Equivalence of Two Least-Squares Estimators for Indirect Mediation Effects

WenWu Wang[a,b,*], Ping Yu[c], Yuejin Zhou[d], Tiejun Tong[e], Zhonghua Liu[b]

[a] *School of Statistics, Qufu Normal University, Qufu, China*
[b] *Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong*
[c] *Faculty of Business and Economics, University of Hong Kong, Hong Kong*
[d] *School of Mathematics and Big Data, Anhui University of Science and Technology, Huainan, China*
[e] *Department of Mathematics, Hong Kong Baptist University, Hong Kong*

Abstract

In social and behavioral sciences, the mediation test based on the indirect effect is an important topic. There are many methods to assess the intervening variable effects. In this paper, we focus on the difference method and the product method in mediation models. Firstly, we analyze the regression functions in the simple mediation model, and provide an expectation-consistent condition. We further show that the difference estimator and the product estimator are numerically equivalent based on the least squares regression regardless of the error distribution. Secondly, we generalize the equivalence result to the three-path model and the multiple mediators model, and prove a general equivalence result in a class of restricted linear mediation models. Finally, we investigate the empirical distributions of the indirect effect estimators in the simple mediation model by simulations, and show that the indirect effect estimators are normally distributed as long as one multiplicand of the product estimator is large.

Keywords: Bootstrap, Difference in coefficients, Empirical distribution, Indirect effects, Least squares regression, Product of coefficients

[*] Corresponding author
Email address: wengewsh@sina.com (WenWu Wang)

1. Introduction

In many disciplines, the effect of an exposure on the outcome variable is often mediated by an intermediate variable. Mediation analysis aims to identify the direct effect of the predictor on the outcome and the indirect effect between the same predictor and outcome via the change in a mediator (MacKinnon, 2008). Since the seminal paper of Baron and Kenny (1986), mediation analysis has become one of the most popular statistical methods in social sciences. For basic information on mediation analysis, one may refer to the recent books including MacKinnon (2008), Hayes (2013), Preacher (2015), and VanderWeele (2015). Empirical applications of mediation analysis have dramatically expanded in sociology, psychology, epidemiology, and medicine (Ogden et al., 2010; Lockhart et al., 2011; Rucker et al., 2011; Newland et al., 2013; Richiardi et al., 2013). Meanwhile, modern scientific investigations require sophisticated models for conducting mediation analysis (VanderWeele and Tchetgen, 2017; Frölich and Huber, 2017; Lachowicz et al., 2018).

One important issue in mediation studies is to infer the mediated effects. Various approaches have been proposed to test the statistical significance of a mediated effect. The first approach is the causal steps approach (Baron and Kenny, 1986), which specifies a series of tests of links in a causal chain. Some variants of this method that test three different hypotheses have been proposed (Allison, 1995; Kenny et al., 1998). The second approach is the difference in coefficients approach (Freedman and Schatzkin, 1992), which takes the difference between a regression coefficient before and after being adjusted by the intervening variable. The third approach is the product of coefficients approach which involves paths in a path model (Sobel, 1982; MacKinnon et al., 1998; MacKinnon and Lockwood, 2001). MacKinnon et al. (2002) gave a summary and comparison of the existing methods, and evaluated their performance via Monte Carlo simulations based on normally distributed data. For more references, see, for example, MacKinnon and Lockwood (2004), Preacher and Hayes (2008), and Preacher and Selig (2012).

In this paper, we focus on the total indirect effect based on the least squares regression. Firstly, we review the simple mediation model and some basic inference methods, provide an expectation-consistent condition for the model, and prove the equivalence between the difference and product estimators using the closed-form expressions. Secondly, we prove the equivalence between the difference and product estimators in the three-path model (Taylor et al., 2008) and in the multiple mediators model (Daniel et al., 2015; Tafuri et al., 2018). Thirdly, we prove a general result on the numerical equivalence between the two estimators in a general linear mediation model with restriction. Finally, we report some empirical distributions of the indirect effect estimators by simulations, and meanwhile point out some limitations of the existing inference methods by analyzing a real data on DNA methylation.

We emphasize that the main result of this paper is Theorem 6 in Section 4. To avoid understanding difficulties, we first use a simple mediation model and two more complex linear models to illustrate this result, and then state the general result in Theorem 6.

2. Simple Mediation Model

The simple mediation model is given in Figure 1, where X is the independent variable, Y is the dependent variable, and M is the mediating variable that mediates the effects of X on Y. Given the observations (X_i, M_i, Y_i) for $i = 1, \ldots, n$, the simple mediation model consists of three regression equations:

$$Y_i = \beta_0 + cX_i + \epsilon_{0,i}, \tag{1}$$

$$M_i = \beta_1 + aX_i + \epsilon_{1,i}, \tag{2}$$

$$Y_i = \beta_2 + c'X_i + bM_i + \epsilon_{2,i}, \tag{3}$$

where c represents the total effect of X on Y, a represents the relation between X and M, c' represents the direct effect of X on Y after adjusting the effect of M, and b represents the relation between M and Y after adjusting the effect of X.

For the simple mediation model, the mediated effect, also called the indirect effect, can be defined in two different forms: ab or $c - c'$. In general, the main goal of mediation analysis is to test whether the null hypothesis $H_0 : ab = 0$ or $H_0 : c - c' = 0$ is true. In this section, we compare the two forms of indirect effect in the least squares regression framework.

2.1. Zero-Mean Error Condition for Model Consistency

Note that the regression Equations (1)-(3) are interrelated in the simple mediation model. We substitute Equation (2) into Equation (3) to obtain

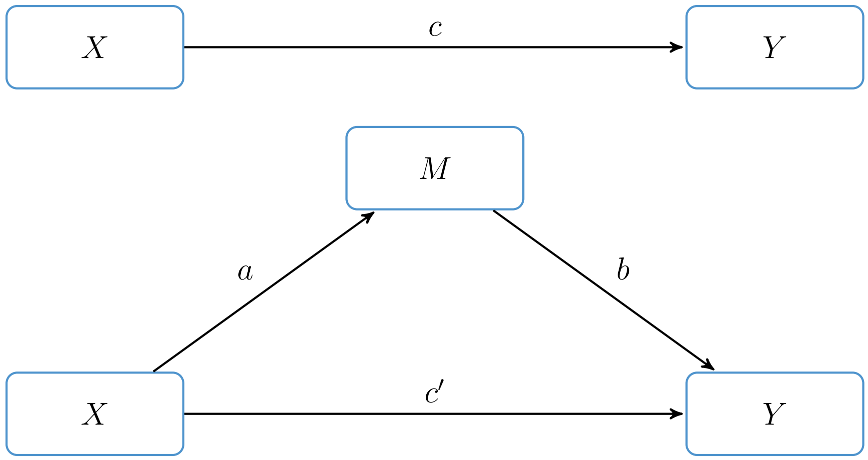

Figure 1: Causal diagram of the simple mediation model, where X is the independent variable, M is the mediating variable, and Y is the dependent variable.

the following equation:

$$\begin{aligned} Y_i &= \beta_2 + c'X_i + b(\beta_1 + aX_i + \epsilon_{1,i}) + \epsilon_{2,i} \\ &= (\beta_2 + b\beta_1) + (c' + ab)X_i + \epsilon_i, \end{aligned} \tag{4}$$

where $\epsilon_i = b\epsilon_{1,i} + \epsilon_{2,i}$. Assume also that $\epsilon_{1,i}$ and $\epsilon_{2,i}$ are zero-mean distributed, where "zero-mean" indicates their conditional means are zero, i.e., $\mathrm{E}[\epsilon_{1,i}|X_i] = 0$ and $\mathrm{E}[\epsilon_{2,i}|X_i] = 0$. Then consequently, ϵ_i is also zero-mean distributed by noting that $\mathrm{E}[\epsilon_i|X_i] = b\mathrm{E}[\epsilon_{1,i}|X_i] + \mathrm{E}[\epsilon_{2,i}|X_i] = 0$. Further by Equations (1) and (4), we have

$$\begin{aligned} \mathrm{E}[Y_i|X_i] &= \beta_0 + cX_i, \\ \mathrm{E}[Y_i|X_i] &= (\beta_2 + b\beta_1) + (c' + ab)X_i. \end{aligned}$$

This shows that $c = c' + ab$. The two expressions of $\mathrm{E}[Y_i|X_i]$ also imply that ϵ_i in (4) is equal to $\epsilon_{0,i}$ in (1).

