To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias

Abstract

"*M*-Bias", as it is called in the epidemiological literature, is the bias introduced by conditioning on a pretreatment covariate due to a particular "*M*-Structure" between two latent factors, an observed treatment, an outcome, and a "collider". This potential source of bias, which can occur even when the treatment and the outcome are not confounded, has been a source of considerable controversy. We here present formulae for identifying under which circumstances biases are inflated or reduced. In particular, we show that the magnitude of *M*-Bias in Gaussian linear structural equation models tends to be relatively small compared to confounding bias, suggesting that it is generally not a serious concern in many applied settings. These theoretical results are consistent with recent empirical findings from simulation studies. We also generalize the *M*-Bias setting to allow for the correlation between the latent factors to be nonzero, and to allow for the collider to also be a confounder between the treatment and the outcome. These results demonstrate that mild deviations from the *M*-Structure tend to increase confounding bias more rapidly than *M*-bias, suggesting that choosing *to* condition on any given covariate is generally the superior choice. As an application, we re-examine a controversial example between Professors Donald Rubin and Judea Pearl.

Key Words: Causality; Collider; Confounding; Controversy; Covariate.

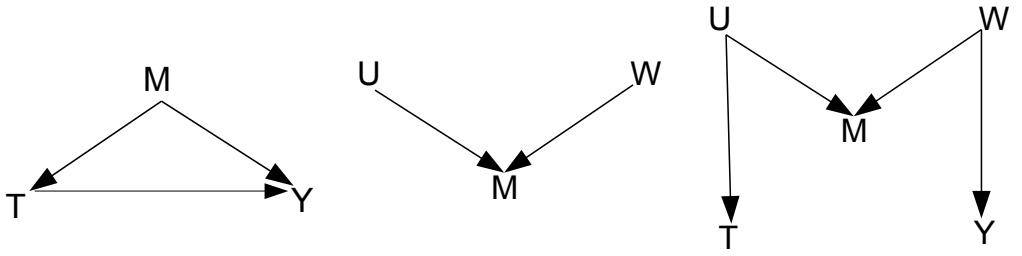
1 Introduction

Many statisticians believe that "there is no reason to avoid adjustment for a variable describing subjects before treatment" in observational studies (Rosenbaum, 2002, pp 76), because "typically, the more conditional an assumption, the more generally acceptable it is" (Rubin, 2009). This advice, recently dubbed the "pretreatment criterion" (VanderWeele and Shpitser, 2011), is widely used in empirical studies, as more covariates generally seem to make the ignorability assumption, i.e., the assumption that conditionally on the observed pretreatment covariates, treatment assignment is independent of the potential outcomes (Rosenbaum and Rubin, 1983), more plausible. As the validity of causal inference in observational studies relies strongly on this (untestable) assumption (Rosenbaum and Rubin, 1983), it seems reasonable to make all efforts to render it plausible.

However, other researchers (Pearl, 2009a,b, Shrier, 2008, 2009, Sjölander, 2009), mainly from the causal diagram community, do not accept this view because of the possibility of a so-called M-Structure, illustrated in Figure 1(c). As a main opponent to Rubin and Rosenbaum's advice, Pearl (2009a) and Pearl (2009b) warn practitioners that spurious bias may arise due to adjusting for a collider M in an M-Structure, even if it is a pretreatment covariate. This form of bias, typically called M-bias, a special version of so-called "collider bias", has since generated considerable controversy and confusion.

We attempt to resolve some of these debates by an analysis of M-bias under the causal diagram or directed acyclic graph (DAG) framework. For readers unfamiliar with the terminologies from the DAG (or Bayesian Network) literature, more details can be found in Pearl (1995) or Pearl (2000). We here use only a small part of this larger framework. Arguably the most important structure in the DAG, and certainly the one at root of almost all controversy, is the "V-Structure" illustrated in Figure 1(b). Here, U and W are marginally independent with a common outcome M, which shapes a "V" with the vertex M being called a "collider". From a data-generation viewpoint, one might imagine Nature generating data in two steps: She first picks independently two values for U and W from two distributions, and then she combines them (possibly along with some additional random variable) to create M. Given this, conditioning on M can cause a spurious correlation between U and W, which is known as the collider bias (Greenland, 2002), or, in epidemiology, Berkson's Paradox (Berkson, 1946). Conceptually, this correlation happens because if one cause of an observed outcome is known to have not occurred, the other cause becomes more likely. Consider an automatic-timer sprinkler system where the sprinkler being on is independent of whether it is raining. Here, the weather gives no information on the sprinkler. However, given wet grass, if one observes a sunny day, one will likely conclude that the sprinklers have recently run.

Correlation has been induced.



(a) A simple DAG (b) V-Structure $U \perp\!\!\!\perp W$ but $U \not\perp\!\!\!\perp W \mid M$ (c) M-Structure $T \perp\!\!\!\perp Y$ but $T \not\perp\!\!\!\perp Y \mid M$

Figure 1: Three DAGs

Where things get interesting is when this collider is made into a *pre-treatment* variable. Consider Figure 1(c), an extension of Figure 1(b). Here U and W are now also causes of the treatment T and the outcome Y, respectively. Nature, as a last, third step generates T as a function of U and some randomness, and Y as a function of W and some randomness. This structure is typically used to represent a circumstance where a researcher observes T, Y, and M in nature and is attempting to derive the causal impact of T on Y. However, U and W are sadly unobserved, or latent. Clearly, the causal effect of T on Y is zero, which is also equal to the marginal association between T and Y. If a researcher regressed Y on T, he or she would obtain a 0 in expectation which is correct for estimating the causal effect of T on Y. But perhaps there is a concern that M, a pretreatment covariate, may be a confounder that is masking a treatment effect. Typically, one would then "adjust" for M to take this possibility into account, e.g., by including M in a regression or by matching units on similar values of M. If we do this in this circumstance, however, then we will not find zero, in expectation. This is the so-called "M-Bias," and this special structure is called the "M-Structure" in the DAG literature.

Previous qualitative analysis for binary variables shows that collider bias generally tends to be small (Greenland, 2002), and simulation studies (Liu, Brookhart, Schneeweiss, Mi, and Setoguchi, 2012) again demonstrate that M-Bias is small in many realistic settings. While mathematically describing the magnitudes of M-Bias in general models is intractable, it is possible to derive exact formulae of the biases as functions of the correlation coefficients in Gaussian linear structural equation models (GLSEMs). The GLSEM has a long history in statistics (Wright, 1921, 1934) to describe dependence among multiple random variables. Sprites (2002) uses linear models to illustrate M-Bias in observational studies, and Pearl (2013a,b) also utilize the transparency of such linear models to examine various types of

causal phenomena, biases, and paradoxes. We here extend these works and provide exact formulae for biases, allowing for a more detailed quantitative analysis of M-bias.

While M-Bias does exist when the true underlying data generating process (DGP) follows the exact M-Structure, it might be rather sensitive to various deviations from the exact M-Structure. Furthermore, some might argue that an exact M-Structure is unlikely to hold in practice. Gelman (2011), for example, doubts the exact independence assumption required for the M-Structure in the social sciences by arguing that there are "(almost) no true zeros" in this discipline. Indeed, since U and W are often latent characteristics of the same individual, the independence assumption $U \perp\!\!\!\perp W$ is a rather strong structural assumption. Furthermore, it might be plausible that the pretreatment covariate M is also a confounder between, i.e., has some causal impact on both, the treatment and outcome. We extend our work by accounting for these departures from a pure M-Structure, and find that even slight departures from the M-Structure can dramatically change the forms of the biases.

