyzing Fully Synthetic Datasets

- The final point estimate for *Q* is given by

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m$$

- The final variance estimate is given by

$$T_f = (1 + 1/m) b_m - \bar{u}_m$$

- Difference in the variance estimate compared to standard multiple imputation is due to the additional sampling step
- Derivations are presented in Raghunathan et al. (2003)

Analyzing Fully Synthetic Datasets

- For large *n* inferences can be based on a *t*-distribution

$$(\bar{q}_m - Q) \sim t_{v_f}(0, T_f)$$

- The degrees of freedom are given by

$$v_f = (m-1)(1-\bar{u}_m/((1+1/m)b_m))^2$$

- Variance estimate can be negative
- Conservative alternative suggested by Reiter (2002) if $T_f<0$

$$T_f^* = \frac{n_{syn}}{n}\bar{u}_m$$

- Negative variances can be avoided by increasing *m*

Analyzing Partially Synthetic Datasets

- The final point estimate for *Q* again is given by

$$\bar{q}_m = \sum_{i=1}^{m} q_i / m$$

- The final variance estimate is given by

$$T_p = \bar{u}_m + b_m / m$$

- Difference in the variance estimate compared to standard multiple imputation is due to the fact that variables are fully observed
- b_m / m is the correction factor because *m* is finite
- Derivations are presented in Reiter (2003)

Analyzing Partially Synthetic Datasets

- For large *n* inferences can be based on a *t*-distribution

$$(\bar{q}_m - Q) \sim t_{v_p}(0, T_p)$$

- The degrees of freedom are given by

$$v_p = (m-1)(1 + \bar{u}_m / (b_m / m))^2$$

- Variance estimate can never be negative

- Inferences for multivariate estimands are derived in Reiter (2005a)

Some Extensions

- Handling item nonresponse and synthesis simultaneously (Reiter, 2004)
- Generate synthetic datasets in two stages to address risk-utility trade-off (Reiter and Drechsler, 2010)
- Sampling with synthesis for Census data (Drechsler and Reiter, 2010)
- Subsampling with synthesis for large datasets (Drechsler and Reiter, 2012)
- Fully synthetic data based on partial synthesis approach (Raab et al., 2017)
- Combining rules differ for the different approaches

Synthetic Data Models

Outline First Day – Theory

- Overview of Data Confidentiality
- Introduction to Synthetic Data
- Synthetic Data Models
  - Modeling Approaches
  - Practical Problems and Modeling Strategies
- Utility Checks
- Disclosure Risk Assessment

Typical Synthesis Strategy

- General approach:
  - Select values to synthesize based on risk considerations
  - Estimate regression models to predict these values from other variables
  - Simulate replacement values from regression models
- To motivate, start with partial synthesis example with no missing values

Illustrative Example of Generating Partially Synthetic Data

- 1989 Survey of Youth in Custody (SYC):
  46 facilities, 2562 youths
- Data comprise facility, race, ethnicity, and 20 crime-related variables
- Stratified sample:
  - 11 large facilities treated as strata
  - Rest grouped into 5 strata based on size
  - 2-stage PPS sample in the 5 strata

Illustrative Example of Generating Partially Synthetic Data

- Replace all values of facility with synthetic data
- Multinomial regressions of stratum indicators on main effects (some dropped due to co-linearities)
- One regression for stratum 1 – 11 and another for stratum 12 – 16

Illustrative Example of Generating Partially Synthetic Data

- For each record, compute vector of predicted probabilities for each facility
- Sample facility according to multinomial distribution with estimated probabilities
- Create m = 5 synthetic implicates
- Recalculate survey weights in each implicate to correspond to implied design

Illustrative Example of Recalculating Weights

- Say stratum 1 has 500 children total

  Synthetic D1: 10 records in stratum 1
  Weight for each: 500/10

  Synthetic D2: 12 records in stratum 1
  Weight for each: 500/12

- See Mitra and Reiter (2006) for details.

Comparison of Observed and Synthetic SYC Inferences

Variable	Obs Est	Obs CI	Syn Est	Syn CI
Avg. age	16.7	(16.6, 16.8)	16.8	(16.7, 16.9)
Avg. age Hisp	13.0	(12.7, 13.2)	13.0	(12.6, 13.2)
Avg. age Others	13.0	(12.9, 13.1)	13.0	(12.8, 13.1)
% age < 15	73.4	(71.3, 75.5)	73.1	(70.8, 75.4)
% age > 18	.39	(.16, .62)	.40	(.15, .64)
% use drugs	25.4	(23.4, 27.3)	25.2	(23.2, 27.1)
% female	7.4	(6.1, 8.6)	7.5	(6.1, 9.0)
Intercept	1.36	(.80, 1.9)	1.33	(.73, 1.9)
Age	-.08	(-.13, -.04)	-.08	(-.13, -.04)
Black	.46	(.25, .67)	.48	(.27, .69)
Asian	.33	(-.72, 1.38)	.76	(-.28, 1.79)
Amer. Indian	-.01	(-.55, .52)	-.09	(-.73, .55)
Other	1.4	(.56, 2.15)	1.2	(.42, 2.0)

Risk note: most likely facility is original for 17%.

Partial Synthesis of Entire Variables

- Suppose Y_1, Y_2, Y_3 (no missing values) to be synthesized.
- Let X represent all variables that are left unchanged.

1) Estimate regression of $Y_1 \mid X$ using all records in original data

2) Simulate synthetic values Y_1^s from this model using X

3) Estimate regression of $Y_2 \mid Y_1, X$ using all records in original data. Simulate synthetic values Y_2^s using (X, Y_1^s)

4) Repeat for Y_3 by estimating the regression $Y_3 \mid Y_1, Y_2, X$

Partial Synthesis of Entire Variables: Software

- Can use models tailored to each variable
- Can adapt free software for multiple imputation of missing data
    - Append copy of entire dataset to the original data
    - Delete all values of variables to be synthesized
    - Run software program to fill in “missing” values *m* times
    - Result is *m* partially synthetic datasets
    - MICE for R and Stata
    - IVEWARE for SAS
- Also can use “synthpop” like we do tomorrow.

