Notes for Causal Mediation Analysis

Xi Chen, Rotterdam School of Management, Erasmus University, chen@rsm.nl

April 14, 2024

Abstract

This note is about causal mediation analysis and what ideas can be developed in this area for behavioral researchers.

Contents

1 Conventional Mediation Analysis

Conventional mediation analysis [Baron and Kenny, 1986] is formulated under a linear structural equation model (LSEM). Let's first introduce some notations. Suppose we have a binary treatment with $D_i = \{0, 1\}$[1]. The focal mediator is represented as M_i and the final outcome is Y_i. The DAG for the conventional mediation analysis is as below.

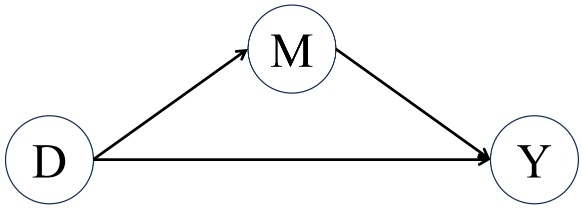

Figure 1: The DAG for Conventional Mediation Analysis

The DAG (implicitly) makes two assumptions that form the basis of the conventional mediation analysis.

Assumption 1. *The treatment D is unconfonded w.r.t the mediator M and outcome Y.*

The first assumption is that the treatment D is unconfounded. Or no confounders exist between D and M and also D and Y. A stronger version of this assumption is D is independent from the $\sigma(M, Y)$, the σ-algebra of M and Y.

Assumption 2. *The focal mediator M, conditional on D, is unconfounded w.r.t. outcome Y.*

The second assumption states that no confounders exist between M and Y, except for the treatment D. Assumption 1 and 2 are known as the "sequential ignorability" assumption in Imai et al. [2010]. Given the assumptions, the LSEM consists of 3 equations:

$$\begin{cases} Y_i &= \alpha_1 + \beta_1 D_i + e_{i1} \\ M_i &= \alpha_2 + \beta_2 D_i + e_{i2} \\ Y_i &= \alpha_3 + \beta_3 D_i + \beta_4 M_i + e_{i3} \end{cases} \tag{1}$$

The test statistics of the mediation effect is then constructed by comparing the total effect vs. the direct effect of the treatment, i.e., $\beta_1 - \beta_3$, or with a mathematically equivalent test statistics $\beta_2\beta_4$. A naive testing approach is to estimate the equations and obtain the β's and then construct a t-test based on the coefficients and their standard errors. A more recent development is to use a bootstrapping procedure to obtain the 95% boostrapped confidence interval for the test statistics [Preacher and Hayes, 2008]. Allegedly, the bootstrapped standard errors are more conservative. This is generally true with small samples as in experimental research.

[1] The binary treatment is easily generalized to multi-level and continuous treatments.

2 Causal Mediation Analysis

A recent advancement in mediation analysis is the use of causal inference framework. The word "causal" stems from two core practices: 1) the formal discussion of identification and explicit presentation of assumptions (i.e., axiomization) and 2) the use of potential outcome languages. Another notable feature of this approach is the development of sensitivity analysis, which enables researchers to relax assumptions required for identification.

2.1 Identification under the sequential ignorability

Let's first define some terms. Given the treatment $D_i = d$, the potential outcome of the mediator is $M_i(d)$. For the outcome variable Y_i, its potential outcome $Y_i(d, m)$ depends on both the treatment $D_i = d$ and the mediator $M_i = m$. The causal mediation effect (or natural indirect effect) for participant i captures the difference between the participant's observed outcome and a counterfactual outcome if the participant's treatment status remains the same but the mediator value equals the value under the other treatment status [Pearl, 2001]:

$$\delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0)) \tag{2}$$

Where $d \in \{0, 1\}$. The term $\delta_i(0)$ is oftentimes called the *pure indirect effect* and the term $\delta_i(1)$ the *total indirect effect*. The objective is to estimate the a*verage causal mediation effect* (ACME) at the finite population level with:

$$\begin{aligned}\bar{\delta}(d) &= E(\delta_i(d)) = E[Y_i(d, M_i(1)) - Y_i(d, M_i(0))] \\ &= E_{d,M_i(1)}[Y_i(d, M_i(1))] - E_{d,M_i(0)}[Y_i(d, M_i(0))]\end{aligned} \tag{3}$$

The formal proof of identification is shown in the appendix. To understand the intuition, suppose $d = 1$, then we aim to identify $E[Y_i(1, M_i(1))]$. This is identified as

$$E[Y_i(1, M_i(1))] = E[Y_i(1, M_i(1)) \mid D_i = 1]. \tag{4}$$

So, we can use the treatment group observations to calculate this value. For the other expectation, we do not observe the potential outcome $Y_i(1, M_i(0))$, as when people are treated, we observe $Y_i(1, M_i(1))$ and when people are not treated, we observe $Y_i(0, M_i(0))$. This is where the sequential ignorability comes into play. As the M_i is unconfounded given D_i, we must have

$$\begin{aligned}E[Y_i(1, M_i(0))] &= E[Y_i(1, M_i(0)) \mid D_i = 1] \\ &= \sum_{M_i(0)} E[Y_i(1, M_i(0) = m) \mid D_i = 1] P(M_i(0) = m \mid D_i = 1)\end{aligned} \tag{5}$$

Note that

$$E[Y_i(1, M_i(0) = m) \mid D_i = 1] = E[Y_i(1, M_i(1) = m) \mid D_i = 1] \tag{6}$$

This is because the mediator is unconfounded w.r.t. Y_i, conditional on the treatment D_i. The result informs us that we can use the estimated distribution of the mediator in the control group and the observed outcome of the treatment group to estimate $E\left[Y_i\left(1, M_i\left(0\right)\right)\right]$.

2.2 Link to the conventional approach

To see the link between the conventional approach and the potential outcome framework, we can re-write the last two equations in the LSEM as:

$$\begin{cases} M_i\left(D_i\right) & = \alpha_2 + \beta_2 D_i + e_{i2}\left(D_i\right) \\ Y_i\left(D_i, M\left(D_i\right)\right) & = \alpha_3 + \beta_3 D_i + \beta_4 M\left(D_i\right) + e_{i3}\left(D_i, M\left(D_i\right)\right) \end{cases} \tag{7}$$

The error terms e_2 and e_3 under the sequential ignorability assumption are independent from the regressors in the two equations. We can use the equation above to calculate the expectations of Y_i, given T_i:

$$\begin{aligned} E\left[Y_i\left(d, M_i\left(d'\right)\right)\right] &= \alpha_3 + \beta_3 d + \beta_4 E\left(M\left(d'\right)\right) \\ &= \alpha_3 + \beta_3 d + \beta_4\left(\alpha_2 + \beta_2 d'\right) \end{aligned} \tag{8}$$

From this, we can compute the ACME as,

$$\bar{\delta}\left(1\right) = \beta_2\beta_4 \text{ and } \bar{\delta}\left(0\right) = \beta_2\beta_4$$

From the result, it is clear that the conventional approach makes an implicit assumption.

