Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

July 28, 2024

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

8 Valid Adjustment Sets from DAGs

"if 'good' is taken to mean 'best' fit, it is tempting to include anything in x that helps predict [treatment]"

– Jeffrey Wooldridge [1].

DAGs give us an intuitive approach to take domain knowledge and turn it into an identification strategy. In this section, we focus on identification by conditioning and discuss graphical criteria that lead to the construction of valid adjustment sets for the identification of average causal effects via regression adjustment. We also discuss how graphical criteria can help us differentiate between "good" and "bad" controls.

8.1 Valid Adjustment Sets

Consider any variable D of an ASEM as a treatment of interest and any of its descendants Y as an outcome of interest. An adjustment set S is said to be valid for identification of the causal effect of D on Y if the conditional exogeneity/ignorability condition holds

$$Y(d) \perp\!\!\!\perp D \mid S.$$

In what follows, we present an exhaustive (complete) approach for finding valid adjustment sets by using SWIGs.

We write down the counterfactual SWIG induced by the

$$\text{fix}(D = d)$$

intervention, which operates on all structural equations defining the descendants of D by setting $D = d$ in these equations.

Then, if we have that the potential outcome $Y(d)$ is d-separated from the (policy) variable D by a set of variables S, conditional exogeneity/ignorability holds:

$$Y(d) \perp\!\!\!\perp D \mid S.$$

Given that conditional exogeneity/ignorability holds, we can identify counterfactual expectations,

$$\mathrm{E}[Y|S = s : \mathrm{do}(d)] := \mathrm{E}[Y(d)|S = s],$$

from expectations of observed variables,

$$\mathrm{E}[Y|S = s, D = d],$$

provided that the positivity condition $p(s, d) > 0$ holds. The agreement between counterfactual and conditional expectations follows because

$$\mathrm{E}[Y(d)|S = s] = \mathrm{E}[Y(d)|D = d, S = s]$$

by exogeneity and

$$\mathrm{E}[Y(d)|D = d, S = s] = \mathrm{E}[Y|D = d, S = s]$$

by consistency.

We can recover unconditional counterfactual means by integration:

$$\mathrm{E}[Y : \mathrm{do}(d)] := \mathrm{E}[Y(d)] = \mathrm{E}[\mathrm{E}[Y|S, D = d]],$$

provided that the positivity condition $p(s, d) > 0$ for each s in the support of $S \mid D = d$ holds.

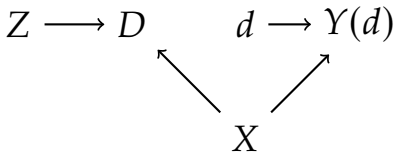


Figure 8.1: CF LS-DAG induced by $\mathrm{fix}(D = d)$ intervention.

Example 8.1.1 (Identification in LS-DAG.) In the SWIG graph in Figure 8.1 corresponding to the LS-DAG model from Example 7.3.1, we see that either $S = X$ or $S = (X, Z)$ d-separates $Y(d)$ from D. Therefore, either choice of S provides a valid adjustment set for identifying counterfactual predictions. Here conditioning on Z is not necessary, though we maintain robustness with respect to the presence of a directed edge from Z to Y by including Z in the conditioning set.

We can identify the entire conditional distribution

$$\mathrm{P}(Y(d) \leq t \mid S = s)$$

from the conditional distribution

$$\mathrm{P}(Y \leq t \mid D = d, S = s).$$

We achieve identification of the distribution by replacing Y with $1(Y < t)$ in all previous statements and applying the same arguments for each $t \in \mathbb{R}$. The unconditional distribution of potential outcomes is retrieved by integrating out S:

$$\mathrm{P}(Y(d) \leq t) := \mathrm{E}[\mathrm{P}(Y(d) \leq t \mid S)].$$

The following theorem, essentially due to [2], records the discussion formally.

Theorem 8.1.1 (A SWIG Criterion for Identification by Conditioning) *Consider any ASEM with DAG* G*. Let us re-label a policy node* X_j *as* D*, and let* Y*, an outcome of interest, be any other descendant of* D*.*

Consider a SWIG DAG $\tilde{\mathsf{G}}(d)$ *which is induced by the* $\mathrm{fix}(D = d)$ *intervention. Consider any other subset of nodes* S *that appears in both* G *and* $\tilde{\mathsf{G}}(d)$*, such that*

$$Y(d) \text{ is d-separated from } D \text{ by } S \text{ in } \tilde{\mathsf{G}}(d).$$

- *Then the following conditional exogeneity/ignorability holds:*

$$Y(d) \perp\!\!\!\perp D \mid S.$$

- *Then*

$$\mathrm{E}[Y(d)|S = s] = \mathrm{E}[Y \mid D = d, S = s]$$

holds for all s such that $p(d, s) > 0$.

The criterion is "complete" in the sense that it can be used to find all valid adjustment sets that don't include descendants of D. It is also "complete" in another sense discussed in [2], in the sense that this approach finds all valid adjustment sets with validity that is verifiable via an implementable intervention; see [2] for further discussion of this fine point.*

Example 8.1.2 (Pearl's Example) Consider the DAG in Figure 8.2, which we introduced as Pearl's Example in Figure 7.14, and the corresponding ASEM, which we don't write out. Here, we are interested in the causal effect $D \to Y$, that is, the effect $d \mapsto Y(d)$. The corresponding SWIG-intervention DAG is shown in Figure 8.3. In this DAG, valid adjustment sets S include

$$\{X_1, X_2\}, \{X_2, X_3\}, \{X_2, Z_2\}, \{X_2, Z_1\},$$

because each d-separates $Y(d)$ and D by blocking all open paths. Conditioning on just X_2 won't work, because it blocks the inner backdoor paths from $Y(d)$ to D, but opens the outer path on which X_2 is a collider. To close this opened path, it suffices to also condition on one of X_1, X_3, Z_1 or Z_2.

