

Mediation analysis with missing data through Mplus and R

Zhiyong Zhang & Lijuan Wang
University of Notre Dame

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 & 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 & 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 & Hauser, 1975) and the surrogate or intermediate endpoint effect (Epidemiology, Freedman & Schatzkin, 1992).

Figure 1 (after Shrout & 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 $\sigma_{e_M}^2$ and $\sigma_{e_Y}^2$. 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 .

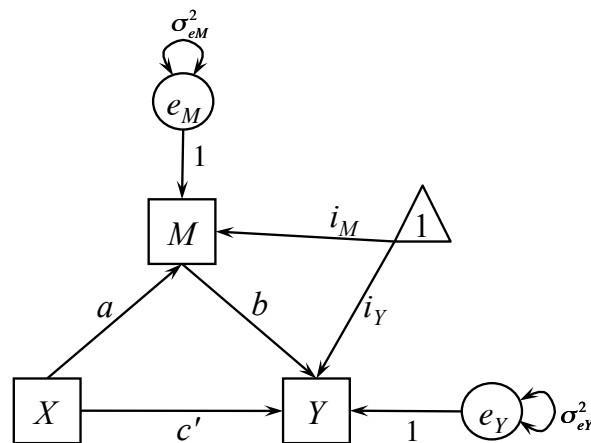


Figure 1: Path diagram demonstration of a mediation model.

Statistical approaches to estimating and testing mediation effects with complete data have been discussed extensively in the psychological literature (e.g., Baron & Kenny, 1986; Bollen & Stine, 1990; MacKinnon et al., 2002, 2007; Shrout & 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 & 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 & Hayes, 2008; Shrout & Bolger, 2002; Zhang & Wang, 2008).

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 & Hayes, 2004). However, with the availability of advanced approaches such as FIML, listwise and pairwise deletion is no longer deemed acceptable (Little & Rubin, 2002; Savalei & Bentler, 2009; Schafer, 1997).

In this study, we discuss how to deal with missing data for mediation analysis using FIML. In the following, we will first present the technical backgrounds of FIML for mediation analysis with missing data. Then, we will discuss how to implement the method using Mplus and R.

Method

Full information maximum likelihood (FIML) method

FIML is a maximum likelihood method to obtain model parameter estimates by maximizing the likelihood of all available data (Little & Rubin, 2002). FIML has gained a lot popularity because of its implementation in programs such as EQS and Mplus. Although the incorporation of auxiliary variables is especially straightforward in MI, Graham (2003) proposed a saturated correlates model that allows the use of auxiliary variables in programs that run FIML. The saturated correlates model imposes a saturated covariance structure on the auxiliary variables by allowing the auxiliary variables to covary with residuals of the variables of interest.

To obtain the FIML mediation effect estimate with auxiliary variables, the model illustrated in Figure 2 can be estimated. In the model, the auxiliary variables $A_i, i = 1, \dots, p$ are included and allowed to correlate with X and residuals e_Y and e_M . If the auxiliary variables are related to the missingness of mediation variables, by estimating such a model in Figure 2, the possible bias in mediation effect estimate using only mediation variables can be corrected.

Testing mediation effects through the bootstrap method

The bootstrap method (Efron, 1979, 1987) was first employed in mediation analysis by Bollen & Stine (1990) and has been studied in a variety of research contexts (e.g., MacKinnon et al., 2004; Mallinckrodt et al., 2006; Preacher & Hayes, 2008; Shrout & Bolger, 2002). This method

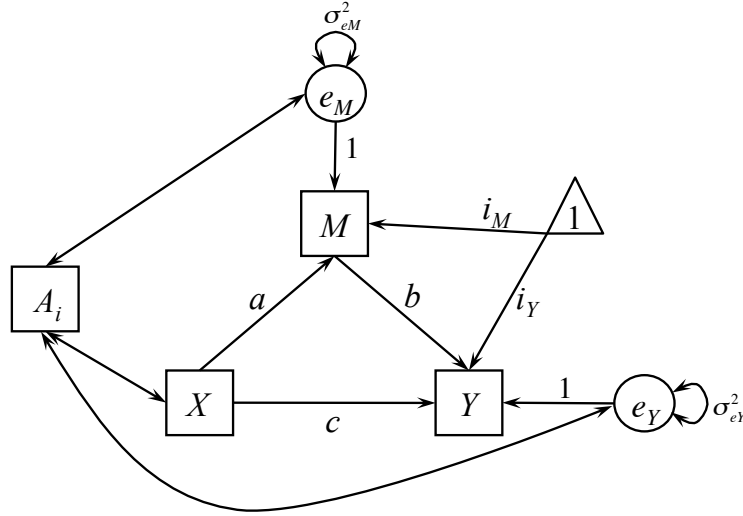


Figure 2: FIML with auxiliary variables. To simplify the path diagram, only one auxiliary variable is portrayed.