Assumption 3 (No-interaction between the Treatment and the ACME)**.** *The pure indirect effect is equal to the total indirect effect.*

This assumption is actually not necessary for the identification of ACME with the LSEM framework. One can easily extend the LSEM to relax this assumption by replacing the third equation in the LSEM with:

$$Y = \alpha_3 + \beta_3 D + \beta_4 M + \gamma DM + e_3 \tag{9}$$

With the addition of the interaction, the ACME is $\bar{\delta}\left(d\right) = \beta_2\left(\beta_4 + \gamma d\right)$.

2.3 Sensitivity analysis to relax the sequential ignorability assumption

In most scenarios, the sequential ignorability assumption is too strong to hold. Even in randomized experiments, Assumption 1 is credible due to random assignment. However, with measured not manipulated mediators, Assumption 2 is non-credible. One remedy is to assume there is a set of control variables that fully block the back-door paths between M and Y, conditional on D. This leads to a revised DAG as below:

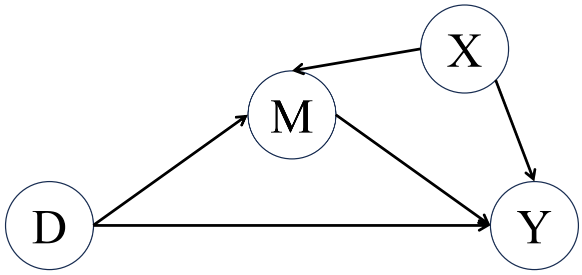

Figure 2: The DAG for Causal Mediation Analysis

The inclusion of control variables X relaxes the original sequential ignorability assumption. However, in experimental research, people rarely collect sufficient number of control variables that credibly relax the assumption. Alternatively, one could use sensitivity analysis to assess the credibility of the estimated ACME. The basic idea of the sensitivity analysis is to quantify the severity of violating the identification assumption. We first simulate data based on the various levels of severity of confounders. With the simulated data, we can re-estimated the ACME and see the change in the ACME. From the analysis, we can obtain the critical level of severity that nullifies the ACME. With the LSEM, we can actually derive a closed-form solution without appealing to the simulation of data. The sensitivity analysis of ACME with the LSEM is based on the following observation:

$$\operatorname{Cov}\left(e_{i3}\left(D_i, M_i\left(D_i\right)\right), M_i\left(D_i\right)\right) = \operatorname{Cov}\left(e_{i3}\left(D_i, M_i\left(D_i\right)\right) \cdot e_{i2}\left(D_i\right)\right) \tag{10}$$

For the sensitivity analysis to be general, the quantity should be normalized and the correlation between M and e_3 suits the purpose. One can easily show that:

$$\rho\left(M, e_3\right) = \frac{\sqrt{\beta_2^2\sigma_D^2 + \sigma_2^2}}{\sigma_2}\rho\left(e_2, e_3\right) \tag{11}$$

In the equation above, σ_D^2 is the variance of the treatment D and σ_2^2 the variance of e_2. The equation implies that one can use the correlation between e_2 and e_3 to assess the endogeneity of e_3 in the third equation of the LSEM. For experimental research, the endogeneity of e_3 is the key issue for the LSEM approach. In addition, a larger magnitude of ρ (i.e., $|\rho|$) implies a more severe endogeneity problem, as e_3 is more associated with M, and therefore a bigger threat to the identification of ACME.

Given ρ, the ACME can be expressed as its function.

$$\text{ACME} = \frac{\beta_2\sigma_1}{\sigma_2}\left[\tilde{\rho} - \rho\sqrt{\frac{1-\tilde{\rho}^2}{1-\rho^2}}\right] \tag{12}$$

In the equation, σ_1 and σ_2 are standard deviation of e_1 and e_2, and $\tilde{\rho}$ is the correlation between e_1 and e_2. All the parameters $\{\sigma_1, \sigma_2, \beta_2, \tilde{\rho}\}$, except for ρ, can be consistently estimated from the first two equations of the LSEM. When $\rho = 0$, the ACME becomes the standard estimation from the LSEM. The partial derivative of ACME with respect to ρ shows that ACME is either monotonically increasing or decreasing in ρ, depending on the sign of β_2. Finally, given all the other parameters $\{\sigma_1, \sigma_2, \beta_2, \tilde{\rho}\}$, varying the value of ρ from -1 to $+1$ will render the value of ACME from $-\infty$ to $+\infty$. This says, without the unconfoundedness of M, the LSEM tells us nothing about the ACME.

3 The Designed-based Approach to Mediation (to be added...)

4 Correcting Bias of the Conventional Mediation Analysis

The core problem in the mediation analysis is with the assumption that the mediator M is unconfounded. This is generally not true even in experimental research, where the treatment D is credibly unconfounded. However, the mediator M is usually measured but not manipulated as the treatment D. Therefore, the core problem is the confounders are unobserved between M and Y as shown in the DAG below.

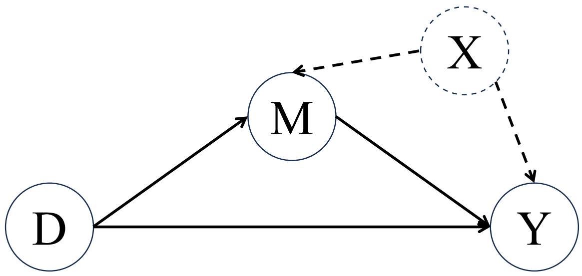

Figure 3: The Core Problem of Causal Mediation Analysis

The core problem leads to an inconsistent estimation of the third equation in the LSEM. Note that the first equation of the LSEM is unnecessary for estimation of the ACME as only β_2 and β_4 are required. Now focus on the last two equations of the LSEM and the core problem shown above:

$$\begin{cases} M_i &= \alpha_2 + \beta_2 D_i + e_{i2} \\ Y_i &= \alpha_3 + \beta_3 D_i + \beta_4 M_i + e_{i3} \end{cases} \tag{13}$$

To consistently estimate the third equation, one common strategy is conditioning. We may appeal to the backdoor criterion and assume we observe a set of control variables that block all the backdoor paths between M and Y. There are at least two issues of this strategy. First, in experimental research, we at most collect a moderate number of pre-treatment variables that can be potentially used as control variables. It is risky to use post-treatment variables as controls as they can be influenced by the mediator M and the outcome Y. Therefore, it is doubtful if we have sufficient controls. Second, even if a rich set of pre-treatment variables are collected, it is always possible that some unobserved confounders exist and bias the estimation of β_4. Therefore, the conditioning strategy is inadequate. Imai et al. [2010] proposed to use sensitivity analysis to assess the adequacy of the control variables. The problems with sensitivity analysis are that 1) the critical value of the sensitivity parameter is hard to interpret and 2) sensitivity analysis provides no fixes to cases where estimation results are sensitive.

