# Mediation analysis with missing data through multiple imputation and bootstrap

# Lijuan Wang, Zhiyong Zhang, and Xin Tong University of Notre Dame

#### Abstract

A method using multiple imputation and bootstrap for dealing with missing data in mediation analysis is introduced and implemented in SAS. 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 works equally well 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.

*Keywords:* Mediation analysis, missing data, multiple imputation, auxiliary variables, bootstrap, SAS

#### Introduction

Mediation models and mediation analysis are widely used in behavioral and social sciences as well as in health and medical research. The influential article on mediation analysis by Baron and Kenny (1986) has been cited more than 8,000 times. 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 1 (after Shrout and Bolger, 2002) 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 residuals or disturbances with variances  $\sigma_{eM}^2$  and  $\sigma_{eY}^2$ . c' is called the direct

effect and the mediation effect or indirect effect is measured by the product term ab. The other parameters in this model include the intercepts  $i_M$  and  $i_Y$ .

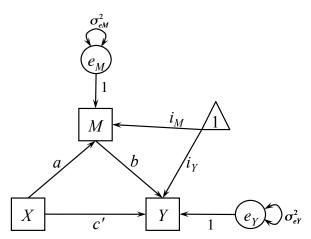


Figure 1. Path diagram demonstration of a mediation model.

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., 2002, 2007; 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 ab through the delta method so that  $s.e.(ab) = \sqrt{\hat{b}^2 \hat{\sigma}_a^2 + 2\hat{a}\hat{b}\hat{\sigma}_{ab} + \hat{a}^2\hat{\sigma}_b^2}$  with parameter estimates  $\hat{a}$  and  $\hat{b}$ , their estimated variances  $\hat{\sigma}_a^2$  and  $\hat{\sigma}_b^2$ , and covariance  $\hat{\sigma}_{ab}$  (e.g., Sobel, 1982, 1986). Many researchers suggested that the distribution of ab may not be normal especially when the sample size is small although with large sample sizes the distribution may 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 ab (MacKinnon et al., 2004; Mallinckrodt et al., 2006; Preacher and Hayes, 2008; Shrout and Bolger, 2002; Zhang and Wang, 2008).

Mediation analysis can be conducted in a variety of programs and software. Notably, the SAS and SPSS macros by Preacher and Hayes (2004, 2008) have popularized the application of bootstrap techniques in mediation analysis. Based on search results from Google scholar, Preacher and Hayes (2004) has been cited more than 900 times and Preacher and Hayes (2008) has already been cited more than 400 times in less than two years after publication.

Missing data problem 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. Particularly, mediation analysis is different from typical path analysis because the focus is on the product of two 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). However, with the availability of advanced approaches such as multiple imputation (MI), listwise and pairwise deletion is no longer deemed acceptable (Little and Rubin, 2002; Savalei and Bentler, 2009; Schafer, 1997).

In this study, we discuss how to deal with missing data for mediation analysis through multiple imputation (MI) and bootstrap using SAS. The rationale of using multiple imputation is that it can be implemented in existing popular statistical software such as SAS and it can deal with different types of missing data. In the following, we will first present the technical backgrounds of multiple imputation for mediation analysis with missing data. Then, we will discuss how to implement the method in SAS. After that, we will present several simulation examples to evaluate the performance of MI for mediation analysis with missing data. Finally, an empirical example will be used to demonstrate the application of the method.

#### Method

In this section, we present the technical backgrounds of mediation analysis with missing data through multiple imputation and bootstrap. First, we will discuss how to estimate mediation model parameters with complete data. Second, we will reiterate the definition of missing data mechanisms by Little and Rubin (2002). Third, we will discuss how to apply multiple imputation to mediation analysis. Finally, we will discuss the bootstrap procedure to obtain the bias corrected confident intervals for mediation model parameters.

#### Complete data mediation analysis

In mathematical form, the mediation model displayed in Figure 1 can be expressed using two equations,

$$M = i_M + aX + e_M$$
  

$$Y = i_Y + bM + c'X + e_Y,$$
(1)

which can be viewed as a collection of two linear regression models. To obtain the parameter estimates in the model, one can maximize the product of the likelihood functions from the two regression models using the maximum likelihood method. Because  $e_M$  and  $e_Y$  are assumed to be independent, maximizing the product of the likelihood functions is equivalent to maximizing the likelihood function of each regression model separately. Thus, parameter estimates can be obtained by fitting two separate regression models in Equation 1. Specifically, the mediation effect estimate is  $\hat{a}\hat{b}$  with

$$\hat{a} = s_{XM}/s_X^2$$

$$\hat{b} = (s_{MY}s_X^2 - s_{XM}s_{XY})/(s_X^2s_M^2 - s_{XM}^2)$$
(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.

