documentclassarticle
usepackageamsmath
usepackagegraphicx
usepackagebooktabs
usepackagearray
usepackagemultirow
usepackagecaption

begindocument

titleEvaluating the Impact of Missing Data Imputation Techniques on Regression Model Validity and Predictive Accuracy
authorMason Davis, Grace Nelson, Jack Williams
date
maketitle

sectionIntroduction

Missing data represents one of the most persistent and challenging problems in statistical analysis and machine learning applications across diverse domains including healthcare, social sciences, and business analytics. The prevalence of incomplete datasets necessitates the development and application of imputation techniques that can generate plausible values for missing observations. While numerous imputation methods have been proposed in the literature, their comparative evaluation has traditionally focused on predictive accuracy metrics, often neglecting the crucial aspect of parameter validity preservation. This research addresses this significant gap by developing a comprehensive evaluation framework that simultaneously assesses both predictive accuracy and statistical validity across multiple imputation approaches.

The fundamental challenge in missing data imputation lies in the inherent tension between generating values that maintain the original data distribution's statistical properties while also enabling accurate predictions in downstream modeling tasks. Conventional evaluation paradigms have predominantly emphasized the latter, potentially leading to the adoption of imputation methods that produce biased parameter estimates or distorted covariance structures. This oversight has profound implications for inferential statistics, where the validity of parameter estimates is paramount for drawing meaningful conclusions from data.

Our research makes several distinctive contributions to the field of missing data analysis. First, we introduce a novel evaluation framework that systematically examines the impact of imputation techniques on both predictive performance and parameter validity across different missing data mechanisms and proportions. Second, we investigate the often-overlooked relationship between impu-

tation method complexity and its ability to preserve statistical properties, challenging the assumption that more sophisticated methods universally outperform simpler alternatives. Third, we provide practical guidance for method selection based on specific analytical objectives, recognizing that different applications may prioritize predictive accuracy or parameter validity differently.

This paper is structured as follows. The Methodology section details our experimental design, including data generation procedures, imputation techniques evaluated, and our novel validity metrics. The Results section presents comprehensive findings across multiple evaluation dimensions, highlighting the complex trade-offs between different performance criteria. The Conclusion discusses the implications of our findings for both methodological development and practical application, along with directions for future research.

sectionMethodology

subsectionExperimental Design

Our experimental framework was designed to systematically evaluate the performance of various imputation techniques across controlled conditions that reflect real-world data scenarios. We generated synthetic datasets with known statistical properties to enable precise assessment of how different imputation methods affect both parameter estimation and predictive accuracy. The experimental factors included three missing data mechanisms (Missing Completely at Random, Missing at Random, and Missing Not at Random), four missingness proportions (10
For each experimental condition, we generated 100 replicate datasets with sample size n=1000 and p=10 variables, including both continuous and categorical types to reflect common analytical scenarios. The data generation process incorporated realistic correlation structures ranging from weak (r=0.2) to strong (r=0.8) associations between variables. The outcome variable was specified as a linear combination of the predictor variables with added Gaussian noise, ensuring known ground truth for both parameter values and predictive relationships.

Missingness was introduced according to the specified mechanisms using carefully controlled probability models. For MCAR conditions, missing values were generated with constant probability across all observations. For MAR conditions, the probability of missingness depended on observed variables through logistic regression models with moderate effect sizes. For MNAR conditions, the missingness mechanism depended on the values of the variables themselves or unobserved latent variables, creating the most challenging scenario for imputation methods.

subsectionImputation Techniques

We evaluated twelve imputation methods representing different methodological

approaches and complexity levels. The methods included simple approaches such as mean/mode imputation and regression imputation; statistical methods including expectation-maximization algorithm, predictive mean matching, and multiple imputation by chained equations; machine learning approaches including k-nearest neighbors imputation, random forest imputation, and support vector imputation; and advanced techniques including generative adversarial imputation networks, Bayesian structural time series, and variational autoencoder-based imputation.

Each method was implemented using established software packages with consistent hyperparameter tuning procedures to ensure fair comparison. For multiple imputation methods, we generated m=20 imputed datasets and used Rubin's rules for parameter pooling. The machine learning and deep learning approaches underwent identical cross-validation procedures for hyperparameter optimization to prevent overfitting and ensure generalizable performance.

subsectionEvaluation Metrics

Our evaluation framework incorporated multiple metrics to comprehensively assess imputation performance across two primary dimensions: predictive accuracy and statistical validity. For predictive accuracy, we employed root mean squared error (RMSE) for continuous variables, proportion of falsely classified entries for categorical variables, and overall predictive accuracy in downstream regression tasks using the imputed data.

For statistical validity assessment, we developed novel metrics that quantify how well each imputation method preserves the underlying data structure. These included parameter bias, measured as the absolute difference between estimated regression coefficients and their true values; covariance preservation, assessed through Frobenius norm differences between original and imputed covariance matrices; distributional similarity, evaluated using Wasserstein distance between original and imputed variable distributions; and inference validity, measured through coverage rates of 95
Additionally, we assessed computational efficiency through execution time measurements and scalability analysis across different dataset sizes. All evaluations were conducted on identical hardware with controlled computational environments to ensure comparability.

sectionResults

subsectionComparative Performance Across Missing Data Mechanisms

The performance of imputation methods varied substantially across different missing data mechanisms, revealing important patterns about their relative strengths and limitations. Under MCAR conditions, most methods performed reasonably well, with multiple imputation and random forest approaches demon-

strating the best balance between predictive accuracy and parameter validity. However, even in this simplest scenario, we observed notable differences in performance, with mean imputation showing the highest parameter bias despite its computational efficiency.