This paper theoretically compares the bias from conditioning on an M to not under several scenarios and finds that M-Bias is indeed small unless there is a strong correlation structure for the variables. We further show that these findings extend to a binary treatment regime as well. This argument proceeds in several stages. First, in Section 2, we examine a pure M-Structure and introduce our GLSEM framework. We then discuss the cases when the latent variables U and W may be correlated and M may also be a confounder between the treatment T and the outcome Y. In Section 3, we generalize the results in Section 2 to a binary treatment. In Section 4, we illustrate the theoretical findings using a controversial example between Professors Donald Rubin and Judea Pearl (Rubin, 2007, Pearl, 2009b). We conclude with a brief discussion and present all technical details in the Appendix.

2 M-Bias and Butterfly-Bias in GLSEMs

We begin by examining pure M-Bias in a GLSEM. As our primary focus is bias, we assume data is ample and that anything estimable is estimated with nearly perfect precision. In particular, when we say we obtain some result from a regression, we implicitly mean we obtain that result in expectation; in practice an estimator will be near the given quantities. We do not compare relative uncertainties of different estimators given the need to estimate more or fewer parameters. There are likely degrees-of-freedom issues that would implicitly advocate using estimators with fewer parameters, but in the circumstances considered here these concerns are likely to be minor as all the models have few parameters.

A causal DAG can be viewed as a hierarchical DGP. In particular, any variable on the graph R can be viewed as a function of its parents and some additional noise, i.e., if R had parents A, B, and C we would have

$$R = f(A, B, C, \varepsilon_R) \text{ with } \varepsilon_R \perp\!\!\!\perp (A, B, C).$$

Generally noise terms such as ε_R are considered to be independent from each other, but they can also be given an unknown correlation structure corresponding to earlier variables not explicitly included in the diagram. This is typically represented by drawing the dependent noise terms jointly from some multivariate distribution. This framework is quite general; we can represent any distribution that can be factored as a product of conditional distributions corresponding to a DAG (which is one representation of the Markov Condition, a fundamental assumption for DAGs).

GLSEMs are special cases of the above with additional linearity, Gaussianity, and additivity constraints. For simplicity, and without loss of generality, we also rescale all primary variables (U, W, M, T, Y) to have zero mean and unit variance. For example, consider this data generating process corresponding to Figure 1(a):

$$\begin{aligned} M, \varepsilon_T, \varepsilon_Y &\overset{\text{iid}}{\sim} N(0,1), \\ T &= aM + \sqrt{1-a^2}\varepsilon_T, \\ Y &= bT + cM + \sqrt{1-b^2-c^2}\varepsilon_Y. \end{aligned}$$

In the causal DAG literature, we think about causality as reaching in and fixing a given node to a set value, but letting Nature take her course otherwise. For example, if we were able to set T at t, the above data generation process would be transformed to:

$$\begin{aligned} M, \varepsilon_T, \varepsilon_Y &\overset{\text{iid}}{\sim} N(0,1), \\ T &= t, \\ Y &= bt + cM + \sqrt{1-b^2-c^2}\varepsilon_Y. \end{aligned}$$

The previous cause, M, of T has been broken, but the impact of T on Y remains intact. This changes the distribution of Y but not M. More importantly, this results in a distribution distinct from that of *conditioning* on $T = t$. Consider the case of positive a, b, and c. If we *observe* a high T, we can infer a high M (as T and M are correlated) and a high Y due to both the bT and cM terms in Y's equation. However, if we *set* T to a high value, M is unchanged. Thus, while we will still have the large bT term for Y, the cM term will be 0 in expectation. Thus, the expected value for Y will be less.

This setting as compared to conditioning is represented with the "do" operator. Given the "do" operator, we define a local causal effect of T on Y at $T = t$ as:

$$\tau_t = \left. \frac{\partial \mathbb{E}\{Y \mid \text{do}(T) = x\}}{\partial x} \right|_{x=t}.$$

For linear models, the local causal effect is a constant, and thus we do not need to specify t. We use "do" here purely to indicate the different distributions. For a more technical overview, see Pearl (1995) or Pearl (2000). Our results, with more formality, can easily be expressed in this more technical notation.

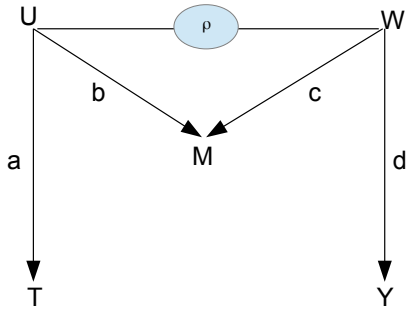


Figure 2: M-Structure with Possibly Correlated Hidden Causes

If we extend the M-Structure in Figure 1(c) by allowing possible correlation between the two hidden causes U and W, we obtain the DAG in Figure 2. This in turn gives the following DGP:

$$\begin{aligned}
\varepsilon_M, \varepsilon_T, \varepsilon_Y &\overset{\text{iid}}{\sim} N(0,1), \\
(U,W) &\sim \boldsymbol{N}_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}, \\
M &= bU + cW + \sqrt{1-b^2-c^2}\varepsilon_M, \\
T &= aU + \sqrt{1-a^2}\varepsilon_T, \\
Y &= dW + \sqrt{1-d^2}\varepsilon_Y.
\end{aligned}$$

Here, the true causal effect of T on Y is zero, namely, $\tau_t = 0$ for all t. The unadjusted estimator for the causal effect obtained by regressing Y onto T is the same as the covariance between T and Y:

$$\text{Bias}_{unadj} = \text{Cov}(T,Y) = \text{Cov}(aU, dW) = ad\text{Cov}(U,W) = ad\rho.$$

The adjusted estimator (see Lemma 2 in Appendix A for a proof) obtained by regressing Y onto (T,M) is

$$\text{Bias}_{adj} = \frac{ad\rho(1-b^2-c^2-bc\rho)-abcd}{1-(ab+ac\rho)^2}.$$

The results above and some of the results discussed later in this paper can be obtained directly from traditional path analysis (Wright, 1921, 1934). However, we provide elementary proofs, which can easily be extended to binary treatment, in the Appendix. In the Gaussian case, if we allowed for a treatment effect, our results would remain essentially unchanged; the only difference would be due to restrictions on the correlation terms needed to maintain unit variance for all variables.

The above can also be expressed in the potential outcomes framework (Neyman, 1923/1990, Rubin, 1974). In particular, for a given unit let Nature draw $\varepsilon_M, \varepsilon_T, \varepsilon_Y, U$, and V as before. Let T be the "natural treatment" for that unit, i.e., what treatment it would receive sans intervention. Then calculate $Y(t)$ for any t of interest using the "do" operator. These are what we would see if we set $T = t$. How $Y(t)$ changes for a particular unit defines that unit's collection of potential outcomes. Then $\mathbb{E}\{Y(t)\}$ for some t is the expected potential outcome over the population for a particular t. We can examine the derivative of this function as above to get a local treatment effect. This connection is exact: the findings in this paper are the same as what one would find using this DGP and the potential outcomes framework. We here examine regression as the estimator. Note that matching would produce identical results as the amount of data grew (assuming the data generating process ensures common support, etc.).

Exact M-Bias. The M-Bias originally considered in the literature is the special case where the correlation coefficient between U and W is $\rho = 0$. In this case, the unadjusted estimator is unbiased and the absolute bias of the adjusted estimator is $|abcd|/\{1-(ab)^2\}$. With moderate correlation coefficients a,b,c,d the denominator $1-(ab)^2$ is close to one, and the bias is close to $-abcd$. Since $abcd$ is a product of four correlation coefficients, it can be viewed as a "higher order bias." For example, if $a = b = c = d = 0.2$, then $1-(ab)^2 = 0.9984 \approx 1$, and the bias of the adjusted estimator is $-abcd/\{1-(ab)^2\} = -0.0016 \approx 0$; if $a = b = c = d = 0.3$, then $1-(ab)^2 = 0.9919 \approx 1$, and the bias of the adjusted estimator is $-abcd/\{1-(ab)^2\} = -0.0082 \approx 0$. Even moderate correlation results in little bias.

