A Crash Course in Good and Bad Controls

Carlos Cinelli* Andrew Forney† Judea Pearl ‡

April 15, 2021

Abstract

Many students of statistics and econometrics express frustration with the way a problem known as "bad control" is treated in the traditional literature. The issue arises when the addition of a variable to a regression equation produces an unintended discrepancy between the regression coefficient and the effect that the coefficient is expected to represent. Avoiding such discrepancies presents a challenge to all analysts in the data intensive sciences. This note describes graphical tools for understanding, visualizing, and resolving the problem through a series of illustrative examples. We have found that the cases presented here can serve as a powerful instructional device to supplement more extended and formal discussions of the problem. By making this "crash course" accessible to instructors and practitioners, we hope to avail these tools to a broader community of scientists concerned with the causal interpretation of regression models.

Introduction

Students trained in the traditional econometrics pedagogy have likely encountered the problem of "bad controls" (Angrist and Pischke, 2009, 2014). The problem arises when an analyst needs to decide whether or not the addition of a variable to a regression equation helps getting estimates closer to the parameter of interest. Analysts have long known that some variables, when added to the regression equation, can produce unintended discrepancies between the regression coefficient and the effect that the coefficient is expected to represent. Such variables have become known as "bad controls," to be distinguished from "good controls" (also known as "confounders" or "deconfounders") which are variables that must be added to the regression equation to eliminate what came to be known as "omitted variable bias"(Angrist and Pischke, 2009; Steiner and Kim, 2016; Cinelli and Hazlett, 2020).

*Department of Statistics, University of California, Los Angeles. Email: carloscinelli@ucla.edu

†Department of Computer Science, Loyola Marymount University, Los Angeles. Email: Andrew.Forney@lmu.edu

‡Department of Computer Science, University of California, Los Angeles. Email: judea@cs.ucla.edu

The problem of "bad controls" however, has not received systematic attention in the standard statistics and econometrics literature. While most of the widely adopted textbooks discuss the problem of omitting "relevant" variables, they do not provide guidance on deciding which variables are relevant, nor which variables, if included in the regression, could induce, or worsen existing biases.[1] Researchers exposed only to this literature may get the impression that adding "more controls" to a regression model is always better. The few exceptions that do discuss the problem of "bad controls" unfortunately cover only a narrow aspect of the problem (e.g. Angrist and Pischke, 2009, 2014; Wooldridge, 2010; Imbens and Rubin, 2015; Gelman et al., 2020). Typical is the discussion found in Angrist and Pischke (2009, p.64)

> Some variables are bad controls and should not be included in a regression model, even when their inclusion might be expected to change the short regression coefficients. Bad controls are variables that are themselves outcome variables in the notional experiment at hand. That is, bad controls might just as well be dependent variables too. Good controls are variables that we can think of having been fixed at the time the regressor of interest was determined.

Here, "good controls" are defined as variables that are thought to be unaffected by the treatment, whereas "bad controls" are variables that could be in principle affected by the treatment. Similar discussion can be found in Rosenbaum (2002) and Rubin (2009), for qualifying a variable for inclusion in propensity score analysis. Some authors (e.g, Wooldridge, 2010; Gelman et al., 2020) briefly warn about the potential of bias amplification of certain pre-treatment variables, but do not elaborate further. Although an improvement over an absence of discussion, these conditions are neither necessary nor sufficient for deciding whether a variable is a good control.

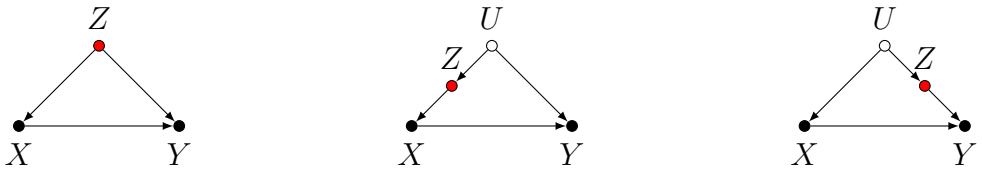
Recent advances in graphical models have produced simple criteria to distinguish "good" from "bad" controls; these range from necessary and sufficient conditions for deciding which set of variables should be adjusted for to identify the causal effect of interest (e.g, the back-door criterion and adjustment criterion in Pearl (1995) and Shpitser et al. (2012)), to, among a set of valid adjustment sets, deciding which ones would yield more precise estimates (Hahn, 2004; White and Lu, 2011; Henckel et al., 2019; Rotnitzky and Smucler, 2019; Witte et al., 2020). The purpose of this note is to provide practicing analysts a concise, simple, and *visual* summary of these criteria through illustrative examples.