Theorem 1. *For the simple mediation model, assume that $\epsilon_{j,i}$ are zero-mean distributed with $\mathrm{E}[\epsilon_{j,i}|X_i] = 0$ for $j = 1, 2$. Then we have the equality*

$$ab = c - c'.$$

In particular, if $\epsilon_{j,i} \overset{i.i.d.}{\sim} N(0, \sigma_j^2)$ for $j = 1, 2$, then they satisfy the zero-mean condition, where *i.i.d.* is an abbreviation of independent and identically distributed. And if we further assume that $\epsilon_{1,i}$ and $\epsilon_{2,i}$ are independent, then $\epsilon_{0,i}$ is also normally distributed with variance $\sigma_0^2 = b^2\sigma_1^2 + \sigma_2^2$.

2.2. Least Squares Regression

The standard mediation analysis uses the least squares regression to estimate the regression parameters. Specifically, by minimizing the sum of squared errors, we have

$$(\hat{\beta}_0, \hat{c})^T = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^TY = \arg\min_{\beta_0,c}\sum_{i=1}^{n}(Y_i - \beta_0 - cX_i)^2, \quad (5)$$

$$(\hat{\beta}_1, \hat{a})^T = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^TM = \arg\min_{\beta_1,a}\sum_{i=1}^{n}(M_i - \beta_1 - aX_i)^2, \quad (6)$$

$$(\hat{\beta}_2, \hat{c}', \hat{b})^T = (\check{X}^T\check{X})^{-1}\check{X}^TY = \arg\min_{\beta_2,c',b}\sum_{i=1}^{n}(Y_i - \beta_2 - c'X_i - bM_i)^2, \quad (7)$$

where $X = (X_1, \ldots, X_n)^T$, $M = (M_1, \ldots, M_n)^T$, $Y = (Y_1, \ldots, Y_n)^T$, $I = (1, \ldots, 1)^T$, $\tilde{X} = (I, X)$, and $\check{X} = (I, X, M)$. Moreover, if we assume that $\epsilon_{1,i} \overset{i.i.d.}{\sim} N(0, \sigma_1^2)$, $\epsilon_{2,i} \overset{i.i.d.}{\sim} N(0, \sigma_2^2)$, and they are independent. Then the least squares estimators in (5)-(7) follow the normal distributions as

$$\begin{aligned}
\hat{c} &\sim N(c, \sigma_0^2 e_{2,2}^T(\tilde{X}^T\tilde{X})^{-1}e_{2,2}),\\
\hat{a} &\sim N(a, \sigma_1^2 e_{2,2}^T(\tilde{X}^T\tilde{X})^{-1}e_{2,2}),\\
\hat{c}' &\sim N(c', \sigma_2^2 e_{2,3}^T(\check{X}^T\check{X})^{-1}e_{2,3}),\\
\hat{b} &\sim N(b, \sigma_2^2 e_{3,3}^T(\check{X}^T\check{X})^{-1}e_{3,3}),
\end{aligned}$$

where $e_{2,2} = (0,1)^T$, $e_{2,3} = (0,1,0)^T$, $e_{3,3} = (0,0,1)^T$, and $\sigma_0^2 = b^2\sigma_1^2 + \sigma_2^2$.

The above results are straightforward and hence is omitted. When the random errors are normally distributed, it is known that the least squares estimator is the most efficient estimator, the minimum variance unbiased estimator, and also the maximum likelihood estimator. In principle, the normality of the errors is too strong for model consistency, and the least squares estimator does not need the normality assumption, but only requires that the expectations of the errors are zero. As an example, the error distribution that is symmetric about zero satisfy the zero-mean condition.

2.3. Equivalence between the Difference and Product Estimators

The indirect effect of X on Y can be estimated by two methods: the difference of the estimated coefficients $\hat{c} - \hat{c}'$, and the product of the estimated coefficients $\hat{a}\hat{b}$. In this section, we show that the two methods provide the same estimate in mediation analysis.

Theorem 2. *In the simple mediation model, the difference estimator is equivalent to the product estimator, i.e.*

$$\hat{a}\hat{b} = \hat{c} - \hat{c}',$$

regardless of the error distribution.

The proof of Theorem 2 is provided in Appendix A. This theorem shows that, no matter what the error distribution is, the two estimators for the indirect effect are exactly the same in the closed expression of the least squares estimators. The equivalence of the two estimators is attributed to three facts: complete data, linear equation and least squares regression. If there are missing data, or if the model is multilevel or logistic, or if we apply the least absolute deviation or the other loss functions, then the equivalence between the two estimators will no longer hold.

MacKinnon et al. (1995) provided the result in Theorem 1, by explicitly deriving the formula of ab and $c - c'$; we provide an alternative proof without explicit derivation. MacKinnon et al. (1995) claimed the numerical equivalence of $\hat{a}\hat{b}$ and $\hat{c} - \hat{c}'$ by examining a few samples, but no rigorous proof was provided; Theorem 2 filled this gap.

2.4. Inference

In mediation analysis, the main aim is to test whether the estimated indirect effect is significantly different from zero. For the difference and product estimators, the test statistics can be constructed as

$$z_p = \frac{\hat{a}\hat{b}}{\hat{\sigma}_p} \tag{8}$$

and

$$z_d = \frac{\hat{c} - \hat{c}'}{\hat{\sigma}_d}, \tag{9}$$

where $\hat{\sigma}_p$ and $\hat{\sigma}_d$ are the standard errors of the two estimators, respectively. Since the two estimators are equivalent, their variances should satisfy $\sigma_p^2 = \sigma_d^2$, where σ^2 without a hat means the variance, and with a hat means an estimator of the variance. Under the assumption of normality, the variance estimation is a key step for inference. There are many works on the variance estimation.