#### Missing 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. If a datum in D is missing, the corresponding element in R is equal to 1. Otherwise, it is equal to 0. Finally, let R denote the auxiliary variables that are related to the missingness of R but not a component of the mediation model in Equation 1.

If the missing mechanism is MCAR, then we have

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

where the vector  $\boldsymbol{\theta}$  represents all model parameters in the mediation model including  $a, b, ab, c', i_M, i_Y, \sigma^2_{eM}$ , and  $\sigma^2_{eY}$ . This suggests that missing data  $D_{miss}$  are a simple random sample of D and missingness is not related to the data of interest D or auxiliary variables A.

If the missing mechanism is MAR, then

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

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 while A are not considered in the data analysis, the missing mechanism is MNAR.

#### 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. Practically, researchers have suggested including auxiliary variables to facilitate MNAR missing data analysis (Graham, 2003; Savalei and Bentler, 2009). Auxiliary variables are variables that are not a component of a model (not model variables) but can explain missingness of variables in the model. After including appropriate auxiliary variables, we may be able to assume that data from both model variables and auxiliary variables are MAR.

The setting for mediation analysis with missing data is described below. Assume that a set of  $p(p \ge 0)$  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. Furthermore, there may or may not be missing data in auxiliary variables. By augmenting the auxiliary variables with the mediation model variables, we have a total of p+3 variables denoted by  $D=(X,M,Y,A_1,\ldots,A_p)$ . To proceed, we assume that the missing mechanism is MAR after including the auxiliary variables. That is

$$Pr(R|D_{obs}, D_{miss}, A_1, \dots, A_p, \boldsymbol{\theta}) = Pr(R|D_{obs}, A_1, \dots, A_p, \boldsymbol{\theta}).$$

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 standard errors of parameter estimates. 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 (K is the number of multiple imputations) 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 Equation 2 to obtain a point mediation effect estimate  $\hat{a}_k \hat{b}_k (j=1,\ldots,K)$ .
- 3. The point estimate for the mediation effect through multiple imputation is the average of the K complete data mediation effect estimates:

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

Parameter estimates for the other model parameters a, b, c',  $i_M$ ,  $i_Y$ ,  $\sigma_{eM}^2$ , and  $\sigma_{eY}^2$  can be obtained in the same way.

#### Testing mediation effects through the bootstrap method

The procedure described above is implemented to obtain point estimates of mediation effects. To test mediation effects, we need to obtain standard errors of the parameter estimates. Because mediation effects are measured by ab, researchers suggest using bootstrap to obtain empirical standard errors as mentioned in a previous section. The bootstrap method (Efron, 1979, 1987) was first employed in mediation analysis by Bollen and Stine (1990) and has been studied in a variety of research contexts (e.g., MacKinnon et al., 2004; Mallinckrodt et al., 2006; Preacher and Hayes, 2008; Shrout and Bolger, 2002). This method has no distribution assumption on the indirect effect ab. Instead, it approximates the distribution of ab using its bootstrap empirical distribution.

The bootstrap method used in Bollen and Stine (1990) 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 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  $\boldsymbol{\theta} = (iM, iY, a, b, c', \sigma_{eM}^2, \sigma_{eY}^2, ab)^t$  denote a vector of model parameters and the mediation effect ab. With data from each bootstrap, we can obtain  $\hat{\boldsymbol{\theta}}^b$ ,  $b=1,\ldots,B$ . The standard error 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  $\hat{\theta}^b$  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 the BC confidence intervals have correct Type I error and largest power among many different evaluated confidence intervals.

The  $1-2\alpha$  BC interval for the pth element of  $\theta$  can be constructed using the percentiles  $\hat{\theta}_p^b(\tilde{\alpha}_l)$  and  $\hat{\theta}_p^b(\tilde{\alpha}_u)$  of  $\hat{\theta}_p^b$ . Here

$$\tilde{\alpha}_l = \Phi(2z_0 + z^{(\alpha)})$$

and

$$\tilde{\alpha}_u = \Phi(2z_0 + z^{(1-\alpha)})$$

where  $\Phi$  is the standard cumulative normal distribution function and  $z^{(\alpha)}$  is the  $\alpha$  percentile of the standard normal distribution and

$$z_0 = \Phi^{-1} \left[ rac{ ext{number of times that } \hat{ heta}_p^b < \hat{ heta}_p}{B} 
ight].$$

# Multiple imputation and bootstrap for mediation analysis with missing data in SAS

To facilitate the implementation of the method described in the above section, we have written a SAS program for mediation analysis with missing data using multiple imputation and bootstrap. The complete SAS program scripts are contained in the Appendix. Now we briefly explain the functioning of each part of the SAS program.

