Chapter 24
Mediation Analysis with Missing Data Through Multiple Imputation and Bootstrap

Zhiyong Zhang, Lijuan Wang, and Xin Tong

Abstract A method using multiple imputation and bootstrap for dealing with missing data in mediation analysis is introduced and implemented in both SAS and R. Through simulation studies, it is shown that the method performs well for both MCAR and MAR data without and with auxiliary variables. It is also shown that the method can work for MNAR data if auxiliary variables related to missingness are included. The application of the method is demonstrated through the analysis of a subset of data from the National Longitudinal Survey of Youth. Mediation analysis with missing data can be conducted using the provided SAS macros and R package bmem.

Keywords Mediation analysis • Missing data • Multiple imputation • Bootstrap

24.1 Introduction

Mediation models and mediation analysis are widely used in behavioral and social sciences as well as in health and medical research. Mediation models are very useful for theory development and testing as well as for identification of intervention points in applied work. Although mediation models were first developed in psychology (e.g., MacCorquodale and Meehl 1948; Woodworth 1928), they have been recognized and used in many disciplines where the mediation effect is also known as the indirect effect (Sociology, Alwin and Hauser 1975) and the surrogate or intermediate endpoint effect (Epidemiology, Freedman and Schatzkin 1992).

Figure 24.1 depicts the path diagram of a simple mediation model. In this figure, X, M, and Y represent the independent or input variable, the mediation variable (mediator), and the dependent or outcome variable, respectively. The e_M and e_Y are

Z. Zhang (✉) • L. Wang
University of Notre Dame, Notre Dame, IN 46556, USA
e-mail: zzhang4@nd.edu; lwang4@nd.edu

X. Tong
University of Virginia, Charlottesville, VA 22904, USA
e-mail: xtong@virginia.edu

L.A. van der Ark et al. (eds.), *Quantitative Psychology Research*, Springer Proceedings in Mathematics & Statistics 140, DOI 10.1007/978-3-319-19977-1_24

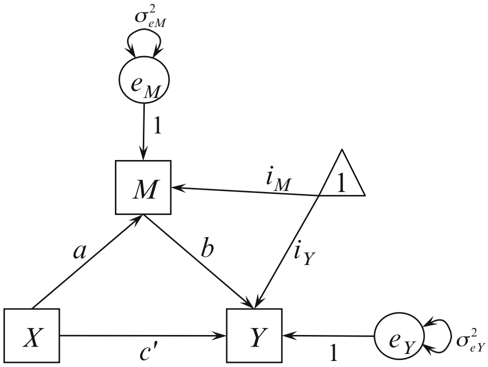

Fig. 24.1 Path diagram demonstration of a mediation model

residuals or disturbances with variances σ^2_{eM} and σ^2_{eY}. The coefficient c' is called the direct effect, and the mediation effect or indirect effect is measured by the product term $ab = a \times b$ as an indirect path from X to Y through M. The other parameters in this model include the intercepts i_M and i_Y.

Statistical approaches to estimating and testing mediation effects with complete data have been discussed extensively in the psychological literature (e.g., Baron and Kenny 1986; Bollen and Stine 1990; MacKinnon et al. 2007, 2002; Shrout and Bolger 2002). One way to test mediation effects is to test $H_0 : ab = 0$. If a large sample is available, the normal approximation method can be used, which constructs the standard error of $\widehat{ab}$ through the delta method so that $s.e.(\widehat{ab}) = \sqrt{\hat{b}^2\widehat{\sigma_a^2} + 2\hat{a}\hat{b}\hat{\sigma}_{ab} + \hat{a}^2\widehat{\sigma_b^2}}$ with the parameter estimates $\hat{a}$ and $\hat{b}$, their estimated variances $\widehat{\sigma_a^2}$ and $\widehat{\sigma_b^2}$, and covariance $\hat{\sigma}_{ab}$ (e.g., Sobel 1982, 1986). Many researchers suggested that the distribution of a mediation effect may not be normal especially when the sample size is small although with large sample sizes the distribution may still approach normality (Bollen and Stine 1990; MacKinnon et al. 2002). Thus, bootstrap methods have been recommended to obtain the empirical distribution and confidence interval of a mediation effect (MacKinnon et al. 2004; Mallinckrodt et al. 2006; Preacher and Hayes 2008; Shrout and Bolger 2002; Zhang and Wang 2008).