In Figure 3, we plot the bias of the adjusted estimator as a function of the correlation coefficients, and let these coefficients change to see how the bias changes. In the first subfigure, we assume all the correlation coefficients have the

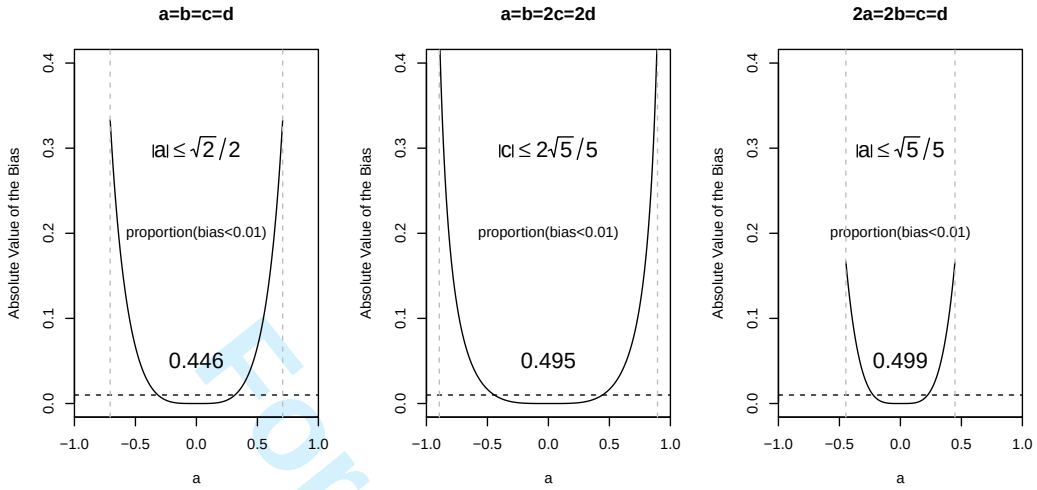


Figure 3: M-Bias with Independent U and W. The three subfigures correspond to the cases when (U,V) are equally/more/less predictive to the treatment than to the outcome. In each subfigure, we show the proportions of the areas where the adjusted estimator has a bias smaller than 0.01.

same magnitude ($a = b = c = d$), and we plot the absolute bias of the adjusted estimator versus a. Here, bias is low unless a is large. Note that the constraints on variance and correlation only allow for some combinations of values for a, b, c and d which limits the domain of the figures. In this case, for example, $|a| \leq \sqrt{2}/2$ due to the requirement that $b^2 + c^2 = 2a^2 \leq 1$. Other figures have limited domains due to similar constraints.

In the second subfigure of Figure 3, we assume that M is more predictive to the treatment T than to the outcome Y, with $a = b = 2c = 2d$. In the third subfigure of Figure 3, we assume that M is more predictive to the outcome Y, with $2a = 2b = c = d$. In all these cases, although the biases do blow up when the correlation coefficients are extremely large, they are generally very small within a wide range of the feasible region of the correlation coefficients. In examining the formula, note that if any correlation is low, the bias will be low.

In Figure 4(a), we assume $a = b$ and $c = d$ and examine a broader range of relationships. Here, the grey area satisfies $|\text{Bias}_{adj}| < \min(|a|, |c|)/20$. For example, when the absolute values of the correlation coefficients are smaller than 0.5 (the square with dashed boundary in Figure 4(a)), the corresponding area is almost grey, implying small bias.

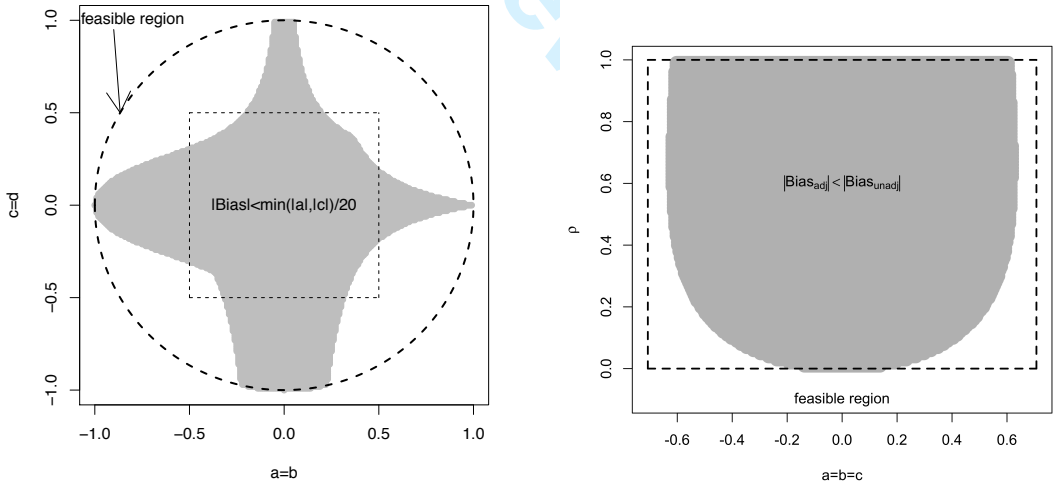
As a side note, Pearl (2013b) noticed a surprising fact: the stronger the correlation between T and M, the larger the absolute bias of the adjusted estimator,

since the absolute bias is monotone increasing in $|ab|$. From the second and the third subfigure of Figure 3, we see that when M is more predictive of the treatment, the biases of the adjusted estimator indeed tends to be larger.

Correlated latent variables. When the latent variables U and W are correlated with $\rho \neq 0$, *both* the unadjusted and adjusted estimators may be biased. The question then becomes: which is worse? The ratio of the absolute biases of the adjusted and unadjusted estimators is

$$\left|\frac{\text{Bias}_{adj}}{\text{Bias}_{unadj}}\right| = \left|\frac{\rho(1-b^2-c^2-bc\rho)-bc}{\rho\{1-(ab+ac\rho)^2\}}\right|,$$

which does not depend on d (the relationship between W and Y). For example, if the correlation coefficients a, b, c, ρ all equal 0.2, the ratio above is 0.714; in this case the adjusted estimator is superior to the unadjusted one by a factor of 1.4. Figure 4(b) compares this ratio to 1 for all combinations of ρ and $a(=b=c)$. Generally, the adjusted estimator has smaller bias except when a, b, and c are quite large.



(a) Pure M-Bias. Within the grey region, the absolute bias of the adjusted estimator is less than $1/20$ of the minimum of $|a|(=|b|)$ and $|c|(=|d|)$.

(b) M-Bias with Correlated U and W. Within the grey region, the adjusted estimator is superior.

Figure 4: M-Bias under different scenarios

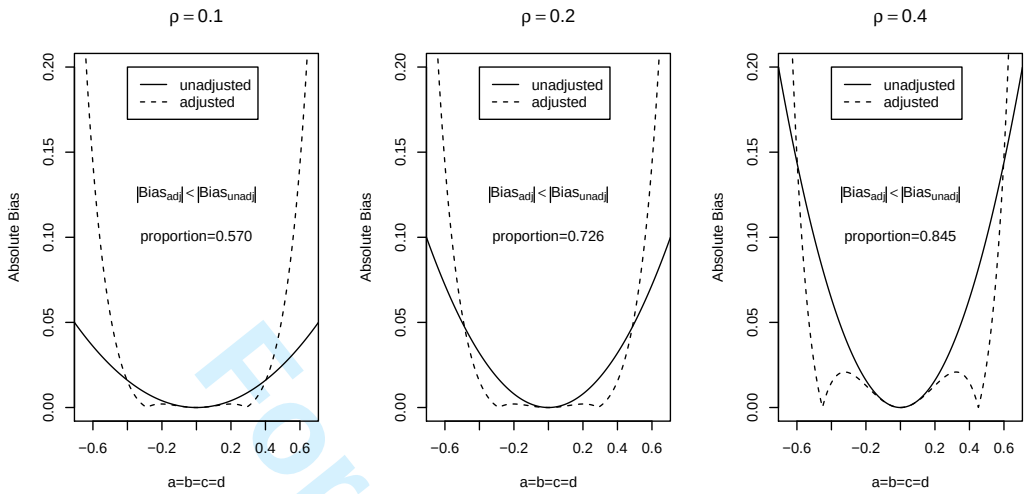


Figure 5: M-Bias with correlated U and W when $\rho \neq 0$ with $a = b = c = d$. In each subfigure, we show the proportion of the areas where the adjusted estimator has a smaller bias than the unadjusted estimator.