Partial Synthesis and Survey Weights

- No strong theory for survey weights in partial synthesis
- When synthesizing only variables not involved in weight construction,
    - Possibly include weights as predictors in synthesis models
    - Leave survey weights at original values
- When synthesizing a variable involved in weight construction,
    - Synthesize from unweighted model
    - Re-calibrate weights as described earlier in SYC example

Full Synthesis

- Same general approach as partial synthesis of entire variables, only there is no X.
- Build chains of conditional distributions to estimate joint distribution of all k variables

$$f(Y_1, Y_2, ..., Y_k) = f(Y_1) f(Y_2 \mid Y_1) ... f(Y_k \mid Y_1, Y_2, ..., Y_{k-1})$$

- Same general strategy: tailor each conditional model to describe the distribution of corresponding outcome

Regression Models for Synthesis

- Common regression models used for synthesis
  - Linear regression for continuous variables
  - Logistic regression for binary variables
  - Multinomial logit for categorical variables
- Other models possible
- Full synthesis theory indicates one should sample from Bayesian posterior predictive distribution
- Not necessary for partial synthesis (Reiter and Kinney, 2012)

Linear Regression Synthesis

$$Y_{obs} = \beta_0 + x_{obs,1}\beta_1 + x_{obs,2}\beta_2 + \ldots + x_{obs,p}\beta_p = X_{obs}\beta + \varepsilon \quad \text{with } \varepsilon \sim \mathrm{N}(0,\sigma^2 I)$$

- Full synthesis: draw new values for the parameters

$$\sigma^2 \mid Y_{obs}, X_{obs} \sim (n_{obs} - (p+1))(y_{obs} - X_{obs}\hat{\beta})'(y_{obs} - X_{obs}\hat{\beta})\chi^{-2}_{n_{obs}-(p+1)}$$

$$\beta \mid \sigma^2, X_{obs} \sim N(\hat{\beta}, \sigma^2 (X'_{obs} X_{obs})^{-1})$$

- Partial synthesis: $\beta = \widehat{\beta},\ \sigma^2 = \widehat{\sigma}^2$
- Draw replacement values $Y_{syn} \mid \beta, \sigma^2, X_{syn} \sim N(X_{syn}\beta, \sigma^2)$
- X_{syn} might contain previously synthesized variables
- Partial synthesis model estimated with observations to be synthesized

Logistic Regression Synthesis

$$P(Y_{i,obs} = 1 \mid X_{i,obs}) = \frac{\exp(X_{i,obs}\beta)}{1 + \exp(X_{i,obs}\beta)}$$

- Full synthesis: draw new values for the parameters
  - Approximation uses $\beta \mid \hat{\Sigma}, X_{obs} \sim N(\hat{\beta}, \hat{\Sigma})$
  - $\hat{\Sigma}$ is estimated variance-covariance matrix from negative inverse of the Fisher information matrix (software output)
- Partial synthesis: $\beta = \hat{\beta}$
- Draw new $Y_{i,syn} \sim Bernoulli\left(\dfrac{\exp(X_{i,syn}\beta)}{1 + \exp(X_{i,syn}\beta)}\right)$

Multinomial Logistic Regression Synthesis

- Categorical variables with more than two categories

$$P(Y_{i,obs} = j \mid X_{i,obs}) = \frac{\exp(X_{i,obs}\beta_j)}{1+\sum_{l=1}^{J-1}\exp(X_{i,obs}\beta_l)}$$

$$P(Y_{i,obs} = J \mid X_{i,obs}) = \frac{1}{1+\sum_{l=1}^{J-1}\exp(X_{i,obs}\beta_l)}$$

- Full synthesis: draw new values $\beta \mid \hat{\Sigma}, X_{obs} \sim N(\hat{\beta}, \hat{\Sigma})$
- Partial synthesis: $\beta = \hat{\beta}$

Multinomial Logistic Regression Synthesis

- Draw replacement values
  - Calculate

$$\pi_l = \exp(X_{syn}\beta_l)/(1+\sum_{l=1}^{J-1}\exp(X_{syn}\beta_l)) \qquad \text{for } l = 1,\ldots,J-1$$

$$\pi_J = 1-\sum_{l=1}^{J-1}\pi_j$$

  - Calculate $R_j^{(i)} = \sum_{l=1}^{j}\pi_l^{(i)}$
  - Draw n_{syn} uniform random numbers, u_1, $u_2,\ldots,u_j,\ldots u_{n_{syn}}$
  - Impute category *j* for $Y_{i,syn}$ when $R_{j-1}^{(i)} \le u_i \le R_j^{(i)}$

Limitations of the Multinomial Model

- Estimating the model parameters can be problematic when
  - number of covariate is large
  - outcome variable has large number of categories
  - multicollinearity between predictor variables
  - some outcome categories are sparse
- ML estimation procedure might not converge
- Estimates unstable and have high variances
- Possible alternatives
  - Multinomial/Dirichlet model (when enough data in all the cells)
  - CART -- classification and regression trees

Full Synthesis: Software and Survey Weights

- Can create fully synthetic data using synthpop in R. Also can use IVEWARE in SAS.
- Handling survey weights
  - Need to estimate population distributions of parameters for all models, not sample distributions
- When frame variables (or weights) for whole population available, can include frame variables as X in synthesis models. Each synthetic record sampled from distribution of X.

Partial Synthesis of Selected Values

- Suppose selected values in Y_1, Y_2 to be synthesized.
- Let X represent all variables that are left unchanged.

1) Estimate regression of $Y_1 \mid Y_2, X$ using **only** records in original data with Y_1 to be replaced.

2) Simulate synthetic values Y_1^s from this model using (X, Y_2)

3) Estimate regression of $Y_2 \mid Y_1, X$ using **only** records in original data with Y_2 to be replaced. Simulate synthetic values Y_2^s using (X, Y_1^s)

Limitations of this Approach

- No principled way to determine order of synthesis
- No guarantee that the sequential models correspond to a proper joint distribution
- Labor intensive modeling tasks
- Use of parametric models can be restrictive

- These issues affect full synthesis too, although to lesser extent

Generating Synthetic Data: Nonparametric Regressions

- Common synthesis scenario

Thousands of units, dozens of variables

-- Numerical and categorical data

-- Skewed or multi-modal distributions

-- Complicated relationships

-- Many public uses

-- Intense synthesis required

Aside: these are not necessary for synthetic data approaches to be useful

Pie-in-the-sky Vision of Synthetic Data Generators

Ideal synthetic data generator would

-- preserve as many relationships as possible while protecting confidentiality

-- handle diverse data types

-- be computationally feasible for large data

-- be easy to implement with little tuning by the agency

Possible Solutions

Build synthesizers using algorithmic methods from machine learning

Regression trees (CART) described here (available in synthpop)

Other approaches based on
-- Random forests
-- Support vector machines

Drechsler and Reiter (2011) find advantages for CART over other machine learning algorithms for generating synthetic data

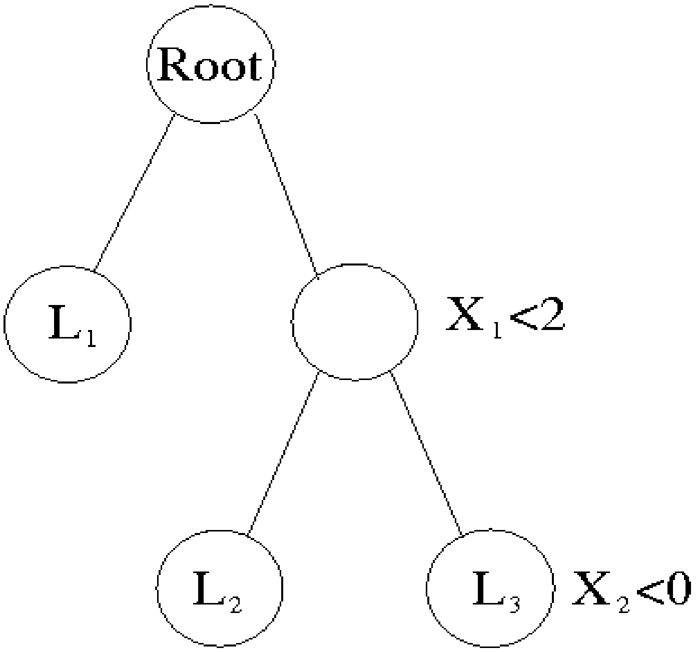
Overview of CART

Goal: Describe $f(Y | X)$.