4.1 The triangular system of equations

In this section, we focus on an alternative approach with insights from econometrics. In econometrics, Equation (13) constitutes a standard system of equation named "triangular system." Triangular system is a special type of simultaneous equation systems where one outcome (Y) is excluded from the other (M) [Newey et al., 1999], hence the name "triangular." The two equations in Equation (13) satisfy the conditions for a triangular system: 1) the treatment D is exogenous, 2) the mediator M is endogenous (or $\rho(e_{i2}, e_{i3}) \neq 0$), and 3) the outcome Y is excluded from the equation of the mediator M.

In general, the identification of the parameters in the second equation (β_3 and β_4) requires an instrumental variable for the mediator M. Such a variable should be exogenous to the error terms and excluded from the outcome Y. In experimental terms, one needs to manipulate the mediator M and randomly assign respondents to treatment conditions. The exclusion restriction requires a treatment that influences M directly and Y indirectly only through M. In psychological experiments, the exclusion restriction usually requires a careful design of the treatment. Given the complexity of human psyche, it is oftentimes difficult to have a treatment that satisfies the exclusion restriction, even if respondents are made unaware of the treatment. Therefore, the methods developed in econometrics that rely on higher-order moments of the errors instead of instruments come in handy.

Here, I will present two approaches based on alternative sets of assumptions on the error terms. The two approaches share similar thoughts. First, observe the triangular system in Equation (13). The first equation can be consistently estimated, as the treatment D is credibly randomized and therefore exogenous. Second, one can make further assumptions about the error terms and combine these additional assumptions with the first observation to identify the parameters in the second equation. Specifically, one approach, named the "constructed IV approach," makes a set assumptions which allow researchers to construct a valid instrument from the estimation of the first equation. The other approach, named the "control function approach," makes an alternative set of assumptions which exploit non-linearity of the variance of error terms to construct valid "proxies" of confounders in the second equation.

4.2 The constructed IV approach

4.2.1 The assumptions of the constructed IV approach

The constructed IV approach [Lewbel, 2012] makes following assumptions for the error terms in the triangular system.

Assumption 4. *The errors e_2 and e_3 have the following factor structure:*

$$\begin{cases} e_2 &= U + V_2 \\ e_3 &= cU + V_3 \end{cases} \tag{14}$$

where c is a constant and U, V_2 and V_3 are unobserved error terms that are mutually independent conditional on D (and other possible control variables X).

This assumption implies M is endogenous because it contains an error component U that appears in the errors of both equations. This assumption is not directly testable and should be justified by appealing to theoretical insights. However, the violation of this assumption, especially the linear-additive specification would not pose a serious threat to the validity of the constructed IV.

Assumption 5. *U^2 is uncorrelated with the treatment D or U is homoskedastic with respect to D.*

This assumption implies that $\text{Cov}\,(De_2, e_3) = 0$, which is straightforward by substituting Equation (14) to the covariance. The assumption ensures the constructed IV is exogenous with respect to e_3. This assumption is partly testable using the coefficients β_3 and β_4 of the constructed IV estimation to obtain e_3. Then, one can use a Pagan-Hall test to check if e_3 is homoskedastic. This test is over-powered, as the rejection of the homoskedasticity of e_3 does not necessarily mean the violation of the assumption. It could be U is still homoskedastic, but V_3 is heteroskedastic. In this case, the assumption still holds, but the test would reject the null hypothesis of homoskedasticity of e_3.

Assumption 6. *e_2^2 is correlated with the treatment D or e_2 is heteroskedastic with respect to D.*

This assumption implies that $\text{Cov}\,(De_2, e_2) \neq 0$, and ensures the relevance of the constructed IV. In fact, the larger the covariance, the stronger the constructed instrument. Conditional on the previous assumption (Assumption 5), this assumption is testable. A Breusch-Pagan test of heteroskedasticity can be used to test the residuals from the regression of the mediator on the treatment. Unlike Assumption 5, here the null hypothesis should be rejected for this assumption to hold.

4.2.2 Identification and estimation procedure

Under Assumption 4, 5 and 6, the following procedure produces consistent estimation of the coefficients β_3 and β_4, and the therefore the indirect effect $\beta_2\beta_4$.

1. Regress the mediator M and the treatment D and obtain the residual $\hat{e}_2$.
2. Construct an instrument with the treatment and the residual $\tilde{D} \cdot \hat{e}_2$, where $\tilde{D} = D - \bar{D}$.
3. Run a 2SLS with the outcome Y as the dependent variable, and M and D as the independent variables, where M is instrumented with the constructed IV.

The formal proof for the identification result is in the appendix.

4.3 The control function approach

4.3.1 The intuition behind the control function approach

Another approach that avoids finding a valid instrumental variable is the control function approach. The control function approach is based on the idea of the sample selection model. The intuition behind it is to construct a valid "proxy" of the confounders to control for the endogeneity of M in the second equation of the triangular system. The control function approach is based on the following observation. The endogeneity of M can be captured by the fact that e_2 and e_3 are correlated. As proved in the sensitivity analysis section, the correlation between M and e_3 is proportional to the correlation between e_2 and e_3. If we know the correlation between e_2 and e_3, we can consistently estimate the second equation. One key observation is that we can construct an error term based on e_2 and e_3 with:

$$a_0 = \arg\min_a E\left(e_3 - ae_2\right) = \frac{\operatorname{Cov}\left(e_2, e_3\right)}{\operatorname{Var}\left(e_2\right)} \tag{15}$$

By construction, the error term $\varepsilon = e_3 - a_0 e_2$ is uncorrelated with e_2 and therefore M. We can decompose $e_3 = \varepsilon + a_0 e_2$ and substitute it into the original equation:

$$Y_i = \alpha_3 + \beta_3 D_i + \beta_4 M_i + a_0 e_{i2} + \varepsilon_i \tag{16}$$

This equation can be consistently estimated provided the data matrix $\begin{bmatrix} D_i & M_i & e_{i2} \end{bmatrix}$ has full rank. However, due to the model specification, e_{i2} is a linear combination of M_i and D_i, and therefore the data matrix is a rank of 2. The main issue is therefore to somehow make e_{i2} not linearly dependent on M_i. To this end, Klein and Vella [2010] exploited the non-linearity in the variance of e_2 and e_3 to apply the control function approach. Overall, the idea is to decompose e_3 into two a linear combination of e_2 and an idiosyncratic error ε, so the correlation between e_2 and e_3 can be separated and used as a control function. This is essentially the insight of the sample selection model, e.g., the inverse Mill's ratio. Therefore, it also has the same restrictions as the sample selection approach - strong distributional / functional assumptions.