Here we assume that readers are familiar with the basic notions of causal inference, directed acyclic graphs (DAGs), and in particular "path-blocking" as well as back-door paths. For those who need to refresh these notions, we provide a gentle introduction in the Appendix. In the following set of models, the target of the analysis is the *average causal effect* (ACE) of a treatment X on an outcome Y, which stands for the expected increase of Y per unit of a *controlled* increase in X. Observed variables will be designated

[1] See Chen and Pearl (2013) for a critical appraisal of econometrics textbooks, and Bollen and Pearl (2013) for eight misconceptions that still prevail in statistics and the social sciences.

by black dots and unobserved variables by white empty circles. Variable Z, highlighted in red, will represent the variable whose inclusion in the regression equation is to be decided, with "good control" standing for bias reduction, "bad control" standing for bias increase, and "neutral control" when the addition of Z neither increases nor decreases the asymptotic bias. For this last case, we will also make brief remarks about how Z could affect the precision of the ACE estimate.

Models 1, 2 and 3 – Good Controls

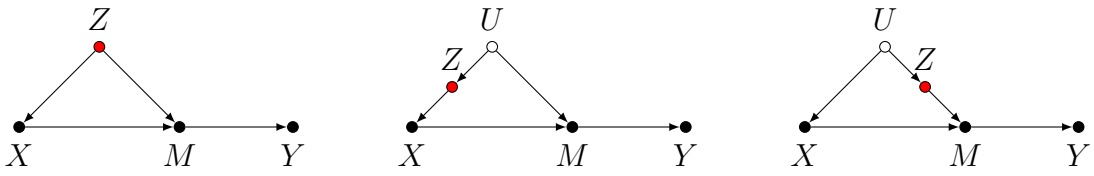


(a) Model 1 (b) Model 2 (c) Model 3

Figure 1: Models 1, 2, and 3

In Model 1, Z stands for a common cause of both X and Y. Once we control for Z, we block the back-door path from X to Y, producing an unbiased estimate of the ACE. In Models 2 and 3, Z is not a common cause of both X and Y, and therefore, not a traditional "confounder" as in Model 1. Nevertheless, controlling for Z blocks the back-door path from X to Y due to the unobserved confounder U, and again, produces an unbiased estimate of the ACE.

Models 4, 5 and 6 – Good Controls



(a) Model 4 (b) Model 5 (c) Model 6

Figure 2: Models 4, 5, and 6

When thinking about possible threats of confounding, modelers need to keep in mind that common causes of X and any mediator (between X and Y) also confound the effect of X on Y. Therefore, Models 4, 5 and 6 are analogous to Models 1, 2 and 3—controlling for Z blocks the back-door path from X to Y and produces an unbiased estimate of the ACE.

Model 7 – Bad Control (M-bias)

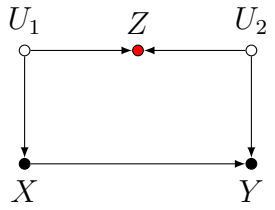


Figure 3: Model 7

We now encounter our first "bad control." Here Z is correlated with the treatment and the outcome and it is also a "pre-treatment" variable. Traditional econometrics textbooks would deem Z a "good control" (Angrist and Pischke, 2009, 2014; Imbens and Rubin, 2015). The back-door criterion, however, reveals that Z is a "bad control." Controlling for Z will induce bias by opening the back-door path $X \leftarrow U1 \rightarrow Z \leftarrow U2 \rightarrow Y$, thus spoiling a previously unbiased estimate of the ACE. This structure is known as the "M-bias," and has spurred several controversies. Readers can find further discussion in Pearl (2009a, p. 186), Shrier (2009), Pearl (2009c,b), Sjölander (2009), Rubin (2009), Ding and Miratrix (2015), and Pearl (2015).

Model 8 – Neutral Control (possibly good for precision)

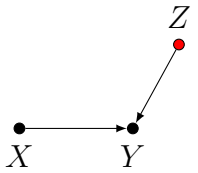


Figure 4: Model 8

In Model 8, Z is not a confounder nor does it block any back-door paths. Likewise, controlling for Z does not open any back-door paths from X to Y. Thus, in terms of asymptotic bias, Z is a "neutral control." Analysis shows, however, that controlling for Z reduces the variation of the outcome variable Y, and helps to improve the precision of the ACE estimate in finite samples (Hahn, 2004; White and Lu, 2011; Henckel et al., 2019; Rotnitzky and Smucler, 2019).

Model 9 – Neutral Control (possibly bad for precision)

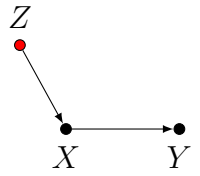


Figure 5: Model 9

Similar to the previous case, in Model 9 Z is "neutral" in terms of bias reduction. However, controlling for Z will reduce the variation of the treatment variable X and so may hurt the precision of the estimate of the ACE in finite samples (Henckel et al., 2019, Corollary 3.4). As a general rule of thumb, parents of X which are not necessary for identification are harmful for the asymptotic variance of the estimator; on the other hand, parents of Y which do not spoil identification are beneficial. See Henckel et al. (2019) for recent developments in graphical criteria for efficient estimation via adjustment in linear models. Remarkably, these conditions also have been shown to hold in non-parametric models for a broad class of non-parametric estimators (Rotnitzky and Smucler, 2019).