Lines 3-9 of the SAS program specifies all global parameters that control multiple imputation and bootstrap for mediation analysis. This part is the one that a user needs to modify according to his/her data analysis environment. Line 3 specifies the directory and name of the data file to be used. Line 4 lists the names of the variables in the data file. Line 5 specifies the missing data value indicator. For example, 99999 in the data file represents a missing datum. Line 6 specifies the number of imputations (K) for imputing missing data. Line 7 defines the number of bootstrap samples (B). A number larger than 1000 is usually recommended. Line 8 and Line 9 specify the confidence level and the random number generator seed, respectively.

Lines 15-22 first read data into SAS from the data file specified on line 3 and then change missing data to the SAS missing data format - a dot. Lines 26-28 impute missing data for the original data set with auxiliary variables and generate K imputed data sets. Lines 30-34 estimate the mediation model parameters for each imputed data set. Lines 37-74 collect the results from the multiple imputed data sets and save the point estimates of model parameters and mediation effect in a SAS data set called "pointest". The SAS codes in this section produce point parameter estimates for the model parameters and the mediation effect based on the original data after multiple imputation.

Lines 77-88 generate B bootstrap samples from the original data set with the same sample size. Lines 91-95 impute each bootstrap sample independently for K times. Lines 98-143 produce point estimates of mediation model parameters and mediation effect for each bootstrap sample and collect the point estimates for all bootstrap samples in the SAS data set named "bootest".

The last part of the SAS program from Line 146 to Line 195 calculates the bootstrap standard errors and the bias-corrected confidence intervals for mediation model parameters and mediation effect. It also generates a table containing the point estimates, standard errors, and confidence intervals in the SAS output window.

To use the SAS program, one only needs to first change the global parameters in Lines 3-9, usually only lines 3 and 4, and then run the whole SAS program from the beginning to the end.

#### Evaluating the method for mediation analysis with missing data

In this section, we conduct several 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 mediation analysis with different proportions of missing data. In the following, we first discuss our simulation design and then present the simulation results.

#### 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 previous literature and set the model parameter values to be a=b=.39, c'=0,  $i_M=i_Y=0$ , and  $\sigma_{eM}^2=\sigma_{eY}^2=\sigma_{eX}^2=1$ . 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 SAS program allows 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. For each of the following simulation studies, results are from R=1,000 sets of simulated data.

For each simulation study, we report point estimate bias, coverage probability, and power or Type I error for evaluations. Let  $\theta$  denote the true parameter value in the simulation and  $\hat{\theta}_r(r=1,\ldots,1000)$  denote the corresponding estimate from the rth replication. The bias is calculated as

$$\text{Bias} = \begin{cases} 100 \times \left[ \frac{\sum_{r=1}^{1000} \hat{\theta}_r}{1000\theta} - 1 \right] & \theta \neq 0 \\ 100 \times \left[ \frac{\sum_{r=1}^{1000} \hat{\theta}_r}{1000} - \theta \right] & \theta = 0 \end{cases}.$$

Note that the bias is rescaled by multiplying 100. Smaller bias indicates the point estimate is less biased. Furthermore, Let  $\hat{l_r}$  and  $\hat{u_r}$  denote the lower and upper limits of the 95% confidence interval in the rth replication. The coverage probability is calculated by

$$coverage = \frac{\#(\hat{l_r} < \gamma < \hat{u}_r)}{1000}$$

where  $\#(\hat{l_r} < \gamma < \hat{u}_r)$  is the total number of replications with confidence intervals covering the true parameter value. Good 95% confidence intervals should give coverage probabilities close to 0.95. Power or Type I error is calculated by

power = 
$$\frac{\#(\hat{l_r} > 0) + \#(\hat{u_r} < 0)}{1000}$$

where  $\#(\hat{l}_i > 0)$  is the total number of replications with the lower limits of confidence intervals larger than 0 and  $\#(\hat{u}_r < 0)$  is the total number of replications with the upper limits smaller than 0. If the population parameter value is not equal to 0, a better method should have greater statistical power. If the population parameter value is equal to 0, a good method should have type I error close to the nominal alpha level.

### Simulation 1. Analysis of MCAR data

The parameter estimate biases, coverage probabilities, and power/Type I errors for MCAR data with 10%, 20%, and 40% missing data are obtained without and with auxiliary variables and are summarized in Table 1. From the results, we can conclude the following. First, biases of the parameter estimates for all conditions under the studied MCAR conditions are smaller than 1.5%. Second, the coverage probabilities are close to the true value .95 except that the coverage probabilities of variance parameters range from .88 to .94 and are slightly underestimated. Third, the inclusion of auxiliary variables in MCAR data mediation analysis does not seem to influence the accuracy of parameter estimates and coverage probabilities although the auxiliary variables are correlated with M and Y(r=.5). The use of auxiliary variables, however, slightly boosters the power of detecting mediation effect especially when the missing proportion is larger (e.g., 40%).

# Simulation 2. Analysis of MAR data

The estimate biases, coverage probabilities, and power for MAR data analysis are summarized in Table 2. 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 are smaller than that from MCAR data given the same proportion of missing data.

