

Synthetic data methods using quantile regression and hot deck with rank swapping

Michael D. Larsen; Jennifer Hockett
mlarsen@bsc.gwu.edu

George Washington University; Battelle

Friday, July 31, 2009
NSF-Census-IRS Workshop on Synthetic Data and
Confidentiality Protection

Outline

- 1 Motivational example
- 2 Sequential quantile regression and synthetic data
- 3 Hot deck matching with rank swapping and synthetic data
- 4 Example
- 5 Future work

Acknowledgments

- Thanks to organizers for invitation.
- Jennifer Hockett supported by U.S. Census Dissertation Fellowship, 2006-2008; Iowa Department of Revenue 2005-2006.
- Method described in
 - SSC 2008 proceedings.
 - ICES III 2007 proceedings,
 - ASA SRMS 2007 proceedings,
 - ASA SRMS 2006 proceedings,

Goals

- Characterize relationships in data; use conditional models
- Simulate values on each record using model estimates
- Release simulated, synthetic data
- Quality record-level data
- Protect confidentiality of respondents

Challenges

- Build good conditional models, represent data and distributions accurately
- Measure data utility – are the data useful
- Measure disclosure risk – are the data safe
- Computationally feasible – large projects

Proposed procedure to generate synthetic data

- Retain values on handful of nonsensitive variables
- Sequential quantile regression models to generate random synthetic values for some variables
- Hot deck imputation and rank swapping to impute and perturb values for other variables

Motivating example: Taxes in Iowa

- Legislative services agency (LSA) needs tax revenue/burden calculations for proposed bills
- Iowa Department of Revenue does calculations and cannot release micro data
- Inefficient for both
- 1.1 million records per year in 99 counties
- Goal: release synthetic data to LSA (and others)

Nonsensitive variables – many in each cell

- County of residence
- Filing status: single, joint, widowed, etc.
- Copy from original data or randomly resample with replacement

Primary sensitive variables

- Age (depends on county and filing status)
- Wage Income (depends on age, county and filing status)
- Other forms of income (depends on other variables; some clusters of zeros)
- Anticipate that some conditional independence assumption can be made – given some variables, others are not useful predictors

Other tax variables

- Harder to model
- Need to have consistency among groups of variables – use donors
- Anticipate that some conditional independence assumption can be made – do not need to match on all other variables

Quantile Regression (Koenker 2005)

- express mean conditional τ^{th} quantile of Y given X as $\mu(X, \beta_\tau) = X\beta_\tau$.
- quantile regression model $Y_\tau = X\beta_\tau + \epsilon_\tau$
- estimate β_τ by solving $\min \sum_i \rho_\tau(Y_i - X_i\beta_\tau)$
- tilted absolute value function:
$$\rho_\tau(Y_i - X_i\beta_\tau) = (Y_i - X_i\beta_\tau)(\tau - I((Y_i - X_i\beta_\tau) < 0)).$$
- $\tau = 0.5$ is least median regression
- R: *rq* function in *quantreg* package: $\hat{\beta}_\tau$ and large sample SE.

Quantile Regression for SDL

For record i , for variable Y ,

- Randomly select τ^* from distribution on Uniform[0, 1]
- Estimate β_{τ^*} in $Y_{\tau^*} = X\beta_{\tau^*} + \epsilon_{\tau^*}$ using original complete data
- Compute predicted value \hat{y}_{τ^*} for case i given X_i .

Sequential Quantile Regression for SDL

For record i , for variables Y_1, Y_2, \dots

- Randomly impute Y_1 given X using random quantile regression generation
- Randomly impute Y_2 given X and Y_1^* using random quantile regression generation
- Etc.
- At some point, consider dropping some variables from the model: e.g., Y_k given X and Y_1 and Y_{k-1} – omit Y_2, \dots, Y_{k-2} .

Hot Deck Imputation: Typical

- Complete records have values recorded on all variables
- Incomplete records have missing values on some variables
- Match complete and incomplete records on variables with recorded values to find potential donors
- Select a donor – random, nearest (Mahalanobis distance), or sequential
- Copy values from matching complete record to fill-in incomplete

Hot Deck for Disclosure Limitation

- Original data set contains complete records, values on all variables
- Synthetic data set contains incomplete records
 - values on nonsensitive variables
 - values on variables with quantile regression predictions
 - missing values on other variables
- Match synthetic incomplete records to original complete records
- Impute values from original to synthetic for other variables

Hot Deck Concern in Disclosure Limitation

- If several variables are imputed together using hot deck, the combination of variables might be identifiable.
- Solution: perturb the hot deck imputations
- You could model the variable and add (Normal) noise
- Another option is 'rank swapping' of imputed values

Rank Swapping of Hot Deck Imputation Values

- Match synthetic incomplete records to original complete records
- Determine rank r of value for a donor value on a given variable
- Randomly draw rank r^* from distribution centered (discrete uniform on ranks) around r
- Impute the value corresponding rank r^* for that variable

Summary of Procedure

- Nonsensitive variables are copied (or resampled)
- Key variables amenable to regression quantile modeling are sequentially, randomly imputed
- Hot deck donors from original complete data are identified
- Random ranks are selected to perturb donor values (to avoid multivariate identifiability)
- Result is a micro data set with realistic values
- Above procedure could be done multiple times for multiple imputations

Example: ACS Veterans

- U.S. Census Bureau: Over 100,000 veterans; records from 2002 American Community Survey
- Goal: share data with other agencies and researchers
- Public Use Microdata Samples: Implement proposed SDL procedure and disclosure risk measures
- Note: data here are already protected

ACS: Nonsensitive Variables and Quantile Regresses

- Retain predictors
 - educational attainment (SCHL)
 - veteran period of service (VPS)
 - – recategorize to SCHL/VPS categories to avoid inconsistencies
- Generate synthetic values using quantile regression:
 - age (AGE)
 - retirement (RET) income
(predict using SCHL/VPS and AGE)
 - social security income (SS)
(predict using SCHL/VPS, AGE, RET)

ACS: hot deck and rank swapping

- Match using AGE and RET within SCHL/VPS categories
- Variables for hot deck donation
 - supplemental social security income (SSI)
 - property tax (TAX)
 - wages (WAGE)

Summary of ACS Results

- Reproduced AGE well
- RET and SS look good for the most part, but perhaps an interaction or quadratic term should have been included
- Excluding one SCHL/VPS category, hot deck variables look OK
- Might want to consider zero versus non-zero values separately as in Iowa tax application
- Utility: relationships generally preserved, but improvement should be possible

Data Utility

- Compare marginal distributions in original and synthetic data
 - empirical cumulative distributions, empirical densities
 - point estimates, standard errors, confidence intervals
- Compare conditional distributions in original and synthetic data
 - empirical distributions within categories
 - regression estimates: coefficients, standard errors, confidence intervals, R^2 , correlations

Disclosure Risk

- Record linkage/probabilistic matching risk
- Intruder formulation: Duncan and Lambert (1989, 1986) and Reiter (2005)
- Extensions to intruder formulation for our synthetic data method: Huckett (2008) thesis
- See proceedings – can assume intruder knows disclosure limitation methods or is naive

Summary and Plans

- A novel method of generating synthetic data when (a) some variables are nonsensitive, (b) some can be sequentially modeled using quantile regression, and (c) others are imputed (with perturbation) from donors.
- Basic ideas for measuring data utility
- Adapt existing 'intruder' framework to method for measuring disclosure risk: disclosure risk in ACS example as measured was quite small.
- Method seems effective for the most part
- Plans: check some details in computing and write articles