Model 10 – Bad Control (bias amplification)

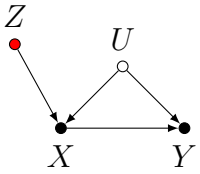


Figure 6: Model 10

We now encounter our second "pre-treatment" "bad control," due to a phenomenon called "bias amplification" (Bhattacharya and Vogt, 2007; Wooldridge, 2009; Pearl, 2011, 2010, 2013; Middleton et al., 2016; Steiner and Kim, 2016). Naive control for Z in this model will not only fail to deconfound the effect of X on Y, but, in linear models, will amplify any existing bias.

Models 11 and 12 – Bad Controls

(a) Model 11 (b) Model 12

Figure 7: Models 11 and 12

If our target quantity is the ACE, we want to leave all channels through which the causal effect flows "untouched." In Model 11, Z is a mediator of the causal effect of X on Y. Controlling for Z will block the very effect we want to estimate (the *total* effect of X on Y), thus biasing our estimates (this is usually known as "overcontrol bias"). In Model 12, although Z is not itself a mediator of the causal effect of X on Y, controlling for Z is equivalent to partially controlling for the mediator M, and will thus bias our estimates. Models 11 and 12 violate the back-door criterion (Pearl, 2009a), which excludes controls that are descendants of the treatment along paths to the outcome. Note the same conclusions would hold if we had an extra direct causal path $X \to Y$.

Total *versus* direct effects

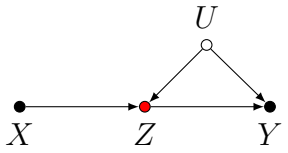


Figure 8: Variation of Model 11

The previous considerations assume the researcher is interested in the *total* effect of X on Y, as given by the ACE. If, instead, interest lies in the *controlled direct effect* (CDE) of X on Y (the effect of X while holding Z constant by intervention), then adjusting for Z in Model 11 (Figure 7(a)) would indeed be appropriate. However, consider a variation of Model 11 with an unobserved confounder of Z and Y, denoted by U, as shown in Figure 8. First notice that U *does not* confound the effect of X on Y, and thus our ACE estimate remains unbiased as it were in Model 11, so long as we do not adjust for Z. On the other hand, here adjusting for Z now opens the colliding path $X \to Z \leftarrow U \to Y$, thus biasing the CDE estimate.

Model 13 – Neutral Control (possibly good for precision)

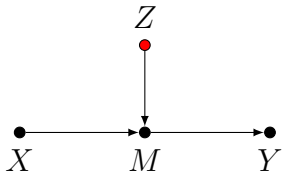


Figure 9: Model 13

At first look, Model 13 might seem similar to Model 12, and one may think that adjusting for Z would bias the effect estimate, by restricting variations of the mediator M. However, the key difference here is that Z is a cause, not an effect, of the mediator (and, consequently, also a cause of Y). Thus, Model 13 is analogous to Model 8, and so controlling for Z will be neutral in terms of bias and may increase the precision of the ACE estimate in finite samples. Readers can find further discussion of this case in Pearl (2013).

Models 14 and 15 – Neutral Controls (possibly helpful in the case of selection bias)

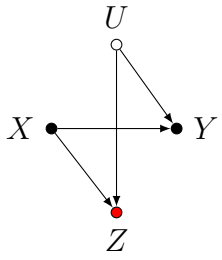
(a) Model 14 (b) Model 15

Figure 10: Models 14 and 15

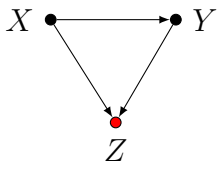
Contrary to econometrics folklore, not all "post-treatment" variables are inherently bad controls. In Models 14 and 15 controlling for Z does not open any confounding paths between X and Y. Thus, Z is neutral in terms of bias. However, controlling for Z does reduce the variation of the treatment variable X and so may hurt the precision of the ACE estimate in finite samples. Additionally, in Model 15, suppose one has only

samples with $W = 1$ recorded (a case of selection bias[2], which we explain next). In this case, controlling for Z can help to obtain the W-specific effect of X on Y by blocking the colliding path due to W.

Models 16 and 17 – Bad Controls (selection bias)



(a) Model 16



(b) Model 17

Figure 11: Models 16 and 17

Contrary to Models 14 and 15, here controlling for Z is no longer harmless, and induces what is classically known as "selection bias" or "collider stratification bias." Adjusting for Z in Model 16 opens the colliding path $X \to Z \leftarrow U \to Y$ and so biases the ACE. In Model 17, adjusting for Z not only opens the path $X \to Z \leftarrow Y$, but also the colliding path due to the latent parents of Y, thus biasing the ACE and motivating our final example.