#### Simulation 3. Analysis of MNAR data

The results from MNAR data analysis are summarized in Table 3. The results clearly show that when auxiliary variables are not included, parameter estimates are highly biased especially when the missing data proportion is larger, e.g., about 67% bias with 40% missing data for the mediation effect. Correspondingly, coverage probabilities are highly underestimated. For example, with 40% of missing data, the coverage probabilities for intercepts and variance parameters are almost zero. However, with the inclusion of appropriate auxiliary variables, the parameter estimate biases dramatically decrease to 3% or below and the coverage probabilities are close to 95%. Thus, multiple imputation can be used to analyze MNAR data and recover true parameter values by including appropriate auxiliary variables that can explain missingness of the variables in the mediation model.

# Simulation 4. Impact of the number of imputations

A potential difficulty of applying multiple imputation is to make an appropriate decision on how many imputations are needed. For example, Rubin (1987) has suggested that five imputations are sufficient in the case of 50% missing data for estimating simple mean. But Graham et al. (2007) recommend that many more imputations than that 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

Table 1
Biases, coverage probabilities, and power/type I error under MCAR situations

		W	ithout Auxi	liary Variables	With Auxiliary Variables			
		Bias	Coverage	Power/Type I error	Bias	Coverage	Power/Type I error	
	$\overline{a}$	0.595	0.938	0.946	0.861	0.943	0.956	
	b	0.055	0.941	0.920	-0.130	0.944	0.927	
	c'	0.487	0.953	0.047	0.304	0.945	0.055	
10%	ab	0.219	0.967	0.900	0.263	0.967	0.920	
10%	$i_Y$	0.116	0.945	0.055	0.304	0.946	0.054	
	$i_M$	0.065	0.956	0.044	-0.072	0.952	0.048	
	$\sigma_{eY}^2$	-0.657	0.935	1.000	-0.494	0.931	1.000	
	$\sigma_{eM}^{2}$	-0.148	0.931	1.000	-0.051	0.930	1.000	
	a	-0.218	0.936	0.907	-0.047	0.938	0.920	
	b	-0.525	0.940	0.829	-0.131	0.937	0.862	
	c'	0.430	0.934	0.066	0.266	0.941	0.059	
20%	ab	-1.222	0.966	0.808	-0.593	0.963	0.845	
20%	$i_Y$	-0.165	0.946	0.054	-0.204	0.944	0.056	
	$i_M$	0.349	0.956	0.044	0.268	0.954	0.046	
	$\sigma_{eY}^2$	-0.818	0.920	1.000	-0.539	0.918	1.000	
	$\sigma_{eM}^2$	-0.244	0.942	1.000	-0.105	0.944	1.000	
	a	0.640	0.938	0.822	0.634	0.930	0.849	
	b	-1.310	0.930	0.565	-0.593	0.935	0.635	
	c'	0.607	0.944	0.056	0.226	0.945	0.055	
40%	ab	-0.716	0.946	0.531	0.112	0.950	0.615	
40%	$i_Y$	-0.007	0.943	0.057	-0.044	0.939	0.061	
	$i_M$	-0.127	0.966	0.034	0.050	0.963	0.037	
	$\sigma_{eY}^2$	-1.484	0.888	1.000	-0.860	0.911	1.000	
	$\sigma_{eY}^2 \ \sigma_{eM}^2$	0.077	0.924	1.000	0.498	0.933	1.000	

because of the amount of computational time involved (In total, K (number of imputations) x B (number of bootstrap samples) mediation models need to be estimated).

In this simulation study, we will briefly investigate the impact of the number of imputations on the point estimates and standard error estimates of mediation effects in mediation analysis with missing data. More specifically, we collect the results from MNAR data analysis with auxiliary variables with the number of imputations from 10 to 100 with an interval of 10. We focus on how the mediation effect estimates and the bootstrap standard error estimates change with the number of imputations. For the purpose of comparison, we calculate the relative deviances of mediation effect estimates and their standard error estimates from those estimates with 100 imputations. Those relative deviances from conditions 10% missing data and 40% missing data are plotted in Figure 2.

Figure 2a portrays the relative deviances from results with 10% missing data. Note

Table 2
Biases, coverage probabilities, and power/type I error under MAR situations

