Imputation of Missing Data for the Pre-Elementary Education Longitudinal Study

Lin Li[1], Hyunshik Lee[1], Annie Lo[1], Greg Norman[1]
[1]Westat, 1650 Research Boulevard, Rockville, MD 20850

Abstract

In the Pre-Elementary Education Longitudinal Study (PEELS), imputation of item missing data was done using AutoImpute (AI) software, which uses semi-parametric modeling to form imputation classes. In this paper, we summarize PEELS experience with AI, investigate the bias aspect of the imputed data for the PEELS teacher questionnaire data, and study the variance estimation of imputed data using multiple imputation by AI. In the study of variance estimation, we look into the bias issue for the multiple imputation method and the performance of AI multiple imputation on domain estimation.

Key Words: Wave missing and imputation, bias due to imputation, variance estimation for imputed data

1. Introduction

PEELS is a longitudinal survey to study the preschool and early elementary school experiences of children with disability and their progression through early special education. The study took a nationally representative sample of about 3,000 children, consisting of three age cohorts, 3, 4, and 5 years of age at the start of the study in 2003–2004. The broad objectives of the study are to learn about the characteristics of children receiving special education and the preschool programs and services they receive. Also of interest are their transition from early intervention to preschool, from preschool to elementary school, their achievement results as they move from one program to another, and the factors that contribute to these results.

Four surveys were taken for the recruited children. The children were assessed by trained assessors to produce assessment data (Child Assessment survey), a computer assisted telephone interviewing was conducted with their parents (Parent survey), and their teachers (Teacher survey) and school principals or program directors (Principal/Program Director survey) were sent a mail questionnaire. The Teacher survey was not about the teacher but about the children they taught. Teachers filled out a questionnaire for each sampled child in their classes. The principal and program director survey was about the school, not specific children. One questionnaire was filled out for their entire school or program. Although different surveys were taken from different people the collected data pertain to the children therefore the analysis unit is always child, not the respondent of a particular survey.

The same children were followed every year beginning in 2003–2004. In following the children, the respondent to the Teacher and Principal and Program Director surveys may differ from year to year, as the child's teacher and/or school may have changed. We recently finished the fourth year of data collection. This was the last wave for all instruments except the child assessment survey, which will be conducted once more. All the sampled units were invited to participate in the study regardless of their response status in the previous wave.

The four surveys had differing level of unit nonresponse. Assessment and Parent surveys had much higher response rates with over 95 percent in wave 1 and maintaining over 80 percent until the fourth wave. The Teacher survey suffered a relatively low response rate at the first wave (74 percent) but improved at later waves to over 80 percent. The Principal and Program Director (PPD) survey response rate was low throughout (below 70 percent), so we supplemented the data using the Quality Education Data (QED). Including the QED cases in the response set, the unit response rates for the four waves were over 90 percent.

Item nonresponse rates in the Assessment, Parent, and Teacher surveys were low. The majority of variables for these surveys had less than 5 percent and almost all had less than 10 percent nonresponse. The Principal and Program Director survey however had considerably higher item nonresponse rates. This was because the QED did not provide compatible data for all the variables in the PPD. Since those variables became item missing, the item missing rates were high with over 20 percent for those variables.

Most of items missing values were imputed using imputation software called AutoImpute, which is a Westat proprietary software package. It performs hot-deck imputation using imputation cells created by regression and generalized linear models (GLM). For each variable, which needs imputation, it builds a regression/GLM model using the variable with missing values as dependent variable and all other variables (marked as predictors) as independent variables. To facilitate this modeling exercise, all missing values that are involved in modeling are temporarily imputed using hot-deck imputation with global imputation classes formed by the specified hard and soft boundaries. Since the same boundaries are used for all variables, the imputed values may not be superior but it helps to build the regression/GLM models to create better imputation classes. Next the temporarily imputed values are used to build better models and imputation classes. This cycle repeats until no more significant improvement is realized. In both preliminary and model based imputation steps, AutoImpute respects skip pattern automatically, which is not a trivial task.

