International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2024: 7.101

Managing Missing Data with Multiple Imputation: A Common - Sense Approach for Enhancing Data Integrity

Dr. Alaina Steele[1], Dr. Daniel Hepworth[2]

[1]Murray State University, Murray, KY
Email: *asteele3[at]murraystate.edu*

[2]Murray State University, Murray, KY
Email: *dhepworth[at]murraystate.edu*

Abstract: ***Missing data is a prevalent issue for those conducting research, with implications that extend beyond statistical inconvenience, especially for studies using longitudinal data, as retaining study participants over time can be challenging. Effectively addressing missing data presents a significant methodological challenge that is exacerbated by the absence of a universally accepted approach within the research community (Schober & Vetter, 2020). However, multiple imputation has emerged as a favored approach for its ability to address missing data without compromising the integrity of the study (Rubin, 1987). Despite this, the improper application of this method is fraught with issues. For example, employing multiple imputation under the assumption that data are missing at random (MAR) without first assessing the validity of that assumption may jeopardize the integrity of study findings. This paper offers researchers a practical approach that seeks to enhance data integrity through the application of diagnostic techniques and the inclusion of auxiliary variables during the imputation process, that strengthen the confidence in the MAR assumption.***

Data Availability Statement: ***The Pathways to Desistance data is available via the Inter - University Consortium for Political and Social Research (ICPSR). It is important to note that there are special restrictions for certain variables in this dataset. This paper utilized the public - use data files, which are available to the general public.***

Keywords: Missing data, multiple imputation, auxiliary variables, longitudinal study, data integrity

1. Introduction

The issue of missing data is a common challenge that researchers encounter, particularly in longitudinal studies where participant retention can be difficult (Engels, & Diehr, 2003; Laird, 1988; Schubert et al., 2004, & Spratt et al., 2010). This challenge is exacerbated by the lack of a consensus among researchers regarding the most appropriate method for addressing this issue (Schober & Vetter, 2020). To protect the integrity and power of study findings, it is essential to utilize an empirically supported method for managing missing data. One such approach is multiple imputation, which has received considerable empirical support. However, this method relies on the assumption that data is missing at random (MAR), a requirement that is difficult to meet, as there is no definitive test for proving MAR (Horton & Kleinman, 2007). This paper offers a practical approach to increase the plausibility of meeting the MAR assumption, thereby making multiple imputation a reasonable and effective tool for addressing missing data. We then explore one application of these solutions on existing research.

2. Solving for Missing Data

Multiple Imputation

Multiple imputation is one of the most respected methods for addressing missing data (Tabachnick & Fidell, 2019). First proposed in Rubin (1978), multiple imputation reduces uncertainty by computing several different options (i. e., imputations). Multiple versions (i. e., iterations) of the same dataset are created and combined to form the best fit for the missing values. This process involves making reasonable or probability estimates based upon the distributions and relationships between variables, maximizing the likelihood that the estimated values are close to the missing values (Morris et al., 2014).

A key assumption of multiple imputation is that data is missing at random (MAR), which can be difficult to meet because there is no test to definitively confirm it (Horton & Kleinman, 2007). Put simply, it is not possible to test whether MAR holds because we lack the missing data values required to compare and assess systematic differences between individuals with and without missing data (Allison, 2001). Despite the sophistication of multiple imputation, the validity of the results can be compromised if the MAR assumption is not met (Potthoff et al., 2006). Thus, ensuring this assumption is satisfied is crucial for maintaining the integrity of study findings.

Consequently, it is important to understand the type of missing data present in the dataset. Rubin (1976) introduced a typology for missing data that distinguishes between random and non - random missing data situations: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). MCAR is where the probability of missing data on a variable is unrelated to both the observed data and the unobserved data. In other words, the missingness is entirely random and does not depend on any factors in the dataset, whether observed or not. This is the least problematic form of missing data because the missingness does not introduce bias into the analysis (Little & Rubin, 2002). MAR occurs when participants with

incomplete data differ from those with complete data, but the pattern of missingness can be predicted or traced using other observed variables in the dataset. As Bennett (2001) explains, MAR means that "the pattern of 'missingness' is traceable or predictable from other variables in the dataset, rather than being due to the specific variable on which the data are missing" (p.464). Finally, MNAR occurs when the missing data cannot be predicted or explained by other variables in the dataset, making it the most difficult to address.