		W	ithout Auxi	liary Variables	With Auxiliary Variables			
			Coverage	Power/type I error	Bias	Coverage	Power/type I error	
	a	0.716	0.944	0.927	0.439	0.938	0.929	
	b	-0.331	0.936	0.896	-0.314	0.946	0.917	
	c'	0.679	0.954	0.046	0.435	0.947	0.053	
10%	ab	-0.119	0.957	0.870	-0.403	0.961	0.893	
10%	$i_Y$	0.369	0.948	0.052	0.294	0.948	0.052	
	$i_M$	-0.010	0.956	0.044	-0.084	0.956	0.044	
	$\sigma_{eY}^2$	-0.574	0.924	1.000	-0.457	0.921	1.000	
	$\sigma_{eY}^2 \ \sigma_{eM}^2$	-0.409	0.932	1.000	-0.234	0.936	1.000	
	a	0.838	0.936	0.862	-1.871	0.935	0.862	
	b	-0.897	0.935	0.807	0.095	0.933	0.833	
	c'	0.320	0.940	0.060	0.027	0.952	0.048	
20%	ab	-0.546	0.962	0.767	-1.940	0.958	0.791	
2070	$i_Y$	-0.294	0.945	0.055	0.035	0.952	0.048	
	$i_M$	0.375	0.951	0.049	-0.120	0.949	0.051	
	$\sigma_{eY}^2$	-0.886	0.918	1.000	-0.650	0.920	1.000	
	$\sigma_{eM}^2$	-0.109	0.941	1.000	-0.063	0.942	1.000	
	a	-0.563	0.937	0.697	-0.135	0.942	0.772	
	b	-1.863	0.929	0.597	-1.372	0.926	0.647	
	c'	0.837	0.940	0.060	0.315	0.945	0.055	
40%	ab	-2.932	0.960	0.511	-1.747	0.955	0.599	
	$i_Y$	0.220	0.950	0.050	0.004	0.943	0.057	
	$i_M$	-0.604	0.945	0.055	-0.202	0.955	0.045	
	$\sigma_{eY}^2$	-1.137	0.908	1.000	-0.452	0.909	1.000	
	$\sigma_{eM}^{2}$	-0.017	0.936	1.000	0.466	0.940	1.000	

that with the number of imputations larger than 50, the relative deviances of point estimates are all close to zero and remain unchanged. Thus, 50 imputations seem to be sufficient for mediation analysis with 10% missing data. With 40% missing data, however, the relative deviances of point estimates do not approach zero until the number of imputations is larger than 80 as shown in Figure 2b. Therefore, the number of imputations required is related to the amount of missing data. In our simulation study, the choice of 100 imputations appears to be enough based on this simulation.

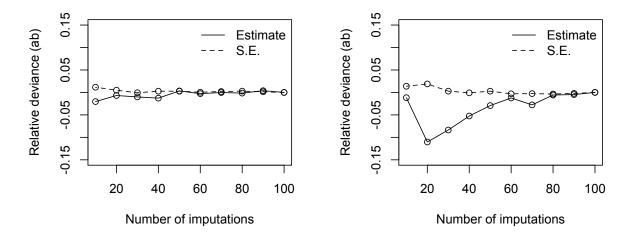
Table 3
Biases, coverage probabilities, and power/type I error under MNAR situations

		Wi	thout Auxil	iary Variables	1	With Auxilia	ary Variables
		Bias	Coverage	Power/type I error	Bias	Coverage	Power/type I error
100	$\overline{a}$	-20.534	0.824	0.891	0.918	0.938	0.956
	b	-15.339	0.888	0.827	-1.099	0.933	0.923
	c'	1.930	0.955	0.045	0.468	0.946	0.054
	ab	-32.633	0.831	0.800	-0.513	0.951	0.925
10%	$i_Y$	11.320	0.739	0.261	-0.076	0.951	0.049
	$i_M$	14.547	0.591	0.409	-0.090	0.948	0.052
	$\sigma_{eY}^2$	-13.029	0.532	1.000	0.004	0.939	1.000
	$\sigma_{eM}^{2}$	-13.121	0.508	1.000	0.248	0.938	1.000
	$\overline{a}$	-29.841	0.728	0.838	0.782	0.941	0.929
	b	-27.443	0.810	0.589	-2.856	0.928	0.826
	c'	2.197	0.943	0.057	0.190	0.947	0.053
20%	ab	-49.117	0.673	0.570	-2.583	0.941	0.815
	$i_Y$	22.228	0.356	0.644	-0.001	0.955	0.045
	$i_M$	27.597	0.145	0.855	0.234	0.956	0.044
	$\sigma_{eY}^2$	-20.661	0.228	1.000	-0.494	0.933	1.000
	$\sigma_{eY}^2 \ \sigma_{eM}^2$	-20.331	0.215	1.000	0.426	0.936	1.000
	a	-45.357	0.525	0.638	-0.044	0.943	0.846
	b	-38.421	0.839	0.355	-1.824	0.934	0.666
40%	c'	3.041	0.936	0.064	1.053	0.947	0.053
	ab	-66.815	0.559	0.305	-2.951	0.951	0.642
	$i_Y$	45.112	0.113	0.887	-1.212	0.950	0.050
	$i_M$	55.439	0.000	1.000	-0.055	0.949	0.051
	$\sigma_{eY}^2$	-31.444	0.086	1.000	0.333	0.923	1.000
	$\sigma_{eY}^2 \ \sigma_{eM}^2$	-31.484	0.048	1.000	1.194	0.921	1.000

# **An Empirical Example**

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



(a) 10% missing data (b) 40% missing data Figure 2. The impact of different numbers of imputations on the accuracy of point estimates and bootstrap standard error estimates.