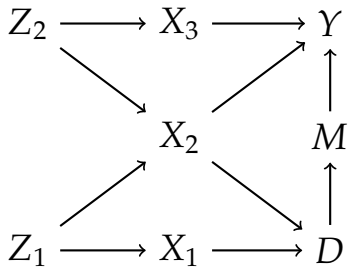


Figure 8.2: A DAG in Pearl's Example

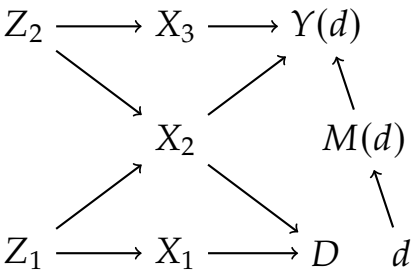


Figure 8.3: The DAG induced by the Fix/SWIG intervention fix($D = d$) in Pearl's Example.

8.2 Useful Adjustment Strategies

Theorem 8.1.1 provides an exhaustive criterion for finding valid adjustment sets. We now discuss other frequently used strategies for obtaining valid adjustment sets which are strictly less general. Some of these strategies are quite helpful because

* The SWIG approach does not identify Z as a valid adjustment set in the DAG $Z \leftarrow D \to Y$ because, under the fixed intervention, Z is replaced by $Z(d)$. In this simple example, we can construct alternative counterfactual systems that keep equations for both Z and $Z(d)$, and these become connected by an undirected edge, and use d-separation to confirm that Z is indeed a valid adjustment set.

they are either very simple to apply or can also be used under partial knowledge of the DAG.[1]

1: See [3] for a more detailed discussion of identification by conditioning under limited knowledge of DAGs.

We consider three approaches that allow us to identify the causal effect of D on Y:

- Conditioning on one of **all parents** of Y (that are not descendants of D), **all parents** of D, or **all parents** of both D and Y is sufficient. This approach provides a valid adjustment set irrespective of the remaining structure of the problem.
- Conditioning using the **backdoor criterion** enables us to find all minimal adjustment sets.
- Conditioning on **all common causes** of D and Y is also sufficient.

Conditioning on Parents

A very simple strategy is conditioning on one of the parents of D, the parents of Y, or the parents of both D and Y.

Example 8.2.1 (Pearl's Example Continued) One simple principle is that conditioning on parents of D, namely X_1 and X_2, is sufficient. Alternatively, conditioning on all parents of Y that are non-descendants of D, namely X_2 and X_3, is also sufficient. We should not condition on M, because it is a descendant of D.

Corollary 8.2.1 (Adjustment for Parents) *Consider any ASEM. Re-label a policy node X_j as D, and let Y, an outcome of interest, be any other descendant of D.*

- *Let Z be all parents of D, and let A be any other set of nodes that are not descendants of D. Then $S = (A, Z)$ is a valid adjustment set.*
- *Let Z be the set of all parents of Y that are non-descendants of D and let A be any other set that are not descendants of D. Then $S = (A, Z)$ is a valid adjustment set.*

Note that A is allowed to be an empty set. Also note that, in the second case, the additional adjustment set A is redundant, since $\mathsf{p}(y \mid a, z, d) = \mathsf{p}(y \mid z, d)$ in this case.

Adjusting for parents is a very useful strategy, because it only requires knowledge of parents in a DAG without precise knowledge of the remaining graph structure. Conditioning on parents

is also behind the propensity score strategies used in many experimental or quasi-experimental empirical analyses. If the propensity score is known, it can be used as a parent of D itself. Finally, conditioning on parents of Y is most useful for attaining maximal statistical efficiency, but may be less robust than conditioning on *both* sets of parents under unforeseen deviations from the given graph structure. See [3] for further detailed discussion of robustness of adjusting for both sets of parents.

Conditioning by Backdoor Blocking

Pearl [4] developed the following powerful criterion.

Corollary 8.2.2 (Backdoor Criterion) *Consider any ASEM. Relabel a policy node X_j as D, and let Y, an outcome of interest, be any other descendant of D. The adjustment set S is valid if the backdoor criterion is satisfied: No element of S is a descendant of D, and all backdoor paths from Y to D are blocked by S.*

In other words, if a collection of random variables S satisfies the backdoor criterion with respect to (D, Y), then conditioning on S identifies the causal effect of D on Y. The basic idea is that if we block the backdoor path, we remove all channels of non-causal association between D and Y.

Example 8.2.2 (Pearl's Example Again, using the Backdoor Criterion) The graph in Figure 8.2 has two backdoor paths from D to Y: the inner path $D \leftarrow X_2 \rightarrow Y$ and the outer path $D \leftarrow X_1 \leftarrow Z_1 \rightarrow X_2 \leftarrow Z_2 \rightarrow X_3 \rightarrow Y$. Conditioning on just X_2 does not allow us to identify the causal effect of D on Y because X_2 blocks the inner backdoor path from Y to D but opens the outer path on which X_2 is a collider. To close this opened path, it suffices to condition on X_1, X_3, Z_1, or Z_2. For example, conditioning sets $S_1 = \{X_1, X_2\}$ or $S_2 = \{X_2, X_3\}$ are valid. Figuring out other valid conditioning sets is left as an exercise. (You can find the answers using the notebook R: Dagitty Notebook or Python: Pgmpy Notebook.) Conditioning on M is obviously not valid – it is a descendant of D, an intermediate outcome.