For the product estimator, Sobel (1982) applied the multivariate delta method and proposed an approximate formula for the standard error as

$$\hat{\sigma}_p = \sqrt{\hat{a}^2\hat{\sigma}_b^2 + \hat{b}^2\hat{\sigma}_a^2},$$

For the difference estimator, Freedman and Schatzkin (1992) developed a method to study binary health measures and proposed to estimate the standard error of $\hat{c} - \hat{c}'$ by

$$\hat{\sigma}_d = \sqrt{\hat{\sigma}_c^2 + \hat{\sigma}_{c'}^2 - 2\sqrt{1-\hat{\rho}^2}\hat{\sigma}_c\hat{\sigma}_{c'}},$$

where $\hat{\rho}$ is the sample correlation between the independent variable X and the mediation variable M. For more estimators of the standard error, one may refer to MacKinnon et al. (2002). Thus the $(1-\alpha)$ confidence intervals of ab and $c - c'$ are

$$[\hat{a}\hat{b} - z_{1-\alpha/2}\hat{\sigma}_p, \hat{a}\hat{b} + z_{1-\alpha/2}\hat{\sigma}_p], \tag{10}$$

$$[\hat{c} - \hat{c}' - z_{1-\alpha/2}\hat{\sigma}_d, \hat{c} - \hat{c}' + z_{1-\alpha/2}\hat{\sigma}_d], \tag{11}$$

respectively, where α is the specified significance level, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution.

Remark 1. Although the estimators $\hat{a}$ and $\hat{b}$ are normally distributed, the product $\hat{a}\hat{b}$ is not normally distributed, no matter whether or not the two estimators are independent (Cui et al., 2016; Nadarajah and Pogány, 2016). Due to the equivalence of the two estimators, the distribution of the difference between $\hat{c}$ and $\hat{c}'$ is not normally distributed either, which implies that $\hat{c}$ and $\hat{c}'$ are not jointly normally distributed.

Since the sampling distributions of z_p and z_d are not normally distributed, but are skewed and leptokurtic in most cases (MacKinnon et al., 2002; MacKinnon and Lockwood, 2004; Preacher and Hayes, 2004, 2008; Sisbu-Sakarya et al., 2014), the tests based on the statistics (8) and (9) have low powers and are often criticized in the literature. Some remedies are proposed by MacKinnon et al. (1998), MacKinnon and Lockwood (2001) and Shrout and Bolger (2002). In general, there are two methods for inference of the indirect effect: the method based on the correct distribution of the product and the resampling method. Firstly, the distribution of the product strategy explores the correct distribution of $\hat{a}\hat{b}$ rather than assumes its normality. The distribution function of the product of two standardized normal variables are presented in Meeker et al. (1981). Preacher and Selig (2012) indicated: "This method performs well in simulation studies, but until recently required recourse to tables with limited availability and knowledge of the population values of either a or b". Secondly, the bootstrap is a popular resampling method to conduct inference (MacKinnon and Lockwood, 2004; Preacher and Hayes, 2008). To improve the finite-sample performance, other bias-corrected and bias-adjusted versions are also provided (MacKinnon, 2008; Preacher and Selig, 2012; Hayes, 2013). The Monte Carlo simulation is an alternative method to the bootstrap, which directly generates sample statistics from their joint distribution, not reasmpling the original data (MacKinnon and Lockwood, 2004; Preacher and Selig, 2012). The drawbacks to the resampling methods include slight inconsistency among replications of the same experiment with the same data due to resampling variability and no theoretical results to guarantee their asymptotic consistency.

3. Beyond Simple Mediation Model

In the previous section, we have focused on the simple mediation model with only one mediator, in which the only mediator transmits the influence of the independent variable to the dependent variable. In applications, the mediation chain with more than two paths or one mediator is also popular (Tein et al., 2000; Allen and Griffeth, 2001; Tekleab et al., 2005; Kim and Cicchetti, 2010; Nübold et al., 2015). In this section, we consider two such mediation models: the three-path mediation model (Taylor et al., 2008) and the multiple mediators model (Preacher and Hayes, 2008), and prove the equivalence of the two least squares estimators for indirect mediation effects.

3.1. Three-Path Mediation Model

In a three-path mediation model, two mediators M_1 and M_2 intervene in a series between an independent variable and a dependent variable (Taylor et al., 2008), which is depicted as a path diagram in Figure 2. It consists of

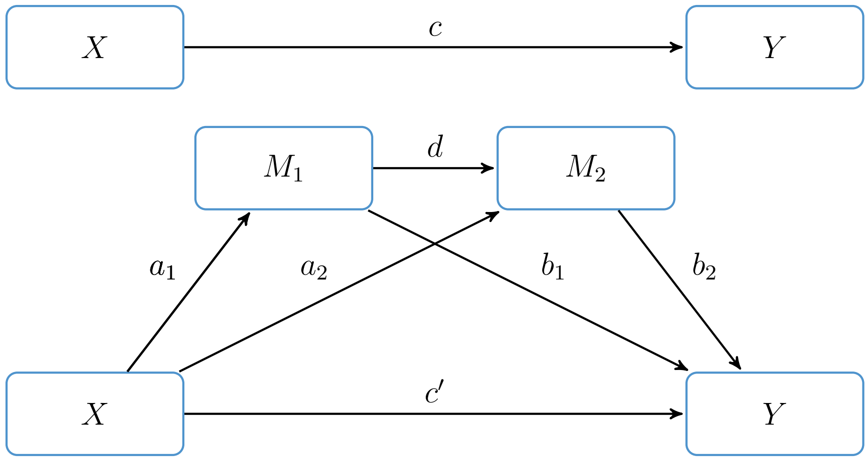

Figure 2: Causal diagram of three-path mediation model: X is the independent variable, M_1 and M_2 are two mediating variables, and Y is the dependent variable.

four regression equations:

$$Y_i = \beta_0 + cX_i + \epsilon_{0,i}, \tag{12}$$
$$M_{1,i} = \beta_1 + a_1 X_i + \epsilon_{1,i}, \tag{13}$$
$$M_{2,i} = \beta_2 + a_2 X_i + dM_{1,i} + \epsilon_{2,i}, \tag{14}$$
$$Y_i = \beta_3 + c' X_i + b_1 M_{1,i} + b_2 M_{2,i} + \epsilon_{3,i}, \tag{15}$$

where the coefficients can be similarly interpreted as in the simple mediation model.

The total indirect mediated effect, the effect passing through all paths, is defined as the sum of the product of the coefficients:

$$a_1 b_1 + a_2 b_2 + a_1 d b_2.$$

Taylor et al. (2008) indicated that: "Although it may be possible to develop a three-path test of mediation based on differences in coefficients, this method would likely be cumbersome in comparison to the product-of-coefficients test." As a result, the difference method is not considered in Taylor et al. (2008). In this subsection, we consider the indirect effect based on both the product and difference methods.

Following the discussion in Section 2, we have the similar equivalent relationship.

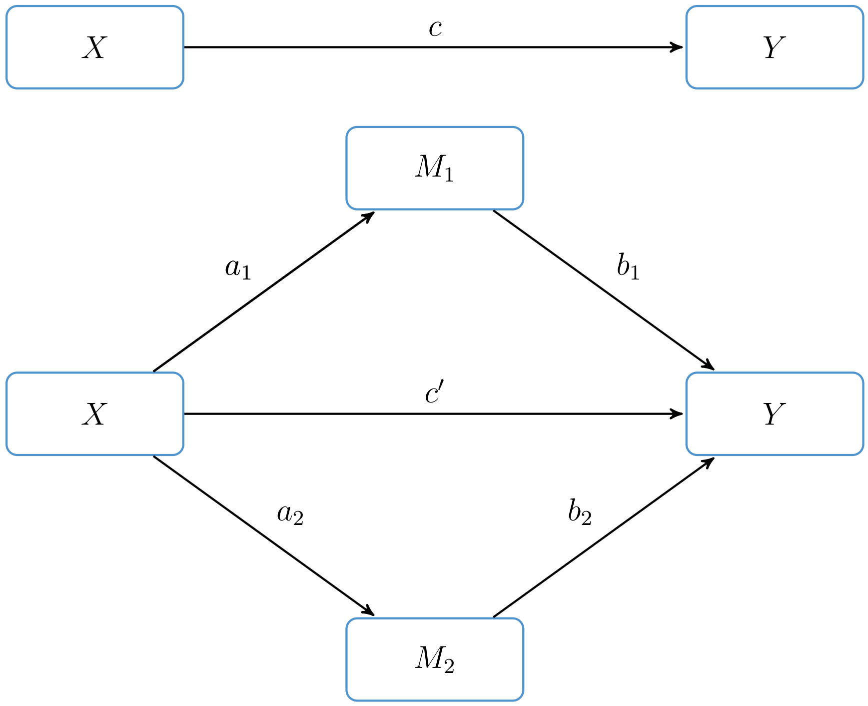

Figure 3: Causal diagram of mediation model with two mediators: X is the independent variable, M_1 and M_2 are two mediating variables, and Y is the dependent variable.