Classification of Missing Data

Assuming data are missing at random (MAR) is a fairly common assumption in multiple imputation and other methods for handling missing data. However, it is crucial to foster the plausibility of this assumption in your specific dataset (Li, 2013). The missing at random assumption states that the probability of the missing data depends on observed data, but not on the missing values themselves. In contrast, when data are missing not at random, the imputed data may be biased, potentially leading to incorrect conclusions. To assess whether MAR assumption is plausible, researchers should perform diagnostic tests. There are two tests that can shed some light on the classification of missing data so that it can properly be addressed: Little's Test of Missing Completely at Random Test – MCAR and a t - test with groups formed by indicator variables (Diggle et al., 1995; Tabachnick & Fidell, 2019).

Little's Test for MCAR is an overall test of randomness that compares the patterns of missing values for all variables with the pattern that would be expected if all missing values were missing completely at random (Diggle et al., 1995; Tabachnick & Fidell, 2019). If data is classified as MCAR, it satisfies the less stringent assumption of MAR, which is often sufficient for many missing data techniques. If the Little's test is not statistically significant, the missing values can be classified as missing completely at random. However, a significant result leads to the rejection of the null hypothesis, warranting further exploration (i. e., a t - test with groups formed by indicator variables) (Enders, 2010). Although this test assumes the dependent variable is normally distributed, the dependent variable in this study is positively skewed. Nevertheless, according to the Central Limit Theorem, the distribution of the mean approaches normality in large samples (n > 30), which mitigates concerns about skewness (LaMonte, 2016).

3. Techniques Applied

Illustrative Example

These methods for managing missing data were applied in a research project which used data from the Pathways to Desistance study (see https: //www.pathwaysstudy. pitt. edu/). This study was a longitudinal two - site study of serious adolescent offenders as they transitioned into adulthood and out of crime. According to Schubert and colleagues (2004), the goal of this study was to enlist serious adolescent offenders with enough heterogeneity to provide valuable insights into the influences of such things as treatment, punishments, and changes in one's life course on criminal trajectories. However, a challenge inherent in longitudinal studies, such as this one, is addressing attrition and missing data through appropriate analytical techniques.

Dependent Variables

The handling of missing data in research studies has been a point of considerable debate among statisticians and researchers due to its potential impact on study findings (Kang, 2013). This debate becomes particularly important when addressing missing values with dependent variables. Central to this debate is the concern over introducing bias. Advocates for inputting the dependent variable cite downward bias, where the association between variables is underestimated, increasing the risk of Type II errors (failing to reject a null hypothesis that is, in fact, false). In other words, one could fail to see an association when there actually is one. On the other hand, imputing the dependent variable could lead to upward bias, resulting in an overestimation of associations and an increased likelihood of Type I errors, or rejecting a null hypothesis that is, in fact, true. While both are concerning, the upward bias resulting from imputing the dependent variable is of much greater concern because it can indicate a relationship when there is not one.

Missing Variables in Example Study

As with any longitudinal study tracking high - risk populations, the Pathways Study faced the issues inherent in tracking serious adolescent offenders with frequently changing social contexts (e. g., changes in residence, entry into and out of correctional facilities, changes in peer groups, new school experiences) for repeated interviews over an extended period of time (Schubert et al., 2004). Ironically, the very nuances of the lives of study participants that made them scientifically interesting created significant issues for Pathways researchers seeking to maintain contact with them.