Model 18 – Bad Control (case-control bias)

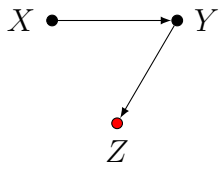


Figure 12: Model 18

In our last example, Z is not in the causal pathway from X to Y, Z is not a direct cause of X, and Z is connected to Y. Thus, one might surmise that, as in Model 8, controlling for Z is harmless for identification, and perhaps beneficial for finite sample

[2]Some economists may denote confounding bias as "selection bias," meaning preferential selection to treatment (Angrist and Pischke, 2009, 2014). Here selection bias means preferential selection into the available data.

efficiency. However, controlling for the effects of the outcome Y will induce bias in the estimate of the ACE, even without the direct arrow $X \to Z$, thus making Z a "bad control." This happens because Z is in fact a descendant of a collider: the outcome Y itself. A visual explanation of this phenomenon using "virtual colliders" can be found in Pearl (2009a, Sec. 11.3), which we also reproduce in the Appendix. Model 18 is special case of selection bias usually known as a "case-control bias" (see Pearl (2013) for a derivation of the bias in linear systems). Finally, although controlling for Z will generally bias numerical estimates of the ACE, it does have an exception when X has no causal effect on Y. In this scenario, X is still d-separated from Y even after conditioning on Z. Thus, adjusting for Z is valid for testing whether the effect of X on Y is *zero*.

Bad controls in applied research

Despite their simplicity, these illustrative examples should provide practitioners with a principled framework to understand many problems found in real world applications. To demonstrate, we now briefly present three cases of bad controls discussed in applied research, coming from diverse areas such as epidemiology, sociology, and economics.

The birth-weight paradox (Hernández-Díaz et al., 2006). Infants born to smokers were found to have higher risks of mortality than infants born to non-smokers. However, among infants with low birth-weight (LBW), this relationship was reversed. This reversal of effects has created many controversies in epidemiology—does it mean that maternal smoking is beneficial for LBW infants? A plausible reason for such a finding could simply be collider stratification bias, as shown in Model 16. Here X is maternal smoking, Y infant mortality, Z birth-weight, and U stands for unobserved risk-factors (such as birth-defects and malnutrition), that could also affect birth-weight. Note that stratifying the analysis by birth-weight would induce a spurious association between smoking and mortality due to the competing risk-factors. LBW infants of non-smokers need to have alternative causes for their LBW (such as malnutrition), and such causes could also lead to higher mortality.

Homophily bias in social network analysis (Elwert and Winship, 2014). An important task in the causal inference of social networks is to estimate the causal effects of social contagion, also known as "interpersonal effects." However, social ties in the analysis of social networks may be pre-treatment colliders as exemplified in the "M-bias" structure of Model 7. Suppose we are interested in assessing whether the civic engagement of individual 1 (X) leads to the civic engagement of individual 2 in the subsequent time period (Y). Let Z denote whether such individuals are friends, and U_1 and U_2 denote the personal characteristics (such as altruism) of individuals 1 and 2, respectively. Here, the social tie Z is a collider, and computing the association of Y

and X between friends ($Z = 1$) would bias the interpersonal causal effects in civic engagement.

The Antebellum Puzzle (Schneider, 2020). An interesting puzzle of economic history is the fact that, during the nineteenth century in Britain and the United States, the average height of adult men fell even though the economic conditions of these countries improved alongside childhood nutrition. One possible explanation for such a paradoxical finding is selection bias in the forms of Models 17 and 18 wherein researchers using data from individuals enlisted in the military or in prison are effectively conditioning on colliders. For military records, consider Model 18, and let X denote childhood nutrition, Y adult height, and Z an indicator of whether the individual was enlisted in the military. The causal path from Y to Z represents the fact that taller men may have better opportunities in the civilian market, and thus shorter men were more likely to enlist. Restricting the analysis to those enlisted in the military is therefore equivalent to controlling for Z, and leads to selection bias. Now for prison records, consider Model 17, and let Z be an indicator of whether the individual was arrested. Here one could argue that both childhood nutrition and adult height have pathways to committing a crime through socio-economic opportunities, thus again leading to selection bias.

These examples are by no means exhaustive. Readers can find other interesting cases across applied sciences, such as: the threats of collider bias in understanding risk factors of COVID-19 (Griffith et al., 2020); the "Obesity paradox," in which obesity appears to benefit individuals who survive heart failure (Banack and Kaufman, 2013); and examples of "bad controls" due to adjustment of mediators and colliders in multigenerational mobility (Breen, 2018), anesthesiology research (Gaskell and Sleigh, 2020) or animal science (Bello et al., 2018). Further discussion of the many instances of bad controls found in applied work is beyond the scope of this crash course.