Theorem 3. *In the three-path mediation model, assume that* $\epsilon_{j,i}$ *for* $j = 1, 2, 3$ *are zero-mean distributed. Then, we have the equality*

(1) the parameters of the regression model satisfy the equality

$$c - c' = a_1 b_1 + a_2 b_2 + a_1 d b_2;$$

(2) the least squares estimates of parameters satisfy the equality

$$\hat{c} - \hat{c}' = \hat{a}_1 \hat{b}_1 + \hat{a}_2 \hat{b}_2 + \hat{a}_1 \hat{d} \hat{b}_2.$$

The proof of Theorem 3 is provided in Appendixes B.

3.2. Multiple Mediators Model

In this subsection, we consider the general mediation model with multiple mediators (Preacher and Hayes, 2008). For simplicity, we consider only two mediators model with mediators M_1 and M_2, which can be expressed in the form of four regression equations:

$$Y_i = \beta_0 + cX_i + \epsilon_{0,i}, \tag{16}$$
$$M_{1,i} = \beta_1 + a_1X_i + \epsilon_{1,i}, \tag{17}$$
$$M_{2,i} = \beta_2 + a_2X_i + \epsilon_{2,i}, \tag{18}$$
$$Y_i = \beta_3 + c'X_i + b_1M_{1i} + b_2M_{2i} + \epsilon_{3,i}. \tag{19}$$

This form of the model is a special case of the three-path model with d in Equation (14) equal to zero. The total indirect mediated effect, the effect passing through either mediator, is defined as the sum of the product of the coefficients:

$$a_1b_1 + a_2b_2.$$

The similar equivalent relationship between the product and difference estimators is obtained.

Theorem 4. *In the two mediators model, assume that $\epsilon_{j,i}$ for $j = 1, 2, 3$ are zero-mean distributed. Then, we have the equality*

(1) the parameters of the regression model satisfy the equality

$$c - c' = a_1b_1 + a_2b_2;$$

(2) the least squares estimates of parameters satisfy the equality

$$\hat{c} - \hat{c}' = \hat{a}_1\hat{b}_1 + \hat{a}_2\hat{b}_2.$$

The proof of Theorem 4 is simpler than that of Theorem 3, and thus is omitted.

In general, the estimation equality still holds for $k > 2$ mediators.

Corollary 5. *In the multiple mediators model with $k > 2$ mediators, we have the estimation equality*

$$\hat{c} - \hat{c}' = \sum_{j=1}^{k} \hat{a}_j\hat{b}_j,$$

where $\hat{a}_j\hat{b}_j$ is the estimated indirect effect through mediator M_j.

This corollary is a special case of the general result in the next subsection.

4. General Linear Mediation Models

There are many linear mediation models with more than two mediators or more than two paths. A mediation graph consists of a set V of vertices and a set E of edges that connect some pairs of vertices (Pearl, 2009). The vertices in mediation graphs correspond to variables including the independent variable X, the dependent variable Y and the mediating variables M, and the edges denote a certain linear relationship between pairs of variables in linear mediation models. A path is defined as a sequence of edges (e.g. $((X, M_1), (M_1, M_2), (M_2, Y)))$ that start from X and end at Y, and each edge starts with the vertex ending the preceding edge. In the general linear mediation models, we assume all pathes start from X; in other words, M_j cannot start a path.

In this section, we consider the cases where each edge is directed, which means that each edge in a path is an arrow that points from the first to the second vertex of the pair. However, the mediation graph is restricted to be acyclic, i.e., contains no directed cycles (e.g., $X \to M$, $M \to X$) and no self-loops ($M \to M$). Now a specific group of linear regression equations is one-to-one to a mediation graph.

Based on the discussions in the previous subsection, we provide a theorem on the equivalence between the difference and product estimators in the general linear mediation model.

Theorem 6. *In a linear mediation model, if*

(i) the mediation graph is acyclic;

(ii) the errors are the zero-mean distributed;

(iii) each M_j equation contains X as a regressor;

(iv) the estimation method is the least squares regression.

Then the estimates of parameters satisfy the equivalence relationship: the difference estimator equals the sum of the product of the estimated parameters in each path.

The assumption (*iii*) that each M_j equation contains X as a regressor cannot be dropped; see the simulation in the following Section 4.2 for an illustration. The proof of Theorem 6 is provided in Appendixes C.

5. Simulation Studies

In this section, we conduct simulations to illustrate the empirical distributions of the indirect effect estimators. The first subsection consider the simple mediation model and the second subsection considers the three-path mediation model in Figure 2 where the edge from X to M_2 is deleted. We do not intend to evaluate test methods under the assumption of normal error distribution.

5.1. Empirical Distribution in Simple Mediation Model

In this simulation study, the independent variable and error are generated from the standard normal distribution independently. The values of ab were chosen to be zero (0), medium (0.02, -0.02), and large (2), corresponding to the cases where (a, b) is equal to $(0, 0)$/ $(0.2, 0.1)$, $(0.02, 1)$/ $(-0.2, 0.1)$, $(-2, 0.01)$ and $(2, 1)$, respectively. The sample size is 100, and the number of replications is set to be 10000 for each case.

Figure 4 shows the empirical distributions of the indirect effect estimators in red color. For comparison, we plot the corresponding normal distributions with the same mean and variance as the empirical distributions. These two distributions may be very different. For the zero indirect effect with $(a, b) = (0, 0)$, the empirical distribution is not normally distributed, because it has a sharper peak than the corresponding normal distribution. For the small indirect effects with $(a, b) = (0.2, 0.1)$ and $(-0.2, 0.1)$, the empirical distributions are skewed, right-skewed for the positive effect and left-skewed for the negative effect. While for the small indirect effects with $(a, b) = (0.02, 1)$ and $(-2, 0.01)$, the empirical distributions are still normally distributed. For the large indirect effect with $(a, b) = (2, 1)$, the empirical distribution is also normally distributed. In one word, the empirical distribution is normally distributed as long as one of the values of a and b is large in the simple mediation model.

5.2. Empirical Distributions when X is Not a Regressor of an M Equation

In the three-path mediation model of Figure 2, suppose there is no edge from X to M_2. Other parameters are set as $\beta_1 = \beta_2 = \beta_3 = a_1 = b_1 = b_2 = c' = d = 1$ for simplicity, and the errors ϵ_1, ϵ_2 and ϵ_3 are independently normal distributed with the same variance 1. The sample sizes are set at $n = 100$ and 1000, and the number of replications is set to be 10000.

