Edit Imputation

Zhenhua Wang

November 8, 2023

1 Automatic Edit and Imputation

- select reported values to change according to some heuristic or loss function
- replace those values with plausible imputations

2 Bayesian Edit Imputation

2.1 Multiple Imputation of Missing or Faulty Values Under Linear Constraints (Kim et al. 2014)

This paper follows a two-step imputation process. It first models the data using Dirichlet Process Mixtures of Normals with Truncation. Then, it uses a hit-and-run sampler for the imputation.

2.2 Simultaneous Edit-Imputation for Continuous Microdata (Kim, Cox, Karr, Reiter and Wang 2015)

In this paper, the author uses a hierarchical model with three levels. It includes a model for the true data with the support on the set of values that satisfy all editing constrain, a model for latent indicators of the variables that are in error, and models (measurement error model) for the reported responses for variables in error.

Specifically, it uses a finite mixture of multivariate normal distributions with a constrained support for true data model, a uniform distribution for error indicator model.

2.3 Bayesian Simultaneous Edit and Imputation for Multivariate Categorical Data (Manrique-Vallier and Reiter 2017)

This paper uses the truncated Bayesian nonparametric latent class model as their response model.

2.4 Simultaneous Edit and Imputation For Household Data with Structural Zeros (Akande et al. 2019)

This paper uses a nested data Dirichlet process mixture of products of multinomial distributions as the model for the true latent values of the data, truncated to allow only households that satisfy all edit constraints.

2.5 Statistical Disclosure Limitation in the Presence of Edit Rules (Kim, Karr and Reiter 2015)

This paper compares edit-after-SDL and edit-preserving SDL. In edit-after-SDL, an agency first applies an SDL method to the collected data. Any post-SDLrecords that violate the constraints are deleted or "repaired". In edit-preserving SDL, we draw candidate masked values repeatedly until they satisfy all edit rules (e.g. through reject sampling).

2.6 Simultaneous edit-imputation and disclosure limitation for business establishment data (Kim et al. 2018)

This paper uses a two-stage process. The first stage (Kim, Cox, Karr, Reiter and Wang 2015) generates m plausible imputations that satisfy all edit rules. The second stage re-estimate the joint probability model on each of the m plausible datasets and generates r synthetic datasets for each plausible edited dataset.

2.7 Synthetic microdata for establishment surveys under informative sampling (Kim et al. 2021)

This paper is built on top of (Kim et al. 2014), and it incorporates pseudo likelihood framework in DP Gaussian mixture model.

3 Complex Survey

3.1 Bayesian estimation under informative sampling (Savitsky and Toth 2016)

3.2 Fully Bayesian estimation under informative sampling (León-Novelo and Savitsky 2019)

This paper derives the posterior population distribution corrected by sampling weights

3.3 Bayesian Data Synthesis and Disclosure Risk Quantification: An Application to the Consumer Expenditure Surveys (Hu and Savitsky 2018)

This paper protects the county label of consumer units in CES. They use two models:

1. Dirichlet Process mixtures of products of multinomials
2. nonparametric version of areal models with Dirichlet Process priors (DP-areal), which models counts of county labels of observations sharing similar characteristics.

3.4 Risk-Efficient Bayesian Data Synthesis for Privacy Protection (Hu et al. 2022b)

This paper generates synthetic data through pseudo posterior where records with high identification risk are down-weighted. The identification risk is defined by the probability that the synthetic values for a record are not close to the truth.

3.5 Bayesian pseudo posterior mechanism under asymptotic differential privacy (Savitsky et al. 2022)

This paper extends (Hu et al. 2022b) and shows a connection between pseudo posterior and differential privacy, by constructing risk-related weights that achieves a formal privacy guarantee.

3.6 Private Tabular Survey Data Products through Synthetic Microdata Generation (Hu et al. 2022a)

1. Fully Bayes model of observed sample: FBS jointly models sampling weighs and observed samples with a bivariate normal distribution
2. Fully Bayes model of population: FBP uses sampling weights to correct population bias and the outcome variable under the population distribution

References

Akande, O., Barrientos, A. and Reiter, J. P.: 2019, Simultaneous edit and imputation for household data with structural zeros, *Journal of Survey Statistics and Methodology* **7**(4), 498–519.

Hu, J. and Savitsky, T. D.: 2018, Bayesian data synthesis and disclosure risk quantification: An application to the consumer expenditure surveys, *arXiv preprint arXiv:1809.10074* .

Hu, J., Savitsky, T. D. and Williams, M. R.: 2022a, Private tabular survey data products through synthetic microdata generation, *Journal of Survey Statistics and Methodology* **10**(3), 720–752.

Hu, J., Savitsky, T. D. and Williams, M. R.: 2022b, Risk-efficient bayesian data synthesis for privacy protection, *Journal of Survey Statistics and Methodology* **10**(5), 1370–1399.

Kim, H. J., Cox, L. H., Karr, A. F., Reiter, J. P. and Wang, Q.: 2015, Simultaneous edit-imputation for continuous microdata, *Journal of the American Statistical Association* **110**(511), 987–999.

Kim, H. J., Drechsler, J. and Thompson, K. J.: 2021, Synthetic microdata for establishment surveys under informative sampling, *Journal of the Royal Statistical Society Series A: Statistics in Society* **184**(1), 255–281.

Kim, H. J., Karr, A. F. and Reiter, J. P.: 2015, Statistical disclosure limitation in the presence of edit rules, *Journal of Official Statistics* **31**(1), 121–138.

Kim, H. J., Reiter, J. P. and Karr, A. F.: 2018, Simultaneous edit-imputation and disclosure limitation for business establishment data, *Journal of Applied Statistics* **45**(1), 63–82.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H. and Karr, A. F.: 2014, Multiple imputation of missing or faulty values under linear constraints, *Journal of Business & Economic Statistics* **32**(3), 375–386.

León-Novelo, L. G. and Savitsky, T. D.: 2019, Fully bayesian estimation under informative sampling.

Manrique-Vallier, D. and Reiter, J. P.: 2017, Bayesian simultaneous edit and imputation for multivariate categorical data, *Journal of the American Statistical Association* **112**(520), 1708–1719.

Savitsky, T. D. and Toth, D.: 2016, Bayesian estimation under informative sampling.

Savitsky, T. D., Williams, M. R. and Hu, J.: 2022, Bayesian pseudo posterior mechanism under asymptotic differential privacy, *The Journal of Machine Learning Research* **23**(1), 2484–2520.