AutoImpute has a few important advantages over other imputation software: (1) it draws prediction power contained in all of the available variables in the data set, not only auxiliary variables without missing data but also survey variables with missing values; (2) it strives to maintain the correlation structure through the modeling process; and (3) it maintains skip patterns in the imputed data. Another feature is its capability to perform multiple imputation. Because it can use many predictors to model imputation classes, AutoImpute is a good tool to impute missing data for longitudinal surveys, where previous waves data can be naturally brought in as predictors.

In this article, we discuss our experience using AutoImpute for imputing missing data in PEELS. In section 2 we describe this experience and highlight the special circumstances of the four surveys and how we dealt with them.

For all of the PEELS surveys, except the Teacher survey, a longitudinal respondent is defined as response in all waves. For the Teacher survey this definition would yield too low of a longitudinal response rate. So any baseline records with one wave missing were also included in the longitudinal data set. To avoid complications due to missing data we imputed the wave missing data, as well as item missing data, for the Teacher survey. However, in doing this there was a concern about the potential for bias. So we examine this issue in section 3.

As mentioned earlier, the PPD survey had high item nonresponse rate. We judged that the imputation variance should not be ignored for the PPD survey and we performed multiple imputation to facilitate variance estimation for the imputed data. In section 4, we present a simulation study that shows how well the multiple imputation variance estimator works in AutoImpute

2. Experience with AutoImpute for Imputation of Missing Data in Longitudinal Data

In this section, we will discuss the following experience with using AutoImpute (AI); preparing for imputation, running AI, checking the results, and finally the strengths and limitations of AI.

2.1 Preparing for Imputation

The most important part of preparation deals with data cleaning. A clean input dataset is important since AI uses the skip patterns built into the questionnaire. Any improper data flow will cause AI to fail. The failure of an AI run will require a rerun on AI. Reruns may take a long time. As a result, data problems may significantly lengthen the imputation process.

In PEELS a substantial amount of data cleaning was required. Data cleaning mostly involved reconciling the inconsistent data within a survey instrument, between instruments, or across waves. A lot of cleaning was even required for the Parent survey which was done by Computer Assisted Telephone Interviewing and should have been a relatively clean dataset.

A classic example of maintaining consistency within a survey instrument is the "reverse control" issue. These are cases where by having any response to a question implies a specific answer to a prior question that had a missing value. For example, in the Parent Interview a child wearing a hearing aid would imply that a hearing problem was diagnosed by a professional.

An example of maintaining consistency between instruments can be found in the Teacher survey. When information about a child was missing in this survey, we first checked to see if it was available from the Parent interview or the Child Assessment. If the equivalent data was available from one of these other surveys we would use this as our imputed value in the Teacher data.

In terms of maintaining cross-wave consistency, data from previous waves could be used to manually impute values for attributes that should not change over time. If for example we know from a previous round that a child has been diagnosed with a hearing problem by a professional, then any future missing values of this question should be consistent with this prior response.

Another important part of preparation is creating the Master Index File (MIF). MIF is a variable-level specification which tells AI how to treat items on the input dataset. For example, it specifies the imputation order, variable types, and skip controllers. MIF is indeed a blueprint for imputation.

2.2 Running AutoImpute

It often requires several runs of AI to get a final set of imputed values because initial runs of AI typically uncover additional data problems (not caught in data checking), problems with the Master Index File, or a lack of imputation donors. Additionally, since each AI run produces a new set of imputed values, each new run of AI can produce its own unique set of problems. What follows are some examples of the problems encountered in imputation for PEELS.

In the Assessment data, assessment scores were used as skip controllers based on the logical relations between subtests. This caused an insufficient donor problem as the scores are continuous variables with wide ranges. To solve the problem, the score skip controllers were categorized.

In the Teacher survey, one question asked the teacher to rank 1 (most important), 2 (second most important), and 3 (third most important) among eight of the approaches to working with the child. To ensure that only three approaches were selected and were ranked 1, 2, and 3, we initially imputed the first approach, then used the first approach as one of the hard boundaries for imputing the second approach, then used the first and second approaches for imputing the third approach, and so on until we had used the first seven approaches as hard boundary for imputing the eighth approach. We refer to this method as the "stacking method". The stacking method was successfully applied to many other questions.

The teacher questionnaire also contains a question where the responses must be a subset of the previous question. The question stated, "Of the items specified earlier what three activities does this child engage in most often in your classroom or program?" To maintain the relationship between this question and the previous question, we switched the order of the questions in the imputation. We imputed the second question initially then used it as a hard boundary for imputing the first question.