Application of the backdoor criterion can produce all minimal adjustment sets. Relative to the complete strategy formalized in Theorem 8.1.1, we exclude the descendants of D from valid adjustment sets when we focus on backdoor paths. A simple

example of a graph where the backdoor criterion does not find all valid adjustment sets is

$$Z \leftarrow D \rightarrow Y.$$

Here conditioning on Z is valid but unnecessary. Conditioning on Z may thus decrease statistical efficiency.[2]

2: We may think that conditioning on Z here could be useful to uncover heterogeneity. However, $Y(d)$ does not depend on Z, so conditioning on Z is not useful for describing heterogeneity and can decrease the efficiency of the estimator.

Conditioning on All Common Causes of D and Y

Another simple and widely used adjustment strategy is conditioning on all common causes of the outcome variable of interest and the treatment variable.

Example 8.2.3 (Pearl's Example Again, using the All Common Causes Criterion) The set of common causes of D and Y is $\{Z_1, Z_2, X_2\}$. This set is a valid adjustment set that differs from the sets found using the parental strategy. We can push the All Common Causes criterion further. For example, we can omit Z_1 and Z_2 from the DAG, and we can create a new node $X = (X_1, X_2, X_3)$ producing the DAG shown in Figure 8.4. This DAG corresponds to a valid ASEM model where X now represents all common causes of D and Y, making it a sufficient adjustment set. This set is bigger than some of the sets found by the previous criteria. It is also tempting to see if the "root common" causes Z_1 and Z_2 in the original DAG, Figure 8.2, form a valid adjustment set – and they actually do not (why?).

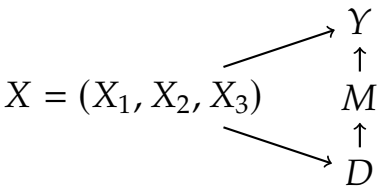


Figure 8.4: Reduced DAG for Pearl's Example

Let $\underline{An}_X$ denote the set of strict ancestors of node X, where strict means that X is excluded. That is,

$$\underline{An}_X = An_X \setminus X.$$

Corollary 8.2.3 (Adjustment for All Common Causes) *Consider any ASEM. Re-label a policy node X_j as D, and let Y, an outcome of interest, be any other descendant of D. Let S be the intersection of the strict ancestors of D and Y, called the common causes:*

$$S = (\underline{An}_D \cap \underline{An}_Y).$$

Then S is a valid adjustment set. Furthermore, the set of variables S' that completely mediates the effects of S on Y and D also constitutes a valid adjustment set.

The strategy above is commonly used in empirical work. However, [3] recommend adjusting for the union S of causes of Y or D (excluding descendants of D) in practice as they formally quantify this strategy as the maximally robust strategy under perturbations of a specified DAG structure that preserves S. This strategy is useful when we don't know the parents of Y or D, but only know that S are their ancestors.

Corollary 8.2.4 (Adjustment for the Union of Causes) *Consider any ASEM. Re-label a policy node X_j as D, and let Y, an outcome of interest, be any other descendant of D. Let S be the union of the ancestors of D and Y that excludes descendants of D other than Y:*

$$S = \underline{An}_D \cup \underline{An}_Y \setminus Ds_D.$$

Then S is a valid adjustment set.

Example 8.2.4 (Pearl's Example Continued) Application of the Union of Causes criterion gives $\{Z_1, Z_2, X_1, X_2, X_3\}$ as a valid adjustment set.

8.3 Examples of Good and Bad Controls

We now present a series of simple example DAGs that might arise in empirical research. Within these examples, we discuss what would be good and bad variables to adjust for in each case (aka good and bad controls), when one is interested in estimating the average treatment effect of a treatment D on an outcome Y.[3] Similar to the collider bias examples we presented in Section 6.3, we will see how adjusting for some of the observed variables can introduce bias and lead to estimating a parameter that is far from the causal effect of interest. In each case, we will denote the candidate control of interest with Z and will denote unobserved variables with U. We depict unobserved variables with a dashed circle in the figures.

3: The content in this section draws heavily from the excellent research paper of Cinelli, Forney and Pearl [5].

We start by analyzing a group of potential control variables that in most empirical applications would correspond to *pre-treatment variables*, i.e. variables whose value was determined prior to the treatment assignment. It is common empirical practice to adjust for as many pre-treatment variables as available in an attempt to ensure that conditional ignorability holds. However, we will see that bias can be introduced by controlling even for pre-treatment variables if one is not careful. Rather than always control for all pre-treatment variables, a better approach

is to adjust only for pre-treatment variables that are ancestors of either the treatment, the outcome, or both. If one is willing to believe that identification by conditioning is feasible, then following this approach is a safe strategy.

We then consider the use of *post-treatment variables*, i.e. variables that correspond to quantities whose value is determined after the treatment assignment. We will see that in this case there are relatively few good control cases. In some cases, controlling for post-treatment variables might not hurt and may even improve precision (reduce variance). However, such settings seem unlikely to be common in empirical practice. Hence, as a high-level rule, controlling for post-treatment variables should be avoided when one is interested in estimating causal effects.

Finally, we provide a separate discussion of post-treatment but *pre-outcome variables*, i.e. variables whose value is determined prior to the determination of the value of the outcome of interest. Pre-outcome variables should be included if one is interested in estimating direct effects of the treatment on the outcome while excluding indirect effects. This type of direct effect is referred to as a *controlled direct effect* to distinguish it from other forms of direct effects appearing in mediation analysis. We will see again that one should be careful that the mediation variables that one conditions on are not themselves confounded through unobserved factors even in this case.