(HE), children's mathematical achievement (Math), children's behavior problem index (BPI), and children's reading recognition and reading comprehension achievement. For the mediation analysis, mothers' education is the independent variable, home environment is the mediator, and children's mathematical achievement is the outcome variable. The missing data patterns and the sample size of each pattern are presented in Table 4. 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 children's mathematical achievement. For the purpose of demonstration, children's behavior problem index (BPI) and children's reading recognition and reading comprehension achievement- are used as auxiliary variables in the data analysis.

In Table 5, the results from empirical data analysis using the proposed method without and with the auxiliary variables are presented. The results reveal that the inclusion of the auxiliary variable 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 from the analysis with auxiliary variables also show that home environment partially mediates the relationship between mothers' education and children's mathematical achievement because both the indirect effect ab and the direct effect c' are significant.

<sup>&</sup>lt;sup>1</sup>For the empirical data analysis, 1000 bootstraps and 100 imputations were used.

Table 4

Missing data patterns of the empirical data set.

Pattern	ME	HE	Math	Sample size
1	O	О	О	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 5
Mediation effect of home environment on the relationship between mothers' education and children's mathematical achievement

	Withou	ıt Auxili	ary Vari	able	With Auxiliary Variable			
Parameter	Estimate	S.E.	95% BC		Estimate	S.E.	95% BC	
$\overline{a}$	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
$\sigma_{eY}^2$	4.532	0.269	4.093	5.211	4.520	0.268	4.075	5.141
$\sigma_{eY}^2 \ \sigma_{eM}^2$	1.660	0.061	1.545	1.789	1.660	0.061	1.542	1.790

*Note.* S.E.: bootstrap standard error. BC: bias-corrected confidence interval.

#### **Discussion**

In this study, we discussed how to conduct mediation analysis with missing data through multiple imputation and bootstrap. We implemented the method by using SAS and the program scripts are also provided and easy to use. 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. The analysis of a subset of data from the NLSY79 revealed that home environment partially mediated the relationship between mothers' education and children's mathematical achievement.

#### 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 commer-

cial software and thus can be extended to mediation analysis. Third, it is natural and easy to include auxiliary variables in multiple imputation for analyzing MNAR data. Fourth, multiple imputation does not assume a specific model for imputing data.

The implementation of multiple imputation and bootstrap in SAS also has its own advantages. First, only a minimum number of parameters usually need to be changed to run the SAS program for mediation analysis with missing data. Second, the SAS program can be easily extended for more complex mediation analysis by taking advantage of available SAS procedures. For example, one can also conduct mediation analysis with moderators through modifying the PROC REG statements. One can conduct mediation analysis with latent variables through the use of SAS PROC CALIS. Third, SAS excels in terms of performance in dealing with large dataset, which is critical for multiple imputation and bootstrap. For example, for a data set with a sample size 100, to generate 1000 bootstrap samples and impute each bootstrap sample 100 times, one needs to deal with a data set with 10,000,000 (ten million) records. Although this seems to be a huge data set, it only took SAS about 7 minutes to conduct such missing data mediation analysis with 20% missing data.

## **Assumptions and limitations**

There are several assumptions and limitations of the current study. First, the study only discusses the mediation model with a single mediator. The current SAS program is also based on this model. 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.

In summary, a method using multiple imputation and bootstrap for mediation analysis with missing data is introduced and the program of implementing this method is developed in SAS. Simulation results show that the method works well in dealing with missing data for mediation analysis under different missing mechanisms. We hope this program can promote the use of advanced techniques in dealing with missing data for mediation analysis in the future.

#### References

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

Baron, R. M. and 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. and 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*. The Ohio State University, Columbus, Ohio.
- Chen, Z. X., Aryee, S., and Lee, C. (2005). Test of a mediation model of perceived organizational support. *Journal of Vocational Behavior*, 66(3):457–470.
- Cole, D. A. and Maxwell, S. E. (2003). Testing mediational models with longitudinal data: Questions and tips in the use of structural equation modeling. *Journal of Abnormal Psychology*, 112:558–57.
- 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. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185.
- Freedman, L. S. and Schatzkin, A. (1992). Sample size for studying intermediate endpoints within intervention trails or observational studies. . 136:1148-1159. *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., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention Science*, 8:206–213.
- Little, R. J. A. and Rubin, D. B. (1987). *Statistical analysis with missing data*. Wiley, New York, N.Y.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical analysis with missing data*. Wiley-Interscience, New York, N.Y., 2nd edition.
- MacCorquodale, K. and 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*. Taylor & Francis, New York, NY.
- MacKinnon, D. P., Fairchild, A. J., and 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., and 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., and Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Be-havioral Research*, 39(1):99–128.
- Mallinckrodt, B., Abraham, T. W., Wei, M., and Russell, D. W. (2006). Advance in testing statistical significance of mediation effects. 53.
- Preacher, K. J. and 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. and 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*. John Wiley & Sons, Inc.
- Savalei, V. and 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. Chapman & Hall/CRC.
- Shrout, P. E. and 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 Leinhardt, S., editor, *Sociological methodology*, pages 290–312. Jossey-Bass, San Francisco.
- Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance structure models. In Tuma, N., editor, *Sociological methodology*, pages 159–186. American Sociological Association, Washington, DC.