In Figure 5, we again assume $a = b = c = d$ and investigate the absolute biases as functions of a for fixed ρ at 0.1, 0.2, and 0.4. When the correlation coefficients $a(= b = c = d)$ are not dramatically larger than ρ, the adjusted estimator has smaller bias than the unadjusted one.

The Disjunctive Cause Criterion. In order to remove biases in observational studies, VanderWeele and Shpitser (2011) propose a new "disjunctive cause criterion" for selecting confounders, which requires controlling for all the covariates that are either causes of the treatment, or causes of the outcome, or causes of both of them. According to the "disjunctive cause criterion", when $\rho \neq 0$, we should control for (U,V) if possible. Unfortunately, neither of (U,V) is observable. However, controlling the "proxy variable" M for (U,V) may reduce bias when ρ is relatively large. In the special case with $b = 0$, the ratio of the absolute biases is

$$\left| \frac{\text{Bias}_{adj}}{\text{Bias}_{unadj}} \right| = \frac{1-c^2}{1-(ac\rho)^2} \leq 1;$$

in another special case with $c = 0$, the ratio of the absolute biases is

$$\left| \frac{\text{Bias}_{adj}}{\text{Bias}_{unadj}} \right| = \frac{1-b^2}{1-(ab)^2} \leq 1. \quad (1)$$

Therefore, if either U or W is not causative to M, the adjusted estimator is always better than the unadjusted one.

Butterfly-Bias: M-Bias with Confounding Bias Models, especially in the social sciences, are approximations. They rarely hold exactly. In particular, for any covariate M of interest, there is likely to be some concern that M is indeed a confounder, even if it is also possible source of M-Bias. If we let M both be a confounder as well as the middle of an M-Structure we obtain a "Butterfly-Structure" (Pearl, 2013b) as shown in Figure 6. In this circumstance, conditioning will help with confounding bias, but hurt with M-Bias. Ignoring M will not resolve any confounding, but avoid M-Bias. The question then becomes that of determining which is the lesser of the two evils.

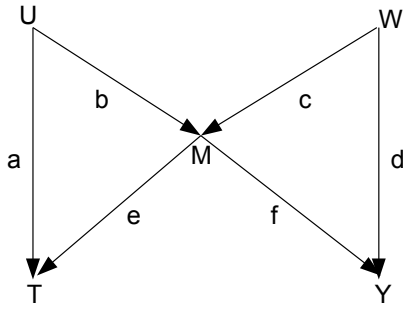


Figure 6: Butterfly-Structure

We now examine this trade-off for a GLSEM corresponding to Figure 6. The DGP is given by the following equations:

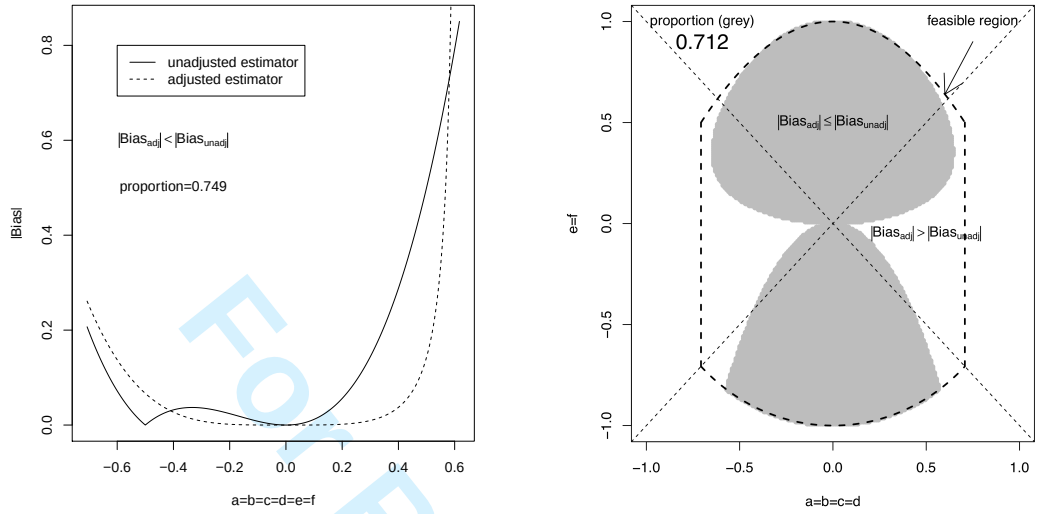
$$\begin{aligned} U, W, \varepsilon_M, \varepsilon_T, \varepsilon_Y &\overset{\text{iid}}{\sim} N(0,1), \\ M &= bU + cW + \sqrt{1-b^2-c^2}\varepsilon_M, \\ T &= aU + eM + \sqrt{1-a^2-e^2}\varepsilon_T, \\ Y &= dW + fM + \sqrt{1-d^2-f^2}\varepsilon_Y. \end{aligned}$$

Again, the true causal effect of T on Y is zero. The unadjusted estimator obtained by regressing Y onto T is the covariance between T and Y:

$$\text{Bias}_{unadj} = \text{Cov}(T,Y) = abf + cde + ef.$$

It is not, in general, 0, implying bias. The adjusted estimator (see Lemma 3 in Appendix for a proof) obtained by regressing Y onto (T,M) has bias

$$\text{Bias}_{adj} = -\frac{abcd}{1-(ab+e)^2}.$$



(a) Absolute biases of both estimators with $a = b = c = d = e = f$.

(b) Comparison of the absolute biases with $a = b = c = d$ and $e = f$. Within 71.2% (in grey) of the feasible region, the adjusted estimator has smaller bias than the unadjusted one.

Figure 7: Butterfly-Bias

If the values of e and f are relatively high (i.e., M has a strong effect on both T and Y), the confounding bias is large and the unadjusted estimator will be severely biased. For example, if a, b, c, d, e, and f all equal 0.2, the bias of the unadjusted estimator is 0.056, but the bias of the adjusted estimator is only -0.0017, an order of magnitude smaller. Generally, the largest term for the unadjusted bias is the second-order term of ef, while the adjusted bias only has, ignoring the denominator, a fourth-order term of $abcd$. This suggests adjustment is generally preferable and that M-bias is in some respect a "higher order bias."

Detailed comparison of the ratio of the biases is difficult, since we can vary six parameters (a, b, c, d, e, f). In Figure 7(a), we assume all the correlation coefficients have the same magnitude, and plot bias for both estimators as a function of the correlation coefficient within the feasible region, defined by the restrictions $-\sqrt{2}/2 < a < (-1+\sqrt{5})/2$, due to the restrictions

$$b^2 + c^2 < 1,\ a^2 + e^2 < 1,\ d^2 + f^2 < 1,\ \text{and}\ |a^2 + e| < 1. \tag{2}$$

Within 74.9% of the feasible region, the adjusted estimator has smaller bias than the

unadjusted one. The unadjusted estimator only has smaller bias than the adjusted estimator when the correlation coefficients are extremely large. In Figure 7(b), we assume $a = b = c = d$ and $e = f$, and compare $|\text{Bias}_{adj}|$ and $|\text{Bias}_{unadj}|$ within the feasible region of (a, e) defined by (2). We can see that the adjusted estimator is superior to the unadjusted one for 71% (colored in grey in Figure 7(b)) of the feasible region. In the area satisfying $|e| > |a|$ in Figure 7(b), where the connection between M to T and Y is stronger than the other connections, the area is almost grey suggesting that the adjusted estimator is preferable. This is sensible because here the confounding bias has larger magnitude than the M-Bias. In the area satisfying $|a| < |e|$, where M-bias is stronger than confounding bias, the unadjusted estimator is superior for some values, but still tends to be inferior when the correlations are roughly the same size.