Pre-Treatment Variables or Proxies of Pre-Treatment Variables

Observed common causes or proxies of common causes. A common example of a good control that we have discussed so far is an observed common cause, Z, of D and Y (Figure 8.5a). Even if the common cause is unobserved, it suffices that we have a proxy control variable that controls all the information flow to either the treatment (complete treatment proxy; Figure 8.5b) or to the outcome (complete outcome proxy; Figure 8.5c). Controlling for such a proxy also blocks the backdoor path $D \leftarrow U \rightarrow Y$. Of course, the proxy blocking the backdoor path only holds if the proxy variable captures *all* the information flow from the unobserved confounder. If, for instance, there are also direct paths from the unobserved variable to the treatment (in the case of a treatment proxy), then controlling for a proxy does not remove confounding bias. In this case, we will see that one can follow more advanced approaches related to proxy controls under additional structure in Chapter 12.

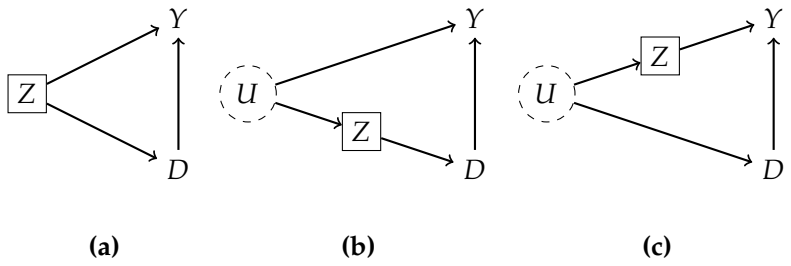


Figure 8.5: Good controls: **(a)** observed common cause, **(b)** complete treatment proxy control of unobserved common cause, **(c)** complete outcome proxy control of unobserved common cause.

Example 8.3.1 (Effect of Multivitamin Consumption on Birth Defects [6]) Suppose we want to estimate the effect of prenatal multivitamin consumption D on birth defects Y. One factor that can potentially influence a mother's decision on multivitamin consumption is prior history of birth defects in the family (Z); see e.g. [7]. Such prior history is possibly due to unobserved genetic factors U that also have a direct effect on the risk of malformation Y; see e.g. [8]. In this case, family medical history Z provides a complete treatment proxy of the unobserved confounder (as in Figure 8.5b) as long as the behavior of a mother is solely driven by the family medical history. Controlling for medical history would thus remove the confounding bias in this scenario.

Confounded mediators with observed common cause or proxies of unobserved common cause. It is important to note that confounding occurs even when there exists a common cause Z of the treatment D and some mediator M in a path from D to Y (Figure 8.6a). In such cases, if we don't condition on the common cause of D and M, there is an open backdoor path $D \leftarrow Z \rightarrow M \rightarrow Y$. In such cases, Z is a good control as it blocks this backdoor path. Similarly, if a common cause U of D and M is unobserved, but some complete treatment proxy control Z (Figure 8.6b) or some complete outcome proxy control Z (Figure 8.6c) is observed, then it suffices to adjust for this proxy Z.

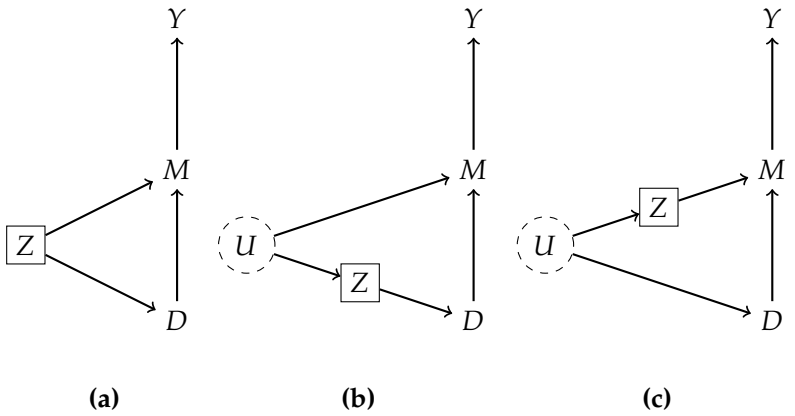


Figure 8.6: Good controls: **(a)** confounded mediator with observed common cause, **(b)** confounded mediator, with observed complete treatment proxy control of unobserved common cause, **(c)** confounded mediator with observed complete outcome proxy control of unobserved common cause.

Causes of only treatment or only outcome. As stated in Corollary 8.2.4, a conservative empirical practice is to include the union of parents of D and Y in the adjustment set. Including variables that are parents of the outcome (Figure 8.7a) can lead to reduced variance during estimation as explained in Chapter 2 where we discuss including pre-treatment covariates in RCTs. Including variables Z that affect the treatment D but have no causal path to the outcome (Figure 8.7b) is potentially more controversial. Including these variables does not introduce bias. However, their inclusion can be detrimental for precision, as such variables can potentially explain away all of the useful variation in the treatment, leaving little variation for the identification of causal effects.

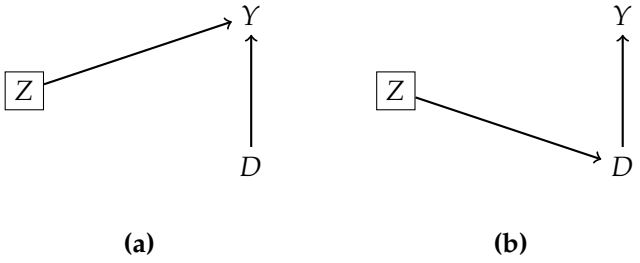


Figure 8.7: Neutral controls: **(a)** Outcome-only cause. Can improve precision; decrease variance. **(b)** Treatment-only cause. Can decrease precision; introduce variance.