Multiple controls

When considering multiple controls, the status of a single control as "good" or "bad" may change depending on the context of the other variables under consideration. Nevertheless, the main lessons from our illustrative examples remain. A set of control variables $\mathbf{Z}$ will be "good" if: (i) it blocks all non-causal paths from the treatment to the outcome; (ii) it leaves any mediating paths from the treatment to the outcome "untouched" (since we are interested in the total effect); and, (iii) it does not open new spurious paths between the treatment and the outcome (e.g., due to colliders). As to efficiency considerations, we should give preference to those variables "closer" to the outcome, in opposition to those closer to the treatment—so long as, of course, this does not spoil identification.

Finally, we remind readers that, when considering models with more complicated

structures, one can always resort to specialized computer programs. Open-source software implementing algorithms for selecting adjustment sets can be found in the `R` packages `pcalg` (Kalisch et al., 2012), `dagitty` (Textor et al., 2016)[3], and `causaleffect` (Tikka and Karvanen, 2017). Users familiar with the software `SAS` may find the procedure `CAUSALGRAPH` useful (Thompson, 2019). A web application implementing the methods discussed Bareinboim and Pearl (2016) is also available.[4] In other words, given a causal diagram, the problem of deciding which variables are good or bad controls has been automatized.

Concluding remarks

In this note, we demonstrated through illustrative examples how simple graphical criteria can be used to decide when a variable should (or should not) be included in a regression equation—and thus whether it can be deemed a "good" or "bad" control. Many of these examples act as cautionary notes against prevailing practices in traditional statistics and econometrics: for instance, Models 7 to 10 reveal that one should be wary of the general recommendation, usually derived from propensity score logic, of conditioning on all pre-treatment predictors of the treatment assignment[5]; whereas Models 14 and 15 show that not all "post-treatment" variables are "bad-controls," and some may even help with identification.

In all cases, structural knowledge is indispensable for deciding whether a variable is a good or bad control, and graphical models provide a natural language for articulating such knowledge, as well as efficient tools for examining its logical ramifications. We have found that an example-based approach to "bad controls," such as the one presented here, can serve as a powerful instructional device to supplement more extended and formal discussions of the problem. By making this "crash course" accessible to instructors and practitioners, we hope to avail these tools to a broader community of scientists concerned with the causal interpretation of regression models.

References

Angrist, J. and Pischke, J.-S. (2009). *Mostly harmless econometrics: an empiricists guide*. Princeton: Princeton University Press.

Angrist, J. D. and Pischke, J.-S. (2014). *Mastering 'metrics: The path from cause to effect*. Princeton University Press.

[3]Also available online in `www.dagitty.net`.

[4]Available online in `www.causalfusion.net`.

[5]For instance, examples of such recommendations can be found in Rosenbaum (2002, p.76), Rubin (2009), Imbens and Rubin (2015, p.265), Dorie et al. (2016, p.3453).

Banack, H. R. and Kaufman, J. S. (2013). The "obesity paradox" explained. *Epidemiology*, 24(3):461–462.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Bello, N. M., Ferreira, V. C., Gianola, D., and Rosa, G. J. (2018). Conceptual framework for investigating causal effects from observational data in livestock. *Journal of animal science*, 96(10):4045–4062.

Bhattacharya, J. and Vogt, W. B. (2007). Do instrumental variables belong in propensity scores? Technical report, National Bureau of Economic Research.

Bollen, K. A. and Pearl, J. (2013). Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer.

Breen, R. (2018). Some methodological problems in the study of multigenerational mobility. *European Sociological Review*, 34(6):603–611.

Chen, B. and Pearl, J. (2013). Regression and causation: a critical examination of six econometrics textbooks. *Real-World Economics Review*.

Cinelli, C. and Hazlett, C. (2020). Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):39–67.

Ding, P. and Miratrix, L. W. (2015). To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference*, 3(1):41–57.

Dorie, V., Harada, M., Carnegie, N. B., and Hill, J. (2016). A flexible, interpretable framework for assessing sensitivity to unmeasured confounding. *Statistics in medicine*, 35(20):3453–3470.

Elwert, F. and Winship, C. (2014). Endogenous selection bias: The problem of conditioning on a collider variable. *Annual review of sociology*, 40:31–53.

Gaskell, A. L. and Sleigh, J. W. (2020). An introduction to causal diagrams for anesthesiology research. *Anesthesiology*, 132(5):951–967.

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

Griffith, G. J., Morris, T. T., Tudball, M. J., Herbert, A., Mancano, G., Pike, L., Sharp, G. C., Sterne, J., Palmer, T. M., Smith, G. D., et al. (2020). Collider bias undermines our understanding of covid-19 disease risk and severity. *Nature communications*, 11(1):1–12.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *Review of Economics and Statistics*, 86(1):73–76.

Henckel, L., Perković, E., and Maathuis, M. H. (2019). Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*.

Hernández-Díaz, S., Schisterman, E. F., and Hernán, M. A. (2006). The birth weight "paradox" uncovered? *American journal of epidemiology*, 164(11):1115–1120.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26.

Middleton, J. A., Scott, M. A., Diakow, R., and Hill, J. L. (2016). Bias amplification and bias unmasking. *Political Analysis*, 24(3):307–323.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2009a). *Causality*. Cambridge University Press.