4.3.2 Assumptions of the control function approach

The following set of assumptions are needed for the control function approach. The key assumption is the heteroskedasticity of e_2 and e_3 with respect to the exogenous variables. In a binary experiment setting, this mean the variance of e_2 and e_3 is a non-linear function of the treatment D. However, given the discrete nature of D, the variation in D is fairly limited. In theory, the discrete nature of D does not impact the identification as long as the non-linear heteroskedasticity holds. In practice, the discrete nature of D prevents us from having a valid estimation of the variance of e_2 and e_3 as a non-linear function of D. Therefore, the control function approach works best if a set of pre-treatment control variables X are observed. D can be used with X to infer the variance functions. Given the set of control variables X, we have following assumptions.

Assumption 7 (Multiplicative Heteroskedasticity)**.** *The error term e_2 and e_3 are heteroskedastic w.r.t. X and can be expressed as $e_2 = S_2(X)\, e_2^*$ and $e_3 = S_3(X)\, e_3^*$, where $S_2(X)$ and $S_3(X)$are non-linear functions of X, e_2^* and e_3^* are idiosyncratic errors, and ${S_2(X)}/{S_3(X)} \neq c$, a constant.*

For this assumption, the multiplicative functional form cannot be tested, but the heteroskedasticity can be partially tested. We can use a Breusch-Pagan test for e_2, and the rejection of the null hypothesis supports the assumption. For the heteroskedasticity of e_3, we can test residuals $\hat{e}_3$ from the control function approach, but the test is under-powered, as the heteroskedasticity could be attributed to e_2 instead of ε. That is, the rejection of the null hypothesis invalidates the assumption, but the failure to reject the null hypothesis does not provide sufficient evidence for the assumption. Lastly, the ratio between $S_2(X)$ and $S_3(X)$ are assumed to be non-constant to rule out corner solutions and ensure the identification of $S_3(X)$.

Assumption 8. *The conditional mean of idiosyncratic errors w.r.t. X are 0, with $E(e_2^* \mid X) = 0$ and $E(e_3^* \mid X) = 0$ and the interaction between e_2^* and e_3^* is independent from X with $E(e_2^* e_3^* \mid X) = E(e_2^* e_3^*) = \lambda$.*

This assumption requires the idiosyncratic errors to have a joint distribution that is unrelated to X. The conditional mean assumption is valid, as long as X are credibly exogenous. For example, X could contain pre-treatment variables that are exogenous to the mediator or the outcome. The constant covariance assumption is more difficult to justify and inherently untestable. It is essentially an exclusion restriction such that the heteroskedasticity of e_2 and e_3 under the multiplicative assumption are identified. Without this assumption, the heteroskedasticity function of e_3 is unidentifiable.

Under Assumption 7 and 8, one can show that the error term e_3 can be decomposed into two parts:

$$e_3 = \lambda \frac{S_3(X)}{S_2(X)} e_2 + \varepsilon \tag{17}$$

Where ε is an error term that is unrelated to e_2. Such a decomposition leads to a control function approach by substituting Equation (17) to the third equation of the

LSEM:

$$Y = \alpha_3 + \gamma X + \beta_3 D + \beta_4 M + \lambda \frac{S_3(X)}{S_2(X)} e_2 + \varepsilon \tag{18}$$

It becomes clear from Equation (18) that the constant covariance assumption is needed to guarantee the identification of $S_3(X)$.

4.3.3 Identification and estimation of the control function approach

Under Assumption 7 and 8, we can decompose the e_3 as shown in Equation (17) and Equation (18) can be consistently estimated. However, for estimation, the main issue is with $S_2(X)$ and $S_3(X)$. Note that $S_2(X)$ can be estimated and plugged into Equation (18), but $S_3(X)$ needs to be estimated along with other parameters in Equation (18). One approach is to use non-parametric estimation, which usually requires a large sample size that is unrealistic in experimental research. A parametric or semi-parametric estimation is therefore preferred. Another difficult arises due to the unknown functional form of $S_3(X)$. Under no further restriction on the functional form of $S_3(X)$, the least square optimization of Equation (18) is not guaranteed to be a convex optimization and thus the optimization can be costly. A general procedure of estimation is as the following:

1. Regression M on D (and other exogenous variables X) to obtain the residual e_2.
2. With the residual e_2, estimate the heteroskedasticity function $S_2(X)$, e.g., a semi-parametric regression of $|e_2|$ on X.
3. With e_2 and $S_2(X)$, obtain the estimation of $(\alpha_3, \gamma, \beta_3, \beta_4, \lambda)$ and $S_3(X)$ as solutions to the following optimization problem:

$$\min_{\{\alpha_3,\gamma,\beta_3,\beta_4,\lambda\}\cup S_3(X)} \left[Y - \left(\alpha_3 + \gamma X + \beta_3 D + \beta_4 M + \lambda \frac{S_3(X)}{S_2(X)} e_2\right)\right]^2$$

The computational cost is substantial, especially with a large sample size and many X's, because the semi-parametric estimation of $S_3(X)$ is nested in the optimization procedure. Moreover, the statistical inference requires subsampling as bootstrapping fails to work due to the direct optimization.

4.4 Comparison of the two approaches

Overall, the constructed IV approach is more suitable for experimental research than the control function approach, despite of more restrictive assumptions on the error terms. There are a few reasons for this verdict. First, the control function approach requires a set of exogenous variables X (at least two continuous variables). Ideally, one needs a rich set of X to better ensure the non-linearity of $S_2(X)$ and $S_3(X)$. However, if researchers have a rich set of X, it is more convenient to use X as direct controls in the LSEM and examine the estimation results with sensitivity analysis. Second, the heteroskedastic assumptions of the constructed IV approach can be conservatively tested. For the control function approach, the functional assumptions are inherently untestable. Moreover, its heteroskedastic assumptions are not refutable because of the under-powered test. Third, the computational costs of the control function approach are much higher than the constructed IV. The statistical inference is relatively easy for the constructed IV approach (i.e., bootstrapping), but more difficult for the control function approach (i.e., subsampling). Overall, the constructed IV approach may require more restrictive assumptions on the heteroskedasticity, but the over-powered test ensures its assumptions can be validated before running the analysis.