3 Extensions to a Binary Treatment

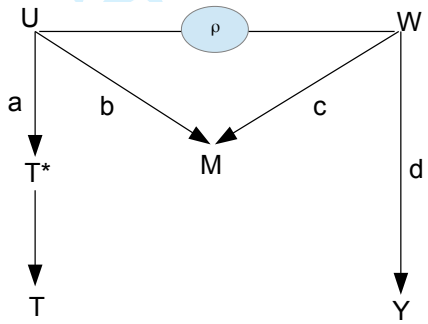


Figure 8: M-Structure with Possibly Correlated Hidden Causes and a Binary Treatment

If the treatment is binary, one might worry that the conclusions in the previous section are not applicable. It turns out, however, that they are. In this section, we extend the results in Section 2 to binary treatments by representing the treatment through a latent Gaussian variable. In particular, extend Figure 2 to Figure 8. Here, T^* is the old T. The generating equations for T and T^* become

$$T = I(T^* \geq \alpha), \text{ and } T^* = aU + \sqrt{1-a^2}\varepsilon_T.$$

Other variables and noise terms remain the same. The intercept α determines the proportion of the individuals receiving the treatment: $\Phi(-\alpha) = P(T = 1)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard Normal distribution. When $\alpha = 0$, the individuals exposed to the treatment and the control are

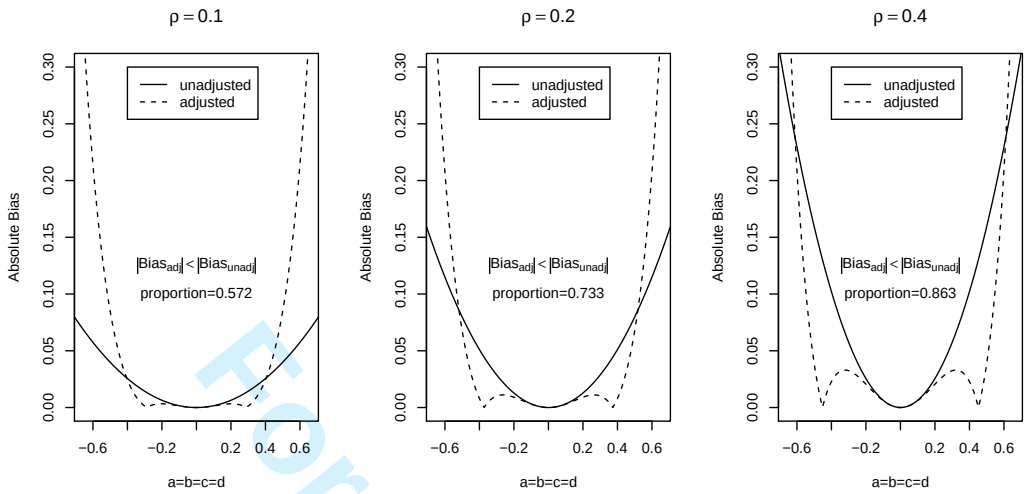


Figure 9: M-Bias with correlated U and W and binary treatment. Compare to Figure 5.

balanced; when $\alpha < 0$, more individuals are exposed to the treatment; when $\alpha > 0$, the reverse.

The true causal effect of T on Y is again zero. Let $\phi(\cdot) = \Phi'(\cdot)$ and $\eta(\alpha) \equiv \phi(\alpha)/\{\Phi(\alpha)\Phi(-\alpha)\}$. Then Lemma 6 in Appendix shows that the unadjusted estimator has bias

$$\text{Bias}_{unadj} = ad\rho\eta(\alpha),$$

and the adjusted estimator has bias

$$\text{Bias}_{adj} = \frac{ad\eta(\alpha)\{\rho(1-b^2-c^2-bc\rho)-bc\}}{\rho\{1-(ab+ac\rho)^2\phi(\alpha)\eta(\alpha)\}}.$$

When $\rho = 0$, the unadjusted estimator is unbiased, but the adjusted estimator has bias

$$-\frac{abcd\eta(\alpha)}{1-(ab)^2\phi(\alpha)\eta(\alpha)}.$$

When $\rho \neq 0$, the ratio of the absolute biases of the adjusted and unadjusted estimators is

$$\left|\frac{\text{Bias}_{adj}}{\text{Bias}_{unadj}}\right| = \left|\frac{\rho(1-b^2-c^2-bc\rho)-bc}{\rho\{1-(ab+ac\rho)^2\phi(\alpha)\eta(\alpha)\}}\right|.$$

The patterns for a binary treatment do not differ much from a Gaussian treatment. As before, if the correlation coefficient is moderately small, the M-Bias

also tends to be small. As shown in Figure 9 (analogous to Figure 5), when $|\rho|$ is comparable to $|a|(=|b|=|c|=|d|)$, the adjusted estimator is less biased than the unadjusted estimator. Only when $|a|$ is much larger than $|\rho|$ is the unadjusted estimator superior.

Butterfly-Bias with a Binary Treatment We can extend the GLSEM Butterfly-Bias setup to binary treatment just as we extended the M-Bias setup. Compare Figure 10 to Figure 6 and Figure 8 to Figure 2. T becomes T^* and T is built from T^* as above. The structural equations for T and T^* for butterfly bias in the binary case are then

$$T = I(T^* \geq \alpha), \text{ and } T^* = aU + eM + \sqrt{1-a^2-e^2}.$$

The other equations and variables remain the same as before.

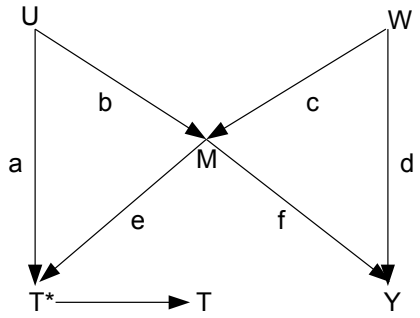


Figure 10: Butterfly-Structure with a Binary Treatment

Although the true causal effect of T on Y is zero, Lemma 7 in Appendix shows that the unadjusted estimator has bias

$$\text{Bias}_{unadj} = (cde + abf + ef)\eta(\alpha), \tag{3}$$

and the adjusted estimator has bias

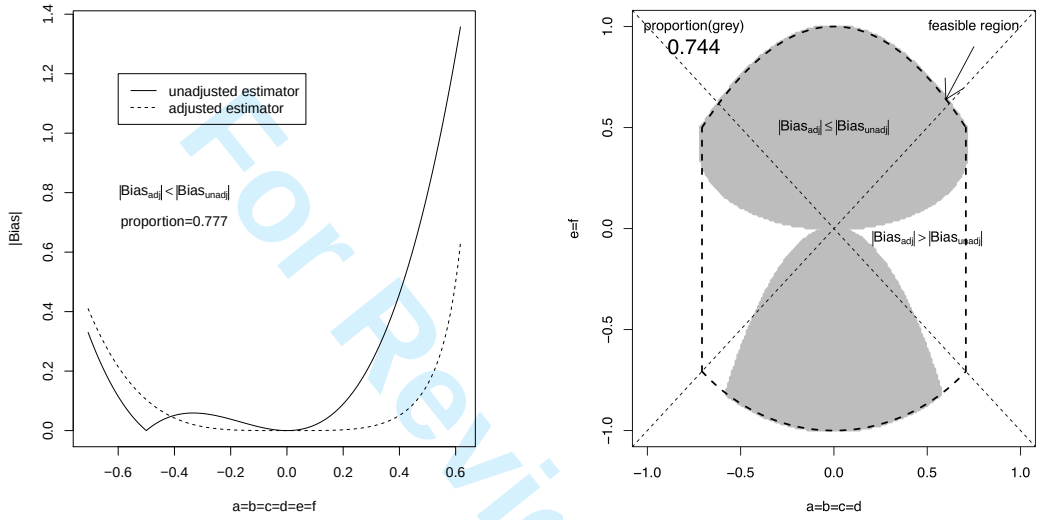
$$\text{Bias}_{adj} = -\frac{abcd\eta(\alpha)}{1-(ab+e)^2\phi(\alpha)\eta(\alpha)}. \tag{4}$$