Pearl, J. (2009b). Letter to the editor: Remarks on the method of propensity score. *Statistics in Medicine*, 28:1420–1423. URL: `https://ucla.in/2NbS14j`.

Pearl, J. (2009c). Myth, confusion, and science in causal analysis. *UCLA Cognitive Systems Laboratory*, Technical Report (R-348). URL: `https://ucla.in/2EihVyD`.

Pearl, J. (2010). On a class of bias-amplifying variables that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 417–424. URL: `https://ucla.in/2N8mBMg`.

Pearl, J. (2011). Invited commentary: understanding bias amplification. *American journal of epidemiology*, 174(11):1223–1227. URL: `https://ucla.in/2PORDX2`.

Pearl, J. (2013). Linear models: A useful "microscope" for causal analysis. *Journal of Causal Inference*, 1(1):155–170. URL: `https://ucla.in/2LcpmHz`.

Pearl, J. (2015). Comment on ding and miratrix:"to adjust or not to adjust?". *Journal of Causal Inference*, 3(1):59–60. URL: `https://ucla.in/2PgOWNd`.

Rosenbaum, P. R. (2002). *Observational studies*. Springer.

Rotnitzky, A. and Smucler, E. (2019). Efficient adjustment sets for population average treatment effect estimation in non-parametric causal graphical models. *arXiv preprint arXiv:1912.00306*.

Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine*, 28(9):1420–1423.

Schneider, E. B. (2020). Collider bias in economic history research. *Explorations in Economic History*, 78:101356.

Shpitser, I., VanderWeele, T., and Robins, J. M. (2012). On the validity of covariate adjustment for estimating causal effects. *arXiv preprint arXiv:1203.3515*.

Shrier, I. (2009). Propensity scores. *Statistics in Medicine*, 28(8):1317–1318.

Sjölander, A. (2009). Propensity scores and m-structures. *Statistics in medicine*, 28(9):1416–1420.

Steiner, P. M. and Kim, Y. (2016). The mechanics of omitted variable bias: Bias amplification and cancellation of offsetting biases. *Journal of causal inference*, 4(2).

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., and Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: the R package 'dagitty'. *International journal of epidemiology*, 45(6):1887–1894.

Thompson, C. (2019). Causal graph analysis with the causalgraph procedure. In *Proceedings of SAS Global Forum*.

Tikka, S. and Karvanen, J. (2017). Identifying causal effects with the R package causaleffect. *Journal of Statistical Software*, 76.

White, H. and Lu, X. (2011). Causal diagrams for treatment effect estimation with application to efficient covariate selection. *Review of Economics and Statistics*, 93(4):1453–1459.

Witte, J., Henckel, L., Maathuis, M. H., and Didelez, V. (2020). On efficient adjustment in causal graphs. *arXiv preprint arXiv:2002.06825*.

Wooldridge, J. (2009). Should instrumental variables be used as matching variables. Technical report, Citeseer.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

A Appendix

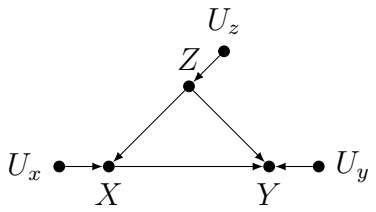
This appendix provides a short introduction to the notions of causal models, causal diagrams and "path-blocking" for the identification of causal effects via adjustment. Readers can find more extensive discussions in Pearl (2009a).

Structural causal models and causal diagrams

In order to decide whether there is a discrepancy between a certain regression equation (an associational quantity), and a target "causal effect" (a causal quantity), we need to mathematically *define* what this causal effect is. And to do that, we first need the concept of a *causal model*. We briefly introduce *structural causal models* (SCM) (Pearl, 2009a) with an example.

$$M = \begin{cases} Z & \leftarrow f_z(U_z) \\ X & \leftarrow f_x(Z, U_x) \\ Y & \leftarrow f_y(X, Z, U_y) \\ \mathbf{U} & \sim P(\mathbf{U}) \end{cases}$$

(a) Structural causal model M



(b) Causal diagram G associated with M

Figure 13: Structural Causal Model M and its associated graph G

Consider the SCM M shown in Figure 13(a). The variables $\mathbf{V} = \{Z, X, Y\}$ are called the *endogenous* variables, and stand for those variables that the investigator chose to model their cause-effect relationships; the variables $\mathbf{U} = \{U_z, U_x, U_y\}$ are called the *exogenous* variables and represent everything else that the investigator chose *not* to explicitly model (these are also usually called *disturbances*). The functions $\mathcal{F} = \{f_z, f_x, f_y\}$ are called *structural equations*, and each function represents a causal process that *assigns* to its respective endogenous variable a value based on the values of the other variables. We use the assignment symbol ($\leftarrow$) to emphasize the *asymmetry* in a causal relationship, flowing from cause to effect. Finally, the exogenous variables have an associated probability distribution $P(\mathbf{U})$ summarizing their uncertainty. In this particular example, we assume the exogenous variables are mutually independent (but in general, this need not be the case). The SCM M induces a joint distribution on the endogenous variables $P(\mathbf{V})$, which we denote by *observational distribution*. In observational studies, the investigator only has access to samples of $P(\mathbf{V})$.

