Effects of Missing Data on Student Growth Estimates

Katherine M. Wright[1], John L. Gatta[1], Terri D. Pigott[2]
[1]Northwestern University, 710 N Lake Shore Drive, Chicago, IL 60611
[2]Loyola University Chicago, 11 E Pearson St, Chicago, IL 60611

Abstract

One property of student growth data that is often overlooked despite widespread prevalence is incomplete or missing observations. As students migrate in and out of school districts, opt out of standardized testing, or are absent on test days, there are many reasons student records are fractured. Missing data in growth models can bias model estimates and growth inferences. This paper presents empirical explorations of how well missing data methodologies recover attributes of would-be complete student data used for teacher evaluation. Missing data methods are compared in the context of a Student Growth Percentiles (SGP) model used by several school systems for accountability purposes. Using a real longitudinal dataset, we evaluate the sensitivity of growth estimates to missing data and compare the following missing data methods: listwise deletion, likelihood-based imputation using an expectation-maximization algorithm, multiple imputation using a Markov Chain Monte Carlo method, multiple imputation using a Predictive Mean Matching method, and inverse probability weighting. Methodological and practical consequences of missing data are discussed.

Key Words: missing data, multiple imputation, growth model, maximum likelihood

1. Introduction

Most statistical procedures are designed for complete datasets. Consequently, analyzing student achievement data prone to missing observations may impact model findings and subsequent data-driven decisions. As students migrate in and out of school systems, leave high school, or are absent from testing, there are many reasons student records are fractured. If not properly accounted for, incomplete student data may be an invisible covariate affecting evaluation inferences in student growth models. Further research is necessary to ensure student growth models mitigate bias due to missing data. The purpose of this study is to present empirical explorations of how well missing data methodologies recover attributes of would-be complete student data used for teacher evaluation.

The Student Growth Percentiles (SGP) model (D. W. Betebenner, 2011) was selected as a focus of this review since over 30 states have chosen to adopt it in some capacity. This model is normative in nature and produces student percentile ranks as its growth metric. Although this study explores missing data in the context of an SGP model, many concepts are applicable to the broader category of growth models.

2. Overview of Missing Student Data

2.1 Motivation

Missing data is a frequent problem for most researchers. In theory, the best way to mitigate the consequences of missing data may be to prospectively design a study that minimizes the potential for incomplete observations. In practice, often the data collection process is a balance of cost, control, and feasibility that results in an imperfect final product with missing observations. Large and small-scale research projects alike are susceptible to missing data due to attrition, participant error, data collection glitches, and data entry problems. Longitudinal data utilized in student growth models is especially vulnerable to missing observations as the reasons above are compounded over multiple years in addition to mobility in and out of the district. As there are likely unobserved covariates in every student achievement data set (e.g. student motivation), missing data methodology is relevant to most educational researchers (D. Rubin et al., 2004).

Concern about missing data is warranted given how prevalent this issue tends to be. In a review of missing data in VAMs, McCaffery found large school districts were missing at least one score from between 42 – 80% of students (D. F. McCaffrey & J. Lockwood, 2011). The distribution of missing student scores was inconsistent across teachers. On average, 37% of teacher rosters contain fully complete student records but this varies from 0 to 100% in every grade. Additionally, missing data occurred in non-random patterns that are especially relevant when selecting a missing data methodology.

Fortunately, statistical packages make many missing data handling techniques readily available to researchers. Unfortunately, the most common default procedure, listwise deletion (or complete case analysis), is only appropriate for specific situations that are unverifiable (Peugh & Enders, 2004; Roth, 1994). This can be troubling as some researchers may not be aware of the bias they introduce by accepting default settings. Either explicitly or implicitly, all researchers account for missing data and should be aware of the consequences of their chosen method.

2.1 Missing Data Mechanisms

Rubin's taxonomy of missing data mechanisms has become the standard classification scheme cited in most research (Donald B. Rubin & Wiley, 1987). He specified three mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Data are MCAR when the probability of missing observations is independent of any other variable (latent or observed). Essentially, data are arbitrarily missing and thus the observed data can be considered a random sample of the complete dataset.

Data are MAR when the probability of missing observations is independent of the missing variable itself, but related to another variable. As the missingness is conditional on another variable in the dataset, there are a variety of methods available to restore attributes of the would-be complete dataset using information from other non-missing variables. More relaxed than the MCAR condition, most missing data procedures require data to be MAR. There are no formal diagnostic tests to detect a MAR mechanism.

Data are MNAR when the probability of missing observations is a function of the missing variable itself. For example, data are missing not at random if all test scores below a specific score were not recorded and thus are missing from the analysis. Because missingness mechanisms cannot be verified, statisticians can conduct sensitivity analyses assuming different mechanisms to determine how robust their findings are.

3. Design and Methods

3.1 Data

To illustrate the consequences of different missing data handling techniques for student growth data, this study analyzes Measures of Academic Progress (MAP) mathematics achievement scores. Study data included 3rd and 4th grade mathematics achievement scores, student demographic characteristics, and classroom rosters. Actual test scores were chosen over simulated data to ensure the relationship between past and future performance accurately reflects what exists in practice. Though simulating scores could provide additional statistical control over missing data patterns/mechanisms, these controls may not translate to practice settings where data is often more complicated.