Therefore, the ratio of the absolute biases of the adjusted and unadjusted estimators is

$$\left|\frac{\text{Bias}_{adj}}{\text{Bias}_{unadj}}\right| = \left|\frac{abcd\eta(\alpha)}{(cde+abf+ef)\{1-(ab+e)^2\phi(\alpha)\eta(\alpha)\}}\right|.$$

Complete investigation of the ratio of the biases is intractable with seven varying parameters (a,b,c,d,e,f,α). However, in the very common case with

$\alpha = 0$, which gives equal-sized treatment and control groups, we again find trends similar to the Gaussian case. See Figure 11. As before, only in the cases with very small $e(= f)$ but large $a(= b = c = d)$, does the unadjusted estimator tend to be superior. Within a reasonable region of α, these patterns are quite similar.



(a) Absolute biases with $a = b = c = d = e = f$.

(b) Comparison of the absolute bias with $a = b = c = d$ and $e = f$. Within 74.4% (in grey) of the feasible region, the adjusted estimator is better than the unadjusted estimator.

Figure 11: Butterfly-Bias with a Binary Treatment

4 Illustration: Rubin-Pearl Controversy

Pearl (2009b) cites Rubin (2007)'s example about the causal effect of smoking habits (T) on lung cancer (Y), and argues that conditioning on the pretreatment covariate "seat-belt usage" (M) would introduce spurious associations, since M could be reasonably thought of as an indicator of a person's attitudes toward society norms (U) as well as safety and health related measures (W). Assuming all the analysis is already conditioned on other observed covariates, we focus our discussion on the five variables (U, W, M, T, Y), of which the dependence structure is illustrated by

Figure 12. Since the patterns with a Gaussian treatment and a binary treatment are similar, we focus our discussion under the assumption of a Gaussian treatment.

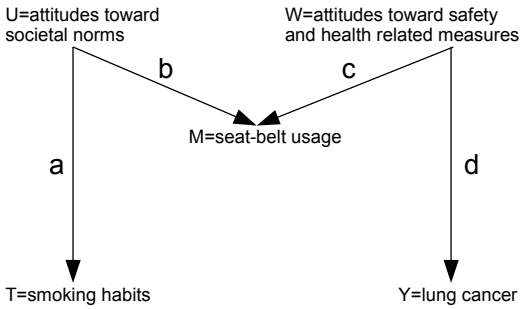


Figure 12: A Potential M-Structure in Rubin's Example (Rubin, 2007, Pearl, 2009b)

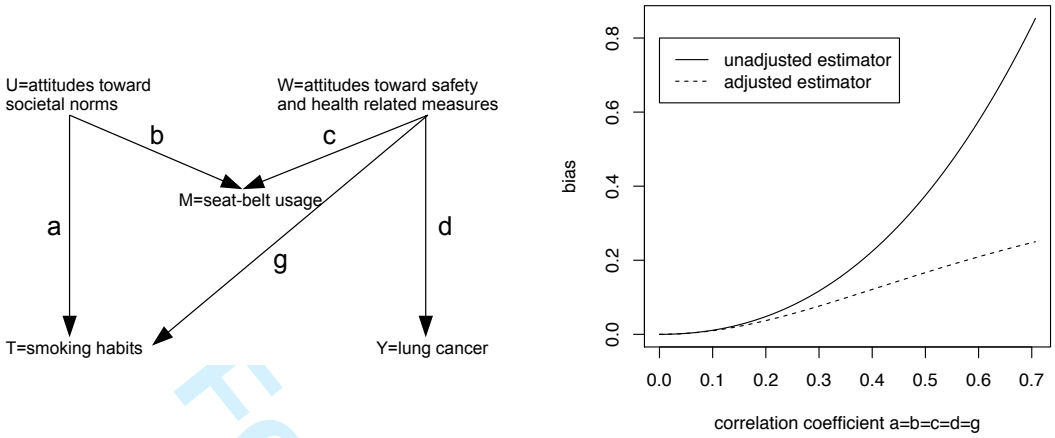
As Pearl (2009b) points out,

> *If we have good reasons to believe that these two types of attitudes are marginally independent, we have a pure M-structure on our hand.*

In the case with $\rho = 0$, conditioning on M will lead to spurious correlation between T and Y under the null, and will bias the estimation of causal effect of T on Y. However, Pearl (2009b) also recognizes that the independence assumption seems very strong in this example, since U and W are both background variables about the habit and personality of a person. Pearl (2009b) further argues:

> *But even if marginal independence does not hold precisely, conditioning on "seat-belt usage" is likely to introduce spurious associations, hence bias, and should be approached with caution.*

Although we believe most things should be approached with caution, our work, above, suggests that even mild perturbations of an M-Structure can switch which of the two approaches, conditioning or not conditioning, is likely to remove more bias. In particular, Pearl (2009b) is correct in that the adjusted estimator indeed tends to introduce more bias than the unadjusted one when an exact M-Structure holds and thus the general advice "to condition on all observed covariates" may not be sensible in this context. However, in the example of Rubin (2007), the exact independence between a person's attitude toward societal norms U and safety and health related measures W is questionable, since we have good reasons to believe that other hidden variables such as income and family background will affect both U and W simultaneously, and thus Pearl's fears may be unfounded.



(a) M-Bias with an Additional Arrow from W to T

(b) Biases of resulting unadjusted and adjusted estimators with $a = b = c = d = g$.

Figure 13: Sensitivity Analysis of Pearl (2009b)'s Critique on Rubin (2007)'s Example

To examine this further, we consider two possible deviations from the exact M-Structure, and investigate the biases of the unadjusted and adjusted estimators for each.

(a) (Correlated U and W) Assume the DGP follows the DAG in Figure 12, with an additional correlation between the attitudes U and W as shown in Figure 2. If we then assume that all the correlation coefficients have the same positive magnitude, earlier results demonstrate that the adjusted estimator is preferable as it strictly dominates the unadjusted estimator except for extremely large values of the correlation coefficients.
Furthermore, in Rubin (2007)'s example, attitudes toward societal norms U are more likely to affect the "seat-belt usage" variable M than safety and health related measures W, which further strengthens the case for adjustment. If we were willing to assume that c is zero but ρ is not, equation (1) again shows that the adjusted estimator is superior.

(b) (An arrow from W to T) Pearl's example seems a bit confusing on further inspection, even if we accept his independence assumption $U \perp\!\!\!\perp W$. In particular, one's "attitudes towards safety and health related measures" likely impact one's decisions about smoking. Therefore, we might reasonably expect an arrow from W to T. In Figure 13(a), we remove the correlation between U and W, but we allow an arrow from W to T, i.e., the generating equation for T becomes $T = aU + gW + \sqrt{1-a^2-g^2}\varepsilon_T$. Lemma 8 in Appendix gives the associated

formulae for biases of the adjusted and unadjusted estimators. Figure 13(b) shows that, assuming $a = b = c = d = g$ (i.e., equal correlations), the adjusted estimator is uniformly better.

