Contemporary Clinical Trials Communications 7 (2017) 130–135

Contents lists available at ScienceDirect

Contemporary Clinical Trials Communications

journal homepage: www.elsevier.com/locate/conctc

Comparison of methods for the analysis of relatively simple mediation models

Judith J.M. Rijnhart[a,*], Jos W.R. Twisk[a], Mai J.M. Chinapaw[b], Michiel R. de Boer[c], Martijn W. Heymans[a]

[a] Department of Epidemiology and Biostatistics, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands
[b] Department of Public and Occupational Health, Amsterdam Public Health Research Institute, VU University Medical Center, Amsterdam, The Netherlands
[c] Department of Methodology and Applied Biostatistics, Faculty of Earth and Life Sciences, Institute for Health Sciences, Amsterdam Public Health Research Institute, VU University, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Mediation analysis
Indirect effect
Ordinary least square regression
Structural equation modeling
Potential outcomes framework
Cross-sectional data

ABSTRACT

Background/aims: Statistical mediation analysis is an often used method in trials, to unravel the pathways underlying the effect of an intervention on a particular outcome variable. Throughout the years, several methods have been proposed, such as ordinary least square (OLS) regression, structural equation modeling (SEM), and the potential outcomes framework. Most applied researchers do not know that these methods are mathematically equivalent when applied to mediation models with a continuous mediator and outcome variable. Therefore, the aim of this paper was to demonstrate the similarities between OLS regression, SEM, and the potential outcomes framework in three mediation models: 1) a crude model, 2) a confounder-adjusted model, and 3) a model with an interaction term for exposure-mediator interaction.
Methods: Secondary data analysis of a randomized controlled trial that included 546 schoolchildren. In our data example, the mediator and outcome variable were both continuous. We compared the estimates of the total, direct and indirect effects, proportion mediated, and 95% confidence intervals (CIs) for the indirect effect across OLS regression, SEM, and the potential outcomes framework.
Results: OLS regression, SEM, and the potential outcomes framework yielded the same effect estimates in the crude mediation model, the confounder-adjusted mediation model, and the mediation model with an interaction term for exposure-mediator interaction.
Conclusions: Since OLS regression, SEM, and the potential outcomes framework yield the same results in three mediation models with a continuous mediator and outcome variable, researchers can continue using the method that is most convenient to them.

1. Introduction

Statistical mediation analysis is an important statistical tool in the field of clinical trials. Many studies use statistical mediation analysis to unravel the pathways underlying the effect of an intervention on a particular outcome variable [1–3]. With statistical mediation analysis the total effect of an intervention on an outcome variable is decomposed into a direct and indirect effect. The indirect effect goes through a mediator variable (a and b paths in Fig. 1), and the remaining effect reflects the direct effect (c' path in Fig. 1) [4]. Therefore, mediation analysis is useful for determining which mediator variables may be targeted by the intervention and thus play a role in the treatment effect.

In 1981, Judd and Kenny proposed the use of the sequence of regression equations (1)–(3) for statistical mediation analysis [5]:

$$\mathrm{Y}=i_1 + c\ \mathrm{X}+ \varepsilon_1 \tag{1}$$

$$\mathrm{M}=i_2 + a\ \mathrm{X}+ \varepsilon_2 \tag{2}$$

$$\mathrm{Y}=i_3 + c'\ \mathrm{X}+ b\ \mathrm{M}+ \varepsilon_3 \tag{3}$$

where in equation (1), c represents the total effect of the exposure variable X on the outcome variable Y. In equation (2), a represents the effect of the exposure variable X on the mediator variable M. In equation (3), c' represents the direct effect of the exposure variable X on the outcome variable Y, and b represents the effect of the mediator

Abbreviations: OLS, ordinary least square; SEM, structural equation modeling; BMI, body mass index; SBC, sweetened beverages consumption; CI, confidence interval; SE, standard error; FIML, full-information maximum likelihood

* Corresponding author. Department of Epidemiology and Biostatistics, VU University Medical Center, PO Box 7057, 1007 MB Amsterdam, The Netherlands.
E-mail address: j.rijnhart@vumc.nl (J.J.M. Rijnhart).

http://dx.doi.org/10.1016/j.conctc.2017.06.005
Received 7 February 2017; Received in revised form 19 June 2017; Accepted 20 June 2017
Available online 22 June 2017

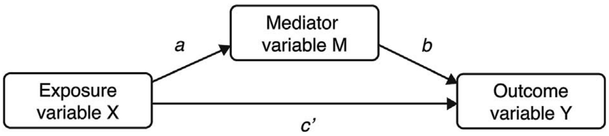

Fig. 1. Path diagram of a relatively simple mediation model.

variable M on the outcome variable Y. In all three equations *i* represents the intercept and *ε* represents the error term. Based on the coefficients from these three equations, the indirect effect can be calculated as the product of the *a* and *b* coefficients or as the difference between the *c* and *c'* coefficients. Furthermore, the proportion mediated can be calculated as either $ab/(ab + c')$, ab/c, or $1-(c'/c)$ [6].