-- Partition X space so that subsets of units formed by partitions have relatively homogenous Y

-- Partitions from recursive binary splits of X

-- Free routines in R



CART for Synthesis

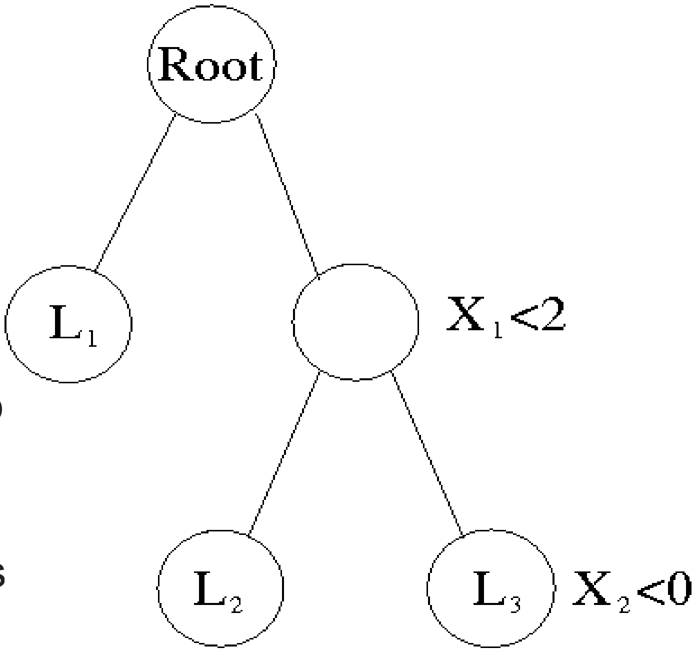
Goal: Synthesize Y | X

-- Grow large tree

-- For any X, trace down tree until reach appropriate leaf

-- Draw Y from Bayes bootstrap or smoothed density estimate

-- Can introduce noise in leaves to improve protection



CART for Synthesis

Synthesize with sequential imputations

a) using genuine data, run CART for each variable conditional on others as appropriate

b) generate new values for each variable using already synthesized data to trace down trees

Reiter (2005b) discusses order of synthesis

Illustration of CART synthesis for entire variables

- 10,000 household heads, March 2000 U.S. Current Population Survey
- Age, race, sex, marital status, education, alimony payments, child support payments, SS payments, income, property taxes
- Cross-tabs of age, race, sex, and marital status: 473 sample uniques, 241 sample duplicates
- Protect data via two scenarios:
  - Synthesize all of marital status and race
  - Synthesize all of marital status, race, and age

Illustration of CART synthesis for entire variables

- Make 5 synthetic datasets using CART
- Obtain confidence intervals using methods in Reiter (2003)
- Compare inferences for regression coefficients in original and synthetic data
- Table labeled “Table 2” indicates reasonable inferences. Problems arise with small sub-pop’s

Illustration of partial synthesis with selected records

Reiter (2005b)

- 51,016 household heads, March 2000 CPS
- Synthesize all values of race, sex, marital status and age for people with

  AP > 0 or CS > 0 or SS > 0 or I > 100,000

  (about 37% of values)

Example of partial synthesis with selected records

- Take random sample of size 10,000 from population
- Make 5 synthetic datasets using CART
- Obtain confidence intervals using formulas in Reiter (2003)
- Repeat process 1000 times
- Table labeled “Table 5” displays repeated sampling properties

Joint Models for Fully Synthetic Data

- Sometimes more effective to specify explicit joint distributions rather than sequences of conditionals
- Kim et al. (2016) – mixture of multivariate normal distributions that simultaneously handles faulty data and generates synthetic data for continuous variables subject to edit constraints
- Hu et al. (2016) – mixture of multinomial distributions for categorical data nested within households, respecting structural zeros
- Papers available from Jerry Reiter

Outline First Day – Theory

- Overview of Data Confidentiality
- Introduction to Synthetic Data
- Synthetic Data Models
  - Modeling Approaches
  - Practical Problems and Modeling Strategies
- Utility Checks
- Disclosure Risk Assessment

Generating Synthetic Data in Practice

- Basic concept of multiple imputation seems to be straight forward to apply:
  - Build a model with the original data (linear, logit, …)
  - Draw new values from this model
  - Impute missing values with drawn values
- But real data applications pose many additional challenges:
  - Semi-continuous variables
  - Skip patterns
  - Imputation within bounds
  - Logical constraints

Semi-continuous Variables

- Some variables have spike at one point of the distribution
- Often spike at $Y=0$
- Use two stage imputation suggested by Raghunathan et al. (2001)
- Build logit model to impute if $Y=0$ or not
- Apply linear model only to records with $Y_{org} \neq 0$ to obtain parameter estimates
- Generate synthetic values only for those records with a positive predicted value from the logit model
- Set all other values to zero

Skip Patterns

- Skip patterns are very common in surveys
- Most SDC techniques cannot deal with skip patterns appropriately
- For synthetic data approach is comparable to the approach for dealing with semi-continuous variables
- Use logit model to decide if filtered questions are applicable
- Impute values only for records with a positive outcome from the logit model

Imputation Within Bounds

- Sometimes it is known that values of a specific variable have to lie within a certain interval
- Imputed values are required to fall into certain bounds
- Simple method: redraw from the model until restriction is fulfilled for all records
- In practice an upper bound *z* needs to be defined for the number of draws
- After *z* unsuccessful draws, the imputed value is set to the closest boundary

Imputation Within Bounds

- Heuristic approach
- Only possible, if truncation point is at the far end of the assumed distribution of the imputation model
- Otherwise, model is mis-specified
- Correction method distorts quantities like the mean that would still have been unbiased under the mis-specified model
- Useful to monitor the number of times the imputed value is set to the closest boundary
- Sometimes better to refine the model
- Implementation: draw from a truncated model

Logical Constraints

- The values of one variable always need to be at least as large as the values of another variable
- E.g., total number of employees>number of part time employees
- Simple approach: redraw from the model until constraint is always fulfilled
- Alternative: transform smaller variable and use transformed variable for imputation

Logical Constraints

- Let $Y_1 > Y_2$
- Synthesize Y_1 with standard imputation model
- Imputation of Y_2 in five steps
  - Generate $x = Y_2/Y_1$
  - Generate $z = logit(x) = log(x/(1-x))$
  - Use standard linear model for z
  - Use inverse logit on z_{syn} to get x_{syn}: $x_{syn} = exp(z_{syn})/(1+exp(z_{syn}))$
  - Multiply by $Y_{1,syn}$ to get final synthetic values for Y_2

Modeling Strategies

- Which variables should be included in the synthesis model?
- Is it wise to condition on all variables?
- What are the consequences of excluding variables from the model?
- Is there a way to automatically select the “right” variables?
- In which order should the variables be synthesized?

Which Variables Should Be Included in the Model?