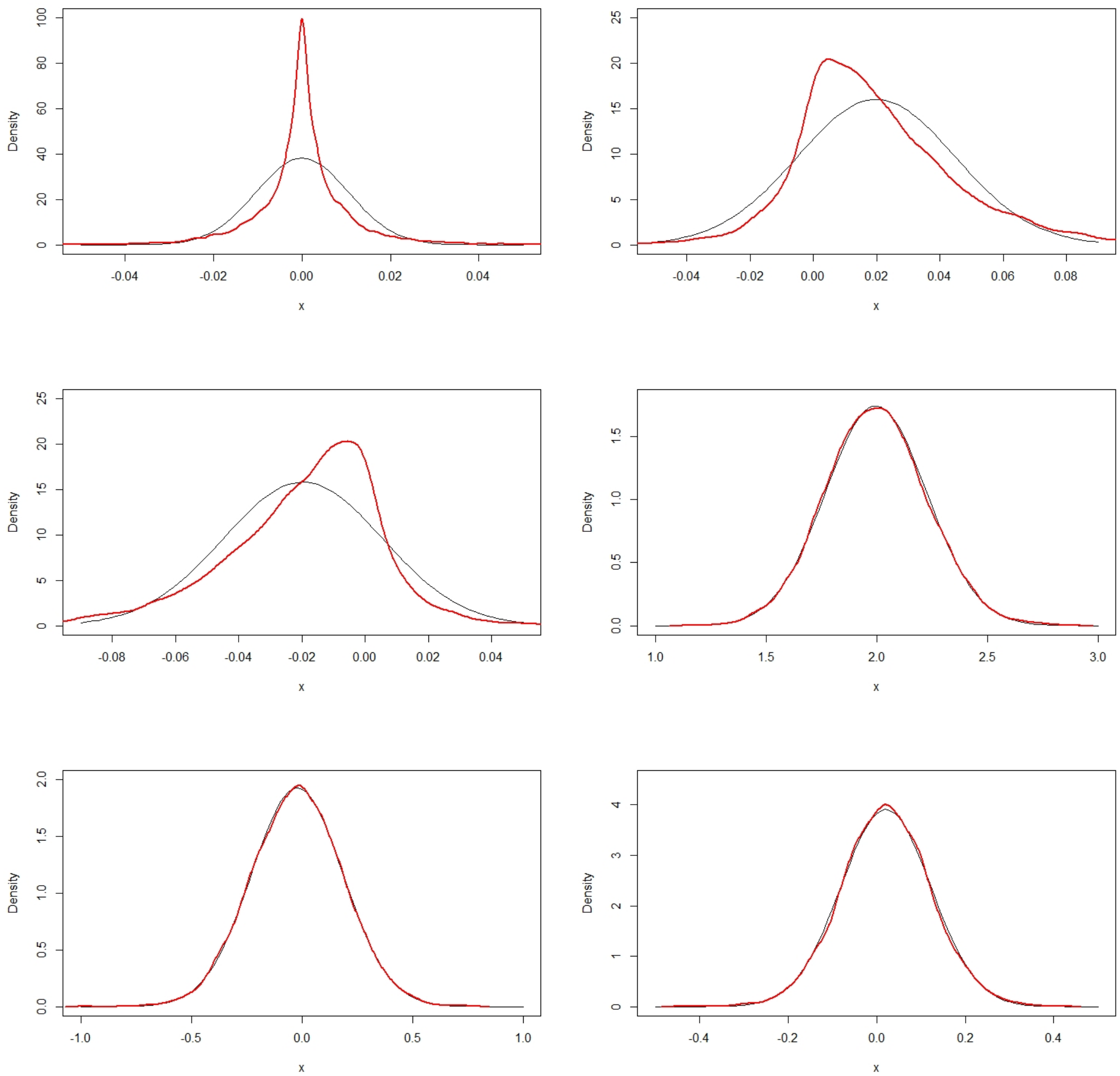

Figure 4: The empirical distribution (red) vs the normal distribution (black) with the same mean and variance as the empirical distribution: the upper-left, upper-right, middle-left, middle-right, lower-left and lower-right figures correspond to the cases where (a, b) is equal to $(0, 0)$, $(0.2, 0.1)$, $(-0.2, 0.1)$, $(2, 1)$, $(-2, 0.01)$ and $(0.02, 1)$, respectively.

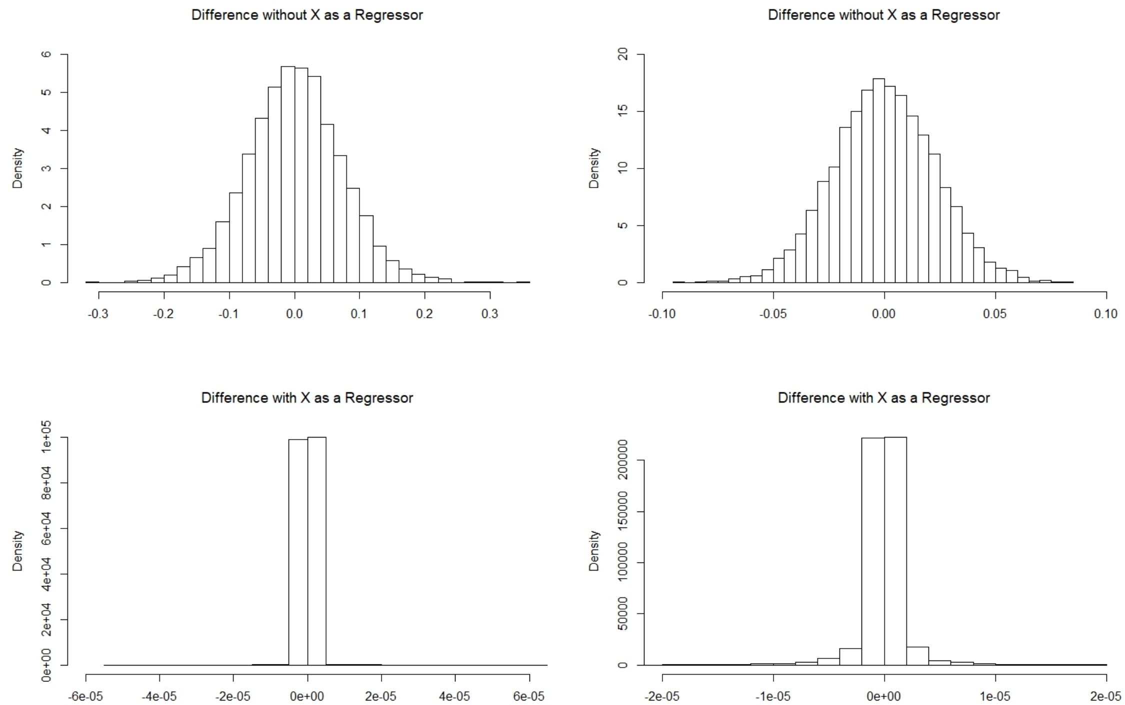

Figure 5: Histograms of the difference between the difference and product estimators with and without X as a regressor: the upper two graphs and lower two graphs correspond to $n = 100$ and $n = 1000$, respectively; the left two graphs and right two graphs is with and without a_2, respectively.

