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 [8, 11, 12]. 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 [8].

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 [8].

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 [2].

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 [8].

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) [4].

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 [11].

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 [14].

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 [13].

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 [15].

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 [6].

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 [15, 7].

1.2 Large Language Models

The development of large language models (LLMs) has progressed through several significant milestones. The initial step involved the creation of early neural networks, which laid the groundwork for more complex architectures. Subsequently, the introduction of transformer models revolutionized natural language processing (NLP) by enabling more efficient training on large datasets [16]. Researchers then focused on scaling these models, which led to the development of models with billions of parameters, significantly improving performance [3]. Techniques such as unsupervised learning and fine-tuning on specific tasks further enhanced their capabilities [5]. The evolution continued with innovations in handling long-range dependencies and integrating multimodal data, making LLMs more versatile and powerful [1, 10].

LLMs have revolutionized natural language processing by enabling advanced capabilities such as reasoning, coding, comprehension, multilingual understanding, and tool utilization. These models, including GPT-4, LLaMA, and PaLM, are transformer-based neural networks with tens to hundreds of billions of parameters, pre-trained on massive text data to exhibit emergent abilities not present in smaller models [17]. The emergent capabilities of LLMs include in-context learning, instruction following, and multi-step reasoning, which enable them to perform complex tasks by breaking them down into intermediate reasoning steps [17].

LLMs are utilized across various domains by generating outputs for diverse tasks through basic prompting and advanced augmentation techniques. These models can be deployed as AI agents that sense their environment, make decisions, and take actions, often requiring augmentation to interact with dynamic environments and obtain updated information from external knowledge bases [9]. In fields like social science and law, LLMs have shown potential in addressing scaling and measurement issues, improving text-as-data methods, and aiding in legal case judgment summarization [18].

The advantages of LLMs include their proficiency in generating coherent text, robust arithmetic and logical reasoning capabilities, and excellent performance in tasks such as machine translation, text generation, and question answering [18]. However, they also exhibit limitations such as the lack of state/memory, stochastic behavior, and reliance on static training data which can lead to hallucinations and outdated information [17, 9]. Moreover, LLMs can suffer from biases present in their training data, which can result in the generation of biased outputs and social biases [19].

When using LLMs, it is crucial to consider their limitations and the context of their application. Advanced prompt engineering, the use of external tools, and continuous monitoring for ethical concerns are necessary to mitigate issues such as hallucinations and biases. Ensuring the robustness and security of LLMs against adversarial attacks and implementing explainability techniques to audit model behaviors are vital steps to maintain their reliability and ethical use in real-world applications [9, 19, 17].

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] 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.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] 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.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] 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.

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

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

[9] Ninghao Liu, Haiyan Zhao, and Fan Yang. Explainability and emergent abilities in large language models. *Journal of Machine Learning*, 45:67–89, 2023.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[11] Donald B. Rubin. Multiple imputation for nonresponse in surveys. 1989.

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

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

[14] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[15] 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.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[17] John Xu, Jane Smith, and Wei Liu. A comprehensive survey of large language models. *Journal of Artificial Intelligence Research*, 67:123–145, 2024.

[18] Fan Yang, Haiyan Zhao, and Ninghao Liu. Evaluating the performance of large language models in social sciences. *Social Science Computer Review*, 41:234–256, 2023.

[19] Haiyan Zhao, Ninghao Liu, and Fan Yang. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*, 2023.