5 Discussion

For objective causal inference, Rubin and Rosenbaum suggest balancing all the pretreamtent covariate in observational studies to parallel with the design of randomized experiments (Rubin, 2007, 2008, 2009, Rosenbaum, 2002), which is called the "pretreatment criterion" (VanderWeele and Shpitser, 2011). However, Pearl and other researchers (Pearl, 2009a,b, Shrier, 2008, 2009, Sjölander, 2009) criticize the "pretreatment criterion" by pointing out that this criterion may lead to biased inference in presence of a possible M-Structure even if the treatment assignment is unconfounded. While we agree that Pearl's warning is very insightful, our theoretical and numerical results show that this conclusion is quite sensitive to various deviations from the exact M-Structure, e.g., to circumstances where latent causes may be correlated or the M variable may also be a confounder between the treatment and the outcome. In particular, results derived from causal DAGs appear to be quite dependent on the absolute zeros that correspond to missing arrows. It is important to then examine the sensitivity to these findings to even modest departures from these pure independence assumption. And indeed, the above results suggest that, in many cases, adjusting for all the pretreatment covariates is in fact a reasonable choice. Many on both sides of this debate argue that scientific knowledge is key for a proper design, either in the generation of the DAG or in the justification of a strong ignorability assumption. We agree, but in either case this scientific knowledge should also shed light on whether a pure M-Structure is a real possibility or not.

Acknowledgment

The authors thank all the participants in the "Causal Graphs in Low and High Dimensions" seminar at Harvard Statistics department in Fall, 2012, and thank Professor Peter Spirtes for sending us his slides presented at WNAR/IMS meeting in Los Angeles CA in 2002.

Appendix: Lemmas and Proofs

The proofs in this paper are based on some simple results from regression used on the structural equations defined by the DAGs. We heavily use the following fact:

Lemma 1 *In the linear regression model* $Y = \beta_0 + \beta_T T + \beta_M M + \varepsilon$ *with* $\varepsilon \perp\!\!\!\perp (T, M)$ *and* $E(\varepsilon) = 0$*, we have*

$$\beta_T = \frac{Cov(Y,T)Var(M) - Cov(Y,M)Cov(M,T)}{Var(T)Var(M) - Cov^2(M,T)}.$$

Proof. Solve for (β_T, β_M) using the following moment conditions

$$\begin{aligned} \mathrm{Cov}(Y,T) &= \beta_T \mathrm{Var}(T) + \beta_M \mathrm{Cov}(M,T), \\ \mathrm{Cov}(Y,M) &= \beta_T \mathrm{Cov}(M,T) + \beta_M \mathrm{Var}(M). \end{aligned}$$

Lemma 2 *Under the model generated by the DAG in Figure 2, the regression coefficient of* T *by regressing* Y *onto* (T,M) *is*

$$\beta_T = \frac{ad\rho(1 - b^2 - c^2 - bc\rho) - abcd}{1 - (ab + ac\rho)^2}.$$

Proof. We simply expand and rearrange the terms in Lemma 1. In particular, all variance terms such as $\mathrm{Var}(M)$ are 1. The covariance terms are easily calculated as well. For example, we have

$$\begin{aligned} \mathrm{Cov}(Y,T) &= \mathrm{Cov}(dW + \sqrt{1-d^2}\varepsilon_Y, aU + \sqrt{1-a^2}\varepsilon_T) \\ &= \mathrm{Cov}(dW, aU) = ad\rho \end{aligned}$$

and

$$\begin{aligned} \mathrm{Cov}(Y,M) &= \mathrm{Cov}(dW + \sqrt{1-d^2}\varepsilon_Y, bU + cW + \sqrt{1-b^2-c^2}\varepsilon_M) \\ &= \mathrm{Cov}(dW, bU) + \mathrm{Cov}(dW, cW) = bd\rho + cd. \end{aligned}$$

Lemma 3 *Using the model generate by the DAG in Figure 7, the regression coefficient of* T *from regressing* Y *onto* (T,M) *is*

$$\beta_T = -\frac{abcd}{1 - (ab + e)^2}.$$

Proof. Follow Lemma 2. Expand Lemma 1 to obtain

$$\beta_T = \frac{(abf+cde+ef)-(cd+f)(ab+e)}{1-(ab+e)^2}.$$

Now rearrange.

Lemma 4 *Assume that* (X_1, X_2) *follows a bivariate Normal distribution:*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \boldsymbol{N}_2 \left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix} \right\}.$$

Then $\mathbb{E}(X_1 \mid X_2 \geq z) - \mathbb{E}(X_1 \mid X_2 < z) = r\eta(z)$, *where* $\eta(z) = \phi(z)/\{\Phi(z)\Phi(-z)\}$.

Proof. Since $X_1 = rX_2 + \sqrt{1-r^2}Z$ with $Z \sim N(0,1)$ and $Z \perp\!\!\!\perp X_2$, we have

$$\mathbb{E}(X_1 \mid X_2 \geq z) = r\mathbb{E}(X_2 \mid X_2 \geq z) = \frac{r}{\Phi(-z)} \int_z^\infty x\phi(x)dx = -\frac{r}{\Phi(-z)} \int_z^\infty d\phi(x) = r\frac{\phi(z)}{\Phi(-z)}.$$

Similarly, we have $\mathbb{E}(X_1 \mid X_2 < z) = \mathbb{E}(X_1 \mid -X_2 > -z) = -r\phi(-z)/\Phi(z) = -r\phi(z)/\Phi(z)$. Therefore, we have

$$\mathbb{E}(X_1 \mid X_2 \geq z) - \mathbb{E}(X_1 \mid X_2 < z) = r\phi(z)\left\{\frac{1}{\Phi(-z)} + \frac{1}{\Phi(z)}\right\} = \frac{r\phi(z)}{\Phi(z)\Phi(-z)}.$$

Lemma 5 *The covariance between* X *and* $B \sim$ *Bernoulli*(p) *is*

$$Cov(X,B) = p(1-p)\{\mathbb{E}(X \mid B=1) - \mathbb{E}(X \mid B=0)\}.$$

Proof. We have $\mathrm{Cov}(X,B) = \mathbb{E}(XB) - \mathbb{E}(X)\mathbb{E}(B) = p\mathbb{E}(X \mid B=1) - p\{p\mathbb{E}(X \mid B=1) + (1-p)\mathbb{E}(X \mid B=0)\} = p(1-p)\{\mathbb{E}(X \mid B=1) - \mathbb{E}(X \mid B=0)\}$.

Lemma 6 *Under the model generated by the DAG in Figure 8, the regression coefficient of* T *from regressing* Y *onto* (T,M) *is*

$$\beta_T = \frac{ad\eta(\alpha)\{\rho(1-b^2-c^2-bc\rho)-bc\}}{\rho\{1-(ab+ac\rho)^2\phi(\alpha)\eta(\alpha)\}}.$$

Proof. We have the following joint Normality of (Y, M, T^*):

$$\begin{pmatrix} Y \\ M \\ T^* \end{pmatrix} = \begin{pmatrix} dW + \sqrt{1-d^2}\varepsilon_Y \\ bU + cW + \sqrt{1-b^2-c^2}\varepsilon_M \\ \alpha + aU + \sqrt{1-a^2}\varepsilon_T \end{pmatrix} \sim \boldsymbol{N}_3 \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & bd\rho + cd & ad\rho \\ bd\rho + cd & 1 & ab + ac\rho \\ ad\rho & ab + ac\rho & 1 \end{pmatrix} \right\}.$$

