Imputation as a Practical Alternative to Data Swapping

FCSM/CDAC Workshop on New Advances in Disclosure Limitation

Saki Kinney

September 27, 2017

Overview

- Background
  - Data swapping and synthetic data
- Approach
- Sample survey project
- Disclosable example
- Future Work
  - Transparency

Data Swapping

- Involves swapping a portion of values of certain variables between records in order to add uncertainty to any attempted re-identification
  - Used by several agencies for demographic, lower risk datasets
  - Precisely preserves marginal distributions but distorts relationships between swapped and unswapped variables
  - Few publicly available routines to facilitate swapping
  - Simple in principle but more difficult to implement for complex data or for very many variables

Data Swapping

- Disclosure protection requires keeping swapping rates, and other details, secret
  - Prevents analysts from accounting for swapping in their analyses
- Generally known that the rate of swapping is limited
  - Swapped data are analyzed as if they are real data; in some cases restricted-use or gold standard data are also swapped
  - Researchers found utility problems even w/very low rates (Drechsler & Reiter 2010)
- Often used in conjunction with coarsening and suppression

Synthetic Data

- Protect confidentiality by replacing values of confidential data with multiple imputations
  - Often most or all of a dataset is replaced with imputed values, generated by modeling the joint distribution of data being imputed conditional on data not being imputed
  - Can provide substantially greater protection than data swapping while allowing analysts to account for disclosure protection, with less need for coarsening and suppression.
  - Multiple implicates allow analysts to account for uncertainty due to imputation using standard methods with simple combining rules
  - Methods are typically quite transparent
  - Can be difficult when modelling large complex datasets

Synthetic Data in a Swapping World

- Identify records and variables for perturbation like you would for swapping
  - Can select more variables than you would for swapping
  - Imputation rate should be ≥ target perturbation rate.
  - Modeling burden reduced since only portion of values replaced; further reduced by using automated routines.
- Instead of swapping, replace values with (single) imputations
  - Include all variables as predictor variables.
- Evaluate risk and utility

Synthetic Data in a Swapping World

- Imputation provides a model-based, flexible, intuitive alternative to swapping
  - Can preserve relationships between perturbed and unperturbed variables
- Improve upon but not eliminate transparency issue
  - Still can't reveal which records have been perturbed, so can't do multiple imputation
  - Without multiple implicates, still no way for analysts to properly account for perturbation
- Like swapping, perturbation rates constrained if data will be analyzed as if unperturbed

CART Synthesis

- Nonparametric methods from machine learning have been adapted for use with synthetic data, starting with CART (Reiter, 2005), with good results
  - Extended to Bagging, Random Forests, and Support Vector Machines but Drechsler and Reiter found CART to be better for general use.
- Perform automatic detection of nonlinear relationships, interaction effects, with minimal tuning
- In default approach, imputed values are actual values, but marginal distributions not precisely retained as in swapping

R package synthpop

- Developed at University of Edinburgh for UK Office of National Statistics
- Original purpose was to generate bespoke fully synthetic datasets for individual research projects using UK Longitudinal Studies data
  - Context in which synthetic data are more for testing and that restricted-use data will be used for validation
- Many customizations are possible but all specifications are optional.
- Does CART and other types of imputation

Example - Sample Survey Project

- Produced public-use and restricted-use files for demographic sample survey
  - Used imputation to perturb restricted-use (RUF) and create consistent public-use files (PUF)
  - A lot of sensitive variables that were not necessarily identifying
  - Took a conservative approach to protection against identity disclosure, particularly with public-use file

Summary of Approach

1. Risk analysis on preliminary public-use file
   a) Finalize coarsening and suppression for both files
   b) Select records for perturbation based on risk for identity disclosure in PUF
2. Imputation of selected values and variables on restricted-use file
3. Evaluate utility of imputed data, finalize RUF
4. Apply additional coarsening and suppression to create final public-use file
5. Evaluate risk