The research conducted using the Pathways Study data utilized 11 waves of analysis; for the sake of simplicity, only Wave 1 of that research will be discussed here (Steele, 2020) [1]. To assess the extent of missing values, a missing data analysis was conducted using SPSS for all analysis variables. Looking at Wave 1, 1382 values were missing from 151 cases (11.15% of the total cases), which computes to 6.805% of the total analysis values. Ultimately, while data for this wave are remarkably complete, it is still too great to consider any method of exclusion. To enhance the plausibility of the missing at random assumption of multiple imputation, it is essential to explore techniques for classifying the missing data.

Classification of Missing Data

A Little's Test of Missing Completely at Random Test (MCAR Test) was conducted for Wave 1 via SPSS. The test was statistically significant ($p = 0.004$), indicating the data did not meet the MCAR assumptions. In total, 245, 302 out of 641, 796 data points were missing. Because the data failed to meet MCAR criteria, further evaluation was needed to assess the plausibility of MAR before proceeding with multiple imputation (Yiran & Chao - Ying, 2013).

A *t - test* with groups formed by indicator variables was

[1] (Steele & Hepworth, in preparation)

performed on all analysis variables for Wave 1 to determine whether significant differences existed between cases with missing values and those with complete data. SPSS does this by first separating cases with complete values for a variable from those with missing values by creating an indicator variable for variables that contain missing values. Then, the group means of the quantitative variables are compared by utilizing the indicator variable as a group variable within the t - test analysis. When a significant result is obtained from a *t* - test, it suggests that the pattern of missing values is not random (IBM Corp., 2019). In other words, it lends plausibility to the assumption that the data are missing at random (MAR).

Missing at random lies on a spectrum (Garson, 2015). In other words, declaring missing data as missing at random (MAR) depends upon how much of the missingness can be explained by other observed variables. For example, "in a large dataset it might happen that missingness on a given variable was significantly related to another observed variable (hence not MCAR) but the relation was so trivial in effect size that missingness could not be predicted from that variable" (Garson, 2015, p.15). Essentially, the point on this spectrum that separates missing at random data from missing not at random data occurs where prediction is no longer useful.

Although multiple imputation assumes MAR, incorporating auxiliary variables related to the cause of missingness can help meet this assumption. When included in the imputation model. These variables can transform what would otherwise be MNAR pattern into a pattern compatible with MAR (Rhoads, 2012; Thoemmes & Rose, 2014). Including auxiliary variables related to the missingness or the variable itself can strengthen the plausibility of the MAR assumption in multiple imputation.

Auxiliary Varialles

When conducting multiple imputation, selecting appropriate auxiliary variables is critical for enhancing the integrity of your results. Auxiliary variables are variables that are not directly related to the outcome of interest but can help explain the missing data or predict the missing values. Including them strengthens the imputation models by providing additional information that can help clarify the missingness or predict missing values (Enders, 2010). However, selecting them requires careful consideration to avoid introducing bias or inefficiencies. There are two key considerations in selecting auxiliary variables: First, the variables must speak to the nature of the missing data, ensuring the plausibility of the MAR assumption. Second, both analysis and auxiliary variables must be included in the imputation model, as they provide insights into the missing values. Estimates of missing values can be improved by incorporating auxiliary variables into the analysis model (Collins et al., 2001).

Arguments for the inclusion of auxiliary variables in the imputation model are premised on their "ability to improve estimates that pertain to analysis variables with missing data" as well as "to reduce error variance and thus increase statistical power and precision of estimates" (Thoemmes & Rose, 2014, p.445). The inclusive strategy argues for the liberal use of auxiliary variables (Collins et al., 2001) According to Collins and colleagues (2001) the liberal use of auxiliary variables reduces the risk of omitting variables that are related to the cause of missingness or the variables with missing values. Omitting such variables could introduce bias into the imputation process by violating the conditional independence of MAR. Additionally, Collins and colleagues (2001) conducted a study which compared the restrictive strategy (i. e., a conservative use of auxiliary variables) and the inclusive strategy. They found that the inclusive strategy, characterized by generous use of auxiliary variables, was preferential. They found that the impacts of incorporating the inclusive strategy were neutral at worst, and significantly beneficial at best.