3.2 Evaluation Design

Using a real longitudinal dataset of student records, the following steps were implemented:

1. **Artificially Censor Observations**: remove observations based on their likelihood of being observed in a reference population of similar students.
2. **Implement Missing Data Methods**: separately implement the following missing data methodologies:
    - Listwise Deletion
    - Imputation using an Expectation Maximization algorithm
    - Multiple Imputation via a traditional Markov Chain Monte Carlo method
    - Multiple Imputation via a Predictive Mean Matching method
    - Inverse Probability Weighting
3. **Analyze Growth using the SGP Framework**: calculate quantile regression estimates for each dataset, compute student growth percentiles for each student
4. **Compare Results**: compare growth estimates to the would-be complete data (benchmark/true) values.

4. Growth Model Specification

4.1 Evaluation Scenario

To isolate the effect of missing data from confounding variables, this analysis concentrates on evaluations across a single subject and grade. For an evaluation scenario with one prior year of data, 4th grade mathematics growth was evaluated from a baseline of 3rd grade mathematics achievement.

4.2 Benchmark Growth Model

The SGP model implements quantile regression techniques to model the complex relationship between historical and future achievement trajectories. This process can be accomplished using the SGP package for the R programming environment (D. V. Betebenner, Adam; Domingue, Ben; Shang,Yi 2014). From this data, a matrix of scale scores and corresponding quantiles can be created for each percentile band. Student growth percentiles are defined as:

$$\text{SGP} = \Pr(\text{Current achievement} \mid \text{Prior Achievement}) * 100$$

A student's growth percentile is determined by identifying the quantile with the value closest to the student's observed score. Before artificially censoring observations, growth estimates were calculated to serve as a benchmark for comparing missing data methods. Five copies of the censored dataset were used to impose each of the 5 missing data methods in this study. After pre-processing data using each missing data method, growth was calculated using a student growth percentiles model. No demographic characteristics are used in the growth analysis; 3^{rd} and 4^{th} grade mathematics scores are the only variables used in the SGP model. To generate student growth estimates, the τ^{th} quantile of 4^{th} grade mathematics achievement, represented as $Q(\tau|\, X = x) = x'_i\beta(\tau)$ is solved by the following:

$$\hat{\beta}(\tau) \; = arg \min_{\beta \in R^P} \sum_{i=1}^{n} P_\tau(y_i - x'_i\beta)$$

where $0 < \tau < 1$ (Chen, 2005; Koneker, 2005). This means $\tau = .25$ represents the 25th percentile, $\tau = .5$ represents the median or 50th percentile, and $\tau = .75$ represents the 75th percentile. The SGP model estimated quantiles 1 through 99 and compared quantile regression estimates to fitted values. A student's SGP was determined by the closest quantile curve to their actual score given their prior test history.

4.3 Analytic Sample

To determine the magnitude of missingness to impose on the would-be complete dataset, missingness was examined in a reference population of students in a different grade at the same school. 7% of students had incomplete mathematics achievement data in the reference population; therefore 7% of current year mathematics scores were also censored in the analytic sample of 4th grade students. Several differences were observed between censored and non-censored students.

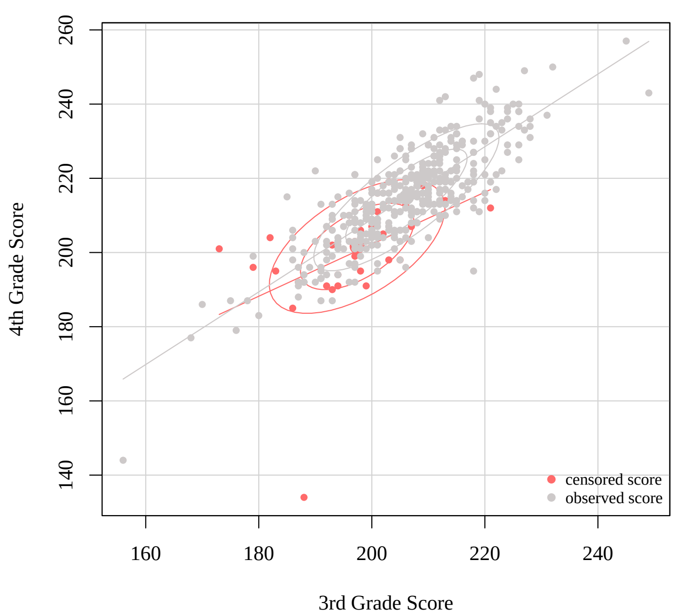

Figure 1: Censored and observed distributions of 3^{rd} and 4^{th} grade scores; concentration ellipses are plotted at .5 and .8 and OLS regression lines are overlaid for censored and observed student cohorts

5. Results

Correlation results in Table 1 show SGPs derived from all 5 missing data methods were comparable to the benchmark SGP values. For the 3 imputation-based missing data methods (EM Imputation, MI via a MCMC method, and MI via a PMM method), imputed values were used to estimate growth quantiles of the overall distribution of 4th grade students. SGPs were not reported for students with imputed scores since predicting individual scores is not the purpose of imputation.

