Beyond the Ballot: Progress Report

Furkan Kadıoğlu

10 Jun 2024

1 Background

1.1 Survey Imputation

Survey imputation is a crucial process in statistical analysis, used to address missing or incomplete data in survey responses. Missing data can introduce bias, reduce the statistical power of analyses, and compromise the validity of research findings [7, 9]. Imputation techniques help provide a complete dataset, enabling more accurate statistical inferences and robust conclusions.

Listwise and pairwise deletion were among the earliest methods used. Listwise deletion involves excluding any cases (respondents) with missing data from the analysis, which can lead to significant data loss and reduced sample size. This reduction can result in biased results because the remaining data may not be representative of the original population. Pairwise deletion involves using only the available data pairs for analysis, which retains more data but can lead to inconsistent results and complications in statistical analysis due to varying sample sizes across different analyses [5].

Simple imputation methods include replacing missing values with the mean, median, or mode of the observed data. For instance, if a survey respondent did not answer a question about their age, the missing value might be replaced with the average age of all other respondents. While simple and easy to use, these methods do not account for the variability in the data. They can underestimate the variance and potentially lead to biased estimates, as they do not consider the underlying distribution or relationships between variables. For example, imputing the mean for a skewed variable would distort the distribution and relationships [7].

Hot deck imputation was developed as a method to replace missing values with observed responses from similar respondents. This technique involves finding a respondent (or several respondents) with similar characteristics to the one with missing data and using their observed value to fill in the missing value. For example, if the income data for a respondent is missing, the method might use the income of another respondent with a similar age, education level, and job type. This method improves upon simple imputation by considering respondent similarity, but it can still introduce bias if the matching process is not well-executed, as it assumes that the chosen respondents are truly representative of the missing cases [1].

Cold deck imputation involves using external sources or prior data for imputation. For example, if a survey is conducted annually, missing values in the current year's survey might be filled in using data from the previous year's survey. This method relies on the availability and relevance of external data, which can sometimes be a limitation if the external data is not comparable to the survey data. Additionally, it assumes that the previous data is accurate and that the relationship between the variables has remained constant over time [7].

Regression imputation uses regression models to predict missing values based on other observed variables. For example, a regression model could be developed to predict income based on variables such as education, age, and job type. This model is then used to estimate the missing income values. This method provides more nuanced imputation than simple mean substitution but assumes that the regression model is correctly specified. If the model is misspecified, it can lead to biased estimates.

Moreover, it tends to reduce the variability in the imputed data because the imputed values are predicted by a deterministic function [5].

The Expectation-Maximization (EM) algorithm, introduced for iteratively estimating parameters in the presence of missing data, refines estimates to improve imputation accuracy. The EM algorithm involves two steps: the Expectation step (E-step), where missing data are estimated based on observed data, and the Maximization step (M-step), where the estimated values are used to update the model parameters. This process is repeated until convergence. Despite its power, the EM algorithm can be computationally intensive and sensitive to initial values. It also assumes that the data are missing at random (MAR) [2].

Multiple imputation (MI) involves creating multiple datasets with different imputed values, analyzing each dataset separately, and then combining the results. This method accounts for the uncertainty inherent in the imputation process and provides more robust statistical inferences. In MI, the missing values are imputed multiple times to create several complete datasets. Each dataset is analyzed separately, and the results are combined to produce estimates and confidence intervals that reflect the uncertainty due to missing data. MI requires careful implementation and computational resources, particularly with large datasets, but it provides a robust framework for dealing with missing data [6].

K-Nearest Neighbors (KNN) imputation imputes missing values based on the average of the nearest neighbors' values, considering the similarity of other variables in the data. For example, if a respondent's income is missing, KNN imputation might find the k respondents with the most similar characteristics (age, education, job type) and use their average income to impute the missing value. This method is computationally feasible with modern technology and handles complex data structures, but it can be sensitive to the choice of the number of neighbors (k) and the distance metric used. Moreover, KNN can struggle with high-dimensional data where distance metrics become less meaningful [8].

Machine learning methods, including decision trees, random forests, and neural networks, have been applied to imputation. These methods can handle large, complex datasets and capture non-linear relationships between variables. For example, a random forest model, which consists of many decision trees, can be trained to predict missing values based on observed data. These methods require significant computational power and expertise in machine learning but offer flexibility and accuracy in handling various types of missing data. However, they can be overfitting and complex to interpret [10].

Advanced statistical techniques like Multiple Imputation by Chained Equations (MICE) and Bayesian methods offer sophisticated ways to handle complex, multivariate missing data patterns. MICE, for instance, imputes missing values by iteratively filling in each variable with missing data using a series of regression models, creating a chained sequence of models. Bayesian methods provide a probabilistic framework for imputation, allowing for the incorporation of prior knowledge and the estimation of uncertainty in a coherent manner. These methods provide flexible and powerful tools for imputation but require careful model specification and can be computationally demanding [11].

Modern imputation methods increasingly leverage AI and deep learning techniques to handle large and complex datasets, providing more accurate and reliable imputed values. For example, denoising autoencoders, a type of deep learning model, can be used to learn a representation of the data that can be used to predict and impute missing values. These methods, while powerful, require extensive computational resources and expertise in AI and deep learning [3].

The development of specialized software and tools, such as the R packages 'mice' and 'Amelia', and Python libraries like 'fancyimpute', has made sophisticated imputation methods more accessible to researchers and practitioners. These tools facilitate the implementation of advanced imputation techniques but still require users to understand the underlying methods to apply them correctly [11, 4].

References

[1] Rebecca R Andridge and Roderick JA Little. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010.

[2] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

[3] Lovedeep Gondara and Ke Wang. Mida: Multiple imputation using denoising autoencoders. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III 22*, pages 260–272. Springer, 2018.

[4] James Honaker, Gary King, and Matthew Blackwell. Amelia ii: A program for missing data. *Journal of Statistical Software*, 45:1–47, 2011.

[5] Hyun Kang. The prevention and handling of the missing data. *Korean journal of anesthesiology*, 64(5):402, 2013.

[6] Peng Li, Elizabeth A Stuart, and David B Allison. Multiple imputation: a flexible tool for handling missing data. *Jama*, 314(18):1966–1967, 2015.

[7] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[8] Utomo Pujianto, Aji Prasetya Wibawa, Muhammad Iqbal Akbar, et al. K-nearest neighbor (k-nn) based missing data imputation. In *2019 5th International Conference on Science in Information Technology (ICSITech)*, pages 83–88. IEEE, 2019.

[9] Joseph L. Schafer. Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:15 – 3, 1999.

[10] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[11] Stef van Buuren and Karin G. M. Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45:1–67, 2011.