The amount of time it takes for a complete AI run varied depending on several factors; the number of data items to be imputed, number of predictors to be used in modeling, number of records in the file, questionnaire complexity (e.g., skip patterns), and computer speed. For example, it took one minute for AI to impute Assessment data that had six variables to be imputed and most imputation rates under 0.4%. It took 12 hours to impute 451 Teacher variables with most imputation rates under 10%. The Assessment input file for AI contained about 2,500 cases and 100 variables; while the Teacher input file has about 3,000 cases and 2,700 variables. In PEELS, AI was run on the Linux server instead of a Windows desktop computer. Imputation is heavy on computing due to the extensive modeling, thus running on the Linux server saves a lot of time. The run time on the server is estimated to be only 10 percent - 20 percent of the time that would need on a desktop computer.

Once a successfully imputed dataset was created, the results were checked further and additional (manual) changes were made as needed. This was often the result of valid imputed data that were considered highly improbable. Data checking was also done to remove any inconsistencies with another instrument or a previous wave's data.

2.3 Limitations and Strengths of AutoImpute

One of the difficulties we encountered is the insufficient donor problem when skip controllers have too many categories. Another limitation is that AI cannot guarantee consistency among imputed items, when the items are not related by skip patterns but do have logical relations. For example, one item is the total enrolment in a school; the other item is the school's special education enrolment. Though there is no skip relation between them, the special education enrolment should not be larger than the total enrolment.

Although it has some limitations, AI has several appealing features for imputing the PEELS data. First is the capability to honor skip patterns. This is important because many PEELS questions have skips. AI also reports on skip-pattern violations so that edit problems can be identified. Second is the capacity to impute the entire questionnaire in a single run. This is important because of the volume of PEELS imputation. For example, each wave of the Teacher data required the imputation of approximately 450 variables. Third is that AI strives to preserve the correlation structure among variables. This is important because of the complex correlation structure of the PEELS cross-sectional data and longitudinal data.

Overall, the use of AI for the PEELS data was successful. Given the volume of imputation that was performed and the complexity of PEELS instruments, AI performed very well in almost all situations.

3. Study on the Bias of Imputed Data

As mentioned previously, to increase the number of cases available for longitudinal analysis for the Teacher survey, we imputed the whole missing wave data for baseline respondents who have only one wave missing. Performing wave imputation raised the concern about a potential for bias in the estimates. A specific question is "Are the cross-sectional estimates produced from the longitudinal file similar to those from cross-sectional files"? The following study was conducted to answer this question.

We selected five key variables in Wave 2 and compared their estimates from the longitudinal file to the estimates from the Wave 2 cross-sectional file. Estimates of similar number of key variables from Waves 3 and 4 were also compared. Table 1 gives the number of cases in the cross-sectional and longitudinal data sets.

Estimates from the longitudinal file and those from the Waves 2-4 cross-sectional files were compared to see if they were equal or not. The equality was tested by the t-test with 5 percent significance level. The results are presented in Table 2.

Table 1: Data sets used to study bias for wave imputed teacher data

Type of data	*Number of cases*	*Response rate (%)*
Longitudinal data set with wave imputation	2,049	66.0
Wave 2 cross-sectional data	2,591	83.5
Wave 3 cross-sectional data	2,514	81.0
Wave 4 cross-sectional data	2,502	80.6

*Note: The response rate is calculated as the number of child records in the data file divided by the total enrolled children (*3,104*) for PEELS in Wave 1.*

Table 2: Results of comparisons between the teacher longitudinal file and cross-sectional files

		Number of variables that are significantly different	
	Overall	*Individual comparison*	*Multiple comparison*
Categorical variable	45	2	0
Continuous variable	8	3	2
Total	53	5	2

For categorical variables, the point estimate for each category was compared between the longitudinal file and cross-sectional files. The small numbers of records (less than 3) in some categories precluded the comparisons of the estimates. Excluding these cases, there were altogether 53 individual comparisons in Table 2. If comparisons using the t-test were made individually, there were five significant results. When we used multiple testing procedures (both the Bonferroni procedure and the Benjamini-Hochberg (1995)) for each variable, none of the categorical variables showed significance, and only two continuous variables showed significance. Overall, the test results indicate no serious difference in the point estimates obtained from the longitudinal and cross-sectional files.