- Wang, L., Zhang, Z., and Estabrook, R. (2009). Longitudinal mediation analysis of training intervention effects. In Chow, S. M., Ferrer, E., and Hsieh, F., editors, *Statistical methods for modeling human dynamics: An interdisciplinary dialogue*, pages 349–380. Lawrence Erlbaum Associates, New Jersey.
- Woodworth, R. S. (1928). Dynamic psychology. In Murchison, C., editor, *Psychologies of 1925*, pages 111–126. Clark Universal Academy Press, Inc., Worcester, MA.
- Zhang, Z. and Wang, L. (2008). Methods for evaluating mediation effects: Rationale and comparison. In Shigemasu, K., Okada, A., Imaizumi, T., and Hoshino, T., editors, *New Trends in Psychometrics*, pages 595–604, Tokyo. Universal Academy Press, Inc.

# Appendix SAS codes for MI and bootstrap

```
1 /*** Setup the global parameters ***/
2 /*The parameters below should be changed accordingly*/
3 %LET filename="c:\mnarmediation\dataname.txt"; \star data file directory and
4 %LET varname=x m y a1 a2; *specify variable names in the data file.
      Please use x for the input variable, m for the mediation variable,
      and y for the output variable. a1 and a2 are two auxiliary variables
       in the data file. You can use any names except for x, m, and y for
      naming the auxiliary variables;
5 %LET missing=99999; *specify the missing data value;
6 %LET nimpute = 100; *define the number of imputations K;
7 %LET nboot = 1000; *define the number of bootstraps B;
8 %LET alpha = 0.95; *define the confidence level;
9 %LET seed = 2010; *random number seed;
10 /*** End of setup of global parameters ***/
11
12
13 /*In general, there is no need to change the codes below*/
14 /*Read data into sas*/
15 DATA dset;
16
    INFILE &filename;
17
     INPUT &varname;
18
     ARRAY nvarlist &varname;
19
     DO OVER nvarlist;
20
       IF nvarlist = &missing THEN nvarlist = .;
21
     END;
22 RUN;
23
24 /*Use multiple imputation to obtain point estimates of the model
      parameters based on the original data set*/
25 /*Imputing the original data set multipe times*/
26 PROC MI DATA = dset SEED = & seed NIMPUTE = & nimpute OUT = imputed NOPRINT;
27
     VAR &varname;
28 RUN; OUIT;
29 /*Estimating model parameters for each imputed data set*/
```

```
30 PROC REG DATA=imputed OUTEST= est NOPRINT;
31
    MODEL y = x m;
32
     MODEL m = x;
33
     BY _Imputation_;
34 RUN; QUIT;
35
36 /*Collecting results from mutiple imputations*/
37 DATA temp;
38
    SET est;
39
     id =INT((N - .1)/2)+1;
40
     modelnum = MOD(_N_+1, 2) + 1;
41 RUN;
42
43 DATA temp1;
44
     SET temp;
45
     ARRAY int[2] iY iM;
46
     ARRAY xpar[2] c a;
47
     ARRAY mpar[2] b tmp1;
48
     ARRAY sigma[2] sy sm;
49
     RETAIN a b c iY iM sy sm;
50
           BY id;
51
           IF FIRST.id THEN DO I = 1 to 2;
52
                   int[I] = .;
53
                   xpar[I] = .;
54
                   mpar[I]=.;
55
                   sigma[I]=.;
56
           END;
57
           int[modelnum] = intercept;
58
           xpar[modelnum] = x;
59
           mpar[modelnum] = m;
60
           sigma[modelnum] = _RMSE_;
61
           IF LAST.id THEN OUTPUT;
62
           KEEP _imputation_ a b c iY iM sy sm;
63 RUN;
64 /*Calcuating mediation effects*/
65 DATA temp2;
66
    SET temp1;
67
     ab=a*b;
68 RUN;
69
70 /*Saving the point estimates of model parameters and mediation effect
      from multiple imputation into a data set named 'pointest' */
71 PROC MEANS DATA=temp2 NOPRINT;
72
     VAR a b c ab iY iM sy sm;
73
     OUTPUT OUT=pointest MEAN(a b c ab iY iM sy sm) = a b c ab iY iM sy sm;
74 RUN;
75
76 /*** Bootstraping data to obtain standard errors and confidence
      intervals ***/
77 DATA bootsamp;
78
     DO sampnum = 1 to &nboot;