Even more importantly, when there are unobserved common causes of D and Y as illustrated in Figure 8.8, adjusting for a treatment-only cause, Z, can exacerbate the bias stemming from unobserved confounding. Essentially, controlling for Z removes exogenous variation in the treatment D that is useful for identifying the causal effect but leaves the confounded variation - as Z is not related directly to the unobserved confounder U. As such, the resulting estimated effect may be essentially driven by the unobserved confounder and thus be heavily biased. For this reason, one should avoid controlling for variables that are

known to have no causal path to the outcome that does not pass through the treatment. As we will see in Chapter 12, such variables are actually what are referred to as *instruments*. These variables can be thought as useful natural experiments that can be leveraged for causal identification even in the presence of unobserved confounding. However, we will need to use alternative identification arguments and estimation strategies to make use of instruments. We introduced these *instrumental variable* approaches in Chapter 12 and Chapter 13. Importantly, instruments *should not* be used in an identification by adjustment strategy.

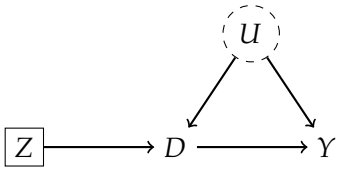


Figure 8.8: Bad control. Bias amplification by adjusting for an *instrument*. Treatment-only cause (*instrument*) that can amplify unobserved confounding bias.

M-bias The DAG in Figure 8.9, typically referred to in the literature as the M structure, is the source of much debate; see e.g. [9, 10]. If such cases were impossible, the high-level strategy of controlling for all pre-treatment variables when attempting to identify causal effects by conditioning would be an unambiguously safe empirical route resulting in no harm other than potentially increasing variance by including an *instrument*. However, this structure shows that there exist settings where adjusting for a pre-treatment covariate Z can lead to a wrong causal effect, while not adjusting for Z would have yielded the correct causal effect. A better high-level strategy is the one highlighted in the prior sections: If we are willing to assume that identification by conditioning is possible, then we should adjust only for pre-treatment variables that are either an ancestor of the treatment, of the outcome, or of both treatment and outcome.

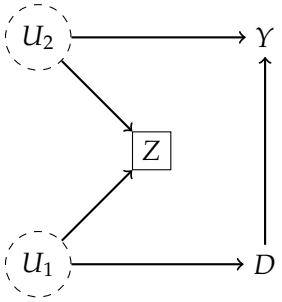


Figure 8.9: Bad control. M-Bias. Pre-treatment variable that introduces Heckman selection bias between two uncorrelated unobserved causes.

More concretely, in the M structure graph (Figure 8.9), D and Y are driven by two independent unobserved causal factors U_1, U_2. The variable Z is a common outcome of these two unobserved causal factors. When conditioning on Z, we introduce collider bias between U_1, U_2, making them correlated factors. Conditioning on Z can thus lead to a causal effect estimate that is solely driven by this spurious correlation between U_1 and U_2, introduced by the collider bias. In graphical terms, adjusting for Z closes the path $D \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y(d)$ in the SWIG DAG $\tilde{\mathsf{G}}(d)$ produced by the fix$(D = d)$ operation. However,

there is no open path connecting D to $Y(d)$ when we do not condition on Z. Hence, the effect identified by not adjusting for any variable is the correct causal effect within this example structure.

Example 8.3.2 (Homophily bias in estimating peer effects) A classical example where M-bias arises in empirical work in social sciences is in the estimation of peer effects on social networks [11, 12]. As a concrete example, suppose that we want to understand the spread of civic engagement among friends. Suppose that we look at data that consist of friendship pairs and let D be the level of civic engagement level of one friend at time t and Y the level of civic engagement of the other friend at time $t + 1$. Note that when we are estimating the correlation of these two variables, we are implicitly conditioning on the friendship variable Z, since we only have data from friendship pairs. Due to homophily, friendship could be driven by the unobserved intrinsic characteristics of each of the two individuals (U_1 and U_2 in Figure 8.9). It is reasonable to assume that these characteristics are independent as they are determined well before any friendship is formed. Moreover, these qualitative characteristics (e.g. levels of altruism) could very well have a direct effect on each individual's civic engagement. Thus, the estimation of peer effects can be heavily biased due to exactly M-bias.

Homophily refers to the tendency of associate with similar individuals - i.e. similar people tend to become friends.

Finally, note that the M-bias argument is very sensitive to the exact independence of the unobserved factors U_1, U_2. In most empirical applications, we expect these unobserved factors that drive the treatment and outcome of interest to be correlated with each other as in Figure 8.10a. In this case, note that even if we don't adjust for Z, the calculated effect is biased due to the backdoor path $D \leftarrow U_1 \rightarrow U_2 \rightarrow Y$. Thus, neither adjusting nor not adjusting for Z gives the correct answer.

Moreover, it is not clear whether adjusting for Z increases or decreases the correlation between U_1 and U_2 and hence exacerbates or ameliorates the confounding bias. Similarly, if Z itself has a direct effect on the outcome (as in Figure 8.10b), on the treatment, or on both (as in Figure 8.10c), then not adjusting for Z opens the backdoor paths $D \leftarrow U_1 \rightarrow Z \rightarrow Y$ and $D \leftarrow Z \rightarrow Y$, correspondingly. Hence, it is not clear that removing the bias induced by these open backdoor paths, by adjusting for Z, is more beneficial than the extra M-bias incurred by closing the path $D \leftarrow U_1 \rightarrow Z \leftarrow U_2 \rightarrow Y$. Work of [9, 13] argues that M-bias in many realistic data generating processes is of lower order than confounding bias and therefore argues

that one should err on the side of adjusting for pre-treatment covariates even in the potential presence of M-bias. [10] provides a counterpoint, arguing that the strength of the different biases will differ in general and thus careful consideration of the strength of each of the causal paths at play should be done on a case-by-case basis.