Missing data is continuously a challenge even for a well-designed study. Although there are approaches to dealing with missing data for path analysis in general (for a recent review, see Graham 2009), there are few studies focusing on the treatment of missing data in mediation analysis. Mediation analysis is different from typical path analysis because the focus is on the product of multiple path coefficients. A common practice is to analyze complete data through listwise deletion or pairwise deletion (e.g., Chen et al. 2005; Preacher and Hayes 2004). Recently, Zhang and Wang (2013b) discussed how to deal with missing data in mediation analysis through multiple imputation and full information maximum likelihood. However, the number of imputations needed to get reliable results remains unclear.

In this study, we discuss how to deal with missing data for mediation analysis through multiple imputation (MI) and bootstrap. We will first present some technical aspects of multiple imputation for mediation analysis with missing data. Then, we will present two simulation studies to evaluate the performance of MI for mediation analysis with missing data. In particular, we investigate the number of imputations needed for mediation analysis. Next, an empirical example will be used to demonstrate the application of the method. Finally, we discuss the limitations of the study and future directions. Instructions on how to use SAS and R to conduct mediation analysis through multiple imputation and bootstrap are provided online as supplemental materials.

24.2 Method

24.2.1 Complete Data Mediation Analysis

We focus our discussion on the simple mediation model to better illustrate the method although the approach works for the general mediation model. In its mathematical form, the mediation model displayed in Fig. 24.1 can be expressed using two equations,

$$\begin{aligned} M &= i_M + aX + e_M \\ Y &= i_Y + bM + c'X + e_Y, \end{aligned} \tag{24.1}$$

which can be viewed as a collection of two linear regression models. To obtain the parameter estimates in the model, the maximum likelihood estimation method for structural equation modeling (SEM) can be used. Specifically for the simple mediation model, the mediation effect estimate is $\widehat{ab} = \hat{a}\hat{b}$ with

$$\begin{aligned} \hat{a} &= s_{XM}/s_X^2 \\ \hat{b} &= (s_{MY}s_X^2 - s_{XM}s_{XY})/(s_X^2 s_M^2 - s_{XM}^2) \end{aligned} \tag{24.2}$$

where $s_X^2, s_M^2, s_Y^2, s_{XM}, s_{MY}, s_{XY}$ are sample variances and covariances of X, M, Y, respectively.

24.2.2 Missingness Mechanisms

Little and Rubin (1987, 2002) have distinguished three types of missing data—missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Let $D = (X, M, Y)$ denote all data that can be potentially observed in a mediation model. D_{obs} and D_{miss} denote data that are actually observed and data that are not observed, respectively. Let R denote an indicator matrix of zeros and ones with the same dimension as D. If a datum in D is missing, the corresponding element in R is equal to 1. Otherwise, it is equal to 0. Finally, let A denote the auxiliary variables that are related to the missingness of D but not a component of the mediation model.

If the missing mechanism is MCAR, then we have

$$\Pr(R|D_{obs}, D_{miss}, \boldsymbol{\theta}) = \Pr(R|\boldsymbol{\theta}),$$

where the vector $\boldsymbol{\theta}$ represents all model parameters in the mediation model. This suggests that missing data D_{miss} are a simple random sample of D and not related to the data observed or auxiliary variables A. If the missing mechanism is MAR, then

$$\Pr(R|D_{obs}, D_{miss}, \boldsymbol{\theta}) = \Pr(R|D_{obs}, \boldsymbol{\theta}),$$

which indicates that the probability that a datum is missing is related to the observed data D_{obs} but not to the missing data D_{miss}.

Finally, if the probability that a datum is missing is related to the missing data D_{miss} or auxiliary variables A but A are not considered in the data analysis, the missing mechanism is MNAR. In particular, we want to emphasize the MNAR mechanism with auxiliary variables where

$$\Pr(R|D_{obs}, D_{miss}, \boldsymbol{\theta}) = \Pr(R|D_{obs}, A, \boldsymbol{\theta}).$$

Note that although the missingness is MNAR if only D is modeled, the overall missingness becomes to MAR if D and A are jointly modeled. Therefore, one way to deal with MNAR is to identify and include the auxiliary variables that are related to missingness.