Equations (1)–(3) can be fitted using ordinary least square (OLS) regression, which is often used within epidemiology, or structural equation modeling (SEM), which is often used within psychology [7]. Another regression-based method for statistical mediation analysis is the potential outcomes framework. The aim of this framework is to enhance causal inferences about the mediation model [8]. Ideally, causal inferences should be based on a comparison of a subjects' value of the mediator and outcome variable under both exposure levels [9]. However, in practice the values of the mediator and outcome variable are only measured under the observed exposure level. The mediator and outcome values under the other exposure level remains unobserved. The potential outcomes framework provides definitions of causal effects that can be used to decompose the total effect of an exposure variable on an outcome variable into causal direct and indirect effects, without requiring the measurement of mediator and outcome values under both exposure levels for each subject [9]. These definitions are based on the coefficients in equations (2) and (3).

With the availability of several methods for statistical mediation analysis, the question arises which method for statistical mediation analysis should be preferred. Although a previous study did compare the results from OLS regression with SEM [10], so far the results from OLS regression and SEM have not been compared with the results from the potential outcomes framework. Therefore, the aim of this paper is to demonstrate the similarities between OLS regression, SEM, and the potential outcomes framework. To do this, we used the three methods to estimate the mediated effect in three mediation models with a continuous mediator and outcome variable: 1) a crude model, 2) a confounder-adjusted model, and 3) a model with an interaction term for exposure-mediator interaction.

2. Methods

2.1. Data example

The data example in this paper comes from a randomized controlled trial assessing the effect of an intervention aiming to prevent unhealthy weight gain among school-aged children [11,12]. In this trial, 546 schoolchildren were randomized to either the experimental ($n = 285$) or control condition ($n = 261$). The main outcome in this trial was the change in body mass index (BMI). The association between the intervention and the change in BMI appeared to be mediated by the change in sweetened beverages consumption (SBC) [13]. The mediator and outcome variable were both measured at baseline and after eight months and for both variables standardized residual change scores were used in the mediation analyses, to be able to take into account the baseline values of these variables.

2.2. Methods for statistical mediation analysis

2.2.1. Ordinary least square regression

With OLS regression, equations (1)–(3) (see Section 1) are fitted as three separate regression models. The regression coefficients in these models are estimated by minimizing the sum of the squared deviations of each observation to the regression line [14]. The indirect effect based on the product of the *a* and *b* coefficients and the indirect effect based on the difference between the *c* and *c'* coefficients will be the same when the mediator and outcome variable are both continuous [15]. Furthermore, also the three methods for calculating the proportion mediated ($ab/(ab + c')$, ab/c, and $1-(c'/c)$) will be the same when the mediator and outcome variable are both continuous [6]. Several methods have been proposed for the calculation of a confidence interval (CI) for the indirect effect. The most often used methods are Sobel's CI, the percentile bootstrap CI, and the distribution of the product CI [16].

2.2.2. Structural equation modeling

With SEM, equations (2) and (3) (see Section 1) are fitted simultaneously as one model. SEM models are based on maximum likelihood estimation, which is an iterative estimation procedure maximizing the agreement between the predicted and the observed covariance matrix [17]. When only equations (2) and (3) are fitted, the indirect effect can be calculated as the product of the *a* and *b* coefficients. Furthermore, the total effect of the exposure variable on the outcome variable can be calculated as the summation of the direct and indirect effect ($ab + c'$), and the proportion mediated as the indirect effect divided by the total effect $ab/(ab + c')$. As in OLS regression, Sobel's CI, the percentile bootstrap CI, and the distribution of the product CI can also be calculated for the indirect effect estimated in SEM [16].

2.2.3. Potential outcomes framework

There are two approaches available for the potential outcomes framework, an analytical and a simulation-based approach [18]. Both approaches use two regression models based on equations (2) and (3) (see Section 1) as input for calculating the causal direct and indirect effect and will generally lead to the same results. The only R package that offers the potential outcomes framework for mediation analysis employs the simulation-based approach [19]. Since we used this R package to analyse the data example in this paper, we will limit our explanation of the potential outcomes framework to the simulation-based approach. Information on the analytical approach can be found elsewhere [18].

Within the simulation-based approach, first, a pre-specified number of bootstrap samples with replacement from the original data set are drawn [8]. After this, two new exposure variables are added to each bootstrap sample; one representing the intervention level, assigning the same value to all subjects, e.g. 1, and one representing the control level, again assigning the same value to all subjects, e.g. 0. Then, an OLS model based on equation (2) is are fitted to each bootstrap sample. Based on this model, the value of the mediator variable is simulated for both the treatment and control level. Where $M(0)$ denotes the simulated value of the mediator variable for the control level, and $M(1)$ denotes the simulated value of the mediator variable for the intervention level. These two simulated values of the mediator variable for each subject for both the treatment and control level are added as new variables to each bootstrap sample.

Then an OLS model based on equation (3) is fitted to each bootstrap sample. Based on this model, the value of the outcome variable is simulated for four combinations of the exposure and mediator values. Where $Y(0, M(0))$ denotes the simulated value of the outcome variable for the control level of the exposure variable and the simulated mediator value for the control level, $Y(0, M(1))$ denotes the simulated value of the outcome variable for the control level of the exposure variable and the simulated mediator value for the intervention level, $Y(1, M(0))$ denotes the simulated value of the outcome variable for the intervention level of the exposure variable and the simulated mediator value for the control level, and $Y(1, M(1))$ denotes the simulated value of the outcome variable for the intervention level of the exposure variable and the simulated mediator value for the intervention level. These four predicted values of the outcome variable are also added as new variable to each bootstrap sample.