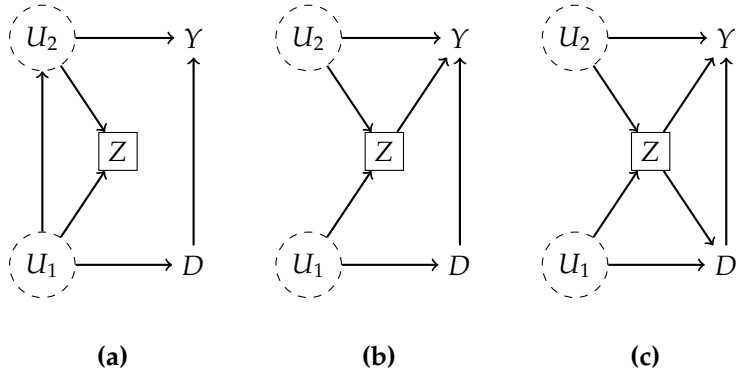


Figure 8.10: No perfect control solutions: **(a)** M-bias with correlated unobserved factors. **(b)** M-Bias with confounding. Pre-treatment variable that introduces Heckman selection between two uncorrelated unobserved causes and is a confounder itself. **(c)** Butterfly Bias. M-bias with direct confounding.

Post-Treatment Variables

Now we turn to adjustment for post-treatment variables. The general message of this section is that explicitly adjusting for post-treatment variables is almost always a bad idea. Importantly, the general message implies that researchers should be careful to avoid implicitly adjusting for post-treatment variables through the way they have structured their observational analysis, data collection, and variable definitions – see e.g. [6] for examples from epidemiology. For instance, when estimating the effect of education on wages using data on *employed* individuals, we are implicitly conditioning on "employment" which is a post-treatment variable and can lead to selection bias.

Mediation. A common way a post-treatment variable can lead to bias in identifying the full causal effect of D on Y is if it lies on a causal path from the treatment to the outcome (Figure 8.11a). In this case, the causal influence that flows through that path is blocked and we are only measuring a partial effect. It is important to note, that the causal influence of such a path can be partially blocked even if one conditions on a descendant of the mediator (Figure 8.11b).

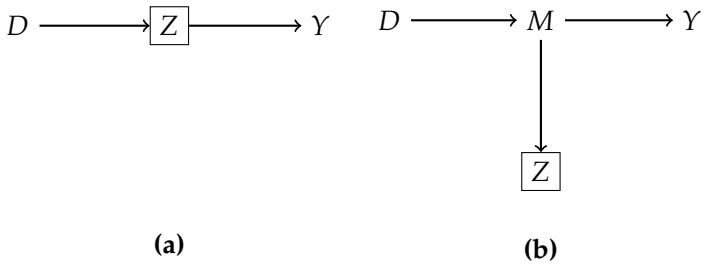


Figure 8.11: Bad controls for learning the full direct effect of D on Y: **(a)** over-control bias, by controlling for a mediator. **(b)** over-control bias, by controlling for an outcome caused by a mediator.

Interestingly, controlling for an ancestor of a mediator (Figure 8.12) does not impede us from learning the full direct effect of D on Y. In this case, the flow through the causal path $D \to M \to Y$ is not blocked by Z. For example, d-separation can be easily checked in the SWIG $\tilde{\mathsf{G}}(d)$ produced by fix$(D = d)$.

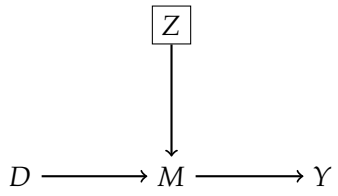


Figure 8.12: Neutral control. Cause of a mediator. Can potentially improve precision.

When we are controlling for a post-treatment variable that mediates the effect of the treatment as in Figure 8.11a, we are only capturing direct effects from the treatment to the outcome that do not work through this mediator. This type of direct effect after controlling for mediators is typically referred to as a *controlled direct effect*. Identifying the controlled direct effect is many times a relevant empirical question, in which case controlling for Z is not problematic. However, even when we are interested in the controlled direct effect, we should pay attention to cases where the mediators are themselves confounded through unobserved factors as illustrated in Figure 8.13. In such settings, by controlling for the mediator, we are opening a collider path $D \to Z \leftarrow U \to Y$ which can lead to severe bias, such as calculating non-zero direct effects even when they are zero.

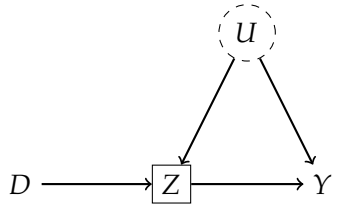
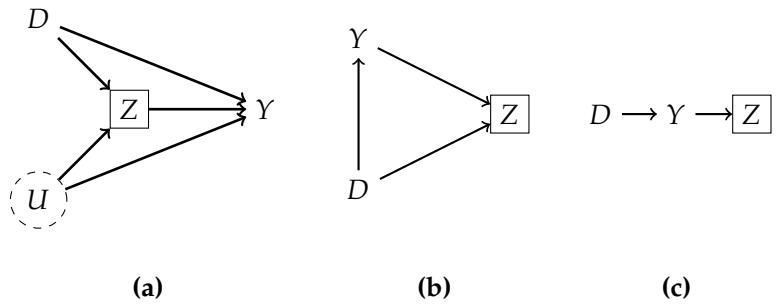


Figure 8.13: Bad control even for the *controlled direct effect*. Confounded mediator bias.

Heckman selection bias Another common way that post-treatment variables can lead to bias is due to collider bias or Heckman selection, as described in Section 6.3. In this case, conditioning on the post-treatment variable introduces spurious correlations between the treatment variable and some other variable which opens new paths of non-causal influence from the treatment to the outcome. For instance, Figure 8.14a corresponds to the low birthweight paradox we presented in Example 6.3.2. Similarly, Figure 8.14b corresponds to the Hollywood Example Example, Example 6.3.1. Finally, Figure 8.14c arises when we are controlling for an outcome of the outcome as might be produced by recall bias in a case-control study.