24.2.3 Multiple Imputation for Mediation Analysis with Missing Data

Most techniques dealing with missing data, including multiple imputation, in general require missing data to be either MCAR or MAR (see also, e.g., Little and Rubin 2002; Schafer 1997). For MNAR, the missing mechanism has to be known to correctly recover model parameters (e.g., Lu et al. 2011; Zhang and Wang 2012). Practically, researchers have suggested including auxiliary variables to facilitate MNAR missing data analysis (Graham 2003; Savalei and Bentler 2009). After including appropriate auxiliary variables, we may be able to assume that data from both model variables and auxiliary variables are MAR.

Assume that a set of $p (p \geq 0)$ auxiliary variables $A_1, A_2, \ldots, A_p$ are available. These auxiliary variables may or may not be related to missingness of the mediation model variables. By augmenting the auxiliary variables with the mediation model variables, we have $D = (X, M, Y, A_1, \ldots, A_p)$, e.g., for the simple mediation model. To proceed, we assume that the missing mechanism is MAR after including the auxiliary variables.

Multiple imputation (Little and Rubin 2002; Rubin 1976; Schafer 1997) is a procedure to fill each missing value with a set of plausible values. The multiple imputed data sets are then analyzed using standard procedures for complete data and the results from these analyses are combined for obtaining point estimates of model parameters and their standard errors. For mediation analysis with missing data, the following steps can be implemented for obtaining point estimates of mediation model parameters:

1. Assuming that $D = (X, M, Y, A_1, \ldots, A_p)$ are from a multivariate normal distribution, generate K sets of values for each missing value. Combine the generated values with the observed data to produce K sets of complete data (Schafer 1997).
2. For each of the K sets of complete data, apply the formula in Eq. (24.2) or use the SEM method to obtain a point mediation effect estimate $\widehat{ab}_k = \hat{a}_k\hat{b}_k$ $(j = 1, \ldots, K)$.
3. The point estimate for the mediation effect through multiple imputation is the average of the K complete data mediation effect estimates:

$$\widehat{ab} = \hat{a}\hat{b} = \frac{1}{K}\sum_{k=1}^{K} \hat{a}_k\hat{b}_k.$$

Parameter estimates for other model parameters can be obtained in the same way.

24.2.4 Testing Mediation Effects Through the Bootstrap Method

The procedure described above is implemented to obtain point estimates of mediation effects. The bootstrap method has been used to test the significance of the mediation effects (e.g., Bollen and Stine 1990). This method has no distribution assumption on the indirect effect. Instead, it approximates the distribution of the indirect effect using its bootstrap empirical distribution. The bootstrap method can be applied along with multiple imputation to obtain standard errors of mediation effect estimates and confidence intervals for mediation analysis with missing data. Specifically, the following procedure can be used.

1. Using the *original data set* (sample size = N) as a population, draw a bootstrap sample of N persons randomly with replacement from the original data set. This bootstrap sample generally would contain missing data.

2. With the bootstrap sample, implement the K multiple imputation procedure described in the above section to obtain point estimates of model parameters and a point estimate of the mediation effect.
3. Repeat Steps 1 and 2 for a total of B times. B is called the number of bootstrap samples.
4. Empirical distributions of model parameters and the mediation effect are then obtained using the B sets of bootstrap point estimates. Thus, confidence intervals of model parameters and the mediation effect can be constructed.

The procedure described above can be considered as a procedure of K multiple imputations nested within B bootstrap samples. Using the B bootstrap sample point estimates, one can obtain bootstrap standard errors and confidence intervals of model parameters and mediation effects conveniently. Let $\boldsymbol{\theta}$ denote a vector of model parameters and the mediation effects. With data from each bootstrap, we can obtain $\hat{\boldsymbol{\theta}}^b$, $b = 1, \ldots, B$. The standard error estimate of the pth parameter $\hat{\theta}_p$ can be calculated as

$$\widehat{s.e.}(\hat{\theta}_p) = \sqrt{\sum_{b=1}^{B} (\hat{\theta}_p^b - \bar{\hat{\theta}}_p^b)^2 / (B-1)}$$

with $\bar{\hat{\theta}}_p^b = \sum_{b=1}^{B} \hat{\theta}_p^b / B$.

