Author(s): Hünermund, Paul; Louw, Beyers; Rönkkö, Mikko

Title: The Choice of Control Variables : How Causal Graphs Can Inform the Decision

Year: 2022

Version: Published version

Please cite the original version:

Hünermund, P., Louw, B., & Rönkkö, M. (2022). The Choice of Control Variables : How Causal Graphs Can Inform the Decision. In Academy of Management Proceedings 2022 (2022). Academy of Management. Academy of Management Annual Meeting Proceedings. https://doi.org/10.5465/AMBPP.2022.294

10.5465/AMBPP.2022.294

THE CHOICE OF CONTROL VARIABLES: HOW CAUSAL GRAPHS CAN INFORM THE DECISION

PAUL HÜNERMUND
Copenhagen Business School
Kilevej 14A, DK-2000 Frederiksberg, E-mail: phu.si@cbs.dk

BEYERS LOUW
Maastricht University School of Business and Economics

MIKKO RÖNKKÖ
University of Jyväskylä School of Business

INTRODUCTION

Control variables have a central role in supporting causal claims with observational data. Yet, how control variables are selected is not always transparent and there is considerable debate in the management literature on how controls should be chosen (Antonakis et al., 2010). We provide a framework based on causal graphs that can settle such debates and allow a more productive discussion on how control variables should be selected and used.

CAUSAL DIAGRAMS AND CAUSAL IDENTIFICATION

Introductory text and guidelines on causal analysis often present three conditions as necessary for demonstrating causality: 1) association between the cause and effect, 2) direction of influence, and 3) elimination of rival explanations (Antonakis et al., 2010; Singleton & Straits, 2018, Chapter 4). Control variables are an important tool for addressing the third condition in observational research. Yet, choosing the right set of control variables in multiple regression analyses is a complex problem. Both including too few or including too many controls can lead to wrong conclusions. Econometrics texts typically focus more on the omitted variable case and how it can bias the estimates (Wooldridge, 2013, pp. 88–92), giving less attention to the issue of controlling for variables that should not be controlled for at all (overcontrolling) (Wooldridge, 2013, pp. 205–207).

Literature on structural causal models provides a more rigorous explanation of when variables should be controlled for and when not. This is done with c*ausal diagrams* (Pearl, 2009, pp. 30), which are nonparametric versions of structural equation models. That is, in a typical structural equation model all relationships are linear but no assumptions of linearity or any other functional form are made in causal diagrams. Causal diagrams, such as the simple examples shown in Table 1, consist of nodes that represent the variables in a model (here: T, Y, and X; Y commonly refers to the outcome) and edges that denote the causal relationship between them. Edges are directed, indicated by arrowheads pointing from a parent node to a child node. Causal diagrams are particularly useful for causal inference because the conditional independence relationships, which are a cornerstone for unbiased effect estimation in econometrics and statistics (Imbens, 2004), can be easily read off from the structure of the graph (Pearl, 1988).

There are three possible configurations between three nodes in a causal diagram:

A *chain*: $A \rightarrow B \rightarrow C$

10.5465/AMBPP.2022.294

A *fork:* $A \leftarrow B \rightarrow C$

And a *collider*: $A \rightarrow B \leftarrow C$

In the first two cases, the variables A and C are (stochastically) dependent. In a chain, A exerts an indirect causal effect on C, which results in a dependence; in a fork, both A and C are influenced by the same parent node and thus partially share the same information. In both cases, the dependence can be broken by conditioning on (controlling) the middle variable B. That is, holding B constant blocks the information transfer between A and C making them *conditionally independent*. In a chain, if B is fixed, A cannot influence C anymore. Likewise, if the common parent B is fixed in a fork, the remaining variation in A and C is independent.

By contrast, a collider behaves exactly the opposite way. Here both A and C are parent nodes of B, but otherwise share no relation and can be expected to be independent. Holding B constant, however, would create a correlation between A and C. A classic example involves a college (Morgan & Winship, 2007, pp. 66-67), where admission (B) depends on SAT score (A) and motivation rating (C), which are independent. However, if we sample only admitted students (condition on admission), we find a negative correlation between SAT and motivation scores. The reason for this is that if a student in the sample has a low SAT score, their motivation must be high to be admitted and vice versa. In the causal modeling language, we say that conditioning on B *unblocks* or *opens up* the path between A and C and renders them stochastically dependent.

A TAXONOMY OF CONTROLS

We next present a taxonomy of controls using causal diagrams based on Cinelli et al. (2022). We focus first on "good" controls that can reduce estimation bias. Then we discuss "bad" controls, which create rather than reduce spurious correlation when included in the analysis.

Table 1 about here

Good controls

Model 1 in Table 1 shows a fully exogenous (since no arrows are pointing into it) and pre-determined control variable X. This is the standard assumption in most textbook treatments of regression analysis. Since X exerts an effect on the treatment T as well as the outcome Y, it acts as a common parent of both. Due to this fork structure, X creates a spurious correlation through the *backdoor path* (Pearl & Mackenzie, 2018, pp. 158) $T \leftarrow X \rightarrow Y$ contaminating the causal effect of T on Y. The spurious correlation can be blocked, however, by conditioning on X to recover the true causal effect.