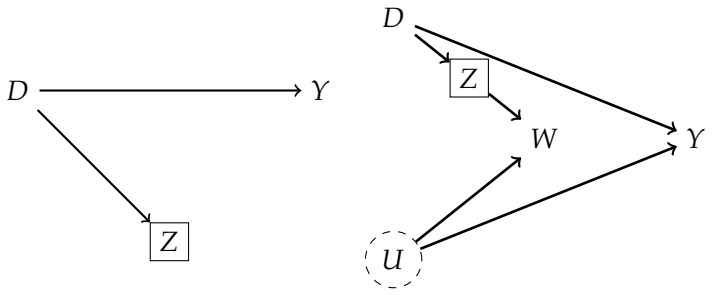


(a) **(b)** **(c)**

Figure 8.14: Bad controls: **(a)** collider stratification bias (e.g. low birth-weight "paradox" example), **(b)** collider stratification bias, **(c)** controlling for an outcome of the outcome of interest.

Example 8.3.3 (The Industrial Growth Puzzle [14]) In a study conducted during the nineteenth century in the US and Britain, it was found that despite nutrition quality D having improved, the height of men Y decreased. One possible explanation of the results of this study is that the subjects of the study were people who were enlisted in the army or in prison. Both of these variables, enlisted in the army and being in prison, are plausibly determined *after* the outcome variable of height is realized. It might, for example, be that taller men had more civilian opportunities growing up and did not end up enlisting in the army. In this case, looking at a sample of enlistees is implicitly controlling for an outcome of the outcome of interest which could lead to a biased estimate of the effect of nutrition on height.

There are of course some edge cases where controlling for a post-treatment variable Z does not lead to selection bias – e.g. Figure 8.15a and Figure 8.15b. In each of these two cases, the post-treatment variable is not a collider on a path from D to Y. However, it is not clear that adjusting for Z improves the analysis in any respect even in these cases, and adjusting for Z could potentially hurt precision.



(a) **(b)**

Figure 8.15: Neutral controls: **(a)** outcome of the treatment that is unrelated to the outcome of interest, **(b)** outcome of the treatment that does not introduce Heckman selection.

Notes

Any empirical study that tries to learn the causal effect of D on Y by conditioning on S must have a thought process that justifies this approach. The DAG/ASEM framework is a rigorous representation of such a thought process which enables explicit incorporation of domain knowledge, automatic checking of identifiability, and automatic deduction of testable restrictions. Graphs also provide an effective way of visualizing and communicating models.

Notebooks

- R: Dagitty Notebook employs the R package "dagitty" to analyze some simple DAGs as well as Pearl's Example. This package automatically finds adjustment sets and also lists testable restrictions in a DAG. Python: Pgmpy Notebook employs the analogue with Python package "pgmpy" and conducts the same analysis.

Study Problems

The study problems ask learners to continue the analysis of Pearl's Example DAG that we started in the Study Problems to Chapter 7. The provided notebooks are a useful starting point. Recall that Pearl's Example is structured as follows:

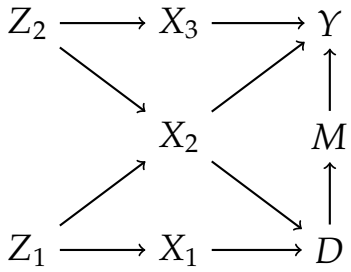


Figure 8.16: Pearl's Example

1. For Pearl's Example, write out the parents, non-parents, descendants, and non-descendants of nodes X_2 and M. List all the backdoor paths between Y and X_2. Can you identify the effect of X_2 on Y by conditioning?

2. (Front-Door-Criterion) For Pearl's Example, show that we can identify the effect $D \to M$ by conditioning on an empty set and the effect $M \to Y$ by conditioning on D. Combining the two results, we can identify the total

effect of D on Y. Solving this exercise analytically is a nice exercise; you can compare your results against causal identification packages. (Identification via this strategy is known as the Front-Door criterion; see Appendix 8.A.

3. Add an arrow $Z_2 \to Z_1$ in Pearl's Example and figure out how to identify the effect of $D \to Y$ by conditioning, of $D \to M$ by conditioning, and of $M \to Y$ by conditioning. (Note that valid conditioning sets may be empty.) Can you identify the effect of $X_2 \to Y$? If so, how? You may solve this analytically or using a causal identification package.

4. Add an arrow $X_1 \to M$ in Pearl's Example and figure out how to identify the effect of $D \to Y$ by conditioning, of $D \to M$ by conditioning, and of $M \to Y$ by conditioning. Can you identify the effect of $X_2 \to Y$? If so, how? You may solve this analytically or using a causal identification package.

5. Try to ask an instruction-following LLM (such as ChatGPT) about identification and valid adjustment sets, both for the original Pearl's Example as well as the variations in the latter two problems. Can you verify or find mistakes in the response? If you find mistakes, how might they be corrected? When mistakes are pointed out to the LLM, is it able to correct them? For example, you can try starting with the following prompt and make variations on it: "I have a causal graph with nodes Z1, Z2, X1, X2, X3, D, M, Y and edges Z1->X1, Z1->X2, Z2->X2, Z2->X3, X1->D, X2->D, X2->Y, X3->Y, D->M, M->Y. Is the effect of D on Y identified? What are the valid adjustment sets?"