5 Extensions to Moderated Mediation

In this section, the sensitivity analysis and the constructed IV approach are further extended to the moderated mediation analysis. The DAG of the moderated mediation analysis is as below:

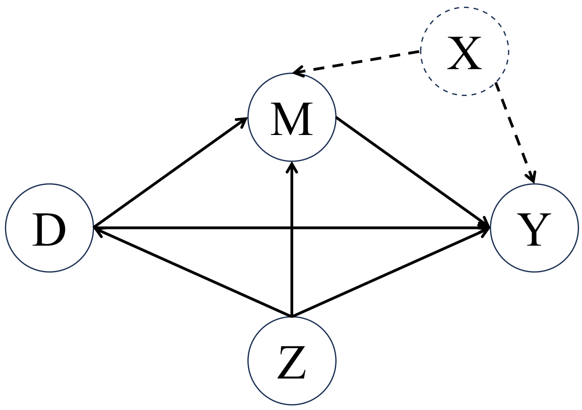

Figure 4: DAG of Moderated Mediation

With the DAG, we can derive the following estimation equations:

$$\begin{cases} Y &= \alpha_1 + \beta_{11} D + \beta_{12} DZ + \gamma_1 Z + e_1 \\ M &= \alpha_2 + \beta_{21} D + \beta_{22} DZ + \gamma_2 Z + e_2 \\ Y &= \alpha_3 + \beta_{31} D + \beta_{32} DZ + \beta_{41} M + \beta_{42} MZ + \gamma_3 Z + e_3 \end{cases} \tag{19}$$

The objective of the estimation is to derive the indirect effect as $(\beta_{21} + \beta_{22} Z)(\beta_{41} + \beta_{42} Z)$. Therefore, the key is to consistently estimate the parameter set $\{\beta_{21}, \beta_{22}, \beta_{41}, \beta_{42}\}$. From the DAG, the equation of $M \sim D + Z + DZ$ or $\{\beta_{21}, \beta_{22}\}$ can be consistently estimated. However, due to the confoundedness of the mediator M, the direct estimation of $\{\beta_{41}, \beta_{42}\}$ is biased. Next, we present the sensitivity analysis for this model and also the constructed IV approach for bias correction.

5.1 Sensitivity analysis for moderated mediation

First, given the covariance between e_2 and e_3, we can decompose the estimated coefficient $b = (X'X)^{-1} X'Y$ as

$$b = \beta + (X'X)^{-1} X' e_3 \tag{20}$$

The explanatory variables include the endogenous variables and exogenous ones with $X_1 = \begin{bmatrix} 1 & D & Z & DZ \end{bmatrix}$ and $X_2 = \begin{bmatrix} M & MZ \end{bmatrix}$. We can partition the regression as

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{pmatrix} X_1'e_3 \\ X_2'e_3 \end{pmatrix} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}^{-1} \begin{pmatrix} 0 \\ Q_{2e} \end{pmatrix} \tag{21}$$

Where Q_{ij} is the covariance matrix between X_i and X_j $(i, j \in \{1, 2\})$ and $Q_{2e} = \begin{pmatrix} \sigma_{23} & \mu_z\sigma_{23} \end{pmatrix}'$. The inverse of the matrix is

$$\begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}^{-1} = \begin{bmatrix} Q_{11\cdot2}^{-1} & -Q_{11\cdot2}^{-1}Q_{12}Q_{22}^{-1} \\ -Q_{22\cdot1}^{-1}Q_{21}Q_{11}^{-1} & Q_{22\cdot1}^{-1} \end{bmatrix}$$

Where $Q_{11\cdot2} = Q_{11} - Q_{12}Q_{22}^{-1}Q_{21}$ and $Q_{22\cdot1} = Q_{22} - Q_{21}Q_{11}^{-1}Q_{12}$. Therefore, the bias of the estimated coefficients is

$$\begin{bmatrix} Q_{11\cdot2}^{-1} & -Q_{11\cdot2}^{-1}Q_{12}Q_{22}^{-1} \\ -Q_{22\cdot1}^{-1}Q_{21}Q_{11}^{-1} & Q_{22\cdot1}^{-1} \end{bmatrix} \begin{pmatrix} 0 \\ Q_{2e} \end{pmatrix} = \begin{pmatrix} -Q_{11\cdot2}^{-1}Q_{12}Q_{22}^{-1}Q_{2e} \\ Q_{22\cdot1}^{-1}Q_{2e} \end{pmatrix}$$

If we focus on the coefficients $\{\beta_{41}, \beta_{42}\}$, we have

$$\begin{pmatrix} b_{41} \\ b_{42} \end{pmatrix} = \begin{pmatrix} \beta_{41} \\ \beta_{42} \end{pmatrix} + \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \begin{pmatrix} \sigma_{23} \\ \mu_z\sigma_{23} \end{pmatrix}$$

From this, we have

$$\begin{cases} \beta_{41} & = b_{41} - (q_{11} + q_{12}\mu_z)\,\sigma_{23} \\ \beta_{42} & = b_{42} - (q_{21} + q_{22}\mu_z)\,\sigma_{23} \end{cases} \tag{22}$$

That is, we can express the true coefficient as a linear term of the covariance between e_2 and e_3. However, the covariance σ_{23} is not conducive for a general sensitivity analysis as the value can be unbounded and also it is difficult to compare across data and studies. Next, we use the insight from Imai et al. [2010] to further transform the values to the correlation term ρ as in Section 2.3. By substituting the M equation in Equation (19) to the Y equation, the error term of the Y equation becomes $(\beta_{41} + \beta_{42}Z)\,e_2 + e_3$. Comparing the total effect equation with e_1, we have[2]:

$$\begin{cases} \mathrm{Var}\,(e_1) & = \mathrm{Var}\,((\beta_{41} + \beta_{42}Z)\,e_2 + e_3) \\ \mathrm{Cov}\,(e_1, e_2) & = \mathrm{Cov}\,((\beta_{41} + \beta_{42}Z)\,e_2 + e_3, e_2) \end{cases} \tag{23}$$

This gives us:

$$\begin{cases} \sigma_1^2 & = \left(x^2 + \left(\tilde{b}_2 - \tilde{q}_2x\right)^2\sigma_z^2\right)\sigma_2^2 + \sigma_3^2 + 2\rho\sigma_2\sigma_3x \\ \tilde{\rho}\sigma_1 & = x\sigma_2 + \rho\sigma_3 \end{cases} \tag{24}$$

[2]Note that, for the sensitivity analysis, e_1 should be obtained from the residuals of the regression $y \sim D + Z + DZ + Z^2 + DZ^2$.