Preliminary Risk Analysis

- Created risk strata using *k*-anonymity principle and R package *sdcMicro*
  - *k*-anonymity is satisfied if all records are identical to at least *k* other records on set of identifying variables
  - Violations of *k*-anonymity on fewer variables, or key identifiers, considered higher risk. Records with higher risk selected with higher probability.
  - Considered possibility of directed attacks. i.e., attacker looking for a certain person known to have participated in the survey.
  - Started with high rate of selection, adjusted as needed.
  - Decided to limit geography on restricted-use and suppressed from public-use

Imputation

- All variables that could be used for record linkage or direct attacks were considered for imputation
  - List pruned for practicality, utility
- Included 200 other variables, and survey weights, as potential predictors
- Only records selected for imputation are used to build models
  - Important since high-risk records can and do differ from full sample in meaningful way
- Used R package *synthpop* to perform imputation
  - CART model "simple synthesis"

Imputation Step

```
imp_dat = syn(
      data = indata,
      method = imp_method,
      models = TRUE,
      m = 1,
      visit.sequence = imp_vars,
      predictor.matrix = predmatrix)
```

- Method: CART for all variables
- Models: Save models to review
- M: Number of imputations
- Visit.sequence: List of variables to impute (in order)
- Predictor.matrix: Indicator matrix of model predictors