8.A Front-Door Criterion via Example

We examine identification in Pearl's Example (Figure 8.2), via the front-door criterion. First note that we can write the potential outcome of interest $Y(d)$ as $Y(M(d))$, since in the SWIG $\tilde{G}(d)$ there is no other path from d to $Y(d)$ other than through $M(d)$.

$$\begin{aligned}
\mathrm{E}[Y(d)] &= \mathrm{E}[Y(M(d))] \\
&= \int \mathrm{E}[Y(M(d)) \mid M(d) = m]\mathrm{P}(M(d) = m)dm \\
&= \int \mathrm{E}[Y(m) \mid M(d) = m]\mathrm{P}(M(d) = m)dm
\end{aligned}$$

Suppose that we make a further surgery to the SWIG graph in Figure 8.3 by adding an intervention on the variable $M(d)$, i.e. take the modified SWIG graph induced by intervention $\text{fix}(D = d)$ and on that graph make a further intervention $\text{fix}(M(d) = m)$. This leads to the new SWIG:

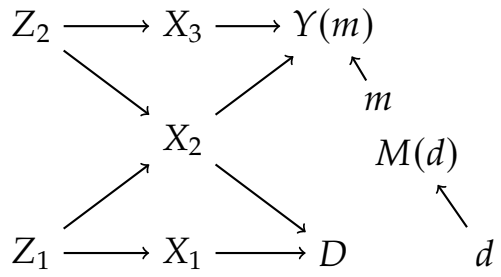


Figure 8.17: The DAG induced by a recursive Fix/SWIG intervention $\text{fix}(M(d) = m)$ on the SWIG in Figure 8.3.

Note that in this SWIG, we have $Y(m) \perp\!\!\!\perp M(d)$. Thus we have:

$$\mathrm{E}[Y(m) \mid M(d) = m] = \mathrm{E}[Y(m)],$$

leading to the front-door formula:

$$\mathrm{E}[Y(d)] = \int \mathrm{E}[Y(m)]\mathrm{P}(M(d) = m)dm$$

The term $\mathrm{E}[Y(m)]$ is the mean counterfactual response of Y when we intervene on M and $\mathrm{P}(M(d) = m)$ is the probability law of the counterfactual response of M when we intervene on D. Both of these interventional quantities can be separately identified via backdoor adjustment. More concretely, $\mathrm{E}[Y(m)] = \mathrm{E}[\mathrm{E}[Y \mid M = m, D]]$, and $\mathrm{P}(M(d) = m) = \mathrm{P}(M = m \mid D = d)$.[4] Note that under linearity assumptions on the CEFs – i.e. $\mathrm{E}[Y \mid M = m, D] = \alpha m + \beta D + c$ and $\mathrm{E}[M \mid D = d] = \gamma d + \delta$ – we get $\mathrm{E}[Y(1) - Y(0)] = \alpha\gamma$.[5] Thus, the average treatment effect $\alpha\gamma$, can be estimated by estimating α via OLS of Y on M, D and γ via OLS of M on D.

4: See Exercise 2.

5: Prove this as a reading exercise.

Bibliography

[1] Jeffrey M Wooldridge. 'Violating ignorability of treatment by controlling for too many factors'. In: *Econometric Theory* 21.5 (2005), pp. 1026–1028 (cited on page 197).

[2] Thomas S. Richardson and James M. Robins. *Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality*. Working Paper No. 128, Center for the Statistics and the Social Sciences, University of Washington. 2013. URL: https://csss.uw.edu/files/working-papers/2013/wp128.pdf (cited on pages 199, 200).

[3] Tyler J. VanderWeele and Ilya Shpitser. 'A new criterion for confounder selection'. In: *Biometrics* 67.4 (2011), pp. 1406–1413 (cited on pages 201, 202, 204).

[4] Judea Pearl. *Causality*. Cambridge University Press, 2009 (cited on page 202).

[5] Carlos Cinelli, Andrew Forney, and Judea Pearl. 'A Crash Course in Good and Bad Controls'. In: *Sociological Methods & Research* (2022) (cited on page 204).

[6] Miguel A Hernán, Sonia Hernández-Díaz, Martha M Werler, and Allen A Mitchell. 'Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology'. In: *American Journal of Epidemiology* 155.2 (2002), pp. 176–184 (cited on pages 206, 210).

[7] A Pastuszak, D Bhatia, B Okotore, and G Koren. 'Preconception counseling and women's compliance with folic acid supplementation.' In: *Canadian Family Physician* 45 (1999), p. 2053 (cited on page 206).

[8] Rolv Terje Lie, Allen J Wilcox, and Rolv Skjærven. 'A population-based study of the risk of recurrence of birth defects'. In: *New England Journal of Medicine* 331.1 (1994), pp. 1–4 (cited on page 206).

[9] Peng Ding and Luke W. Miratrix. 'To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias'. In: *Journal of Causal Inference* 3.1 (2015), pp. 41–57 (cited on pages 208, 209).

[10] Judea Pearl. 'Comment on Ding and Miratrix: "To Adjust or Not to Adjust?"' In: *Journal of Causal Inference* 3.1 (2015), pp. 59–60. DOI: doi:10.1515/jci-2015-0004 (cited on pages 208, 210).

[11] Cosma Rohilla Shalizi and Andrew C Thomas. 'Homophily and contagion are generically confounded in observational social network studies'. In: *Sociological Methods & Research* 40.2 (2011), pp. 211–239 (cited on page 209).

[12] Felix Elwert and Christopher Winship. 'Endogenous selection bias: The problem of conditioning on a collider variable'. In: *Annual Review of Sociology* 40 (2014), pp. 31–53 (cited on page 209).

[13] Wei Liu, M Alan Brookhart, Sebastian Schneeweiss, Xiaojuan Mi, and Soko Setoguchi. 'Implications of M bias in epidemiologic studies: a simulation study'. In: *American Journal of Epidemiology* 176.10 (2012), pp. 938–948 (cited on page 209).

[14] Eric B Schneider. 'Collider bias in economic history research'. In: *Explorations in Economic History* 78 (2020), p. 101356 (cited on page 212).