- Modeling step essential for the quality of the synthetic data
- If model is mis-specified, results from the synthetic data will be biased
- Only those relationships that are incorporated in the model will be reflected in the synthetic data
- At least include all variables that will be part of the analysis model
- Include those variables that explain a considerable amount of variance of the target variable

Is it Wise to Condition on All Variables?

- All relationships of interest should be reflected in the synthetic data
- Impossible to know all potentially interesting analyses in advance
- Relationship with variables not included in the model biased towards zero
- Good advice to condition on all variables in the dataset if possible
- Similar to multiple imputation for nonresponse if imputer and analyst are different persons
- Not conditioning on some variables can lead to uncongeniality (Meng, 1994)

Imputation and Uncongeniality

- The theory of multiple imputation is based on the assumption that all three models, the data generating model, the imputation model, and the analyst's model are identical.
- Meng (1994) coined the term uncongeniality if the imputation model differs from the analyst's model.
- Two scenarios possible:
  (1) The analyst's model is based on more information (less free parameters) than the imputation model.
  (2) The analyst's model is based on less information (more free parameters) than the imputation model.

Imputation and Uncongeniality

- In the first scenario, the results are still unbiased but some efficiency is lost.
- In the second scenario, results are potentially biased unless the more restrictive assumptions in the imputation model are correct (super-efficiency).
- If the imputer and the analyst are the same person, the analyst should include all those variables in the imputation model that he wants to use in his analysis (No endogeneity possible!).
- If the imputer and the analyst are different persons, the general advise is to include as many variables in the imputation model as possible.

Is There a Way to Automatically Select the "Right" Variables?

- Typical variable selection procedures based on forward or backward selection
- Final model somewhat arbitrary
- Decision based on *p* values
- Not really helpful in the multiple imputation context
- Especially not, if aim is to avoid multicollinearity
- Area of future research

In Which Order Should the Variables be Synthesized?

- In theory ordering doesn't matter
- In practice all models are wrong
- Different orderings can lead to different levels of disclosure risk and analytical validity
- Ordering variables from largest amount of synthesis to lowest might increase analytical validity
- Ordering variables from lowest amount of synthesis to largest might increase disclosure protection
- If same amount of data is synthesized ordering could be based on evaluations of the model fit

Utility Checks

Outline First Day – Theory

- Overview of Data Confidentiality
- Introduction to Synthetic Data
- Synthetic Data Models
- Utility Checks
- Disclosure Risk Assessment

Evaluating the Analytical Validity

- Quality of the imputation models is essential
- Evaluating the quality of the model easier for synthetic data than for imputed data
- Model evaluation criteria for imputation models still useful
- Four different dimensions
  - Imputation model evaluation
  - Exploratory comparisons
  - Global validity measures
  - Model specific validity measures

Imputation Model Evaluation

- Most imputation models are parametric models
- Model assumptions and fit of the model should be evaluated
- Possible evaluation methods
  - Q-Q plots
  - Plots of the residuals from the regression against the fitted values
  - Binned residual plots
- All methods that should be used to check the assumptions in an applied analysis can be used.

Exploratory Comparisons

- Should always be the first step
- Easy to carry out
- Compare quantiles, means, histograms etc.
- Check for reasonability
  - No small scale grocery store with 100 Mio euro turnover
  - No negative income
  - No pregnant fathers
  - Etc.
- Evaluations by subject matter experts

Global Analytical Validity Measures

- Ideally it would be possible to compare the original and synthetic data directly using one global measure
- Difficult in practice
- Global measures like the Kullback-Leibler or Hellinger distance too general
- Might not identify important differences
- Only suitable to compare different synthetic datasets
- Useful alternative: evaluations based on propensity score matching

Global Analytical Validity Measures

- Proposed by Woo et al. (2009)
- Idea is to measure how well one can discriminate between the original data and synthetic data
- Based on the literature on causal inference for observational studies (Rosenbaum and Rubin, 1983)
- Main steps
  - Stack the original and the synthetic data
  - Include an indicator for the data source
  - Calculate the propensity of being “assigned” to the original data
  - Compare the distributions of the estimated propensity score in the two datasets

Global Analytical Validity Measures

- Distribution should be similar (ideally close to 0.5 for all the records)
- Significant coefficients in the propensity models identify variables for which the synthesis didn’t work
- Can only be applied to each synthetic dataset separately

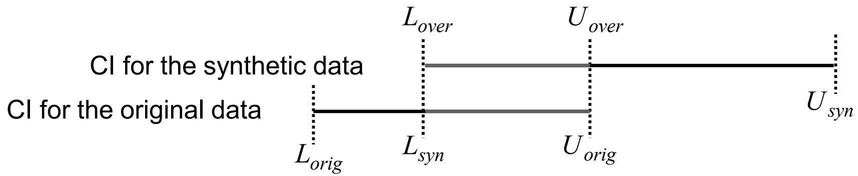
Model Specific Validity Measures

- Directly compare the validity of specific analysis of interest
- Comparing point estimates (means, regression coefficients) not sufficient
- Point estimates might look substantially different but statistical inference still comparable because of high uncertainty in the estimates
- Point estimates might look similar but statistical inference different because parameter estimates are very precise
- Useful measure: confidence interval overlap

Confidence Interval Overlap

- Suggested by Karr et al. (2006)
- Measure the overlap of CIs from the original data and CIs from the synthetic data
- The higher the overlap, the higher the data utility
- Compute the average relative CI overlap for any estimate of interest

$$J_k = \frac{1}{2}\left[\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}}\right]$$



Output Example Based on the IAB Establishment Panel

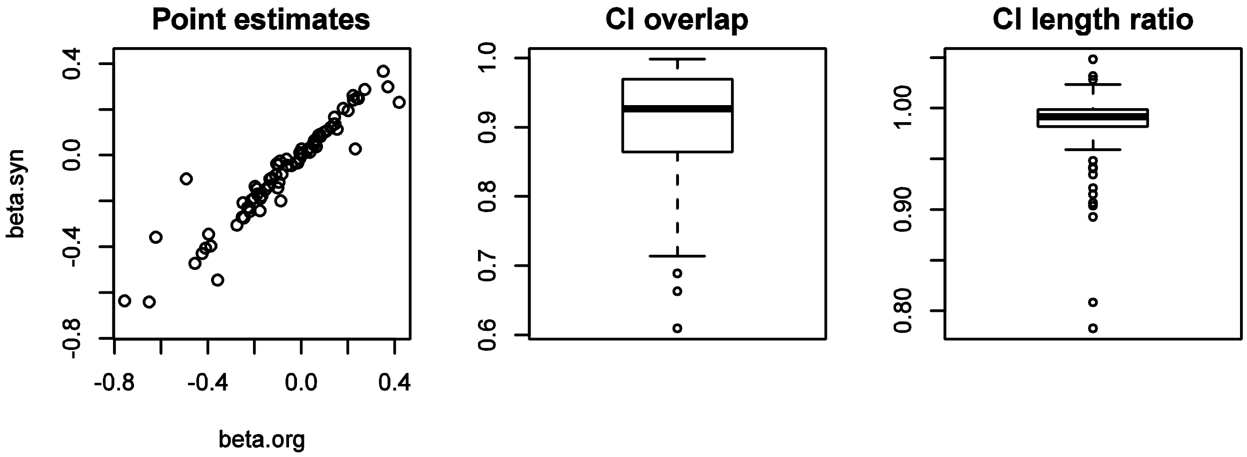
- Dependent variable: part-time employment (yes/no)