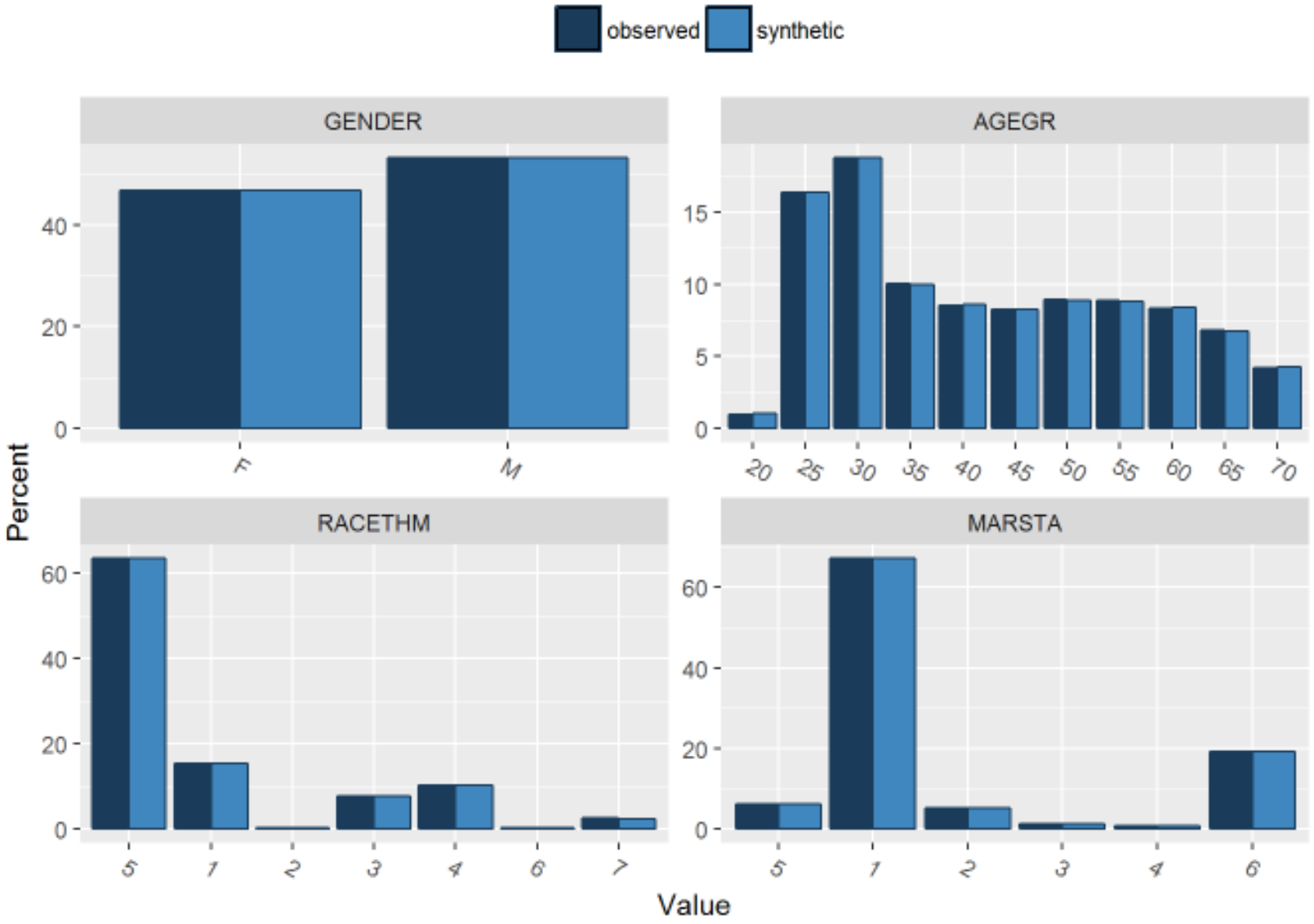
Utility evaluation

- Used synthpop global utility measures and functions to compare original and imputed data
- Additional comparisons for weighted data and conditional distributions
  - Results were generally quite good. All proportions compared were within .01
- Logical checks for gate-nest variables
  - A handful of skips needed to be manually enforced
  - Can also specify rules for logical consistency in imputation function

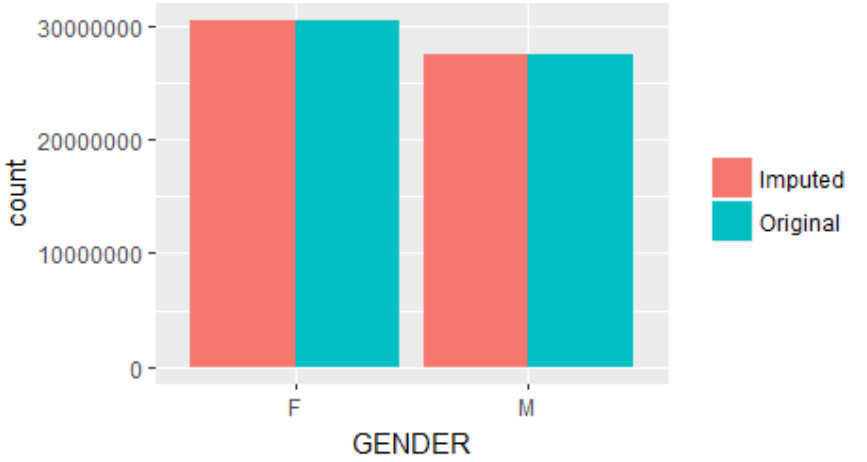
Disclosable Example

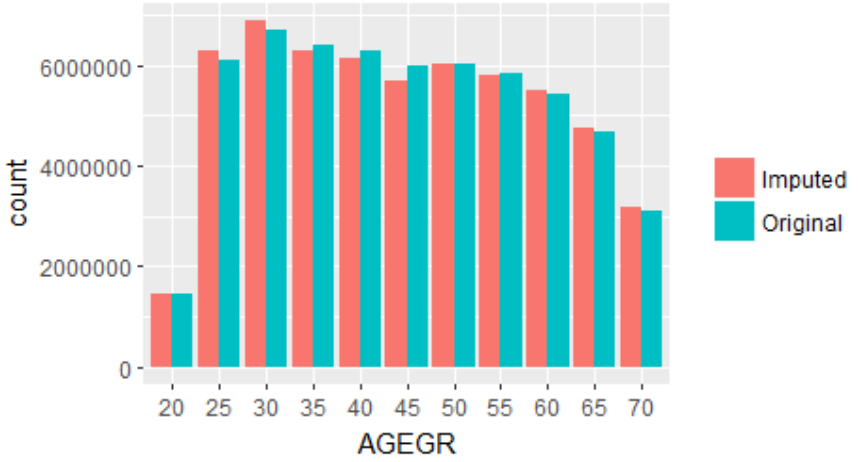
- Extracted 11 variables from NSF's 2015 National Survey of College Graduates Public Use File. Treated this as a confidential dataset.
- Imputed 18.9% of records for 7 variables
- 18.7% of records had at least one value perturbed
- Perturbation rates by variable ranged from 3% to 16%.