However, Thoemmes and Rose (2014) conducted a study to assess the impact of including auxiliary variables during the imputation process, with the following finding: "The overarching picture that emerged from our study is that there are situations in which auxiliary variables can induce bias or increase existing bias" (pp., 453 - 453). Furthermore, Thoemmes and Rose (2014) counter Collins and colleagues (2001) by arguing that while the benefits of the inclusive strategy may be acknowledged, its practical application has limits. In particular, multiple imputation using the inclusive strategy on large datasets with hundreds of variables would "likely encounter convergence problems" (p.445).

Consistent with this finding, multiple imputation on the Pathways to Desistance data using an inclusive strategy resulted in a MAXMODELPARAM warning for numerous variables, indicating they contained more than 100 parameters, which resulted in a failure to impute. The MAXMODELPARAM was increased which allowed for the imputation to run. However, as warned by IBM, it ran for twelve hours without moving past iteration zero. As noted by Thoemmes and Rose (2014), large datasets can cause convergence issues because the more variables included in the imputation model, the more parameters the model must estimate. One solution is to simplify the model by omitting unnecessary variables. As such, appropriate auxiliary variables were included in the imputation model instead of all variables within the dataset.

Auxiliary variables, while not analysis variables, can be used during the multiple imputation process to help predict missing values for analysis variables (Rhoads, 2012). Auxiliary variables incorporated into the multiple imputation model included scale level variables that fell under the umbrella of procedural justice measures, deterrence measures, socioeconomic status measures, drug and/or alcohol treatment, and the Weinberger Adjustment Inventory (WAI).

Missing Data Thresholds for Imputation

While there is no universally acceptable threshold for the percentage of missing values at which multiple imputation should not be used, Madley - Dowd and colleagues (2019) argue that the proportion of missing data should not be the sole criterion for deciding whether to use multiple imputation. However, the proportion of missing data should inform the researcher's choice regarding what auxiliary variables to include in the imputation model. Madley - Dowd and colleagues (2019) further argue that if the imputation model is properly specified and the data are missing at random, even large percentages of missing data (up to 90%) can be imputed without introducing significant bias.

Iterations

In our study utilizing the Pathways data, we used 20 iterations to increase the likelihood of getting a model that fits (i. e., convergence). Following Rubin's (1987) guidelines, five imputations became the default. According to Schafer (1999) " [u]nless rates of missing information are unusually high, there tends to be little or no practical benefit to using more than five to ten imputations" (p.7). However, Allison (1999) argues that such estimates, which are based on efficiency, are insufficient for standard error estimates, confidence intervals, and p - values. As software improves and becomes more powerful, adding iterations becomes more reasonable. Based on simulations, Graham and colleagues (2007) found that twenty iterations were needed for data that was missing between 10% and 30%. When data loss reached 50%, they found that forty imputations were needed.

4. Discussion

This article presented a common - sense approach for addressing missing data that centered around the use of multiple imputation, a powerful and popular approach for addressing missingness (Li et al., 2015). It is difficult to meet the assumption that data are missing at random, as there is no definitive test to confirm this (Horton & Kleinman, 2007). Fortunately, there are techniques that can help classify missing data and make the MAR assumption more plausible. Such techniques include Little's Test of Missing Completely at Random Test – MCAR and a *t* - test with groups formed by indicator variables. Additionally, the inclusion of auxiliary variables can help change a pattern of missing not at random to missing at random. To add clarity to the discussion, the Pathways to Desistance Study was used to illustrate this approach.

This example study provides extensive longitudinal data for analyzing the impact of various factors on the criminal trajectories of serious adolescent offenders. To help foster an appropriate classification of the missing data, Little's Test of Missing Completely at Random Test – MCAR and a *t* - test with groups formed by indicator variables were utilized. A Little's Test MCAR Test was conducted on Wave 1 of the Pathways to Desistance dataset, indicating that the data were not missing completely at random.