Where we define $q_1 = q_{11} + q_{12}\mu_z$ and $q_2 = q_{21} + q_{22}\mu_z$.

$$\begin{cases} x &= \beta_{41} + \beta_{42}\mu_z \\ \tilde{b}_2 &= \left(b_2 - \dfrac{q_2 \left(b_1 + b_2\mu_z \right)}{q_1 + q_2\mu_z} \right) \\ \tilde{q}_2 &= \dfrac{q_2}{q_1 + q_2\mu_z} \end{cases} \tag{25}$$

We then solve for x in Equation (27) with:

$$x = \frac{\tilde{\rho}\sigma_1}{\sigma_2} \left[1 - \frac{\rho \left(\sqrt{A} - \rho\tilde{q}_2\sigma_2 \left(b_2\sigma_z^2 + \tilde{\rho}\tilde{q}_2\sigma_1 \right) \right)}{\sigma_1\tilde{\rho} \left(\sigma_z^2\tilde{q}_2^2\rho^2 - \rho^2 + 1 \right)} \right] \tag{26}$$

Where

$$A = \sigma_z^2 \left[\rho^2 \left(\tilde{b}_2\sigma_2 - \tilde{q}_2\sigma_1 \right)^2 - \left(\tilde{b}_2\sigma_2 - \tilde{\rho}\tilde{q}_2\sigma_1 \right)^2 \right] + \sigma_1^2 \left(\rho^2 - 1 \right) \left(\tilde{\rho}^2 - 1 \right)$$

Note that x has the same form as in the sensitivity analysis. If there is no bias and $\rho = 0$, then the expected value of $E(x) = (\beta_{41} + \beta_{42}\mu_z) = \frac{\tilde{\rho}\sigma_1}{\sigma_2}$. With x, we can solve for the covariance σ_{23} and calculate the true coefficient values β_{41} and β_{42}. Alternatively, one can derive the solution of σ_{23} numerically based on the system equations, given the correlation ρ.

5.2 Constructed IV for moderated mediation

To apply the constructed IV, we first make the same assumption to decompose the errors into a common unobservable and an idiosyncratic term

$$\begin{cases} e_2 &= U + V_2 \\ e_3 &= cU + V_3 \end{cases} \tag{27}$$

Similar to the base model, we make two assumptions. First, we need to assume U is homoskedastic with D. Second, e_2 or V_2 should be heteroskedastic with respect to D. To this end, one can follow this procedure for the estimation:

1. Regress $M \sim D + Z + DZ$ and obtain the residual $\hat{e}_2$.
2. Test the heteroskedasticity of $\hat{e}_2$ with respect to D or $\mathrm{Cov}\left(\hat{e}_2^2, D\right)$. If $\hat{e}_2$ is heteroskedastic with respect to D, continue to the next step.
3. Construct the IV from Step 2 with $\tilde{D}\hat{e}_2 = \left(D - \bar{D} \right) \hat{e}_2$.
4. Run a 2SLS version of $Y \sim D + Z + M + DZ + MZ$ with the constructed IV:
   (a) The first stage for M is: $M \sim D + Z + DZ + \tilde{D}\hat{e}_2$.
   (b) The first stage for MZ is: $MZ \sim D + Z + DZ + Z^2 + DZ^2 + \tilde{D}\hat{e}_2$.

Note that MZ is a non-linear term with respect to D and Z. Substituting the predicted values from Step 4a would lead to a forbidden regression. Therefore, we need to run a different first stage for the interaction term MZ with explanatory variables include non-linear terms of D and Z.

6 Extensions to Serial Mediation

The DAG of the serial mediation under confounded mediator M_1 and M_2 is

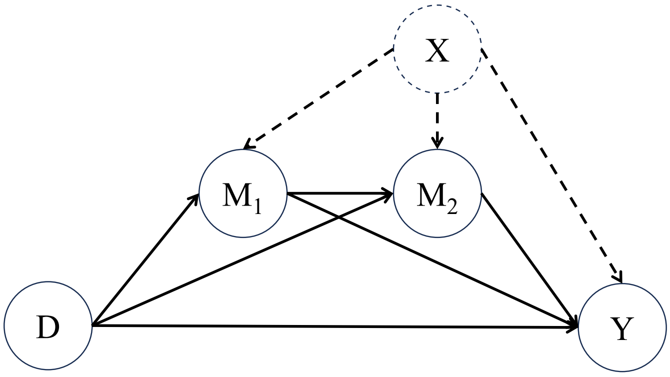

Figure 5: DAG for the Serial Mediation (2 mediators)

The system of equations that capture this model is

$$\begin{cases} Y & = \alpha_1 + \beta_1 D + e_1 \\ M_1 & = \alpha_{21} + \beta_{21} D + e_2 \\ M_2 & = \alpha_{22} + \beta_{22} D + e_3 \\ M_2 & = \alpha_{31} + \beta_{31} D + \gamma M_1 + e_4 \\ Y & = \alpha_{32} + \beta_{32} D + \beta_{41} M_1 + \beta_{42} M_2 + e_5 \end{cases} \tag{28}$$

6.1 Sensitivity analysis for serial mediation

In the serial mediation, we aim to quantify 3 indirect paths between D and Y in the DAG: $D \to M_1 \to Y$, $D \to M_2 \to Y$, and $D \to M_1 \to M_2 \to Y$. The indirect effects are captured by $\beta_{21}\beta_{41}$, $\beta_{22}\beta_{42}$ and $\beta_{21}\gamma\beta_{42}$, respectively. Due to the confoundedness, $\{\beta_{41}, \beta_{42}, \gamma\}$ cannot be consistently estimated. The biases of the direct estimation depends on the spurious correlations between M_1, M_2 and Y. Or equivalently, he correlation between error terms $\{e_2, e_4\}$, $\{e_2, e_5\}$ and $\{e_3, e_5\}$. We denote the correlation between two error terms $\{e_i, e_j\}$ as ρ_{ij} and the standard deviation of e_i as σ_i. Applying the same insights in Imai et al. [2010], the true value of γ is

$$\gamma = \frac{\sigma_3}{\sigma_2}\left[\rho_{23} - \rho_{24}\sqrt{\frac{1-\rho_{23}^2}{1-\rho_{24}^2}}\right] \tag{29}$$

Given ρ_{25} and ρ_{35}, the true coefficient β_{41} and β_{42} can be solved from the following system of equations

$$\begin{cases} \mathrm{Cov}\left(e_1, e_2\right) & = \mathrm{Cov}\left(\beta_{41} e_2 + \beta_{42} e_3 + e_5, e_2\right) \\ \mathrm{Cov}\left(e_1, e_3\right) & = \mathrm{Cov}\left(\beta_{41} e_2 + \beta_{42} e_3 + e_5, e_3\right) \\ \mathrm{Var}\left(e_1\right) & = \mathrm{Var}\left(\beta_{41} e_2 + \beta_{42} e_3 + e_5\right) \end{cases} \tag{30}$$

Expanding the equation and write covariance as the correlation multiplied by standard deviations, we have

$$\begin{cases} \rho_{12}\sigma_1 & = \beta_{41}\sigma_2 + \beta_{42}\rho_{23}\sigma_3 + \rho_{25}\sigma_5 \\ \rho_{13}\sigma_1 & = \beta_{41}\rho_{23}\sigma_2 + \beta_{42}\sigma_3 + \rho_{35}\sigma_5 \\ \sigma_1^2 & = \beta_{41}^2\sigma_2^2 + \beta_{42}^2\sigma_3^2 + \sigma_5^2 + 2\beta_{41}\beta_{42}\rho_{23}\sigma_2\sigma_3 + 2\beta_{41}\rho_{25}\sigma_2\sigma_5 + 2\beta_{42}\rho_{35}\sigma_3\sigma_5 \end{cases} \tag{31}$$

The equation has a closed-form solution for $\{\beta_{41}, \beta_{42}, \sigma_5\}$. Or, it is possible to solve the system equations numerically. Note that for β_{41} and β_{42}, the bias depends on both ρ_{23} and ρ_{25}, which increases the dimensionality of the computation for finding the critical values, as we need to specify the tuple $\langle\rho_{23}, \rho_{25}\rangle$.

