Online Supplementary Material to: Causal Mediation Analysis in the Presence of a Mismeasured Outcome

Results with parametric model

We investigate the asymptotic bias of the natural direct and indirect effects based on two regression models for the outcome and the mediator.

When the outcome is continuous, we model both the mediator and the outcome using the linear link, and allow the outcome model to include exposure-mediator interaction.

$$E(M \mid A = a, \boldsymbol{C} = \boldsymbol{c}) = \beta_0 + \beta_1 a + \boldsymbol{\beta_2'} \boldsymbol{c}$$

$$E(Y \mid A = a, M = m, \boldsymbol{C} = \boldsymbol{c}) = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta_4'} \boldsymbol{c}.$$

However, we can observe only Y^*, and thus we can operate only with the observed regression model for the outcome:

$$E(Y^* \mid M = m, A = a, \boldsymbol{C} = \boldsymbol{c}) = \theta_0^* + \theta_1^* a + \theta_2^* m + \theta_3^* am + \boldsymbol{\theta_4^{*'}} \boldsymbol{c}.$$

Let $\hat{\boldsymbol{\theta}}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, \hat{\theta}_3^*, \hat{\boldsymbol{\theta}}_4^*)$ be the naive MLE. Let $X = (A, M, AM, \boldsymbol{C})$ and assume that X is centered. Let $\Delta = E[X^{'} X]$ and δ^{ij} denote an element of the inverse of Δ. We can obtain the asymptotic bias of parameters of the outcome regression:

$$ABIAS(\hat{\boldsymbol{\theta}}^*) = \lim E(\hat{\boldsymbol{\theta}}^*) - \boldsymbol{\theta} = \lim E\{(X^{'} X)^{-1} X' Y^* - (X^{'} X)^{-1} X' Y\}$$

$$= \quad limE\{(X^{'} X)^{-1} X' U\} = \Delta^{-1} E(X' U) = 0.$$

Therefore, we can consistently estimate the direct and indirect effects using the observed outcome.

When the outcome is binary and is modeled with a logit link, the outcome model can be replaced by :

$$\text{logit}\{P(Y=1 \mid M=m, A=a, \boldsymbol{C}=\boldsymbol{c})\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4' \boldsymbol{c},$$

and the observed regression model is:

$$\text{logit}\{P(Y^*=1 \mid M=m, A=a, \boldsymbol{C}=\boldsymbol{c})\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \boldsymbol{\theta}_4' \boldsymbol{c},$$

Following the reasoning of Neuhaus[1], we can deduce that the parameter estimates from the observed regression for a general link function will approximately converge to $\boldsymbol{\theta} H(\theta_0)$, where

$$H(\theta_0) = \frac{(SN+SP-1)g'\{(SN+SP-1)g^{-1}(\theta_0)+1-SP\}}{g'\{g^{-1}(\theta_0)\}}, \text{and } g(x) = \log\{x/(1-x)\}.$$

The above formula provides some intuition for the bias. If there is no misclassification, then $SP = SN = 1$, hence $H(\theta_0) = 1$. The observed regression model can consistently estimate the true parameters. When misclassification presents, we have

$$\begin{aligned} & \frac{1}{g'\{(SN+SP-1)g^{-1}(\theta_0)+\theta_0\}+1-SP} \\ = & \frac{1}{g'\{(SN+SP-1)g^{-1}(\theta_0)+(2-SP-SN)(1-SP)/(2-SP-SN)\}} \\ \geq & \frac{SN+SP-1}{g'\{g^{-1}(\theta_0)\}} + \frac{2-SN-SP}{g'\{(1-SP)/(2-SP-SN)\}} \geq \frac{SN+SP-1}{g'\{g^{-1}(\theta_0)\}}. \end{aligned}$$

The first inequality is due to the concavity of $1/g'(x) = x(1-x)$. Thus, we obtain that $0 \leq H(\theta_0) \leq 1$, which means that using the observed outcome leads to attenuated estimates of the outcome regression model. It is straightforward to obtain the asymptotic biases of the direct and indirect effects by Plugging the regression models into (1) to (3).

Correction approach for direct and indirect effect estimators

We use the EM algorithm to correct the bias result from the misclassified outcome. For the correction approach, we should first specify the sensitivity and specificity parameters, then implement the EM algorithm to obtain the parameter estimates of the regression models. Estimates of direct and indirect effects are recovered by plugging the estimated parameters in (1) to (3).

Dempster, Laird and Rubin[2] propose the Expectation-Maximization (EM) algorithm to obtain the maximum likelihood estimation when there is missing data. In our problem misclassification of the true outcome Y can be seen as a type of missing data problem, and we can write the log-likelihood as:

$$\begin{aligned} & l(\boldsymbol{\theta}) = l_{Y|A,M,C}(\boldsymbol{\theta}|Y_i, A_i, M_i, C_i) \\ = \;& Y_i l_{Y|A,M,C}(\boldsymbol{\theta}|Y_i = 1, A_i, M_i, C_i) + (1 - Y_i) l_{Y|A,M,C}(\boldsymbol{\theta}|Y_i = 0, A_i, M_i, C_i). \end{aligned}$$

The EM algorithm consists of two steps, the E-step and the M-step. At the E-step, we calculate the expectation of the log-likelihood conditioning on the current parameters and the observed data $\{A, M, Y^*, \boldsymbol{C}\}$. Denote this expectation by $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$, we have that:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)}) \;=\; \sum_{i=1}^{n} w_{it}(\boldsymbol{\theta}^{(k)}) l_{Y|A,M,C}(\boldsymbol{\theta}|Y_i, A_i = t, M_i, C_i),$$

where

$$\begin{aligned} & w_{it}(\boldsymbol{\theta}^{(k)}) = P(Y_i = y|Y_i^*, A_i, M_i, C_i) \\ = \;& \frac{P(Y_i^*|Y_i = y) P(Y_i = y|A_i, M_i, C_i)}{\sum_{y'=0,1} P(Y_i^*|Y_i = y') P(Y_i = y'|A_i = t, M_i, C_i)}. \end{aligned}$$

In the above formula, $P(Y_i^*|Y_i)$ depends on the pre-specified SN and SP. In practice, we can set a plausible range of values for sensitivity analysis.

At the M-step, we update the parameters by maximizing $Q(\boldsymbol{\theta})$ using weighted logistic regression.

References

[1] Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 1999;86(4):843--855.

[2] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal statistical Society: Series B* 1977;39(1):1--38.