This study addressed the pervasive problem of missing data, especially in longitudinal studies where participant attrition is common. Multiple imputation is a respected method for handling missing data while protecting the integrity of study results. However, it is critical to foster the plausibility of multiple imputation's MAR requirement. Utilizing diagnostic techniques before imputation and incorporating auxiliary variables during the imputation process strengthens the MAR assumption. Collectively, these strengthen the data analysis and provide a more accurate picture of the population being studied.

As previously mentioned, there is no universal test available to definitively differentiate between data missing at random (MAR) and missing not at random (MNAR) (see Horton & Kleinman, 2007). However, because the data did not meet the assumptions of missing completely at random (MCAR), it became crucial to conduct a more thorough evaluation of the missingness to determine whether it was plausible that the data are missingness at random (MAR) before moving forward with multiple imputation (Yiran & Chao - Ying, 2013). To further examine the significance of the differences between individuals with missing values and those with complete values, a *t* - test was performed on all within - wave analysis variables for (IBM Corp., 2019). This test indicated statistically significant differences between the missing and completed groups for many study variables.

Missing at random is best understood as a spectrum (Garson, 2015). Classifying missing data as MAR is depends on the extent to which observed variables can account for the missingness. The point on this spectrum that separates missing at random data from missing not at random data occurs where prediction is no longer useful (Garson, 2015). Fortunately, there is an additional avenue for supporting the use of multiple imputation: the incorporation of auxiliary variables into the imputation model. Auxiliary variables are variables believed to be correlated with the underlying cause of missing data. By including such variables, researchers can strengthen the plausibility of the MAR assumption and transform the missingness process from MNAR to MAR (Rhoads, 2012; see also Thoemmes & Rose, 2014).

As the extant literature (as summarized above) and the Pathways example study illustrate, researchers conducting large - scale longitudinal research can address missing data by utilizating the practical tools outlined in this paper to strengthen the plausibility of the often - overlooked MAR assumption of multiple imputation. By incorporating diagnostic checks and leveraging auxiliary variables, researchers can better support the missing data at random assumption, which is vital for trustworthy and unbiased study results. Ultimately, this paper offers actionable strategies that researchers can apply to protect the integrity of their findings when confronting missing data challenged.

References

[1] Allison, P. D. (2001). *Missing data: Quantitative applications in social sciences* (vol.136). Thousand Oaks, CA: Sage Publications.
[2] Bennett, D. A. (2001). How can I deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, 25 (5), 464 - 469. https: //doi. org/10.1111/j.1467 - 842X.2001. tb00294. x
[3] Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6 (4), 330 - 351. https: //doi. org/10.1037/1082 - 989X.6.4.330
[4] Diggle, P., Liang, K., & Zeger, S. L. (1995). *Analysis of longitudinal data*. Clarendon Press; Oxford University Press. (Original work published 1994).
[5] Enders, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.
[6] Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: a comparison of methods. *Journal of Clinical Epidemiology*, 56 (10), 968 - 976. https: //doi. org/10.1016/S0895 - 4356 (03) 00170 - 7
[7] Horton, N. J., & Kleinman, K. P. (2007). Much ado about nothing: A comparison of missing data methods

and software to fit incomplete data regression models. *The American Statistician*, 61 (1), 79 - 90. https: //doi. org/ 10.1198/000313007X172556

[8] IBM Corp. (2019*). IBM SPSS Missing Values 26*. https: //www.ibm. com/docs/en/SSLVMB_26.0.0/pdf/en/IBM_SPSS_Missing_Values. pdf

[9] Kang, H. (2013). The prevention and handling of the missing data. Korean journal of anesthesiology, 64 (5), 402 - 406.