The upper two graphs in Figure 5 show the histograms of the difference between the difference and product estimators when $n = 100$ and 1000. For comparison, we also report in the lower two graphs the histograms of the difference between the two estimators when the edge from X to M_2 is added in and a_2 is set at zero. It shows that the two estimators with X as a regressor are numerically equivalent, while they are not without X as a regressor. Simulation results match the prediction of Theorem 6. Specifically, the mean and standard error for the difference between the two estimators are computed as $(-1.17 \times 10^{-7}, 5.29 \times 10^{-6})$, $(2.22 \times 10^{-3}, 7.24 \times 10^{-2})$, $(1.05 \times 10^{-9}, 3.00 \times 10^{-6})$ and $(1.45 \times 10^{-4}, 2.27 \times 10^{-2})$, which correspond to the cases: $n = 100$ without X, $n = 1000$ and without X, $n = 100$ and with X, $n = 1000$ and with X, respectively.

6. Real Data Analysis

In this section, we apply the least squares methods to a real data set to estimate the indirect effects (IE) of socioeconomic status (SES) on body mass index (BMI) that might be mediated by DNA methylation CpG sites on chromosome 17, where SES is quantified by a scalar index ranging from 0 to 100, and BMI is a body fatness index of an individual as in Loucks et al. (2016). To compare the two estimators, we choose three possible continuous mediators from DNA methylation: cg05157340, cg05156120 and cg05157970, take SES as the exposure X, and BMI as the outcome Y.

For the simple mediation analysis, the IEs are estimated using the least squares method in Section 2, the standard errors (SE) are estimated by the formulas $\hat{\sigma}_p$ and $\hat{\sigma}_d$ for product and difference estimators, and the 95% confidence intervals (CI) based on variance estimation (which are denoted by CI_1) are constructed using formulas (10) and (11), and the bootstrap CIs (which are denoted by CI_2) are constructed as

$$[2\hat{a}\hat{b} - (\hat{a}^*\hat{b}^*)_{0.975}, 2\hat{a}\hat{b} - (\hat{a}^*\hat{b}^*)_{0.025}],$$

where $(\hat{a}^*\hat{b}^*)_q$ is the qth quantile of the bootstrap distribution of $\hat{a}\hat{b}$, based on 1000 bootstrap samples. Table 1 summarizes the estimated IE values, SEs and the two 95% CIs. It can be seen that the difference and product estimators are numerically equivalent, while their SEs are different because different formulas are employed, which results in different CIs and different inference conclusions, e.g., when the mediator is cg05156120, we cannot reject the null that the indirect effect is zero based on IE_1 but can reject based on IE_2. Bootstrap CIs are different from those based on variance estimate, although the inference conclusions are the same in this example.

7. Discussions

In the literature, there are two estimation methods for the indirect mediated effect: the difference method and the product method. Most researchers recommend using the product form as the measure of the indirect effect, because it is in line with the causal interpretation of the mediation effect (MacKinnon et al., 2007; Pearl, 2012; Yuan and MacKinnon, 2014).

In this paper, we provide an identification condition for expectation consistency in the simple mediation model, prove that the difference estimator

	Estimate	SE	95% CI_1	95% CI_2
		cg05157340		
IE_1	-0.0056	0.0056	[-0.0166, 0.0054]	[-0.0337, 0.0220]
IE_2	-0.0056	0.0205	[-0.0457, 0.0345]	[-0.0381, 0.0222]
		cg05156120		
IE_1	-0.0227	0.0104	[-0.0373, -0.0080]	[-0.0465, 0.0106]
IE_2	-0.0227	0.0156	[-0.0532, 0.0078]	[-0.0462, 0.0110]
		cg05157970		
IE_1	-0.0922	0.0174	[-0.1264, -0.0580]	[-0.1431, -0.0265]
IE_2	-0.0922	0.0272	[-0.1455, -0.0388]	[-0.1410, -0.0303]

Table 1: The indirect effects in the SES-BMI data: IE_1 and IE_2 are the difference and product estimators, CI_1 and CI_2 are the confidence intervals based on variance estimation and the bootstrap, respectively.

is numerically equivalent to the product estimator in the least squares regression, and summarize the statistical theories. One interesting finding is that the equivalence of the two estimators depends only on the estimation method-least squares, not on the error distribution. Furthermore, the equivalence can be generalized to the three-path mediation model and the multiple mediators model.

Since the two estimators are equivalent, they should have the same distribution. However, inferences based on the two estimators may be different, and our real data analysis indicates this phenomenon. MacKinnon et al. (2002), MacKinnon and Lockwood (2004) and Preacher and Hayes (2004) made extensive simulations based on normal errors to assess their Type I error rate and the power, and recommended to use the distribution of the product estimator. These empirical results depend on the assumption of normal error distribution. As far as we know, there are not asymptotic results for either estimator, and thus the performance of empirical studies is not well understood. The mathematical expressions for the indirect effect estimators pave a way to develop the asymptotic theory, which is helpful to the statistical inference.

In practical applications, violations of normality commonly encountered include heavy tails, skewness, outliers, contamination, and multimodality. Micceri (1989) examined 440 data sets from the psychological and educational literature, including 125 psychometric measures such as scales measuring personality, anxiety, and satisfaction. None of these data sets are normally distributed at the $\alpha = 0.01$ significance level; rather, the distributions were often heavy-tailed and skewed. In order to further improve the estimation efficiency and the power for the hypothesis testing, we can apply the composite quantile regression of Zou and Yuan (2008), the weighted quantile average regression of Zhao and Xiao (2014), and the difference method of Wang et al. (2019) to the mediation model. These works deserve further investigation.

References

Allen, D., Griffeth, R., 2001. Test of a mediated performance-turnover relationship highlighting the moderating roles of visibility and reward contingency. Journal of Applied Psychology 86, 1014–1021.

Allison, P., 1995. The impact of random predictors on comparison of coefficients between models: Comment on Clogg, Petkova, and Haritou. American Journal of Sociology 100, 1294–1305.

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

Cui, G., Yu, X., Iommelli, S., Kong, L., 2016. Exact distribution for the product of two correlated gaussian random variables. IEEE Signal Processing Letters 23, 1662–1666.

Daniel, R., De Stavla, B., Cousens, S., Vansteelandt, S., 2015. Causal mediation analysis with multiple mediators. Biometrics 71, 1–14.

Freedman, L., Schatzkin, A., 1992. Sample size for studying intermediate endpoints within intervention trials of observational studies. American Journal of Epidemiology 136, 1148–1159.

Frölich, M., Huber, M., 2017. Direct and indirect treatment effects-causal chains and mediation analysis with instrumental variables. Journal of the Royal Statistical Society: Series B 79, 1645–1666.

Hayes, A., 2013. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York: Guilford Press.

Kenny, D., Kashy, D., Bolger, N., 1998. Data analysis in social psychology. In: Gilbert, D., Fiske, S., Lindzey, G. (Eds.), The Handbook of Social Psychology. Boston: McGraw-Hill, pp. 233–265.

Kim, J., Cicchetti, D., 2010. Longitudinal pathways linking child maltreatment, emotion regulation, peer relations, and psychopathology. Journal of Child Psychology and Psychiatry 51, 706–716.

Lachowicz, M., Preacher, K., Kelley, K., 2018. A novel measure of effect size for mediation analysis. Psychological Methods 23, 244–261.

Lockhart, G., MacKinnon, D., Ohlrich, V., 2011. Mediation analysis in psychosomatic medicine research. Psychosomatic Medicine 73, 29–43.