Table 1. SGP Correlations

	Benchmark	Listwise Deletion	EM Imputation	MI via a MCMC method	MI via a PMM method	Inverse Probability Weighting
Benchmark		0.996***	0.997***	0.997***	0.996***	0.979***
Listwise Deletion	0.996***		0.997***	0.997***	0.997***	0.983***
EM Imputation	0.997***	0.997***		0.995***	0.998***	0.979***
MI via a MCMC Method	0.997***	0.997***	0.995***		0.995***	0.980***
MI via a PMM method	0.996***	0.997***	0.998***	0.995***		0.981***
Inverse Probability Weighting	0.979***	0.983***	0.979***	0.980***	0.981***	

****p<.001*

The figure below shows the relationship between true/benchmark student growth percentiles derived using the full data and estimates derived using the censored analytic samples.

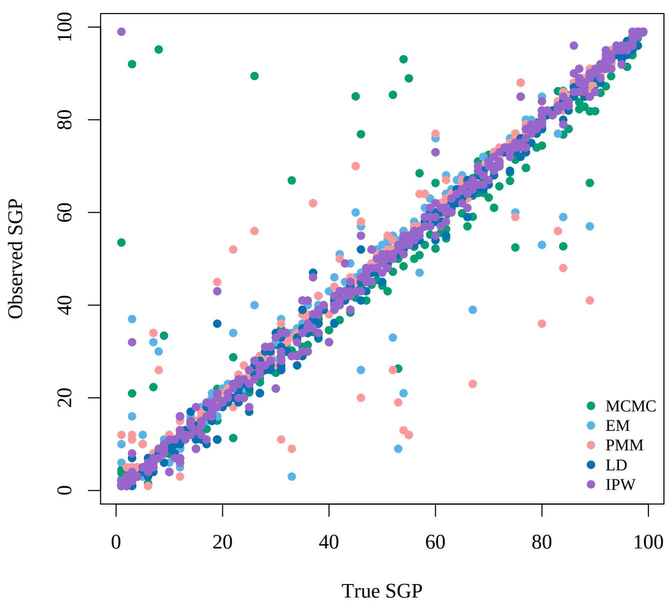

Figure 2: True (benchmark) and observed SGPs for each missing data method

Correlations are not the only criteria for comparing growth estimates derived under each missing data method. By itself, a correlation does not indicate how many percentile values students change (e.g. the same correlation value could represent a shift from the 1st to 2nd growth percentile values or a shift from the 1st to 52nd growth percentile). Framing model differences using the actual SGP metric provides additional context. To supplement correlation findings, Table 2 provides he average absolute values for SGP residuals for each missing data method.

Table 2. Differences between Benchmark SGPs and Missing Data SGP Estimates

	Mean Absolute Residual
Listwise Deletion	1.764
EM Imputation	1.203
MI via a MCMC method	1.558
MI via a PMM method	0.968
Inverse Probability Weighting	1.948

In this study, listwise deletion data produced the largest mean absolute error, followed by inverse probability weighting, while multiple imputation via the semi-parametric PMM method produced the smallest mean absolute error in growth estimates.

6. Conclusions and Future Work

The literature on missing data methodologies for student growth models is sparse; therefore, patterns of missingness were explored using a real dataset of mathematics achievement scores and student characteristics. In general, the results favored imputation methods over deletion and weighting approaches when the criteria were 1) the correlation to benchmark SGP values, 2) the smallest mean absolute residuals. Similarities between models are not surprising since a relatively small amount of missingness was imposed. Still, this study demonstrates the utility of missing data methods in improving growth estimates when the amount of missing observations is relatively small.

A logical extension of this work would be to manipulate the percent of missingness imposed on the analysis to determine if missing data methods perform differently. This information can guide practitioners as they choose a method for their specific context. Given the complex nature of school systems, results from this study may not generalize across different growth models, assessments, grades, or magnitudes of missingness. However, these findings may generate starting points for discussion when implementing student growth models that rely on incomplete student records.

References

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. The National Center for the Improvement of Educational Assessment.

Betebenner, D. V., Adam; Domingue, Ben; Shang,Yi (2014). SGP: An R Package for the Calculation and Visualization of Student Growth Percentiles & Percentile Growth Trajectories.: R package version 1.2-0.0.

Rubin, D., Stuart, E., & Zanutto, E. (2004). A potential outcomes view of value-added assessment in education. Journal of Educational and Behavioral Statistics, 103-116.

McCaffrey, D. F., & Lockwood, J. (2011). Missing data in value-added modeling of teacher effects. *The Annals of Applied Statistics, 5*(2A), 773-797.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. Review of Educational Research, 74(4), 525-556.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. Personnel psychology, 47(3), 537-560.

Rubin, D. B., & Wiley, I. (1987). Multiple imputation for nonresponse in surveys.

Chen, C. (2005). An introduction to quantile regression and the QUANTREG procedure. Paper presented at the Proceedings of the Thirtieth Annual SAS Users Group International Conference.

Koenker, R. (2005). Quantile regression: Cambridge university press.