	beta org.	*beta syn.*	*J.k.beta*	*z-score org.*	*z-score syn.*	*CI length ratio*
Intercept	-0.809	-0.752	0.87	-7.23	-6.85	0.99
5-10 employees	0.443	0.437	0.97	8.52	7.99	1.06
10-20 employees	0.658	0.636	0.90	11.03	10.88	0.98
20-50 employees	0.797	0.785	0.95	13.02	12.36	1.04
100-200 employees	0.892	0.908	0.96	9.23	9.48	0.99
200-500 employees	1.131	1.125	0.99	9.99	9.87	1.01
>500 employees	1.668	1.641	0.97	8.22	8.33	0.97
growth in employment exp.	0.010	0.006	0.98	0.18	0.12	0.99
decrease in emp. expected	0.087	0.100	0.96	1.11	1.27	1.00
share of female workers	1.449	1.366	0.73	17.63	18.71	0.89
share of employees with university degree	0.319	0.368	0.91	2.18	2.59	0.97
share of low qualified workers	1.123	1.148	0.93	12.17	11.87	1.05
share of temporary employees	-0.327	-0.138	0.75	-1.74	-0.71	1.05
share of agency workers	-0.746	-0.856	0.88	-3.09	-4.24	0.84
employment in the last 6 month	0.394	0.369	0.87	8.33	7.82	1.00
dismissal in the last 6 months	0.294	0.279	0.92	6.38	6.03	1.00
foreign ownership	-0.113	-0.117	0.99	-1.33	-1.38	0.99
good or very good profitability	0.029	0.033	0.98	0.72	0.82	0.99
salary above collective wage agreement	0.020	0.031	0.95	0.35	0.54	0.99
collective wage agreement	0.016	0.007	0.95	0.31	0.13	0.97

- Average overlap: 0.92

Graphical Alternative

- Based on 78 point estimates



- Average CI overlap: 0.91   minimum CI overlap: 0.61

Disclosure Risk Assessment

Outline First Day – Theory

- Overview of Data Confidentiality
- Introduction to Synthetic Data
- Synthetic Data Models
- Utility Checks
- Disclosure Risk Assessment
- The Future of Synthetic Data

Evaluating the Risk of Disclosure

- Every data release increases the risk of disclosure
- Risks should be evaluated before any release
- But this is not easy…
  - Not clear what intruders know about the released records
  - Not clear how they will attack the released data
- Possible solution: Evaluate risks under different scenarios of intruder knowledge

Disclosure Risk for Fully Synthetic Data

- Risk of identification disclosures should be low
  - When generated properly, released units cannot be matched meaningfully to external data sources
  - But risks possible with over-saturated synthesizers
    - Suppose data have 4 binary variables
    - Only one case in confidential data with x = (0, 0, 0, 0)
    - Use multinomial synthesizer with probabilities equal to empirical frequencies
    - When synthetic data include a case at (0, 0, 0, 0), someone in confidential data must have those values

Disclosure Risk for Fully Synthetic Data

- Risk of attribute disclosures not zero
- Synthesizer model may perfectly predict some *x* for a certain type of individual, so that synthetic *x* for individuals of this type always match actual *x*
  - In regression tree, all individuals in some group have same race, and tree splits to make this perfect prediction. For individuals in that group, synthetic race equals true race.
- Related, synthetic data models may be too accurate in predicting some values, particularly outliers
  - In regression tree, smoothed draws may not have enough noise, especially when multiple datasets released

Disclosure Risk Measures of Reiter et al. (2014)

- Reiter et al. (2014) use conservative assumption: intruder knows all but one target record (value)
- Evaluate the posterior distribution of possible original values of target given the released data and information about the data generation mechanism

$$\Pr(Y_i \mid D, X, d_{-i}^{org}, M)$$

with:

D	set of released synthetic datasets
X	unchanged original data (for partial synthesis)
M	any additional information about the generation of *D*
d_{-i}^{org}	the original data excluding record *i*

How do we estimate $\Pr(Y_i \mid D, X, d_{-i}^{org}, M)$?

$$\Pr(Y_i \mid D, X, d_{-i}^{org}, M) \propto \Pr(D \mid Y_i, X, d_{-i}^{org}, M) \Pr(Y_i \mid X, d_{-i}^{org}, M)$$

- Prior beliefs $\Pr(Y_i \mid X, d_{-i}^{org}, M)$ unknown to the agency
- Evaluate risks under reasonable prior distributions
- Reasonable options
  - use uniform prior over a sensible range
  - use priors based on d_{-i}^{org}
  - use a prediction model, for example the one used in *M*

- Examples: Paiva et al. (2014), Hu et al. (2014)

Measuring Disclosure Risk for Partially Synthetic Data

- Risk of disclosure generally higher, since possible to match cases (confidential and synthetic data consist of the same records)
- Distinguish two scenarios
  - Intruder knows that record of interest is in the sample
  - Intruder doesn't know that record of interest is in the sample
- First scenario is conservative and computationally easier
- Second approach takes additional uncertainty from sampling into account

General Setting for Both Approaches

- Identification disclosure risk measures based on Reiter & Mitra (2009) and Drechsler and Reiter (2008)
- Compute probabilities of re-identification for each record *j, (j=1,…,n)* in the released dataset
- Individual level risk measures useful on their own. Can be aggregated to file level risks.
- Assumptions:
  - Intruder has exact information for some target record *t*
  - Target record may not correspond to unit in released data

General Setting for Both Approaches

- Let t_0 be the unique identifier for the target record
- Let d_{j0} be the identifier for record *j* in the released data $D = \{D^{(1)},\ldots,D^{(m)}\}$, $j=1,\ldots,n$
- Intruder: match when $t_0 = d_{j0}$; no match when $t_0 \neq d_{j0}$
- Let *J* be a random variable with

$$J = \begin{cases} j & \text{for } d_{j0} = t_0 \text{ and } j \in D \\ n+1 & \text{for } d_{j0} = t_0 \text{ and } j \notin D \end{cases}$$

General Setting for Both Approaches

- Intruder seeks to calculate

$$\Pr(J = j \mid t, D, M) = \int \Pr(J = j \mid t, Y_{rep}, D, M)\,\Pr(Y_{rep} \mid t, D, M)\,dY_{rep}$$

with: *D* set of released synthetic datasets

M any additional information about the generation of *D*

- Intruder can decide whether or not any identification probabilities are large enough to declare a match
- Intruder does not know actual values in Y_{rep}
- Integrate over its possible values

General Setting for Both Approaches

$$\Pr(J = j \mid t, D, M) = \int \Pr(J = j \mid t, Y_{rep}, D, M) \Pr(Y_{rep} \mid t, D, M) dY_{rep}$$