Derive by computing all covariance terms between Y and M, etc., and then plugging them into the above matrix. From Lemma 4, we have

$$\begin{aligned} \mathbb{E}(M \mid T=1) - \mathbb{E}(M \mid T=0) &= \mathbb{E}(M \mid T^* \geq \alpha) - \mathbb{E}(M \mid T^* < \alpha) = (ab + ac\rho)\eta(\alpha), \\ \mathbb{E}(Y \mid T=1) - \mathbb{E}(Y \mid T=0) &= \mathbb{E}(Y \mid T^* \geq \alpha) - \mathbb{E}(Y \mid T^* < \alpha) = ad\rho\eta(\alpha). \end{aligned}$$

Therefore, from Lemma 5, the covariances are

$$\begin{aligned} \mathrm{Cov}(M, T) &= \Phi(\alpha)\Phi(-\alpha)(ab + ac\rho)\eta(\alpha), \\ \mathrm{Cov}(Y, T) &= \Phi(\alpha)\Phi(-\alpha)ad\rho\eta(\alpha). \end{aligned}$$

According to Lemma 1, the regression coefficient β_T is

$$\begin{aligned} \beta_T &= \frac{\Phi(\alpha)\Phi(-\alpha)ad\rho\eta(\alpha) - (bd\rho + cd)\Phi(\alpha)\Phi(-\alpha)(ab + ac\rho)\eta(\alpha)}{\Phi(\alpha)\Phi(-\alpha) - \Phi^2(\alpha)\Phi^2(-\alpha)(ab + ac\rho)^2\eta^2(\alpha)} \\ &= \frac{ad\rho\eta(\alpha) - (bd\rho + cd)(ab + ac\rho)\eta(\alpha)}{1 - (ab + ac\rho)^2\phi(\alpha)\eta(\alpha)} \\ &= \frac{ad\eta(\alpha)\{\rho(1 - b^2 - c^2 - bc\rho) - bc\}}{\rho\{1 - (ab + ac\rho)^2\phi(\alpha)\eta(\alpha)\}}. \end{aligned}$$

Lemma 7 *Under the model generated by the DAG in Figure 10, the regression coefficient of T from regressing Y onto (T, M) is*

$$\beta_T = -\frac{abcd\eta(\alpha)}{1 - (ab + e)\phi(\alpha)}.$$

Proof. We have the following joint Normality of (Y, M, T^*):

$$\begin{pmatrix} Y \\ M \\ T^* \end{pmatrix} = \begin{pmatrix} dV + fM + \sqrt{1-d^2}\varepsilon_Y \\ bU + cV + \sqrt{1-b^2-c^2}\varepsilon_M \\ \alpha + aU + eM + \sqrt{1-a^2-e^2}\varepsilon_T \end{pmatrix} \sim \boldsymbol{N}_3 \left\{ \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & cd + f & cde + abf + ef \\ cd + f & 1 & ab + e \\ cde + abf + ef & ab + e & 1 \end{pmatrix} \right\}.$$

From Lemma 4, we have

$$\begin{aligned}\mathbb{E}(M \mid T=1)-\mathbb{E}(M \mid T=0) &= \mathbb{E}(M \mid T^* \geq \alpha)-\mathbb{E}(M \mid T^* < \alpha) = (ab+e)\eta(\alpha),\\ \mathbb{E}(Y \mid T=1)-\mathbb{E}(Y \mid T=0) &= \mathbb{E}(Y \mid T^* \geq \alpha)-\mathbb{E}(Y \mid T^* < \alpha) = (cde+abf+ef)\eta(\alpha).\end{aligned}$$

From Lemma 5, we obtain their covariances

$$\begin{aligned}\mathrm{Cov}(M,T) &= \Phi(\alpha)\Phi(-\alpha)(ab+e)\eta(\alpha),\\ \mathrm{Cov}(Y,T) &= \Phi(\alpha)\Phi(-\alpha)(cde+abf+ef)\eta(\alpha).\end{aligned}$$

According to Lemma 1, the regression coefficient β_T is

$$\begin{aligned}\beta_T &= \frac{\Phi(\alpha)\Phi(-\alpha)(cde+abf+ef)\eta(\alpha)-(cd+f)\Phi(\alpha)\Phi(-\alpha)(ab+e)\eta(\alpha)}{\Phi(\alpha)\Phi(-\alpha)-\Phi^2(\alpha)\Phi^2(-\alpha)(ab+e)^2\eta^2(\alpha)}\\ &= \frac{(cde+abf+ef)\eta(\alpha)-(cd+f)(ab+e)\eta(\alpha)}{1-(ab+e)^2\phi(\alpha)\eta(\alpha)}\\ &= -\frac{abcd\eta(\alpha)}{1-(ab+e)^2\phi(\alpha)\eta(\alpha)}.\end{aligned}$$

Lemma 8 *Under the model generated by the DAG in Figure 13(b), the unadjusted estimator has bias $ad\rho+dg$, and the adjusted estimator has bias*

$$\frac{dg-(cd)(ab+cg)}{1-(ab+cg)^2}.$$

Proof of Lemma 8. The unadjusted estimator is the covariance $\mathrm{Cov}(T,Y)=ad\rho+dg$. Expanding Lemma 1 gives the above as the regression coefficient of T from regressing Y onto (T,M).

References

Berkson, J. (1946): "Limitations of the application of fourfold table analysis to hospital data," *Biometrics Bulletin*, 2, 47–53.

Gelman, A. (2011): "Causality and statistical learning," *American Journal of Sociology*, 117, 955–966.

Greenland, S. (2002): "Quantifying biases in causal models: classical confounding vs collider-stratification bias," *Epidemiology*, 14, 300–306.

Liu, W., M. A. Brookhart, S. Schneeweiss, X. Mi, and S. Setoguchi (2012): "Implications of M bias in epidemiologic studies: a simulation study," *American Journal of Epidemiology*, 176, 938–948.

Neyman, J. (1923/1990): "On the application of probability theory to agricultural experiments. essay on principles. section 9," *Statistical Science*, 5, 465–472.

Pearl, J. (1995): "Causal diagrams for empirical research," *Biometrika*, 82, 669–688.

Pearl, J. (2000): *Causality: Models, Reasoning and Inference*, Cambridge University Press.

Pearl, J. (2009a): "Letter to the editor: Remarks on the method of propensity score," *Statistics in Medicine*, 28, 1415–1416.

Pearl, J. (2009b): "Myth, confusion, and science in causal analysis," *Technical Report*.

Pearl, J. (2013a): "Discussion on 'surrogate measures and consistent surrogates'," *Biometrics*, 69, 573–577.

Pearl, J. (2013b): "Linear models: A useful microscope for causal analysis," *Journal of Causal Inference*, 1, 155–170.

Rosenbaum, P. R. (2002): *Observational Studies*, Springer.

Rosenbaum, P. R. and D. B. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rubin, D. B. (1974): "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology*, 66, 688.

Rubin, D. B. (2007): "The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials," *Statistics in Medicine*, 26, 20–36.

Rubin, D. B. (2008): "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 808–840.

Rubin, D. B. (2009): "Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups?" *Statistics in Medicine*, 28, 1420–1423.

Shrier, I. (2008): "Letter to the editior," *Statistics in Medicine*, 27, 2740–2741.

Shrier, I. (2009): "Propensity scores," *Statistics in Medicine*, 28, 1315–1318.

Sjölander, A. (2009): "Propensity scores and M-structures," *Statistics in Medicine*, 28, 1416–1420.

Sprites, P. (2002): "Presented at: WNAR/IMS Meeting. Los Angeles, CA, June 2002," .

VanderWeele, T. J. and I. Shpitser (2011): "A new criterion for confounder selection," *Biometrics*, 67, 1406–1413.

Wright, S. (1921): "Correlation and causation," *Journal of Agricultural Research*, 20, 557–585.

Wright, S. (1934): "The method of path coefficients," *The Annals of Mathematical Statistics*, 5, 161–215.