Under MAR conditions, the performance hierarchy shifted significantly. Methods that incorporated the relationship between observed variables and missingness patterns, such as multiple imputation by chained equations and expectation-maximization algorithm, outperformed simpler approaches. The machine learning methods, particularly random forest imputation, showed strong predictive performance but exhibited higher variability in parameter preservation across different missingness proportions.

The MNAR scenario presented the greatest challenges for all imputation methods. Even advanced techniques struggled to recover the true data structure, with most methods showing substantial parameter bias and distorted covariance estimates. Generative adversarial imputation networks demonstrated some advantage in predictive accuracy under high missingness conditions, but at the cost of increased computational requirements and occasional convergence issues.

subsectionTrade-offs Between Predictive Accuracy and Parameter Validity

A central finding of our study concerns the inherent trade-offs between predictive accuracy and parameter validity across different imputation methods. We observed that methods optimizing for one dimension often compromised performance in the other. For instance, machine learning approaches like random forest and support vector imputation consistently achieved high predictive accuracy in downstream modeling tasks but frequently introduced bias in parameter estimates and distorted covariance structures.

Conversely, statistically principled methods like multiple imputation and Bayesian approaches better preserved parameter validity and distributional properties but sometimes underperformed in pure prediction tasks, particularly when the analysis model differed from the imputation model. This divergence highlights the importance of aligning imputation method selection with analytical objectives—whether the primary goal is prediction or inference.

We quantified these trade-offs through a novel composite metric that balances both dimensions, revealing that no single method dominated across all conditions. The optimal choice depended on the specific combination of missing data mechanism, missingness proportion, and analytical priorities.

subsectionImpact of Missingness Proportion

The proportion of missing data emerged as a critical factor influencing imputation performance across all methods. As missingness increased from 10
Simple methods like mean imputation showed the most rapid performance deterioration, becoming essentially unusable at high missingness levels due to se-

vere parameter bias and variance inflation. More sophisticated methods demonstrated better robustness to increasing missingness, with multiple imputation and generative approaches maintaining reasonable performance even at 60
Notably, the relationship between method complexity and performance was not monotonic. Some intermediate-complexity methods, particularly random forest imputation and expectation-maximization algorithm, often achieved performance comparable to more complex deep learning approaches while requiring significantly less computational resources.

subsectionDistributional Preservation and Covariance Structure

Our analysis of distributional preservation revealed important insights about how different imputation methods affect the underlying data structure. Methods that explicitly model the joint distribution, such as multiple imputation and Bayesian approaches, generally better preserved covariance structures and distributional characteristics. In contrast, methods focusing on conditional distributions or nearest-neighbor relationships often introduced subtle distortions in higher-order moments and dependence structures.

These distributional distortions had practical consequences for downstream analyses. Even when predictive accuracy remained high, distorted covariance structures could lead to invalid inferences, particularly in applications requiring accurate estimation of relationship strengths or variance components. This finding underscores the importance of evaluating imputation methods beyond simple accuracy metrics when statistical inference is a primary concern.

sectionConclusion

This research has provided a comprehensive evaluation of missing data imputation techniques with a specific focus on the dual objectives of predictive accuracy and parameter validity. Our findings challenge several conventional assumptions in the field and offer practical guidance for method selection based on specific analytical needs and data characteristics.

The central contribution of our work lies in demonstrating that the choice of imputation method involves fundamental trade-offs between different performance dimensions. Methods that excel in predictive tasks may compromise statistical validity, while approaches preserving parameter accuracy may underperform in pure prediction contexts. This insight necessitates a more nuanced approach to imputation method selection that explicitly considers the analytical objectives and inferential requirements of each application.

Our results also question the prevailing trend toward increasingly complex imputation methods. While advanced techniques like generative adversarial networks and variational autoencoders show promise in specific scenarios, their advantages are not universal, and they come with substantial computational costs and implementation complexity. In many practical situations, well-established

methods like multiple imputation offer a favorable balance of performance, interpretability, and computational efficiency.

Several limitations of our study warrant mention. Our evaluation focused on synthetic data with known ground truth, which enabled precise performance assessment but may not fully capture the complexities of real-world datasets. Future research should extend this evaluation framework to empirical data across diverse domains. Additionally, our study considered primarily cross-sectional data; extending this work to longitudinal and time-series contexts represents an important direction for future investigation.

From a practical perspective, our findings suggest that analysts should carefully consider their primary analytical goals when selecting imputation methods. For prediction-focused applications, machine learning approaches may offer advantages, while inference-oriented analyses may benefit from statistically principled methods that better preserve parameter validity. In all cases, multiple imputation emerges as a robust default choice, particularly when both prediction and inference are important.

In conclusion, this research contributes a more sophisticated understanding of how missing data imputation affects subsequent analyses and provides a comprehensive framework for method evaluation. By highlighting the critical trade-offs between different performance dimensions and challenging assumptions about method superiority, we hope to inform more thoughtful and effective practices in missing data handling across diverse research and application contexts.

section*References

Little, R. J. A., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). John Wiley & Sons.

Van Buuren, S. (2018). Flexible imputation of missing data (2nd ed.). Chapman and Hall/CRC.

Enders, C. K. (2022). Applied missing data analysis (2nd ed.). Guilford Publications.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. Statistics in Medicine, 30(4), 377-399.

Stekhoven, D. J., & Buhlmann, P. (2012). MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics, 28(1), 112-118.

Yoon, J., Jordon, J., & van der Schaar, M. (2018). GAIN: Missing data imputation using generative adversarial nets. In International Conference on Machine Learning (pp. 5689-5698).

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. Statistical Science, 33(2), 142-159.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work?. International Journal of Methods in Psychiatric Research, 20(1), 40-49.

Graham, J. W. (2012). Missing data: Analysis and design. Springer Science & Business Media.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147-177.

enddocument