Edit Imputation

Zhenhua Wang

October 28, 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. Kim et al. (2014)

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.

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.

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.