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 & Stine (1990) can be applied to the 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.
2. With the bootstrap sample, estimate model parameters and the mediation effect through FIML.
3. Repeat Steps 1 and 2 for a total of B times. B is the number of bootstrap samples.
4. Empirical distributions of model parameters and the mediation effect are then obtained using the B sets of bootstrap estimates. Thus, confidence intervals of model parameters and mediation effect can be constructed.

Using the bootstrap sample estimates, one can obtain the bootstrap standard errors and confidence intervals of model parameters conveniently. Let $\theta = (iM, iY, a, b, c', c, \sigma_{eM}^2, \sigma_{eY}^2, ab)^t$ denote a vector of model parameters and the mediation effect ab . With data from each bootstrap, we can obtain $\hat{\theta}^b$, $b = 1, \dots, B$. The standard error of the p th parameter $\hat{\theta}_p$ can be calculated as

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

with

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

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 confidence intervals evaluated.

The $1 - 2\alpha$ BC interval for the p th element of θ 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 Φ is the standard normal cumulative distribution function and $z^{(\alpha)}$ is the α percentile of a standard normal distribution and

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

Implementation FIML and bootstrap using Mplus and R

To facilitate the application of FIML and bootstrap, we implement the method using Mplus (Muthén & Muthén, 1998-2007) and R (R Development Core Team, 2009). The latest version of Mplus (V5.21) allows the direct inclusion of auxiliary variables. However, with auxiliary variables, the implementation of bootstrap is not allowed. To obtain the bootstrap confidence intervals for FIML with auxiliary variables, we provide an R function which resamples data in R, obtains FIML estimates using Mplus, and constructs BC confidence intervals within R.

To use the R function, we first need to have a working Mplus program that can conduct mediation analysis with missing data using auxiliary variables. Figure 3 presents such an example Mplus program with two auxiliary variables (saved in a file called `mplus.inp`). In this Mplus program, line 3 specifies the name of the data file `data.txt`. Line 5 lists the variables in the data file. Line 6 selects the variables - X, M, and Y - in the mediation model. Line 7 specifies the auxiliary variables - U and V. Line 8 gives the missing data indicator, the value 99999 represents a missing datum. The key for this Mplus program is to save the parameter estimates into a file `est.txt` that is given on lines 20-21. When the Mplus program is run, the parameter estimates will be saved into the file specified and then be read into R for further analysis.

The Mplus program in Figure 3 only conducts data analysis once for a data set supplied. To obtain the bootstrap standard errors and confidence intervals, we need to resample the data and repeat the mediation analysis for each bootstrap sample. This job is conducted in an R program shown in Figure 4.

The R program consists of two R functions, `bc.ci` on lines 1-9 and `boot.mplus` on lines 11-40. The second function is the main function that calls the first function to construct the bootstrap confidence intervals. There is no need to make any change on the first function and thus we focus on discussing the second function.

The main functionality of `boot.mplus` is to conduct bootstrap, call `mplus` for data analysis, and then construct the bootstrap standard errors and confidence intervals. There are 5 input parameters for the function. The first is the number of bootstraps (B), the second is the name of

```

1 TITLE: Mediation analysis with auxiliary variables
2 DATA:
3     FILE = data.txt ;
4 VARIABLE:
5     NAMES ARE X M Y U V;
6     USEVARIABLES ARE X M Y;
7     AUXILIARY = (m) U V;
8     MISSING = ALL(99999);
9 ANALYSIS:
10    COVERAGE = .01;
11
12 MODEL:
13    M ON X ;
14    Y ON M X;
15 MODEL INDIRECT:
16    Y IND X;
17
18 OUTPUT: tech1;
19 SAVEDATA:
20    ESTIMATES = est.txt;

```

Figure 3: Mplus program for FIML with auxiliary variables

the data file to be used (*datafile*), the third is the command to run Mplus for mediation analysis using FIML (*runmplus*), the fourth is the name of the file that the parameter estimates from Mplus are saved (*estfile*), and the fifth is the confidence level for the calculation of confidence intervals (*alpha*).

To use the R function, one only need to supply the corresponding parameters. For example, line 43 conducts the analysis with $B=1000$ and confidence interval $\alpha=0.95$ for data in the file “*inputdata.txt*”. The Mplus program is called “*mplus.inp*” and the Mplus software is installed in the folder “*c:/progra~1/mplus/*”. Furthermore, the parameter estimates from Mplus are saved in the file “*est.txt*”.

When using the R program, several things should be paid special attention.