- Monte Carlo approach to estimate $\Pr(J = j \mid t, D, M)$

  (1) generate Y_{new}, a sample of Y_{rep} drawn from $\Pr(Y_{rep} \mid t, D, M)$

  (2) compute $\Pr(J = j \mid t, Y_{rep} = Y_{new}, D, M)$ using exact or, for continuous synthesized variables, distance-based matching assuming Y_{new} are collected values

- Iterate the two steps large number of times.
- Estimate $\Pr(J = j \mid t, D, M)$ as average over all

  $\Pr(J = j \mid t, Y_{rep} = Y_{new}, D, M)$

- Default: simulated values as plausible draws of Y_{rep}

Example – Ignoring the Uncertainty from Sampling

- Let age, race, and sex be the only quasi-identifiers in a survey (notice weakness of this risk evaluation – unverifiable assumption about intruder knowledge)
- Suppose agency releases no information about the imputation models
- Intruder seeks to identify a white male aged 45 and **knows the target is in the sample**
- Intruder matches on age, race and sex
- Calculate average matching probability based on the released synthetic values

Ignoring the Uncertainty from Sampling

- Average matching probability

$$p_{match,j} = \Pr(J = j \mid t, D, M) = (1/m)\sum_k (1/N^{(k)}) I_j^{(k)}$$

with $N^{(k)}$ = number of records that fulfill the matching criteria in dataset *k, k=1,…,m*

$I_j^{(k)}$ = 1 if record *j* is among the $N^{(k)}$ records in dataset *k*, 0 otherwise

m = number of synthetic datasets

- Probability that target record is not in the sample: zero

Example – Accounting for the Uncertainty from Sampling

- Intruder seeks to identify a white male aged 45 and **does not know the target is in the sample**
- Replace $N^{(k)}$ with F_t, the number of records in the population that match on age, race and sex
- F_t usually unknown and estimated (e.g., using log-linear models of Elamir & Skinner, 2006)

Accounting for the Uncertainty from Sampling

$$p_{match,j} = \Pr(J = j \mid t, D, M) = (1/m)\sum_k \min(1/F_t, 1/N^{(k)}) I_j^{(k)}$$

with F_t = number of records that fulfill matching criteria in the population

$N^{(k)}$ = number of records that fulfill matching criteria in dataset *k, k=1,…,m*

$I_j^{(k)}$ = 1 if record *j* is among the $N^{(k)}$ records in dataset *k*, 0 otherwise

m = number of synthetic datasets

- Estimated probability that target not in released data

$$\Pr(J = s+1 \mid t, D, M) = 1 - \sum_{j=1}^{s} \Pr(J = j \mid t, D, M)$$

Accounting for the Uncertainty from Sampling

- In many cases $\Pr(J = s+1 \mid t, D, M)$ is highest probability
- Reasonable to assume that intruder will not match when this is the case
- Alternatively agency can define a threshold *γ* and assume that the intruder only matches when

$$\Pr(J = s+1 \mid t, D, M) < \gamma$$

Summarizing the Identification Probabilities

- Prudent to assume that intruder selects record *j* with highest value of $\Pr(J = j \mid t, D, M)$
- Identification probabilities are calculated for each target record
- Summaries helpful to quantify the overall risk
- Further definitions:

c_j = number of records with $max(p_{match,i})$ for target t_j
I_j = 1 if true match is among the c_j units, 0 otherwise
K_j = 1 if c_jI_j=1, 0 otherwise
F_j = 1 if $c_j(1-I_j)$=1, 0 otherwise
s = number of records with c_j=1

File Level Identification Disclosure Risk Measures

- Three measures of disclosure risk for partially synthetic datasets
- Expected match risk $\sum_j (1/c_j) I_j$
- True match rate $\sum K_j / s$
- False match rate $\sum F_j / s$

Attribute Disclosure Risk Measures for Partial Synthesis

- Conservative approach is to assume intruder identifies the correct record, then uses released data to estimate unknown sensitive value
- Continuous variables: can use relative squared error based on (average of imputed values – true value)
- Categorical variables: can use most frequently occurring value as best guess

Differential Privacy

- Recently, agencies have started trying to create synthetic data that satisfy differential privacy
  - Proposed by Dwork et al. (2006) and others
- Criterion that mathematically encodes the following heuristic
  - When publishing a statistic S, make it difficult for users to tell whether any particular individual was in data used to make S
  - Released value of S plausibly could have been generated from a dataset that includes any (hypothetical) individual or excludes that individual
- Differential privacy is a criterion, not a technique.

Formal Definition

- *A randomized function* κ *gives ε-differential privacy if and only if for all datasets* D_1 *and* D_2 *differing on at most one element, and for all* $S \subset Range(\kappa)$

$$P(\kappa(D_1) \in S) \leq \exp(\varepsilon) P(\kappa(D_2) \in S)$$

- The probability of obtaining a specific result does not change significantly, if one uses D_2 instead of D_1
- Implies that the amount of information that can be obtained about a single individual is bounded

Basic Implementations of Differential Privacy

- Building blocks are simple mechanisms like the Laplace mechanism or geometric mechanism

  published statistic = true statistic + random noise

- Random noise has a standard deviation that can be tuned to offer more or less privacy
  - Degree of privacy depends on a tunable parameter called ε (known as privacy-loss budget)
  - Higher values mean less noise, which means less privacy protection
  - Literature recommends ε < 1, but acknowledges that this may not always be feasible or reasonable

The Laplace Mechanism

- Define the sensitivity of an output of function f as

$$\Delta f = \max_{D_1, D_2} \left\| f(D_1) - f(D_2) \right\|_1$$

- Add noise to value computed with confidential data based on Laplace distribution

$$\kappa(x) = f(x) + (\text{Lap}(\mu = 0, b = \Delta f / \varepsilon))^k$$

Example of Laplace Mechanism Outputs

True Count	ε = .1	ε = .25	ε = 1	ε = 10
1,000,000	1,000,005.8	999,996.5	999,999.5	1,000,000.1
100,000	99,993.9	100,002.1	99,999.3	100,000.0
10,000	9,978.0	9,999.4	9,997.4	9,999.8
1,000	9,996.5	1,002.4	999.2	1,000.1

Why Differential Privacy is so Attractive

- Offers formal privacy guarantees
  - Bounds risk even for intruder with very strong background knowledge
  - Avoids ad hoc assumptions about intruder knowledge
- Privacy leakage can be quantified and cumulated
  - Release one product with ε = 1 and another with ε = 2, then total privacy loss is bounded by ε = 3
  - Can bound total privacy loss for all uses of data
- Immune to future risk of data release

Approaches for Generating DP Synthetic Data

- Add DP noise to each cell of fully cross-classified table
  - Create as many microdata records as noisy counts
  - Measurement error model to propagate uncertainty
- Add noise to sufficient statistics of model, and generate data from noisy sufficient statistics
  - Again, measurement error model could help
- Use DP machine learning methods
  - GANs, random trees
- Bayesian models (with certain priors) can be DP

Why isn’t everyone doing this?