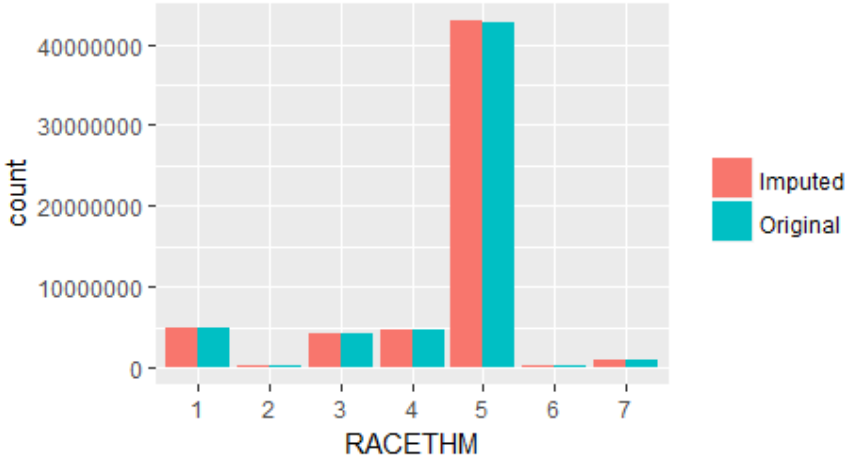
Example Results – synthpop output

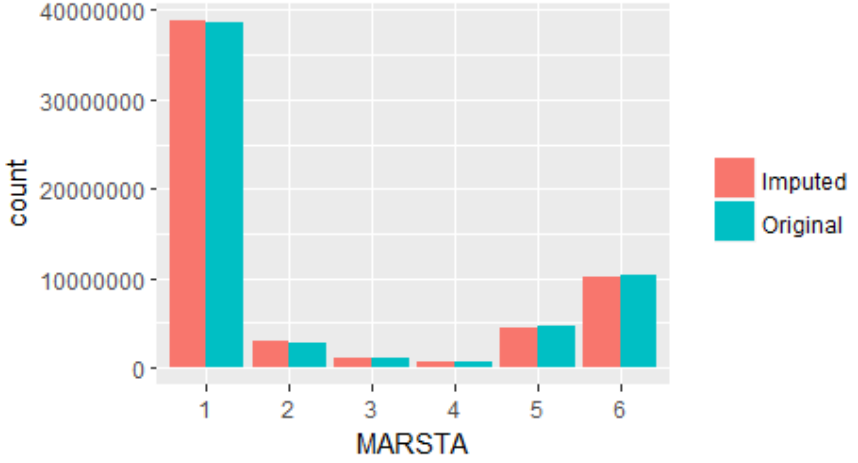


Example Results - Weighted

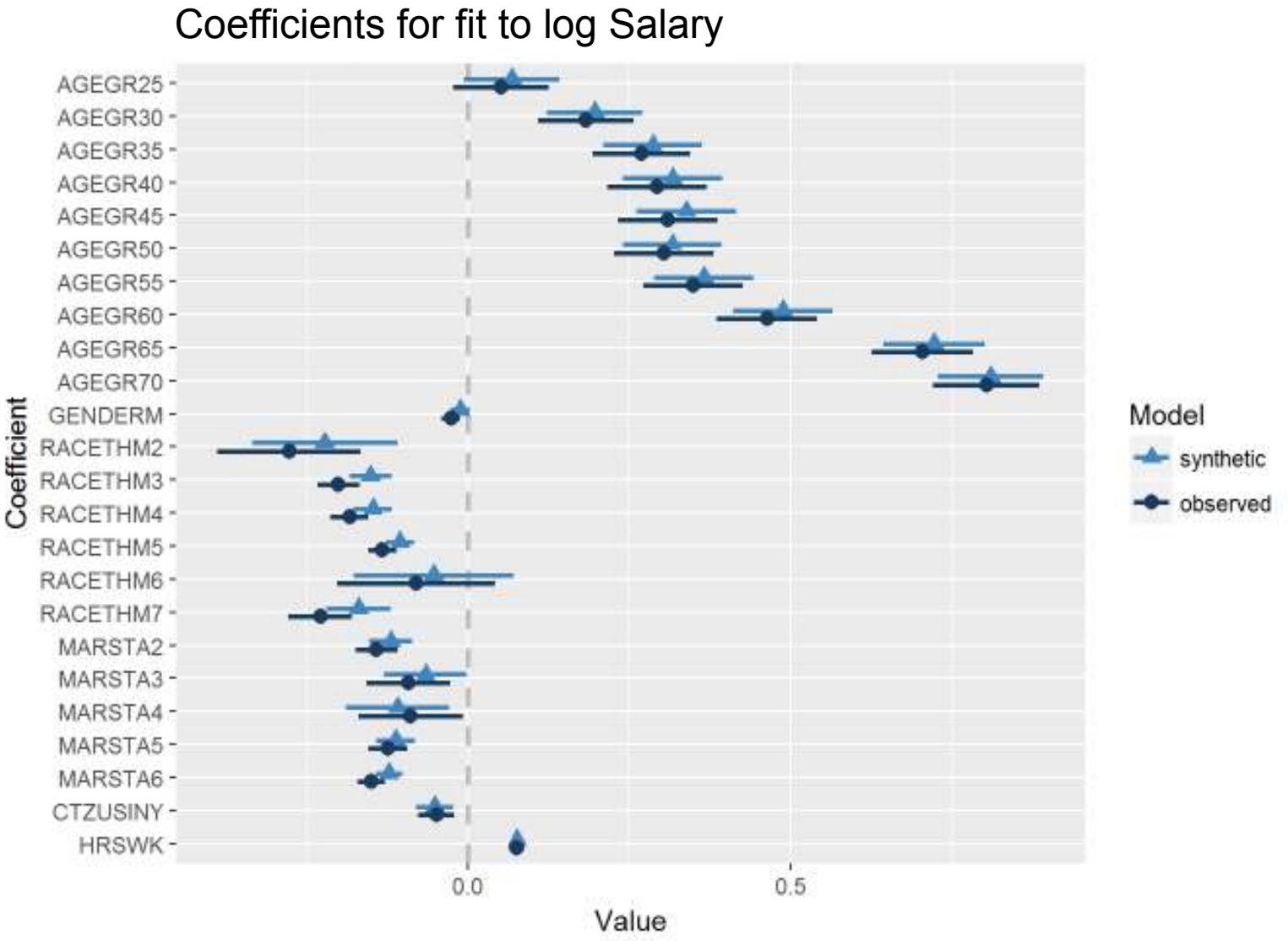




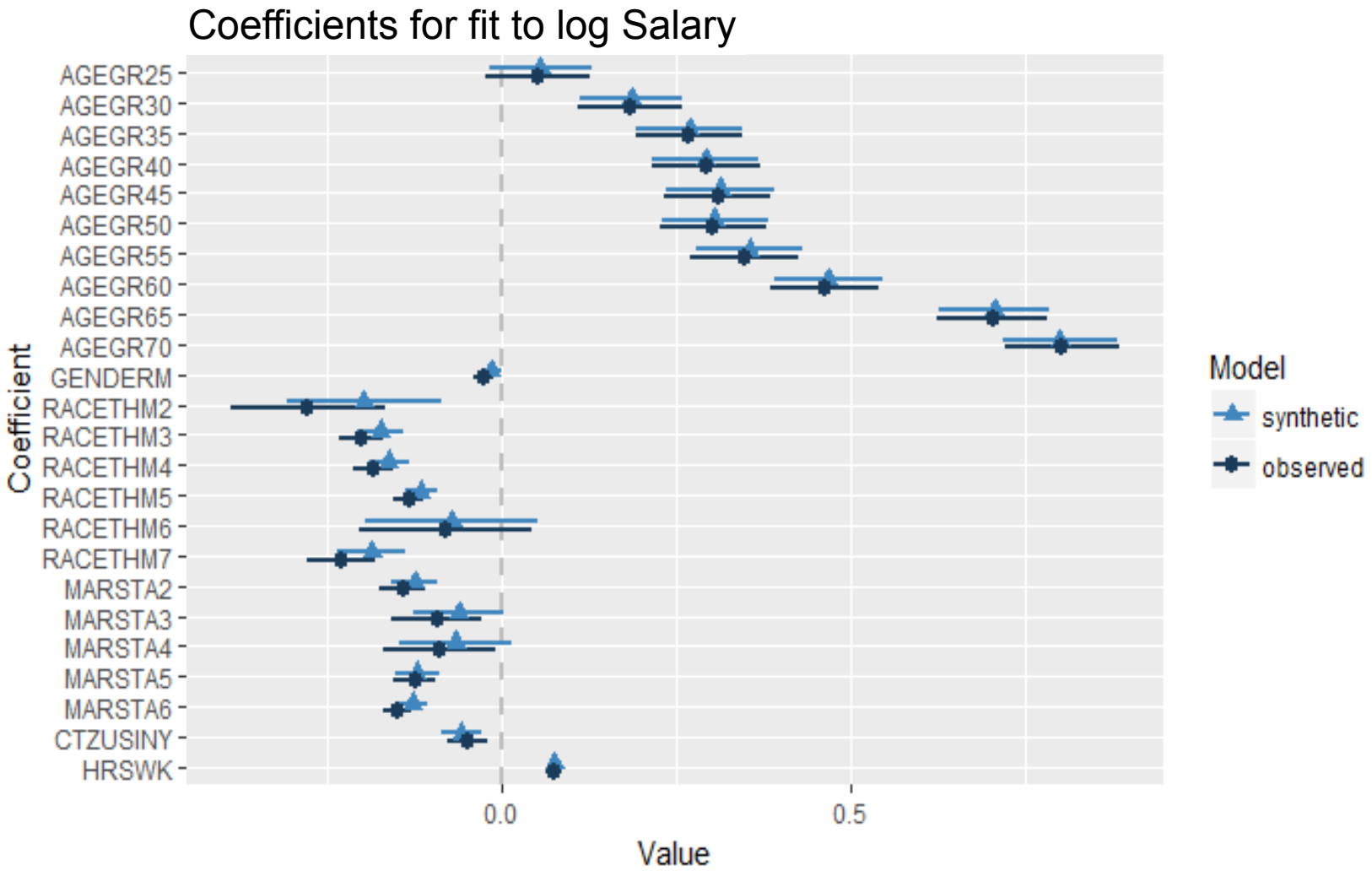




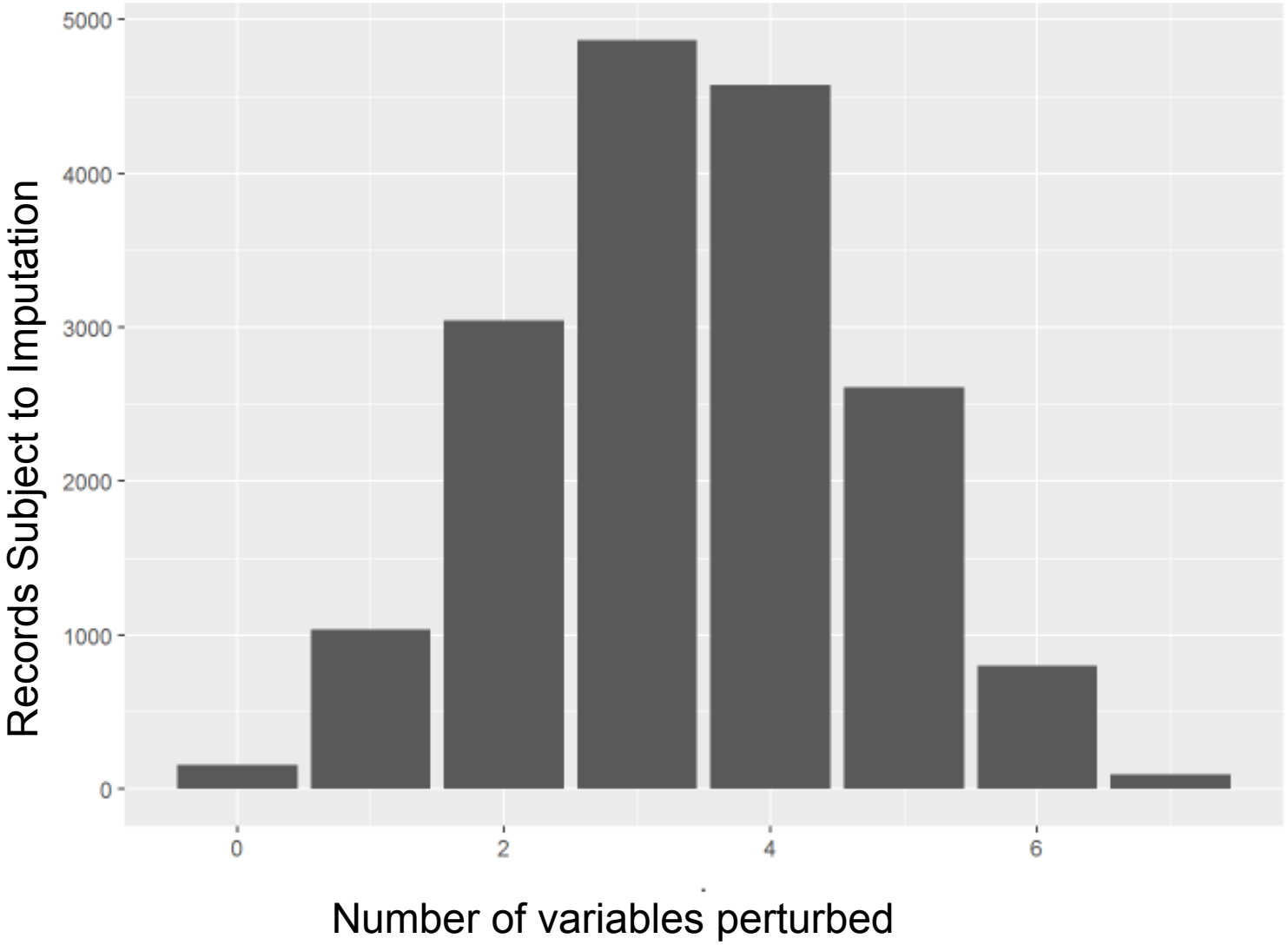
Example Results - Multivariate



Example – Multivariate w/MI



Example Results – Perturbation Rate



Transparency

- When imputing or swapping only a portion of risky records, disclosure protection relies on mystery of which records have been perturbed
  - Following the swapping paradigm, we did not disclose perturbation details. Can we provide more transparency?
  - Synthetic data methods and model specifications are typically reported, as well what records and variables were imputed.
  - Can fix by increasing perturbation rate but this may not provide desired results in current context; suggests shift toward synthetic data
- Imputation provides flexibility for different types of dissemination models

Thank You

- References
  - Nowok B, Raab GM, Dibben C, 2016. synthpop: Bespoke Creation of Synthetic Data in R. *J Statistical Software*.
  - Templ, M. 2008. Statistical Disclosure Control for Microdata Using the R-Package sdcMicro. *Trans. Data Privacy*.
  - Drechsler & Reiter, 2011. An Empirical Evaluation of Easily Implemented, Nonparametric Methods for Generating Synthetic Datasets" *Comp. Statistics & Data Analysis*.
  - Drechsler & Reiter, 2010. “Sampling with Synthesis: A New Approach to Releasing Public Use Microdata Samples of Census Data”, *JASA*
- Contact
  - Saki Kinney (skinney@rti.org)