Assuming that the estimates from cross-sectional files are nearly unbiased, the study indicates that the bias due to imputation is not serious in the longitudinal file[1]. An important advantage of the wave imputation option is that it enhances the usability of existing data for longitudinal analysis of the teacher data. Almost 20 percent more cases are available for longitudinal analysis than the longitudinal data set created with the strict criterion of longitudinal records (i.e., records with no wave missing).

4. Study on the Variance Estimation of Imputed Data

Variance estimation of the imputed data was investigated for the multiply imputed Principal/Program Director (PPD) data. As mentioned earlier, the imputation rates for several key PPD items are over 20%. Since the ordinary variance estimator underestimates the true variance of imputed data, the underestimation may be non-negligible with such high imputation rates. To address this issue, we created multiply (5 times) imputed PPD data using AI. Data users can then compute multiple-imputation variance estimates. However, the evidence that the AI version of multiple imputation estimates provides valid variance estimates is scarce. So we conducted a simulation study on the bias of the variance estimates produced by multiple imputation method using AI.

4.1 Variance Estimation of Data Multiply Imputed by AI

The multiple imputation method proposed by Rubin (1987) provides a very simple variance estimator. Let $\hat{\theta}_i$ be the estimate for θ from the *i*-th multiple imputation, and $\hat{V}_i(\hat{\theta})$ be an ordinary variance estimator of $\hat{\theta}_i$. Then the multiple imputation estimate for θ and its variance estimate are, respectively, given by

$$\hat{\theta}_M = \frac{1}{K}\sum_{i=1}^{K}\hat{\theta}_i ,$$

$$\hat{V}_M(\hat{\theta}) = \frac{1}{K}\sum_{i=1}^{K}\hat{V}_i(\hat{\theta}) + \frac{K+1}{K(K-1)}\sum_{i=1}^{K}(\hat{\theta}_i - \hat{\theta}_M)^2 ,$$

where *K* is the number of imputations.

The multiple imputation variance estimator consists of two parts: the within-imputation variance, which estimates the sampling variance, and the between-imputation variance, which estimates the variance due to imputation. Due to time constraints, we focused on the bias of the between-imputation variance estimate.

The simulation study used the imputed PPD data, so there is no missing value in the simulation study data. Treating the PPD sample data as the population data, no sampling experiment was done but just nonresponse was simulated. So there was no sampling variance. In total twenty-two PPD items were investigated. For each study variable, 20 percent of data were set to missing by one of two non-response mechanisms: missing completely at random (MCAR) and missing at random (MAR). Determination of the nonresponse mechanism for each variable was done on the basis of the R^2 statistic of the imputation model used in the actual imputation. Then AI was run five times independently to generate five sets of multiply imputed data. These two steps were repeated 1,000 times.

[1] One limitation of the study is that we focused on univariate analysis rather than multivariate analysis for which different results could have emerged.

The variance of 1,000 multiple imputation point estimates gives the simulated between-imputation variance. The relative bias of the between-imputation variance estimated by Rubin's formula is calculated by

$$\text{Relative Bias} = \frac{\text{Between imputation variance by Rubin's formula} - \text{Simulated between imputation variance}}{\text{Simulated between imputation variance}}.$$

The findings are summarized in Table 3. For the overall population, all of the items we looked at had an underestimate of their between-imputation variance of more than 20 percent. A different picture emerged for domains defined by MSA status (urban, suburban, and rural areas). For the urban and suburban areas, although for most of the items the bias is still negative, it is generally smaller, while in rural area, which is the smallest domain, for one half of the items the bias is negative, and for the other half, it is positive.