Loucks, E., Huang, Y., Agha, G., Chu, S., Eaton, C., Gilman, S., Buka, S., Kelsey, K., 2016. Epigenetic mediators between childhood socioeconomic disadvantage and mid-life body mass index: the new england family study. Psychosomatic Medicine 78, 1053–1065.

MacKinnon, D., 2008. Introduction to Statistical Mediation Analysis. New York: Taylor & Francis Group.

MacKinnon, D., Fairchild, A., Fritz, M., 2007. Mediation analysis. Annual Review of Psychology 58, 593–614.

MacKinnon, D., Lockwood, C., 2001. Distribution of products tests for the mediated effect. Technical Report, Arizona State University, USA.

MacKinnon, D., Lockwood, C., Hoffman, J., 1998. A new method to test for mediation. The Annual Meeting of the Society for Prevention Research, Park City, USA.

MacKinnon, D., Lockwood, C., Hoffman, J., West, S., Sheets, V., 2002. A comparison of methods to test mediation and other intervening variable effects. Psychological Methods 7, 83–104.

MacKinnon, D., Lockwood, C.M. Williams, J., 2004. Confidence limits for the indirect effect: Distribution of the product and resampling methods. Multivariate Behavioral Research 39, 99–128.

MacKinnon, D., Warsi, G., Dwyer, J., 1995. A simulation study of mediated effect measures. Multivatiate Behavioral Research 30, 41–62.

Meeker, W., Cornwell, L., Aroian, L., 1981. Selected Table in Mathematical Statistics: The Product of Two Normally Distributed Random Variables. Providence, RI: American Mathematical Society.

Micceri, T., 1989. The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin 105, 156–166.

Nadarajah, S., Pogány, T., 2016. On the distribution of the product of correlated normal random variables. Comptes Rendus de l'Académie des Sciences, Series I 354, 201–204.

Newland, R., Crnic, K., Cox, M., Mills-Koonce, W., Investigators, F. L. P. K., 2013. The family model stress and maternal psychological symptoms: Mediated pathways form economic hardship to parenting. Journal of Family Psychology 27, 96–105.

Nübold, A., Dörr, S., Maier, G., 2015. Considering the orphan: Personal identification and its relations with transformational leadership, trust, and performance in a three-path mediation model. Leadership 11, 230–254.

Ogden, C., Carroll, M., Curtin, L., Lamb, M., Flegal, K., 2010. Prevalence of high body mass index in U.S. children and adolescents, 2007-2008. Journal of the American Medical Association 303, 242–249.

Pearl, J., 2009. Causality: Models, Reasoning and Inference, 2nd Edition. Cambridge University Press, New York, USA.

Pearl, J., 2012. In Causality: Wiley Series in Probability and Statistics. Wiley Blackwell, pp. 151–179.

Preacher, K., 2015. Advances in mediation analysis: A survey and synthesis of new developments. Annual Review of Psychology 66, 825–852.

Preacher, K., Hayes, A., 2004. SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behavior Research Methods, Instruments, & Computers 36, 717–731.

Preacher, K., Hayes, A., 2008. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behavior Research Methods 40, 879–891.

Preacher, K., Selig, J., 2012. Advantages of Monte Carlo confidence intervals for indirect effects. Communication Methods and Measures 6, 77–98.

Richiardi, L., Bellocco, R., Zugna, D., 2013. Mediation analysis in epidemiology: Methods, interpretation and bias. International Journal of Epidemiology 42, 1511–1519.

Rucker, D., Preacher, K., Tormala, Z., Petty, R., 2011. Mediation analysis in social psychology: Current practices and new recommendations. Social and Personality Psychology Compass 5, 359–371.

Shrout, P., Bolger, N., 2002. Mediation in experimental and nonexperimental studies: New procedures and recommendations. Psychological Methods 7, 422–445.

Sisbu-Sakarya, Y., MacKinnon, D., Miočević, 2014. The distribution of the product explains normal theory mediation confidnce interval estimation. Multivariate Behavioral Research 49, 261–268.

Sobel, M., 1982. Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology 13, 290–312.

Tafuri, M., Featherstone, J., Cheng, J., 2018. Causal mediation analysis with multiple causally non-ordered mediator. Statistical Methods in Medical Research 27, 3–19.

Taylor, A., MacKinnon, D., Tein, J., 2008. Tests of the three-path mediated effect. Organizational Research Methods 11, 241–269.

Tein, J., Sandler, I., Zautra, A., 2000. Stressful life events, psychological distress, coping, and parenting of divorced mothers: A longitudinal study. Journal of Family Psychology 14, 27–41.

Tekleab, A., Bartol, K., Liu, W., 2005. Is it pay level or pay raises that matter to fairness and turnover? Journal of Organizational Behavior 26, 899–921.

VanderWeele, T., 2015. Explanation in Causal Inference: Methods for Mediation and Interaction. Oxford: Oxford University Press.

VanderWeele, T., Tchetgen, E., 2017. Mediation analysis with time varying exposures and mediators. Journal of the Royal Statistical Society: Series B 79, 917–938.

Wang, W., Yu, P., Lin, L., Tong, T., 2019. Robust estimation of derivatives using locally weighted least absolute deviation regression. Journal of Machine Learning Research 20 (60), 1–49.

Yuan, Y., MacKinnon, D., 2014. Robust mediation analysis based on median regression. Psychological Methods 19, 1–20.

Zhao, Z., Xiao, Z., 2014. Efficient regression via optimally combining quantile information. Econometric Theory 30, 1272–1314.

Zou, H., Yuan, M., 2008. Composite quantile regression and the oracle model selection theory. The Annals of Statistics 36, 1108–1126.

Appendix A: Proof of Theorem 2

Proof. We first consider the simple case where $\beta_0 = \beta_1 = \beta_2 = 0$. The least squares estimators for the simplified models are

$$\begin{aligned}
\hat{c} &= \arg\min_c \sum_{i=1}^n (Y_i - cX_i)^2 = \frac{X^T Y}{X^T X}, \\
\hat{a} &= \arg\min_a \sum_{i=1}^n (M_i - aX_i)^2 = \frac{X^T M}{X^T X}, \\
(\hat{c}', \hat{b})^T &= \arg\min_{c',b} \sum_{i=1}^n (Y_i - c'X_i - bM_i)^2 = \frac{Y\begin{pmatrix} M^T M X^T - X^T M M^T \\ X^T X M^T - X^T M X^T \end{pmatrix}}{X^T X M^T M - X^T M X^T M}.
\end{aligned}$$

where $X = (X_1, \ldots, X_n)^T$, $M = (M_1, \ldots, M_n)^T$, and $Y = (Y_1, \ldots, Y_n)^T$. By the above least squares estimators, the difference estimator is

$$\hat{c} - \hat{c}' = \frac{X^T M (X^T X M^T - X^T M X^T) Y}{X^T X (X^T X M^T M - X^T M X^T M)},$$

and the product estimator is

$$\hat{a}\hat{b} = \frac{X^T M (X^T X M^T - X^T M X^T) Y}{X^T X (X^T X M^T M - X^T M X^T M)}.$$

This shows that $\hat{c} - \hat{c}' = \hat{a}\hat{b}$. That is, the difference estimator is equivalent to the product estimator for the linear regression models with zero intercept.

The proof can readily be generalized to the models with non-zero intercept and so is omitted.