Every SCM has an associated graph G, usually called its *causal diagram*. In the types of models we consider here, which do not exhibit cycles, the causal diagram will be a directed acyclic graph (DAG). The causal diagram of our example is shown in Figure 13(b). The graph G contains one node for each variable in M, and a directed arrow $V_i \rightarrow V_j$ whenever V_i appears in the structural equation of V_j, meaning that V_i is a

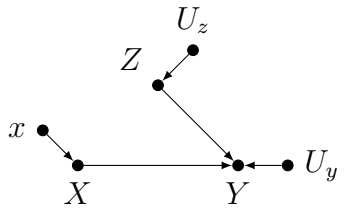
direct cause of V_j. Here we explicitly show the exogenous variables, but, conventionally, these are omitted from the graph for brevity. When the exogenous variables are omitted from the diagram, a dashed bidirected arrow $V_i \leftrightarrow V_j$ should be added whenever the exogenous variables entering f_{v_i} and f_{v_j} are *not* independent.

Interventions and causal effects

Interventions are modeled by modifying mechanisms of the SCM. For example, the act $do(X = x)$ in the model of Figure 13(a) amounts to replacing the original mechanism $X \leftarrow f_x(Z, U_x)$ with a new mechanism in which X is externally forced to attain the value x, i.e., $X \leftarrow x$. This results in the modified SCM M_x of Figure 14(a).

$$M_x = \begin{cases} Z & \leftarrow f_z(U_z) \\ X & \leftarrow x \\ Y & \leftarrow f_y(X, Z, U_y) \\ \mathbf{U} & \sim P(\mathbf{U}) \end{cases}$$

(a) Modified SCM M_x



(b) Modified causal diagram G_x

Figure 14: Effect of intervention $do(X = x)$

The model M_x induces an *interventional distribution* on the endogenous variables, denoted by $P(\mathbf{V} \mid do(X = x))$.With the concept of an intervention in mind, we can now define the average causal effect (i.e, the expected increase of Y per unit of a *controlled* increase in X) as the average contrast of Y under two distinct interventions:

$$ACE(x) = E[Y \mid do(x + 1)] - E[Y \mid do(x)]$$

In general the ACE varies depending on levels of x, but in linear models, as we show below, the ACE reduces to a single number. Other causal effects can be defined with the same model modification logic. For instance, the controlled direct effect (or CDE, i.e, the expected increase of Y per unit of a controlled increase in X, while holding Z constant) is defined as the difference:

$$CDE(x, z) = E[Y \mid do(x + 1), do(z)] - E[Y \mid do(x), do(z)]$$

Potential outcomes

Potential outcomes $\mathbf{V}_x$ are defined as the solution of the endogenous variables $\mathbf{V}$ in the modified model M_x. Thus, $P(\mathbf{V} \mid do(X = x))$ can be equivalently written as $P(\mathbf{V}_x)$ (and we could have written all variables in M_x and G_x as Z_x, X_x and Y_x). Likewise, the ACE can be equivalently written as $ACE(x) = E[Y_{x+1}] - E[Y_x]$.

Causal and non-causal paths: chains, forks and colliders

To make things concrete, let us suppose that the structural equations of our example are linear, that is, $Z \leftarrow U_z$, $X \leftarrow \lambda_{zx} Z + U_x$, $Y \leftarrow \lambda_{xy} X + \lambda_{zy} Z + U_y$. Further assume that the disturbances $\mathbf{U}$ are multivariate Gaussian. Then the ACE evaluates to:

$$ACE(x) = E[Y \mid do(x+1)] - E[Y \mid do(x)] = \lambda_{xy}$$

To contrast, now let us compute the regression coefficient of Y on X, denoted by β_{yx}

$$\beta_{yx} = \frac{Cov(Y, X)}{Var(X)} = \lambda_{xy} + \lambda_{zx}\lambda_{zy}$$

Note how the regression coefficient $\beta_{yx} = \lambda_{xy} + \lambda_{zx}\lambda_{zy}$ differs from the $ACE = \lambda_{xy}$. This happens because the observed association of X and Y mixes both the *causal* association (the path $X \rightarrow Y$), and the *non-causal* association due to the confounder Z (the path $X \leftarrow Z \rightarrow Y$). We call such confounding paths, that start with an arrow pointing to X, "back-door paths." Note, however, that the regression coefficient of Y on X adjusting for Z (denoted by $\beta_{yx.z}$) evaluates to (after some algebra)

$$\beta_{yx.z} = \lambda_{xy}$$

That is, controlling for Z in this model effectively blocks the back-door path, and recovers the ACE.