- Most methods have not been shown to produce synthetic data with acceptable utility for reasonable values of ε.
  - Curse of dimensionality
- No way yet to handle survey weights, nor calibration, editing, and nonresponse adjustments
- General issues with DP affect synthetic data apps too
  - Can be difficult to interpret and establish ε
  - Engineering is challenging
- Very active area of research, so stay tuned….

The Future of Synthetic Data

Outline Second Day – Practical Applications

- Overview of Data Confidentiality
- Introduction to Synthetic Data
- Synthetic Data Models
- Utility Checks
- Disclosure Risk Assessment
- The Future of Synthetic Data

Chances for Synthetic Datasets

- Try to preserve high analytical validity
- Low risk of disclosure
- Easy to use for the analyst compared to other sophisticated methods
- Allow to address many real data problems
- Highly sensitive datasets might be released
- Users can decide if the data is suitable for their analysis if information about the imputation models is released

Obstacles for Synthetic Datasets

- Burdensome for the agency
- Good modeling skills and knowledge of the data required
- New models need to be developed for each dataset
- Only relationships that are included in the model will be reflected in the released data
- Analysts skeptical to use fake data

Glimpse into the Future

- Might be possible to develop generally applicable synthesizers
- Verification servers
  - See recent paper by Barrientos *et al.* (2018) in *Annals of Applied Statistics*
- Promise to run analysis code on the original data in the end as an incentive to use synthetic data
- Software to simplify generating synthetic data is starting to become available

The Role of Synthetic Data

- Synthetic data as one piece in the data access toolkit
- Most useful approach to disseminate data
- Other methods might not sufficiently protect the data in the future given the ever increasing availability of digital information
- Alternatives: on-site use, remote access
- Useful to get familiar with the data
- Provide access to data that might otherwise not be available
- Still much research necessary but critical step from pure theoretical concept to practical implementation has been managed successfully

References

References

- Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies, 215–277. Amsterdam: North-Holland.
- Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., Privacy in Statistical Databases, 290–297. New York: Springer.
- Barrientos, A. F, Bolton, A. Balmat, T., Reiter, J. P.,de Figueired, J. M., Machanavajjhala, A., Chen, Y., Kneifel, C., DeLong, M. (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the U. S. federal government. Annals of Applied Statistics, 12, 1124-1156.
- Drechsler, J. and Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin, eds., Privacy in Statistical Databases, 227–238. New York: Springer.
- Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. Journal of the American Statistical Association 105, 1347–1357.
- Drechsler, J. and Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets, Computational Statistics and Data Analysis, 55, 3232 –3243.
- Dwork, C., Mcsherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In Proceedings of the 3rd Theory of Cryptography.
- Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical report, Department of Statistics, Carnegie-Mellon University.
- Fienberg, S. E., Makov, U. E., and Steele, R. J. (1998). Disclosure limitation using perturbation and related methods for categorical data. Journal of Official Statistics 14, 485–502.
- Hu, J., Reiter, J. P., and Wang, Q. (2014). Disclosure risk evaluation for fully synthetic data, in Privacy in Statistical Databases, edited by J. Domingo-Ferrer, Lecture Notes in Computer Science 8744, Heidelberg: Springer, 185–199.

References

- Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. The American Statistician 60, 224–232.
- Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., Record Linkage Techniques, 1997, 248–267.Washington, DC: National Academy Press.
- Kinney, S. K. and Reiter, J. P. (2010). Tests of multivariate hypotheses when using multiple imputation for missing data and disclosure limitation. Journal of Official Statistics 26, 301–315.
- Little, R. J. A. (1993). Statistical analysis of masked data. Journal of Official Statistics 9, 407–426.
- Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). Statistical Science 9, 538–558.
- Mitra, R. and Reiter, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data, in Privacy in Statistical Databases 2006, Lecture Notes in Computer Science, New York: Springer-Verlag, 177–188
- Paiva, T., Chakraborty, A., Reiter, J. P., Gelfand, A. E., (2014). Imputation of confidential data sets with spatial locations using disease mapping models, Statistics in Medicine, 33, 1928–1945.
- Raab, G. M., Nowok, B., and Dibben, C. (2017) Practical Data Synthesis for Large Samples. Journal of Privacy and Confidentiality: Vol. 7 : Iss. 3 , Article 4.
- Raghunathan, T. E., Lepkowski, J. M., van Hoewyk, J., and Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a series of regression models. Survey Methodology 27, 85–96.
- Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. Journal of Official Statistics 19, 1–16.

References

- Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. Journal of Official Statistics 18, 531–544.
- Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. Survey Methodology 29, 181–189.
- Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. Survey Methodology 30, 235–242.
- Reiter, J. P. (2005a). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. Journal of Statistical Planning and Inference 131, 365–377.
- Reiter, J. P. (2005b). Using CART to generate partially synthetic, public use microdata. Journal of Official Statistics 21, 441–462.
- Reiter, J. P. and Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. Statistica Sinica 20, 405–421.
- Reiter, J. P. and Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary, Journal of Official Statistics, 28, 583–590.
- Reiter, J. P. and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. Journal of Privacy and Confidentiality 1, 99–110.
- Reiter, J. P., Wang, Q., and Zhang, B. (2014). Bayesian estimation of disclosure risks in multiply imputed, synthetic data, Journal of Privacy and Confidentiality, 6:1, Article 2.
- Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
- Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. Journal of Official Statistics 9, 462–468.
- Schafer, J. L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman and Hall.

References

- Schenker, N., Raghunathan, T. E., Chiu, P. L., Makuc, D. M., Zhang, G., and Cohen, A. J. (2006). Multiple imputation of missing income data in the National Health Interview Survey. Journal of the American Statistical Association 101, 924–933.
- Winkler, W. E. (2007). Examples of easy-to-implement, widely used methods of masking for which analytic properties are not justified. Technical report, Statistical Research Division, U.S. Bureau of the Census, Washington, DC.
- Woo, M. J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. Journal of Privacy and Confidentiality 1, 111–124.

Thank you for your attention

Handouts for Short Course on Synthetic Data

Handouts for Short Course on Synthetic Data

This handout describes results of repeated sampling simulations for synthetic data generation with the CART synthesizer. Results taken from Reiter (2005, *Journal of Official Statistics*).

Table 1: Description of variables used in the empirical studies

Variable	Label	Range	Notes
Sex	*X*	male, female	
Race	*R*	white, black, Amer. Indian, Asian	
Marital status	*M*	7 categories	
Highest attained education level	*E*	16 categories	
Age (years)	*G*	15 – 90	integers
Household alimony payments ($)	*A*	0 – 54,008	0.4% have A>0
Child support payments ($)	*C*	0 – 23,917	3.3% have C>0
Social security payments ($)	*S*	0 – 50,000	23.6% have S>0
Household property taxes ($)	*P*	0 – 99,997	64.8% have P>0
Household income ($)	*I*	-21,011 – 768,742	11.7% have I>100,000

Table 2. Simulation results when imputing sensitive variables: Simple estimands and a multiple regression involving child support payments