Appendix B: Proof of Theorems 3

Proof. By substituting Equations (13) and (14) into Equation (15), it follows that

$$\begin{aligned}
Y_i &= \beta_3 + c'X_i + b_1 M_{1,i} + b_2(\beta_2 + dM_{1,i} + a_2 X_i + \epsilon_{2,i}) + \epsilon_{3,i} \\
&= (\beta_3 + b_2\beta_2) + (c' + a_2 b_2)X_i + (b_1 + b_2 d)M_{1,i} + (b_2\epsilon_{2,i} + \epsilon_{3,i}) \\
&= (\beta_3 + b_2\beta_2) + (c' + a_2 b_2)X_i + (b_1 + b_2 d)(\beta_1 + a_1 X_i + \epsilon_{1,i}) + (b_2\epsilon_{2,i} + \epsilon_{3,i}) \\
&= (\beta_3 + b_2\beta_2 + (b_1 + b_2 d)\beta_1) + (c' + a_2 b_2 + a_1 b_1 + a_1 d b_2)X_i + \epsilon_i, \qquad (20)
\end{aligned}$$

where $\epsilon_i = (b_1 + b_2 d)\epsilon_{1,i} + b_2\epsilon_{2,i} + \epsilon_{3,i}$. Since $\epsilon_{j,i}$ for $j = 1, 2, 3$ are zero-mean distributed, ϵ_i is also zero-mean distributed with $\mathrm{E}[\epsilon_i|X_i] = 0$. Taking expectation of Equations (12) and (20), we have

$$
\begin{aligned}
&E[Y_i|X_i] = \beta_0 + cX_i,\\
&E[Y_i|X_i] = (\beta_3 + b_2\beta_2 + (b_1 + b_2 d)\beta_1) + (c' + a_1 b_1 + a_2 b_2 + a_1 d b_2)X_i.
\end{aligned}
$$

This leads to $\beta_0 = \beta_3 + b_2\beta_2 + (b_1 + b_2 d)\beta_1$ and $c - c' = a_1 b_1 + a_2 b_2 + a_1 d b_2$.

For the simplified models with $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$, the least squares estimators are

$$
\begin{aligned}
\tilde{c} &= \arg\min_{c}(Y_i - cX_i)^2 = \frac{A_7}{A_1},\\
\tilde{a}_1 &= \arg\min_{a}(M_{1,i} - a_1 X_i)^2 = \frac{A_2}{A_1},\\
(\tilde{a}_2, \tilde{d})^T &= \arg\min_{a_2,d}(M_{2,i} - a_2 X_i - dM_{1,i})^2 = \frac{\begin{pmatrix} A_3A_4 - A_2A_{10} \\ A_1A_{10} - A_2A_3 \end{pmatrix}}{A_1A_4 - A_2A_2},\\
(\tilde{c}', \tilde{b}_1, \tilde{b}_2)^T &= \arg\min_{c',b_1,b_2}(Y_i - c'X_i - b_1M_{1,i} - b_2M_{2,i})^2\\
&= \frac{\begin{pmatrix} (A_4A_6 - A_5A_5)A_7 + (A_3A_5 - A_2A_6)A_8 + (A_2A_5 - A_3A_4)A_9 \\ (A_3A_5 - A_2A_6)A_7 + (A_1A_6 - A_3A_3)A_8 + (A_2A_3 - A_1A_5)A_9 \\ (A_2A_5 - A_3A_4)A_7 + (A_2A_3 - A_1A_5)A_8 + (A_1A_4 - A_2A_2)A_9 \end{pmatrix}}{A_1(A_4A_6 - A_5A_5) + A_2(A_3A_5 - A_2A_6) + A_3(A_2A_5 - A_3A_4)}.
\end{aligned}
$$

where $A_1 = X^TX$, $A_2 = X^TM_1$, $A_3 = X^TM_2$, $A_4 = M_1^TM_1$, $A_5 = M_1^TM_2$, $A_6 = M_2^TM_2$, $A_7 = X^TY$, $A_8 = M_1^TY$, $A_9 = M_2^TY$, and $A_{10} = M_1^TM_2$. By the above least squares estimators, it is easy to verify that $\tilde{c} - \tilde{c}' = \tilde{a}_1\tilde{b}_1 + \tilde{a}_2\tilde{b}_2 + \tilde{a}_1\tilde{d}\tilde{b}_2$. That is, the difference estimator is equivalent to the product estimator for the linear regression models with zero intercept.

The proof can readily be generalized to the models with non-zero intercept and so is omitted.

Appendix C: Proof of Theorem 6

Proof. Suppose there are k mediators:

$$
\begin{aligned}
M_j &= \beta_j + R_j^T a_j + \epsilon_j, \quad j = 1, \ldots, k,\\
Y &= \beta_{k+1} + Xc' + R_Y^T b + \epsilon_{k+1},
\end{aligned}
$$

where R_j contains the non-constant regressors in the equation for M_j, which may include X and/or other M_j's, and other symbols can be similarly understood. After substituting the equations for $M_j, j = 1, \ldots, k$, into the equation for Y, suppose we have

$$Y = \alpha_0 + (c' + \alpha_1)X + \epsilon_0 \equiv \alpha_0 + cX + \epsilon_0,$$

where (α_0, α_1) are functions of the coefficients in the equations for $\{M_j\}_{j=1}^k$ and Y, i.e.,

$$\begin{aligned}\alpha_0 &= f(\beta_1, \ldots, \beta_{k+1}; a_1, \ldots, a_k, b),\\ \alpha_1 &= f(a_1, \ldots, a_k, b),\end{aligned}$$

α_1 does not depend on $\beta_1, \ldots, \beta_{k+1}$ because it measures the sensitivity of Y to X while $\beta_1, \ldots, \beta_{k+1}$ does not contain such information, c' is the coefficient of X in the equation for Y, and ϵ_0 is a linear combination of the error terms in these $(k+1)$ equations, so it satisfies $\mathrm{E}[\epsilon_0|X] = 0$.

Because all the coefficients are estimated by least squares regression, they employ the moment conditions

$$\mathrm{E}\left[\begin{pmatrix}1\\ R_j\end{pmatrix}(M_j - \beta_j + R_j^T a_j)\right] = 0,$$
$$\mathrm{E}\left[\begin{pmatrix}1\\ X\\ R_j\end{pmatrix}(Y - \beta_{k+1} - Xc' - R_Y^T b)\right] = 0.$$

If these moment conditions imply

$$\mathrm{E}\left[\begin{pmatrix}1\\ X\end{pmatrix}(Y - \alpha_0 - cX)\right] = 0,$$

then our result follows since we just replace $\mathrm{E}[\cdot]$ by $1/n\sum_{i=1}^n$ in the least squares estimation. However, this indeed holds because ϵ_0 is a linear function of $\{\epsilon_j\}_{j=1}^{k+1}$ so that

$$\mathrm{E}\left[\begin{pmatrix}1\\ X\end{pmatrix}\epsilon_j\right] = 0$$

implies

$$\mathrm{E}\left[\begin{pmatrix}1\\ X\end{pmatrix}\epsilon_0\right] = 0.$$

Here, note that X must be a regressor in the M_j equation, otherwise $\mathrm{E}\left[\begin{pmatrix} 1 \\ X \end{pmatrix} \epsilon_j\right] = 0$ cannot hold such that $\mathrm{E}\left[\begin{pmatrix} 1 \\ X \end{pmatrix} \epsilon_0\right] = 0$ cannot hold and the equivalence result fails.