Many methods for constructing confidence intervals from $\hat{\boldsymbol{\theta}}^b$ have been proposed such as the percentile interval, the bias-corrected (BC) interval, and the bias-corrected and accelerated (BCa) interval (Efron 1987; MacKinnon et al. 2004). In the present study, we focus on the BC interval because MacKinnon et al. (2004) showed that, in general, the BC confidence intervals have performed better in terms of Type I error and statistical power among many different confidence intervals. The $1 - 2\alpha$ BC interval for the pth element of $\boldsymbol{\theta}$ can be constructed using the percentiles $\hat{\theta}_p^b(\tilde{\alpha}_l)$ and $\hat{\theta}_p^b(\tilde{\alpha}_u)$ of $\hat{\theta}_p^b$ with $\tilde{\alpha}_l = \Phi(2z_0 + z^{(\alpha)})$ and $\tilde{\alpha}_u = \Phi(2z_0 + z^{(1-\alpha)})$ where Φ is the standard cumulative normal distribution function and $z^{(\alpha)}$ is the α percentile of the standard normal distribution and

$$z_0 = \Phi^{-1}\left[\frac{\text{number of times that } \hat{\theta}_p^b < \hat{\theta}_p}{B}\right].$$

24.3 Simulation Studies

In this section, we conduct two simulation studies to evaluate the performance of the proposed method for mediation analysis with missing data. We first evaluate its performance under different missing data mechanisms including MCAR, MAR, and MNAR without and with auxiliary variables. Then, we investigate how many imputations are needed for different proportions of missing data.

24.3.1 Simulation Study 1: Estimate of Mediation Effects Under MCAR, MAR, and MNAR Data

24.3.1.1 Simulation Design

For mediation analysis with complete data, simulation studies have been conducted to investigate a variety of features of mediation models (e.g., MacKinnon et al. 2002, 2004). For the current study, we follow the parameter setup from the previous literature and set the population parameter values to be $a = b = .39$, $c' = 0$, $i_M = i_Y = 0$, and $\sigma^2_{eM} = \sigma^2_{eY} = \sigma^2_{eX} = 1$. Furthermore, we fix the sample size at $N = 100$ and consider three proportions of missingness with missing data percentages at 10, 20, and 40 %, respectively. To facilitate the comparisons among different missing mechanisms, missing data are only allowed in M and Y although our software programs also allow missingness in X. Two auxiliary variables (A_1 and A_2) are also generated where the correlation between A_1 and M and the correlation between A_2 and Y are both 0.5.

Missing data are generated in the following way. First, 1000 sets of complete data are generated. Second, for MCAR, each data value has the same probability of missing for M and Y. Third, for the MAR data, the probability of missingness in Y and M depends only on X. Specifically, if X is smaller than a given percentile, M is missing and if X is larger than a given percentile, Y is missing. Finally, to generate MNAR data, we assume that missingness of M depends on A_1 and missingness of Y depends on A_2. If A_1 is smaller than a given percentile, M is missing, and if A_2 is smaller than a percentile, Y is missing. Clearly, if auxiliary variables A_1 and A_2 are included in an analysis, the missing mechanism becomes MAR. However, if the auxiliary are not considered, the missing mechanism is MNAR.

The generated data are analyzed using the R package bmem. To estimate the mediation effects, the sample covariance matrix of the imputed data is first estimated. Then, the mediation model is fitted to the estimated covariance matrix to obtain the model parameters through the SEM maximum likelihood estimation method (Bollen 1989). Finally, the mediation effects are calculated as the product of the corresponding direct effects.

24.3.1.2 Results

The parameter estimate bias, coverage probability, and power for MCAR, MAR, and MNAR data with 10, 20, and 40 % missing data were obtained without and with auxiliary variables and are summarized in Table 24.1. Based on the results, we can conclude the following. First, the bias of the parameter estimates under the

Table 24.1 Bias, coverage probability, and power under MCAR

	Without auxiliary variables			With auxiliary variables		
	Bias	Coverage	Power	Bias	Coverage	Power
MCAR						
10 %	0.219	0.967	0.900	0.263	0.967	0.920
20 %	−1.222	0.966	0.808	−0.593	0.963	0.845
40 %	−0.716	0.946	0.531	0.112	0.950	0.615
MAR						
10 %	−0.119	0.957	0.870	−0.403	0.961	0.893
20 %	−0.546	0.962	0.767	−1.940	0.958	0.791
40 %	−2.932	0.960	0.511	−1.747	0.955	0.599
MNAR						
10 %	−32.633	0.831	0.800	−0.513	0.951	0.925
20 %	−49.117	0.673	0.570	−2.583	0.941	0.815
40 %	−66.815	0.559	0.305	−2.951	0.951	0.642