			95% CI Coverage	
Estimand	Q	Avg. $\bar{q}_5$	Observed	Synthetic
Average income	52632	52893	96.4	92.6
Average social security	2229	2225	94.9	94.8
Average child support	139	137	93.9	92.6
Average alimony	41	42	92.5	92.4
% of households with income > 200,000	2.10	2.10	95.3	95.9
% of households with social security > 10,000	10.53	10.25	96.5	85.4
Coefficient in regression of *A* on:				
Intercept	4315	6087	89.6	88.6
Income	.14	.08	67.7	73.8
Coefficient in regression of *A* on:				
Intercept	9846	10046	92.2	92.9
Child support	.078	.065	97.2	96.4
Coefficient in regression of *S* on:				
Intercept	2999	3017	93.7	92.0
Income	-.015	-.015	93.0	91.0
Coefficient in regression of $\sqrt{C}$ on:				
Intercept	-93.28	-64.91	94.7	79.8
Indicator for sex=female	13.30	1.57	96.0	38.1
Indicator for race=black	-9.69	-6.49	96.9	93.4
Education	3.37	3.01	95.2	89.8
Number of youths in house	2.95	1.69	93.1	82.5

Population means and percentages calculated using all records. See Table 1 for percentages of imputed values.
Alimony regressions fit using records with A>0. 100% of these records have imputed *A*.
Social security regression fit using all records. 33% of these records have imputed *S* or *I*.

Table 3. Simulation results when imputing sensitive variables: Multiple regressions involving incomes and social security payments

			95% CI Coverage	
Estimand	Q	Avg. $\bar{q}_5$	Observed	Synthetic
Coefficient in regression of $\sqrt{S}$ on:				
Intercept	79.87	82.97	93.7	84.6
Indicator for sex=female	-13.30	-12.94	94.2	89.5
Indicator for race=black	-5.85	-4.68	95.5	84.7
Indicator for race=American Indian	-7.00	-5.01	94.3	96.7
Indicator for race=Asian	-3.27	-2.11	90.2	96.2
Indicator for marital status=married in armed forces	2.08	-0.71	92.6	84.2
Indicator for marital status=widowed	7.30	6.47	95.2	88.4
Indicator for marital status=divorced	-0.88	-1.12	95.1	91.3
Indicator for marital status=separated	-5.44	-4.67	96.6	97.0
Indicator for marital status=single	-1.54	-1.05	93.9	91.2
Indicator for education=high school	5.49	5.60	95.3	92.3
Indicator for education=some college	6.77	7.13	96.3	93.9
Indicator for education=college degree	8.28	9.10	93.7	88.3
Indicator for education=advanced degree	10.67	11.90	89.2	90.6
Age	0.21	0.17	94.1	85.1
Coefficient in regression of $\log(I)$ on				
Intercept	4.92	4.90	92.9	93.2
Indicator for race=black	-0.17	-0.17	94.5	94.4
Indicator for race=American Indian	-0.25	-0.25	89.5	89.0
Indicator for race=Asian	-0.0064	-0.010	92.5	92.8
Indicator for sex=female	0.0035	-0.0011	96.9	96.4
Indicator for marital status=married in armed forces	-0.52	-0.52	94.5	95.5
Indicator for marital status=widowed	-0.31	-0.30	96.5	96.6
Indicator for marital status=divorced	-0.31	-0.30	94.1	93.8
Indicator for marital status=separated	-0.52	-0.52	88.8	89.0
Indicator for marital status=single	-0.32	-0.31	92.7	92.7
Education	0.11	0.11	93.0	92.9
Indicator for household size > 1	0.50	0.50	93.0	93.2
Interaction for females married in armed forces	-0.52	-0.52	92.5	92.4
Interaction for widowed females	-0.31	-0.30	95.6	95.8
Interaction for divorced females	-0.31	-0.30	94.6	94.5
Interaction for separated females	-0.52	-0.52	91.1	91.0
Interaction for single females	-0.32	-0.31	90.8	91.0
Age	0.044	0.044	93.1	93.2
Age^2	-0.00044	-0.00044	93.4	93.3
Property tax	0.000037	0.000040	52.3	53.1

Social security regression fit using records with S>0 and G>54. 100% of these records have imputed S.

Table 5. Simulation results when imputing key variables

			95% CI Coverage	
Estimand	Q	Avg. $\bar{q}_5$	Observed	Synthetic
Avg. education for married black females	39.44	39.46	94.4	94.1
Coefficient in regression of $\sqrt{C}$ on:				
Intercept	-93.28	-88.11	94.5	93.8
Indicator for sex=female	13.30	7.46	96.2	81.3
Indicator for race=black	-9.69	-5.26	94.3	88.2
Education	3.37	3.38	94.2	94.5
Number of youths in house	2.95	2.67	93.9	93.6
Coefficient in regression of $\sqrt{S}$ on:				
Intercept	79.50	83.79	94.6	81.3
Indicator for sex=female	-13.34	-12.94	93.8	91.3
Indicator for race=black	-6.04	-6.12	94.5	94.2
Indicator for race=American Indian	-7.12	-4.48	94.7	95.0
Indicator for race=Asian	-3.22	-2.19	89.3	94.7
Indicator for marital status=widowed	7.37	7.20	94.5	94.2
Indicator for marital status=divorced	-0.79	-0.96	93.7	96.4
Indicator for marital status=single	-1.46	0.18	93.8	92.3
Indicator for education=high school	5.51	5.53	94.8	95.8
Indicator for education=some college	6.78	6.77	94.5	94.8
Indicator for education=college degree	8.31	8.12	92.7	92.4
Indicator for education=advanced degree	10.72	10.99	89.1	90.6
Age	0.22	0.16	93.8	80.6
Coefficient in regression of $\log(I)$ on				
Intercept	4.92	4.95	91.2	90.2
Indicator for race=black	-0.17	-0.17	94.9	94.3
Indicator for race=American Indian	-0.25	-0.25	88.6	91.0
Indicator for race=Asian	-0.0064	-0.0045	92.5	92.0
Indicator for sex=female	0.0035	-0.0018	96.2	95.5
Indicator for marital status=married in armed forces	-0.028	-0.091	94.9	90.4
Indicator for marital status=widowed	-0.015	-0.057	96.6	89.4
Indicator for marital status=divorced	-0.16	-0.16	93.5	93.9
Indicator for marital status=separated	-0.24	-0.23	87.3	88.5
Indicator for marital status=single	-0.17	-0.17	93.3	94.1
Education	0.11	0.11	93.0	92.2
Indicator for household size > 1	0.50	0.50	93.5	92.1
Interaction for females married in armed forces	-0.52	-0.43	92.2	88.9
Interaction for widowed females	-0.31	-0.27	96.8	90.0
Interaction for divorced females	-0.31	-0.30	92.8	93.1
Interaction for separated females	-0.52	-0.48	89.0	89.1
Interaction for single females	-0.32	-0.31	92.2	92.7
Age	0.044	0.043	94.1	91.3
Age^2	-0.00044	-0.00043	94.4	92.8
Property tax	0.000037	0.000040	51.8	51.8

Average education calculated using all black females. 29.2% of these records have imputed *G*, *M*, *X*, and *R*.
Child support regression fit using records with $C>0$. 100% of these have imputed *G*, *M*, *X*, and *R*.
Social security regression fit using records with $S>0$ and $G>54$. 100% of these have imputed *G*, *M*, *X*, and *R*.