```

```
79
         DO i = 1 TO nobs;
            ran = ROUND(RANUNI(&seed) * nobs);
80
81
            SET dset
82
           nobs = nobs
83
            point = ran;
84
            OUTPUT;
85
         END;
86
   END;
87
     STOP;
88 RUN; QUIT;
90 /*** Imputing K data sets for each bootstrap sample ***/
91 PROC MI DATA-bootsamp SEED-&seed NIMPUTE-&nimpute OUT-imputed NOPRINT;
92
    EM MAXITER = 500;
93
    VAR &varname;
94 BY sampnum;
95 RUN; QUIT;
96
97 /*Estimate model parameters for each imputed data set (in total, there
       are B*K imputed data sets.) */
98 PROC REG DATA=imputed OUTEST= est NOPRINT;
99
    MODEL y = x m;
100
   MODEL m = x;
     BY sampnum _Imputation_;
102 RUN; QUIT;
103
104 /*Collecting results from different imputed data sets*/
105 DATA temp;
106
     SET est;
107
     id =INT((_N_-.1)/2)+1;
108
   modelnum = MOD(N_+1, 2) + 1;
109 RUN;
110
111 DATA temp1;
112 SET temp;
113
     ARRAY int[2] iY iM;
114
     ARRAY xpar[2] c a;
115
     ARRAY mpar[2] b tmp1;
116
     ARRAY sigma[2] sy sm;
     RETAIN a b c iY iM sy sm;
117
118
            BY id;
119
            IF FIRST.id THEN DO I = 1 to 2;
                    int[I] = .;
120
121
                    xpar[I] = .;
122
                    mpar[I]=.;
123
                    sigma[I]=.;
124
            END;
125
            int[modelnum] = intercept;
126
            xpar[modelnum] = x;
127
            mpar[modelnum] = m;
128
            sigma[modelnum] = _RMSE_;
```

```
IF LAST.id THEN OUTPUT;
129
130
            KEEP sampnum _imputation_ a b c iY iM sy sm;
131 RUN;
132
133 DATA temp2;
134
      SET temp1;
      ab=a*b;
135
136 RUN;
137
138 /*Compute point estimates of model parameters and mediation effect for
       each bootstrap sample and the results are saved in the data file
       named 'bootest'. */
139 PROC MEANS DATA=temp2 NOPRINT;
140
     BY sampnum;
141
      VAR a b c ab iY iM sy sm;
142
      OUTPUT OUT=bootest MEAN(a b c ab iY iM sy sm) = a b c ab iY iM sy sm;
143 RUN;
144
145 /*** Calculate the BC intervals based on the point estimates from
       different bootstrap samples and produce a table containing the
       points estimates, standard errors, confidence intervals in the
       output window.***/
146 PROC IML;
147
      START main;
148
      USE pointst;
149
      READ ALL INTO Y;
150
     USE bootest;
151
      READ ALL INTO X;
152
153
      n=NROW(X);
154
      m=NCOL(X);
155
156
      bc lo=J(1, m-3, 0);
157
      bc_{up}=J(1, m-3, 0);
158
      se=J(1, m-3, 0);
159
160
      alphas=1-(1-&alpha)/2;
161
      zcrit = PROBIT(alphas);
162
163
      DO j=1 TO m-3;
164
            se[j] = SQRT((SSQ(X[,j+3]) - (SUM(X[,j+3]))**2/n)/(n-1));
165
            number=0;
166
            DO i=1 TO n;
167
                     IF X[i,j+3] < Y[j+2] THEN number=number+1;
168
            END;
169
            p=number/n;
170
            z0hat=PROBIT(p);
171
172
            q1=z0hat+(z0hat-zcrit);
173
            g2=z0hat+(z0hat+zcrit);
174
            alpha1=PROBNORM(q1);
```

```
175
            alpha2=PROBNORM(q2);
176
177
            vec=X[,j+3];
            CALL SORT (vec, {1});
178
179
180
            low=int(alpha1*(n+1));
181
            up=int(alpha2\star(n+1));
182
            IF low<1 THEN low=1;</pre>
183
            IF up>n THEN up=n;
184
            bc_lo[j]=vec[low];
185
            bc_up[j]=vec[up];
186
      END;
187
188
      result=Y[3:10]||se'||(bc_lo')||(bc_up');
189
      MATTRIB result ROWNAME=({a, b, c, ab, iy, im, sy, sm})
190
                      COLNAME=({estiamtes se CI_lo CI_up})
191
                      LABEL='MEDIATION_ANALYSIS_RESULTS' FORMAT=f10.5;
192
      PRINT result;
193
      FINISH;
194
      RUN main;
195 QUIT;
```