Note: We have also investigated other conditions where the sample size, population parameters, and correlation between auxiliary variables and model variables are different. We observed the same patterns in the results.

studied MCAR conditions was small enough to be ignored. Second, the coverage probability was close to the nominal level 0.95. Third, the inclusion of auxiliary variables in MCAR did not seem to influence the accuracy of parameter estimates and coverage probability. The use of auxiliary variables, however, boosted the power of detecting the mediation effect especially when the missing proportion was large (e.g., 40 %). The findings from MAR data are similar to those from MCAR data and thus are not repeated here. However, the power of detecting mediation effects from MAR data were smaller than that from MCAR data given the same proportion of missing data.

The results for MNAR data clearly showed that when auxiliary variables were not included, parameter estimates were highly underestimated especially when the missing data proportion was large, e.g., about 67 % bias with 40 % missing data for the mediation effect. Correspondingly, coverage probability was highly underestimated, too. For example, with 40 % of missing data, the coverage probability was only about 56 %. However, with the inclusion of auxiliary variables, the parameter estimate bias dramatically decreased to less than 3 % and the coverage probabilities were close to 95 %. Thus, multiple imputation can be used to analyze MNAR data and recover true parameter values by including auxiliary variables that can explain missingness of the variables in the mediation model. This is because the inclusion of the auxiliary variables converts the missingness mechanism to MAR. However, this does not mean that the inclusion of auxiliary variables can always address the non-ignorable problems and more discussion on this can be found in Zhang and Wang (2013b)

24.3.2 Simulation Study 2: Impact of the Number of Imputations

One difficulty in applying multiple imputation is to decide on how many imputations are sufficient. For example, Rubin (1987) has suggested that five imputations are sufficient in the case of 50 % missing data for estimating the simple mean. But Graham et al. (2007) recommend that many more imputations than what Rubin recommended should be used. Although one may always choose to use a very large number of imputations for mediation analysis with missing data, this may not be practically possible because of the amount of computational time involved.

In this simulation study, we investigate the impact of the number of imputations on point estimates and standard error estimates of mediation effects with different proportions of missing data. The same model in the first simulation study is used. The data are generated in the following way. First, two groups of 100 sets of complete data with two auxiliary variables are generated so that the correlation between the auxiliary variables and the mediation model variables is ρ =0.1 for the first group and $\rho = 0.4$ for the second group, respectively. Second, 10 and 40 % of missing data are generated, respectively, for each group of the 100 sets of complete data, where the missingness is related to the auxiliary variables as in the first simulation study. Therefore, in total, we have 4 groups of 100 sets of missing data.

For each data set, we obtain the results from the data analysis including auxiliary variables with the number of imputations ranging from 10 to 100 at intervals of 10. Note that the overall missingness is MAR. After that, we calculate the average mediation effect and standard error estimates from the 100 sets of data. For better comparison, we calculate the relative deviance of mediation effect estimates and their standard error estimates from those estimates with 100 imputations. The relative deviance from the simulation is plotted in Fig. 24.2. Since the results were based on 100 replications of simulation, the absolute difference was small as a result. Therefore, we rescaled the relative deviance by 1000 times to focus on the relative change of the deviance corresponding to the number of imputations. We have found that the analysis of individual data sets showed the similar pattern but with much larger deviance.

Comparing the results with 10 % missing data in Fig. 24.2a, c and the results with 40 % missing data in Fig. 24.2b, d, it is clear that there are more fluctuations in both mediation effect estimates and their standard errors with more missing data regardless of $\rho = 0.1$ or 0.5. Therefore, a greater number of imputations is needed with more missing data. More specifically, with 10 % missing data, the parameter estimates, especially the standard estimates, seem to become stable with more than 40 or 50 imputations. With 40 % missing data, however, the relative deviance of point estimates and standard error estimates does not appear stabilized until with more than 80 imputations. In our simulation study, the choice of 100 imputations appears to be adequate based on this simulation. The conclusion on the specific number of imputations here only applies to the simple mediation model. For more complex mediation analysis, more imputations might be needed.