1. The data file name specified in *mplus.inp* should be the same as the file name specified on Lines 14 and 22 in Figure 4, in the above example, “*data.txt*”.
2. It is recommended that the file name specified in *mplus.inp* is different with the file name used on Line 43. For example, on Line 43, “*inpdata.txt*” is used while in *mplus.inp*, “*data.txt*” is used.
3. The fourth input parameter in the function *boot.mplus* should be the same as the file name given on Line 20 in *mplus.inp*.
4. On Line 42, for *runmplus*, the first part is the path to to run mplus (“*c:/progra~1/mplus/mplus mplus.inp*”) and the second part is the name of the Mplus program file (“*mplus.inp*”).

```

1 bc.ci<-function(res, est, alpha){ ## function to obtain BC CI
2   iB<-length(res)
3   alpha<-c((1-alpha)/2, (1+alpha)/2)
4   z0<-qnorm(sum(res<est)/iB)
5   zalpna<-qnorm(alpha)
6   bcalpha<-pnorm(2*z0+zalpha)
7   BC<-quantile(res,bcalpha)
8   c(est, sd(res), BC)
9 }
10
11 boot.mplus<-function(B, datafile, runmplus, estfile, alpha=.95){
12   temp.data<-read.table(datafile)
13   N<-dim(temp.data)[1]
14   write.table(temp.data, 'data.txt', row.names=F, col.names=F)
15   system(runmplus,show.output.on.console=F)
16   res<-scan(estfile)
17
18   boot.res<-NULL
19   for (j in 1:B){
20     index<-sample(1:N, replace=T)
21     dset<-temp.data[index, ]
22     write.table(dset, 'data.txt', row.names=F, col.names=F)
23     system(runmplus,show.output.on.console=F)
24     temp<-scan(estfile)
25     boot.res<-rbind(boot.res, temp)
26     cat(j, '\n')
27   }
28
29   res<-res[c(19:21,24,25,27,31,33,36)]
30   res<-c(res, res[4]*res[5])
31   boot.res<-boot.res[,c(19:21,24,25,27,31,33,36)]
32   boot.res<-cbind(boot.res, boot.res[,4]*boot.res[,5])
33   BC.matrix<-NULL
34   for (i in 1:10){
35     BC.matrix<-rbind(BC.matrix, bc.ci(boot.res[,i], res[i], alpha))
36   }
37   colnames(BC.matrix)<-c('Estimate', 'S.E.', '2.5%', '97.5%')
38   rownames(BC.matrix)<-c('iM', 'iY', 'iX', 'a', 'b', "c", 'eM2', 'eY2', 'eX2', 'ab')
39   list(res=res, boot.res=boot.res, BC=BC.matrix)
40 }
41
42 runmplus<-"c:/progra~1/mplus/mplus mplus.inp"
43 allres<-boot.mplus(1000, 'inpdata.txt', runmplus, 'est.txt', .95)
44 allres

```

Figure 4: R program to run Mplus and construct bootstrap standard errors and BC confidence intervals

	Estimate	S.E.	2.5%	97.5%
iM	88.0140	3.357033160	81.39459716	94.5456534
iY	9.9410	1.515817917	6.88690878	13.0092261
iX	73.6330	0.111654259	73.41299045	73.8479043
a	-0.7890	0.045118015	-0.88100000	-0.7030317
b	0.3000	0.007448508	0.28400000	0.3130000
c'	0.0010	0.018999314	-0.03789028	0.0380000
eM2	150.0020	4.529848722	141.60967712	159.1515632
eY2	16.5510	0.547681011	15.56587657	17.7281524
eX2	34.8600	0.856623597	33.08364084	36.4339514
ab	-0.2367	0.014665061	-0.26823523	-0.2095433

Figure 5: An example output from the running of Mplus and R

The R program has three parts of output. The first part consists of the parameter estimates and mediation effect estimate of the original data. The second part consists of the parameter estimates and mediation effect estimate of each bootstrap sample. The third part includes the bootstrap standard errors and BC confidence intervals. An example output of the third part of the results is given in Figure 5.

Discussion

There are many advantages of using Mplus and R for mediation analysis with missing data.

1. R is very flexible for data manipulation.
2. Mplus is very flexible and powerful for path analysis and structural equation modeling. Thus more complex mediation analysis can be conducted.

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., & Stine, R. A. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115–140.
- 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.
- 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., & Schatzkin, A. (1992). Sample size for studying intermediate endpoints within intervention trials or observational studies. . 136:1148-1159. *American Journal of Epidemiology*, 136, 1148–1159.
- Graham, J. W. (2003). Adding missing-data-relevant variables to fimo-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.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, N.Y.: Wiley-Interscience.
- MacCorquodale, K., & Meehl, P. E. (1948). On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55(2), 95–107.
- 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. , 53.
- Muthén, L. K., & Muthén, B. O. (1998-2007). *Mplus user's guide* (Fifth ed.). Los Angeles, CA: Muthén and Muthén. (<http://www.statmodel.com>)
- 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.
- R Development Core Team. (2009). *R: A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Available from <http://www.R-project.org> (ISBN 3-900051-07-0)
- 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*. 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.
- Woodworth, R. S. (1928). Dynamic psychology. In C. Murchison (Ed.), *Psychologies of 1925* (pp. 111–126). Worcester, MA: Clark Universal Academy Press, Inc.
- 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.