Table 3: Distribution of the relative bias of between-imputation variance estimates

Relative bias of between-imputation variance estimate %	*Number of PPD items*			
		Domain (MSA status)		
	Overall (n = 2,433)	*Urban* (*n* = 920)	*Suburban* (*n* = 1,139)	*Rural* (*n* = 374)
< –40	10	1	1	2
–40 ~ –20	12	12	8	2
–20 ~ 0	0	4	9	7
0 ~ 20	0	5	3	5
>= 20	0	0	1	6
Total	22	22	22	22

The underestimation of the between-imputation variance in AI multiple imputation is due to the hot-deck method used by AI. As pointed out by Rubin and Schenker (1986), the hot-deck method does not adjust for the uncertainty due to parameter estimation, thus underestimates the true variance. The underestimation for domains is less serious because of selecting donors from outside the domain during imputation, which increases the estimated variance, thus offsets the underestimation of AI multiple imputation. In general the multiple imputation method overestimates the domain variance as shown by Kim et al. (2006), and the underestimation of the AI multiple imputation dampens this effect. In addition, the between-imputation variance constitutes only a small portion of the total variance, ranging from 2 to 32 percent with majority of them less than 20 percent in our study. So the underestimation for the total variance is not as serious as the between-imputation variance suggests.

4.2 Variance Estimation of Data Imputed by Pseudo-ABB Method

Rubin and Schenker (1986) proposed several multiple imputation methods that can account for the variance not captured by hot-deck imputation. One of them is Approximate Bayesian Bootstrap (ABB) imputation. However, since ABB has to be done within imputation cells, which means it has to be included within the AI macro, it is not feasible to implement correct ABB under the current AI structure. So we experimented a pseudo-ABB approach, i.e. before running AI, we drew simple random samples with replacement from the non-missing records. This approach implies that AI was run with a bootstrapped response set instead of the original data to impute missing items. One limitation of this experiment is that in one AI run, all the variables to be imputed have to be missing for the same cases, so variables with different missing patterns have to be imputed in separate runs.

Figure 1 illustrates the relative bias of between-imputation variance estimates by the hot-deck method versus the pseudo-ABB method. The findings are based on six PPD items (V1, V2, …, V6) with 200 simulations. As can be seen, the relative bias of the pseudo-ABB method is much closer to 0 than that of the hot-deck method. Even though the scale of the study is small, the pseudo-ABB method shows considerable improvement on variance estimation over the hot-deck method.

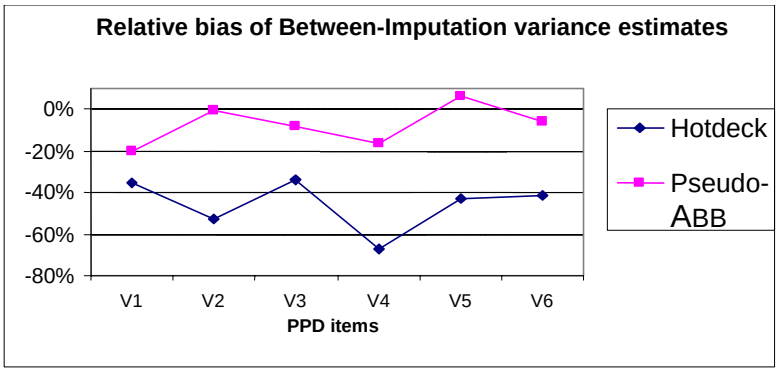

Figure 1: Comparing the relative biases of between-imputation variance estimates by hot-deck method and those by the pseudo-ABB method.

Conclusion

Imputation of missing data is widely used to enhance the usability of survey data and to avoid undesirable complications when analyzing data with missing values. In this paper, we summarized our experience of conducting semi-parametric imputation of PEELS data with AutoImpute software. The use of AI improved the quality of PEELS imputed data and reduced the imputation time. We also reported two studies on PEELS imputed data. The bias study on Teacher data indicates no serious bias in cross-sectional estimates in the longitudinal file due to wave imputation. The study on the PPD data shows that AI multiple imputation underestimates the between-imputation variance, although the underestimation is negligible in terms of the total variance, while the underestimation is much smaller for domains. If multiple imputation is to be correctly implemented in AI, it should have the ability to use ABB.

References

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.

Judkins, D., Krenzke, T., Piesse, A., Fan. Z., and Haung, W. (2007). Preservation of skip patterns and covariance structure through semi-parametric whole-questionnaire imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association* (pp. 3211-3218).

Kim, J.K., Brick, J.M., Fuller, W.A., and Kalton, G. (2006). On the bias of the multiple imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society, Series B*, 68, 509–521.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons.

Rubin, D., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.