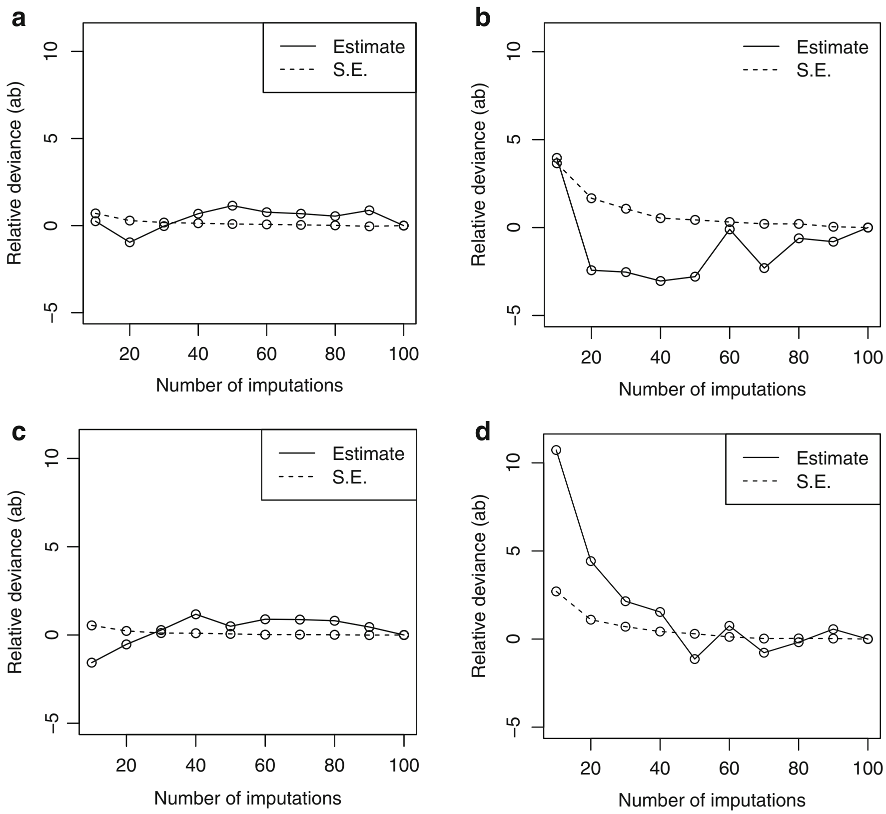

Fig. 24.2 The impact of different numbers of imputations on the accuracy of point estimates and bootstrap standard error estimates. (**a**) $\rho = 0.1$, 10 % missing data, (**b**) $\rho = 0.1$, 40 % missing data, (**c**) $\rho = 0.5$, 10 % missing data, (**d**) $\rho = 0.5$, 40 % missing data

24.4 An Empirical Example

In this section, we apply the proposed method to a real study to illustrate its application. Research has found that parental education levels can influence adolescent mathematical achievement directly and indirectly. For example, Davis-Kean (2005) showed that parental education levels are related to child academic achievement through parental beliefs and behaviors. To test a similar hypothesis, we investigate whether home environment is a mediator in the relation between mother's education and child mathematical achievement.

Data used in this example are from the National Longitudinal Survey of Youth, the 1979 cohort (NLSY79, Center for Human Resource Research 2006). Data were collected in 1986 from $N = 475$ families on mother's education level (ME), home environment (HE), child mathematical achievement (Math), child behavior problem index (BPI), and child reading recognition and reading comprehension achievement.

Table 24.2 Missing data patterns of the empirical data set

Pattern	ME	HE	Math	Sample size
1	O	O	O	417
2	O	X	O	36
3	O	O	X	14
4	O	X	X	8
Total				475

Note: *O* observed, *X* missing, *ME* mother's education level, *HE* home environments, *Math* mathematical achievement

Table 24.3 Mediation effect of home environment on the relationship between mother's education and child mathematical achievement