6.2 Constructed IV for serial mediation

For the constructed IV, we first made assumptions about the error terms to decompose them into the common unobservable U and the idiosyncratic errors V. Due to the fact that we only observe one exogenous variable D in the system, we cannot test a fully relaxed model where both M_1 and M_2 are assumed to be endogenous. We may test a partial model where either M_1 or M_2 are assumed to be endogenous.

6.2.1 Endogenous M_1 and exogenous M_2

If we assume M_1 is endogenous and M_2 exogenous, the following assumptions hold:

$$\begin{cases} e_2 & = U + V_2 \\ e_4 & = V_4 \\ e_5 & = c_5 U + V_5 \end{cases} \tag{32}$$

Where V_2 is heteroskedastic with respect to D, but U and V_5 is homoskedastic with respect to D. Given the assumptions, the procedure below gives consistent estimates of β_{41} and β_{42}:

1. Regress $M_1 \sim D$ and obtain the residual $\hat{e}_2$.
2. Regress $M_2 \sim D + M_1$ to obtain γ and the residual $\hat{e}_4$.
3. Construct the IV for M_1 as $\tilde{D}\hat{e}_2$.
4. Run a 2SLS regression of $Y \sim D + M_1 + \hat{e}_4$, with the first stage $M_1 \sim D + \tilde{D}\hat{e}_2$.
   (a) The coefficient of $\hat{e}_4$ is β_{42}.
   (b) The corrected coefficient of M_1 is $\beta_{41} + \gamma\beta_{42}$.

6.2.2 Exogenous M_1 and endogenous M_2

If we assume M_1 is exogenous and M_2 endogenous, the following assumptions hold:

$$\begin{cases} e_2 & = V_2 \\ e_4 & = U + V_4 \\ e_5 & = c_5 U + V_5 \end{cases} \tag{33}$$

Where V_4 is heteroskedastic with respect to D, but U andV_5 is homoskedastic with respect to D. Given the assumptions, the procedure below gives consistent estimates of β_{41} and β_{42}:

1. Regress $M_2 \sim D + M_1$ to obtain the residual $\hat{e}_4$.
2. Construct the IV for M_2 as $\tilde{D}\hat{e}_4$.
3. Run a 2SLS regression of $Y \sim D + M_1 + M_2$, with the first stage $M_2 \sim D + M_1 + \tilde{D}\hat{e}_2$.

The proof for the identification is omitted here as it is the same as the standard mediation model (extending the matrix of explanatory variables to $\left[\begin{array}{cc} D & M_1 \end{array}\right]$).

7 Appendix

7.1 The definitions of various effects in causal mediation analysis

Effects	Definitions	Relationships
Natural direct effect	$\zeta_i(d) = Y_i(1, M_i(d)) - Y_i(0, M_i(d))$	
Pure direct effect	$\zeta_i(0)$	
Total direct effect	$\zeta_i(1)$	
Natural indirect effect	$\delta_i(d) = Y_i(d, M_i(1)) - Y_i(d, M_i(0))$	
Pure indirect effect	$\delta_i(0)$	
Total indirect effect	$\delta_i(1)$	
Total effect	$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0))$	$\tau_i = \delta_i(d) + \zeta_i(1-d)$

Table 1: Definitions of Various Effects

7.2 Identification of the natural indirect effect

Under sequential ignorability, the natural indirect effect (NIE) is nonparametrically identified. Using the DAG framework, we have

$$\begin{gathered} P\left(Y_i \mid do\left(D_i\right) = d, do\left(M_i(1)\right) = m, do\left(M_i(0)\right) = m'\right) = \\ P\left(Y_i \mid D_i = d, M_i(1) = m, M_i(0) = m'\right) \end{gathered} \tag{34}$$

First, observe that the treatment D_i is unconfounded w.r.t. to Y_i. So, by the rule of the do-calculus, we can remove the do-operator of D_i. In addition, we have M_i unconfounded w.r.t. Y_i given the treatment D_i. Therefore, both $M_i(1)$ and $M_i(0)$ are unconfounded, as they are expressed the as the conditional distribution of M_i on the treatment. So, we can remove the do-operators for both both $M_i(1)$ and $M_i(0)$. Therefore, the NIE is identified.

7.3 The covariance between M and e_3

The covariance of M_i and e_{i3} is:

$$\begin{aligned}\mathrm{Cov}\left(e_{i3}\left(D_i, M_i\left(D_i\right)\right) \cdot M_i\left(D_i\right)\right) &= E\left(M_i\left(D_i\right) \cdot e_{i3}\left(D_i, M_i\left(D_i\right)\right)\right) \\ &= E\left(e_{i3}\left(D_i, M_i\left(D_i\right)\right) \cdot \left(\alpha_2 + \beta_2 D_i + e_{i2}\left(D_i\right)\right)\right) \\ &= E\left(e_{i3}\left(D_i, M_i\left(D_i\right)\right) \cdot e_{i2}\left(D_i\right)\right) \\ &= \mathrm{Cov}\left(e_{i3}\left(D_i, M_i\left(D_i\right)\right) \cdot e_{i2}\left(D_i\right)\right)\end{aligned} \tag{35}$$

The first equality is because the expectation of e_{i3} is 0, the third is because the treatment D_i is unconfounded, and the fourth is because the expectation of error terms are 0.