[10] Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine*, 7 (1-2), 305 - 315. https: //doi. org/10.1002/sim.4780070131

[11] LaMonte, W. W. (2016, July 24). *Central limit theorem*. The Role of Probability. http: //sphweb. bumc. bu. edu/otlt/MPH - Modules/BS/BS704_Probability/BS704_ Probability12. html

[12] Li, C. (2013). Little's Test of Missing Completely at Random. The Stata Journal, 13 (4), 795–809. https: //doi. org/10.1177/1536867X1301300407

[13] Li, P., Stuart, E. A., & Allison, D. B. (2015). Multiple imputation: a flexible tool for handling missing data. *Jama*, 314 (18), 1966 - 1967 https: //doi. org/10.1001/jama.2015.15281

[14] Little, R. J., & Rubin, D. B. (2002). Bayes and multiple imputation. *Statistical Analysis with Missing Data*, 200 - 220. https: //doi. org/10.1002/9781119013563. ch10

[15] Madley - Dowd, P., Hughes, R., Tilling, K., & Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*, 110, 63 - 73. https: //doi. org/10.1016/j. jclinepi.2019.02.016

[16] Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. G. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70 (2), 371 - 388. https: //doi. org/10.1111/j.1467 - 9868.2007.00640. x

[17] Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14 (1), 75. https: //doi. org/10.1186/1471 - 2288 - 14 - 75

[18] Potthoff, R. F., Tudor, G. E., Pieper, K. S., & Hasselblad, V. (2006). Can one assess whether missing data are missing at random in medical studies? *Statistical Methods in Medical Research*, 15 (3), 213 - 234. https: //doi. org/10.1191/0962280206sm448oa

[19] Rhoads, C. H. (2012). Problems with tests of the missingness mechanism in quantitative policy studies. *Statistics, Politics, and Policy*, 3 (1). https: //doi. org/10.1515/2151 - 7509.1012

[20] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63 (3), 581 - 592. https: //doi. org/10.1093/biomet/63.3.581

[21] Rubin, D. B. (1978, January). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, *(*1), 20 - 34. http: //www.asasrms. org/Proceedings/papers/1978_004. pdf

[22] Schubert, C. A., Mulvey, E. P., Steinberg, L., Cauffman, E., Losoya, S. H., Hecker, T., Chassin, L., & Knight, G. P. (2004). Operational Lessons from the Pathways to Desistance Project. *Youth Violence and Juvenile Justice*, 2, 237 - 255. https: //doi. org/10.1177/1541204004265875

[23] Schober, P., & Vetter, T. R. (2020). Missing data and imputation methods. *Anesthesia & Analgesia*, 131 (5), 1419 - 1420. http: // doi. org/10.1213/ANE.0000000000005068

[24] Spratt, M., Carpenter, J., Sterne, J. A., Carlin, J. B., Heron, J., Henderson, J., & Tilling, K. (2010). Strategies for multiple imputation in longitudinal studies. *American Journal of Epidemiology*, 172 (4), 478 - 487. https: //doi. org/10.1093/aje/kwq137

[25] Steele, A. D. (2020). A theoretical integration of procedural justice and deterrence: a tale of two theories. [Doctoral dissertation, Southern Illinois University Carbondale]. ProQuest Dissertations & Theses Global.

[26] Steele, A. D. & Hepworth, D. P. (manuscript submitted for publication). Crime and consequence: analyzing the deterrent effect on high - risk juvenile offenders.

[27] Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

[28] Thoemmes, F., & Rose, N. (2013). Selection of auxiliary variables in missing data problems: Not all auxiliary variables are created equal. *Technical Report Technical Report* R - 002, Cornell University. https: //www.researchgate. net/publication/260059479_Selection_of_auxiliary_variables_in_missing_data_problems_Not_all_auxiliary_variables_are_created_equal

[29] Thoemmes, F., & Rose, N. (2014). A cautious note on auxiliary variables that can increase bias in missing data problems. *Multivariate Behavioral Research*, 49 (5), 443 - 459. https: //doi. org/10.1080/00273171.2014.931799

[30] 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. https: //doi. org/10.1002/sim.4067

[31] Yiran, D., & Chao - Ying, J. P. (2013). Principled missing data methods for researchers. *Springer Plus*, 2 (1), 1 - 17. https: //link. springer. com/content/pdf/10.1186/2193 - 1801 - 2 - 222. pdf