	Without auxiliary variable				With auxiliary variable			
Parameter	Estimate	S.E.	95 % BC		Estimate	S.E.	95 % BC	
a	0.035	0.049	0.018	0.162	0.036	0.049	0.018	0.163
b	0.475	0.126	0.252	0.754	0.458	0.125	0.221	0.711
c'	0.134	0.191	0.071	0.611	0.134	0.188	0.072	0.609
ab	0.017	0.021	0.005	0.071	0.016	0.021	0.005	0.067
i_Y	7.953	2.047	3.530	9.825	8.045	2.025	3.778	10.006
i_M	5.330	0.556	3.949	5.641	5.327	0.558	3.945	5.646
σ^2_{eY}	4.532	0.269	4.093	5.211	4.520	0.268	4.075	5.141
σ^2_{eM}	1.660	0.061	1.545	1.789	1.660	0.061	1.542	1.790

Note: The results are based on 1000 bootstrap samples and 100 imputations
S.E. bootstrap standard error, *BC* bias-corrected confidence interval

For the mediation analysis, mother's education is the independent variable, home environment is the mediator, and child mathematical achievement is the outcome variable. The missing data patterns and the sample size of each pattern are presented in Table 24.2. In this data set, 417 families have complete data and 58 families have missing data on at least one of the two model variables: home environment and child mathematical achievement. In this study, BPI and child reading recognition and reading comprehension achievement are used as auxiliary variables because they have been found to be related to mathematical achievement in the literature (e.g., Grimm 2008; Wu et al. 2014). In addition, it is reasonable to believe that it is more difficult to collect data from children with behavior problems and children with reading problems can have a harder time to complete tests on mathematics, which, therefore, could lead to more missing data.

In Table 24.3, the results from the empirical data analysis using the proposed method without and with the auxiliary variables are presented. The results reveal that the inclusion of the auxiliary variables only slightly changed the parameter estimates, standard errors, and the BC confidence intervals. This indicates that the auxiliary variables may not be related to the missingness in the mediation model variables. The results also show that home environment partially mediates the relationship between mother's education and child mathematical achievement because both the indirect effect ab and the direct effect c' are significant.

24.5 Software

We have developed both SAS macros and an R package bmem (Zhang and Wang 2013a) for mediation analysis with missing data through multiple imputation and bootstrap. Instructions on how to use the SAS macros and the R package can be found on our website at http://psychstat.org/imps2014. The SAS macros are designed for the simple mediation model and have computational advantages that make them run faster than bmem. The R package bmem, however, can handle more complex mediation analysis with multiple mediators and latent variables. Both programs can utilize auxiliary variables to potentially handle MNAR data.

24.6 Discussion

In this study, we discussed how to conduct mediation analysis with missing data through multiple imputation and bootstrap. Through simulation studies, we demonstrated that the proposed method performed well for both MCAR and MAR without and with auxiliary variables. It is also shown that multiple imputation worked equally well for MNAR if auxiliary variables related to missingness were included, because the overall missingness becomes essentially MAR. The analysis the NLSY79 data revealed that home environment partially mediated the relationship between mother's education and child mathematical achievement.

24.6.1 Strength of the Proposed Method

The multiple imputation and bootstrap method for mediation analysis with missing data has several advantages. First, the idea of imputation and bootstrap is easy to understand. Second, multiple imputation has been widely implemented in both free and commercial software and thus can be extended to mediation analysis relatively easily. Third, it is natural and easy to include auxiliary variables in multiple imputation. Fourth, unlike the full information maximum likelihood method, multiple imputation does not assume a specific model when imputing data.

24.6.2 Assumptions and Limitations

There are several assumptions and limitations of the current study. First, the current SAS program is based on a simple mediation model. In the future, the program should be expanded to allow more complex mediation analysis. Second, in applying multiple imputation, we have assumed that all variables are multivariate normally distributed. However, it is possible that one or more variables are not normally distributed. Third, the current mediation model only focuses on the cross-sectional data analysis. Some researchers have suggested that the time variable should be considered in mediation analysis (e.g., Cole and Maxwell 2003; MacKinnon 2008; Wang et al. 2009). Fourth, in dealing with MNAR data, we assume that useful auxiliary variables that can explain missingness in the mediation model variables are available. Therefore, the true missingness mechanism is actually MAR. However, sometimes the auxiliary variables may not be available. Other methods for dealing with MNAR data can be investigated in the future.

In summary, a method using multiple imputation and bootstrap for mediation analysis with missing data is introduced. Simulation results show that the method works well in dealing with missing data for mediation analysis under different missing mechanisms. Both SAS macros and an R package are provided to conduct mediation analysis with missing data, which is expected to promote the use of advanced techniques in dealing with missing data for mediation analysis in the future.