7.4 ACME as a function of ρ

In a randomized experiment, the treatment D_i is unconfounded. Therefore, we must have $E\left(e_{ij} \mid D_i\right) = 0$. We can consistently estimate the first two equations and obtain the values of $\{\alpha_1, \alpha_2, \beta_1, \beta_2\}$, and the variance and correlation terms $\{\sigma_1^2, \sigma_2^2, \tilde{\rho}\}$. Replace M in the third equation with the second equation in the LSEM and compare with the first equation:

$$\begin{cases} Y_i &= \alpha_1 + \beta_1 D_i + e_1 \\ Y_i &= \left(\alpha_3 + \alpha_2\beta_4\right) + \left(\beta_3 + \beta_2\beta_4\right) D_i + \left(\beta_4 e_2 + e_3\right)\end{cases} \tag{36}$$

For the error terms, we must have:

$$\begin{cases} \mathrm{Var}\left(e_1\right) &= \mathrm{Var}\left(\beta_4 e_2 + e_3\right) \\ \mathrm{Cov}\left(e_1, e_2\right) &= \mathrm{Cov}\left(\beta_4 e_2 + e_3, e_2\right)\end{cases} \tag{37}$$

The equations give us:

$$\begin{cases} \sigma_1^2 &= \beta_4^2\sigma_2^2 + \sigma_3^2 + 2\beta_4\rho\sigma_2\sigma_3 \\ \tilde{\rho}\sigma_1\sigma_2 &= \beta_4\sigma_2^2 + \rho\sigma_2\sigma_3\end{cases} \tag{38}$$

Solve for σ_3 with the second equation of Equation (18) and substitute it to the first equation, we have a quadratic equation:

$$\beta_4^2 - 2\frac{\tilde{\rho}\sigma_1}{\sigma_2}\beta_4 + \frac{\sigma_1^2\left(\tilde{\rho}^2 - \rho^2\right)}{\sigma_2^2\left(1 - \rho^2\right)} = 0 \tag{39}$$

Then β_4 can be solved from the quadratic equation:

$$\beta_4 = \frac{\sigma_1}{\sigma_2}\left[\tilde{\rho} - \rho\sqrt{\frac{1 - \tilde{\rho}^2}{1 - \rho^2}}\right] \tag{40}$$

The ACME is equal to $\beta_2\beta_4$.

7.5 Identification of the constructed IV approach

7.5.1 The standard mediation model

Observe that β_2 is identified from the regression of M on D as D is exogenous. Therefore, a consistent sample analog of the error term e_2 can be obtain by calculating the residual of this regression. Define the reduced-form errors as W with:

$$\begin{cases} W_2 & = M - D'E\left(DD'\right)^{-1}E\left(DM\right) \\ W_3 & = Y - D'E\left(DD'\right)^{-1}E\left(DY\right) \end{cases} \tag{41}$$

Substitute the equations in the triangular system into the equation, we have:

$$\begin{cases} W_2 & = e_2 \\ W_3 & = e_3 + e_2\beta_4 \end{cases} \tag{42}$$

By Assumption 5, we have $\operatorname{Cov}\left(D, e_2e_3\right) = 0$, which implies $\operatorname{Cov}\left(D, W_2\left(W_3 - \beta_4W_2\right)\right) = 0$. From this equation or moment condition, we can identify β_4, with

$$\beta_4 = \frac{\operatorname{Cov}\left(D, W_2W_3\right)}{\operatorname{Cov}\left(D, W_2^2\right)}$$

With the identified β_4, we can identify β_3 and also the indirect effect $\beta_2\beta_4$.

7.5.2 The moderated mediation model

Observe that β_{21} and β_{22} are identified from the regression of $M \sim D + Z + DZ$ as D and Z are both exogenous. Therefore, a consistent sample analog of the error term e_2 can be obtain by calculating the residual of this regression. Let the data matrix $D_1 = \begin{bmatrix} D & Z & DZ \end{bmatrix}$ and $D_2 = \begin{bmatrix} D & Z & DZ & Z^2 & DZ^2 \end{bmatrix}$Define the reduced-form errors as W with:

$$\begin{cases} W_1 & = M - D_1'E\left(D_1D_1'\right)^{-1}E\left(D_1Y\right) \\ W_2 & = MZ - D_2'E\left(D_2D_2'\right)^{-1}E\left(D_2MZ\right) \\ W_3 & = Y - D_2'E\left(D_2D_2'\right)^{-1}E\left(D_2Y\right) \end{cases} \tag{43}$$

From this, we have the following equations:

$$\begin{cases} W_1 & = e_2 \\ W_2 & = Ze_2 \\ W_3 & = \beta_{41}e_2 + \beta_{42}Ze_2 + e_3 \end{cases} \tag{44}$$

Using the assumptions on the instrumental variables, we have the following:

$$\begin{cases} \operatorname{Cov}\left(DW_1, W_3 - \beta_{41}W_1 - \beta_{42}W_2\right) & = 0 \\ \operatorname{Cov}\left(DW_2, W_3 - \beta_{41}W_1 - \beta_{42}W_2\right) & = 0 \end{cases} \tag{45}$$

Solving the system of equations gives us β_{41} and β_{42}.

7.6 Proof for the sensitivity analysis for the moderated mediation effect

$$\begin{cases} \beta_{41} & = b_{41} - q_1\sigma_{23} \\ \beta_{42} & = b_{42} - q_2\sigma_{23} \end{cases} \text{with} \begin{cases} q_1 & = q_{11} + q_{12}\mu_z \\ q_2 & = q_{21} + q_{22}\mu_z \end{cases}$$

$$\begin{aligned} x = \beta_{41} + \beta_{42}\mu_z \Longrightarrow \sigma_{23} &= \frac{b_{41} + b_{42}\mu_z - x}{q_1 + q_2\mu_z} \\ \Longrightarrow \beta_{42} = b_{42} - q_2\frac{b_{41} + b_{42}\mu_z - x}{q_1 + q_2\mu_z} &= \left(b_{42} - \frac{q_2\left(b_{41} + b_{42}\mu_z\right)}{q_1 + q_2\mu_z}\right) + \frac{q_2}{q_1 + q_2\mu_z}x \\ \Longrightarrow \beta_{42} = \tilde{b}_2 - \tilde{q}_2 x & \end{aligned}$$

Where

$$\begin{cases} \tilde{b}_2 & = \left(b_{42} - \dfrac{q_2\left(b_{41} + b_{42}\mu_z\right)}{q_1 + q_2\mu_z}\right) \\ \tilde{q}_2 & = \dfrac{q_2}{q_1 + q_2\mu_z} \end{cases}$$

References

Kosuke Imai, Luke Keele, and Teppei Yamamoto. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*, 25(1):51–71, 2010.

J Pearl. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence, 2001*, pages 411–420. Morgan Kaufman, 2001.

Roger Klein and Francis Vella. Estimating a class of triangular simultaneous equations models without exclusion restrictions. *Journal of Econometrics*, 154(2):154–164, 2010.

Reuben M Baron and David A Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51(6):1173–1182, 1986.

Whitney K Newey, James L Powell, and Francis Vella. Nonparametric estimation of triangular simultaneous equations models. *Econometrica*, 67(3):565–603, 1999.

Arthur Lewbel. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. *Journal of Business & Economic Statistics*, 30(1): 67–80, 2012.

Kristopher J Preacher and Andrew F Hayes. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3):879–891, 2008.