In general, how does path blocking work in a graphical model? To answer this question, we need to understand the three main patterns of a causal diagram, which helps us characterize when paths (consisting of sequences of the following triplets) of the graph are blocked or open.

- **Chains (mediators).** Chains are patterns of the form $X \rightarrow Z \rightarrow Y$, meaning that X causally affects Y through the mediator Z. Conditioning on Z in a chain blocks this flow of association.

- **Forks (common causes).** Forks are patterns of the form $X \leftarrow Z \rightarrow Y$, meaning X and Y share a common cause (a confounder) Z, thus inducing a *non-causal* association between both variables. Conditioning on Z in a fork blocks this flow of association.

- **Colliders (common effects).** Colliders are patterns of the form $X \rightarrow Z \leftarrow Y$, meaning that both X and Y share a common effect Z. Contrary to the other two patterns, this path is closed by default—conditioning on Z *opens* the path and induces a *non-causal* association between X and Y.

A final rule to keep in mind is that controlling for a descendant of a variable is equivalent to "partially" controlling for that variable. Thus, controlling for a descendant of a

mediator partially blocks the flow of association, whereas controlling for a descendant of a collider partially opens the flow of association.

We can now judge whether any path p in a graph, no matter how complicated, is blocked by a set $\mathbf{Z}$. This happens if, and only if: (i) p contains a chain or a fork, such that the middle node is in $\mathbf{Z}$; *or*, (ii) p contains a collider, such that neither the middle node, nor any of its descendants, are in $\mathbf{Z}$.

The back-door (or adjustment) criterion

Armed with these tools, the DAG reveals which set of variables $\mathbf{Z}$ blocks the correct paths for valid estimation of the ACE. We would like to find a set $\mathbf{Z}$, such that

- it blocks all spurious paths from X to Y;
- it *does not* (partially) block any of the causal paths from X to Y; and,
- it does not open other spurious paths.

That's the essence of what is known as the *back-door criterion*. If we can find such a set of controls $\mathbf{Z} = \{Z_1, \ldots, Z_k\}$, then the interventional expectation of Y can be computed from the observational distribution as

$$E[Y \mid do(X = x)] = E[E[Y \mid X = x, \mathbf{Z}]] \tag{1}$$

Linear *versus* non-linear models

The previous identification result is non-parametric, and it involves *two* expectations. First we compute the conditional expectation $E[Y \mid X = x, \mathbf{Z} = \mathbf{z}]$, then we *average* this conditional expectation over $P(\mathbf{Z})$. If, however, the conditional expectation function $E[Y \mid X = x, \mathbf{Z} = \mathbf{z}]$ is linear, the expression simplifies to

$$E[E[Y \mid X = x, \mathbf{Z}]] = \beta_{yx.\mathbf{z}}x + \sum_{j=1}^{k} \beta_{yz_j.x\mathbf{z}_{-\mathbf{j}}} E[Z_j]$$

Where $\mathbf{Z}_{-\mathbf{j}}$ denotes the set $\mathbf{Z}$ excluding Z_j. Therefore, under the parameteric assumption of linearity, the ACE simply equals the regression coefficient $\beta_{yx.\mathbf{z}}$, and no averaging over the distribution of $\mathbf{Z}$ is necessary (similar result can be obtained if the conditional expectation is linearly separable on X). If, however, the conditional expectation is not linear, the regression coefficient $\beta_{yx.\mathbf{z}}$ targets a different causal quantity, which may be an incomplete summary of the ACE (see, e.g., Angrist and Pischke, 2009). In such cases, users should resort back to the proper adjustment formula as given by Equation 1.

Virtual colliders and d-separation

Finally, we explain both d-separation and virtual colliders using the case of Model 18. Rewrite Model 18 showing the exogenous variables explicitly, as in Figure 15.

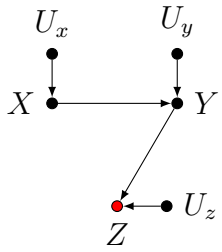


Figure 15: Model 18 showing exogenous variables

We can now clearly see the colliding path $X \rightarrow Y \leftarrow U_y$. Conditioning on Z, a descendant of Y, thus partially opens this path, and creates a spurious association between X and the disturbance of Y, making Z a "bad control."

Now let us consider the case in which the arrow $X \rightarrow Y$ is removed (zero causal effect of X on Y). First recall that two nodes X and Y are *d-separated* conditional on $\mathbf{Z}$ if the set $\mathbf{Z}$ blocks every path from X to Y in the graph. If X and Y are *d-separated* conditional on $\mathbf{Z}$, this implies the conditional independence $Y \perp\!\!\!\perp X \mid \mathbf{Z}$. In Model 18, when there is no path from X to Y, conditioning on Z also does not open any other paths between these two variables. Hence, X is still d-separated from Y even after conditioning on Z, and the conditional independence $Y \perp\!\!\!\perp X \mid \mathbf{Z}$ holds.