References

Alwin, D. F., & Hauser, R. M. (1975). The decomposition of effects in path analysis. *American Sociological Review, 40*, 37–47.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bollen, K. A., & Stine, R. A. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology, 20*, 115–140.

Center for Human Resource Research. (2006). *NLSY79 CHILD & YOUNG ADULT DATA USERS GUIDE: A Guide to the 1986–2004 Child Data*. Columbus, OH: The Ohio State University

Chen, Z. X., Aryee, S., & Lee, C. (2005). Test of a mediation model of perceived organizational support. *Journal of Vocational Behavior, 66*(3), 457–470.

Cole, D. A., & Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology, 112*, 558–577.

Davis-Kean, P. E. (2005). The influence of parent education and family income on child achievement: The indirect role of parental expectations and the home environment. *Journal of Family Psychology, 19*, 294–304.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171–185.

Freedman, L. S., & Schatzkin, A. (1992). Sample size for studying intermediate endpoints within intervention trails or observational studies. *American Journal of Epidemiology, 136*, 1148–1159.
Graham, J. W. (2003). Adding missing-data-relevant variables to fiml-based structural equation models. *Structural Equation Modeling, 10*, 80–100.
Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology, 60*, 549–576.
Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science, 8*, 206–213.
Grimm, K. J. (2008). Longitudinal associations between reading and mathematics. *Developmental Neuropsychology, 33*, 410–426.
Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York, NY: Wiley.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley-Interscience.
Lu, Z., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with non-ignorable missing data. *Mulitvariate Behavioral Research, 46*, 567–597.
MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review, 55*(2), 95–107.
MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. New York, NY: Taylor & Francis.
MacKinnon, D. P., Fairchild, A. J., & Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology, 58*, 593–614.
MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83–104.
MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research, 39*(1), 99–128.
Mallinckrodt, B., Abraham, T. W., Wei, M., & Russell, D. W. (2006). Advance in testing statistical significance of mediation effects. *Journal of Counseling Psychology, 53*(3), 372–378.
Preacher, K. J., & Hayes, A. F. (2004). Spss and sas procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*, 717–731.
Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavioral Research Methods, 40*, 879–891.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581–592.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Savalei, V., & Bentler, P. M. (2009). A two-stage approach to missing data: Theory and application to auxiliary variables. *Structural Equation Modeling, 16*, 477–497.
Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC.
Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods, 7*, 422–445.
Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 290–312). San Francisco: Jossey-Bass.
Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. In N. Tuma (Ed.), *Sociological methodology* (pp. 159–186). Washington, DC: American Sociological Association.

Wang, L., Zhang, Z., & Estabrook, R. (2009). Longitudinal mediation analysis of training intervention effects. In S. M. Chow, E. Ferrer, & F. Hsieh (Eds.), *Statistical methods for modeling human dynamics: An interdisciplinary dialogue* (pp. 349–380). New Jersey: Lawrence Erlbaum Associates.

Woodworth, R. S. (1928). Dynamic psychology. In C. Murchison (Ed.), *Psychologies of 1925* (pp. 111–126). Worcester, MA: Clark Universal Academy Press, Inc.

Wu, S. S., Willcutt, E., Escovar, E., & Menon, V. (2014). Mathematics achievement, anxiety and their relation to internalizing and externalizing behaviors. *Journal of Learning Disorders, 47*(6), 503–514.

Zhang, Z., & Wang, L. (2008). Methods for evaluating mediation effects: Rationale and comparison. In K. Shigemasu, A. Okada, T. Imaizumi, & T. Hoshino (Eds.), *New trends in psychometrics* (pp. 595–604). Tokyo: Universal Academy Press, Inc.

Zhang, Z., & Wang, L. (2012). A note on the robustness of a full Bayesian method for non-ignorable missing data analysis. *Brazilian Journal of Probability and Statistics, 26*(3), 244–264.

Zhang, Z., & Wang, L. (2013a). *bmem: Mediation analysis with missing data using bootstrap. R package version 1.5.* https://cran.r-project.org/web/packages/bmem/index.html.

Zhang, Z., & Wang, L. (2013b). Methods for mediation analysis with missing data. *Psychometrika, 78*, 154–184.