The same holds if X is simply correlated with either T or Y, as in model 2 of Table 1. In this case, there is an unobserved common influence factor U that affects X and Y, which otherwise share no genuine causal connection. The path $T \leftarrow X \leftarrow U \rightarrow Y$ is a combination of a fork and a chain and can be blocked by conditioning on X

Bad controls

We now turn attention to bad controls that, in contrast to good controls, compromise causal identification when included in the analysis. Model 3 in Table 1 depicts a graph in which

10.5465/AMBPP.2022.294

X exerts no causal effect, neither on T nor on Y. X is correlated with both, however, due to the presence of two unobserved confounders U_1 and U_2. Since both unobservables emit arrows that point into X, the node is a collider on the path $T \leftarrow U_1 \rightarrow X \leftarrow U_2 \rightarrow Y$. X thus blocks this path, which means that the relationship between T and Y is currently not confounded. By contrast, if the analyst decided to include X as a control in the regression, this conditioning would open up the path and lead to estimation bias.

In the cases discussed this far, X has been assumed to be a *pre-treatment* variable, which means that it either causally precedes the treatment T or is codetermined with T by other influence factors. We now turn attention to the mediator case in model 4 of Table 1. Here, X is itself causally affected by T, which renders it a *post-treatment* variable. This case is interesting because mediation analysis has a long tradition in management and leadership research (Baron and Kenny, 1986). The mediation path $T \rightarrow X \rightarrow Y$ is a chain thus controlling for X would close the path and block one of the mechanisms by which T exerts its influence on Y. This phenomenon is often described as "controlling away" part of the effect of a treatment.

The mediator X can itself be confounded by an unobserved variable U, as in model 5 of Table 1. Here, X is a collider on the path $T \rightarrow X \leftarrow U \rightarrow Y$, which means that controlling for X unblocks the path and leads to a spurious correlation between the treatment and outcome. In this case, it is possible to estimate the total causal effect of T on X, but decomposing it into directed and mediated effects is not possible.

Tricky cases

The correct way to deal with bad controls is to leave them out of the analysis. In the bad control cases we described so far, the (total) causal effect was already identified by regressing Y and T, and including X in the regression only made things worse. Unfortunately, things are not always quite so simple and solutions to the causal identification problem cannot easily be found. The most obvious case is the one depicted in model 6, Table 1, where an important confounder (or a set of confounders), affecting both T and Y, is not included in the data. In this case, covariate adjustment is not sufficient for identification and the analyst must use other techniques, such as instrumental variables (Antonakis et al., 2010).

Model 7 of Table 1 presents a much less obvious case. Here, X is both a confounder on the path $T \leftarrow X \rightarrow Y$, as well as a collider on the path $T \leftarrow U_1 \rightarrow X \leftarrow U_2 \rightarrow Y$. This means that while we would, in principle, like to control for X to close the confounding path, this automatically opens up the second path, which will in turn create collider bias. Unfortunately, there is no way out of this dilemma. As in the previous case, instead of covariate adjustment, other methods that can account for unobservables might be applicable though.

Unnecessary controls

Finally, we present two cases in which controlling for a third variable is not necessary for causal identification but might affect estimation precision (efficiency). In models 8 and 9 of Table 1, X only influences either the treatment or the outcome respectively. This means that X is not a confounder and therefore needs not be controlled. However, controlling for X influences the variance of estimates. In model 8, including X increases the precision of the estimate of the causal effect. Model 9 is the "irrelevant regressor" case that is often discussed in introductory econometric texts (Wooldridge, 2013, pp. 88) and leads to decreased efficiency.

10.5465/AMBPP.2022.294

GUIDELINES ON CHOOSING CONTROL VARIABLES

We will now present a set of guidelines for choosing control variables. We recommend that control variables are chosen based on a causal graph. A systematic application of a graph serves three purposes: 1) It helps in identifying those variables that should be controlled, 2) it makes it clear what variables should not be controlled, and 3) it makes the reporting of these decisions more transparent. In the full paper, we provide an empirical example of this process.

How to come up with a causal diagram?

First, start by longlisting potential variables focusing on the treatment and outcome of interest based on a thorough literature review. The review should focus on two aspects: 1) theory or what kinds of causal mechanisms have been proposed in the literature and 2) empirics or what variables other researchers have controlled for when studying the variable. While prior studies might have used controls that are unnecessary or even bad, longlisting all controls used in prior studies is important because 1) there may be a clear reason to use a control, but this is simply left unreported and 2) if a control is not included, particularly if it is determined to be a bad control, this decision should be documented to inform future research that can then avoid the control as well. In this stage, the focus should be on variables that are related to both the treatment and the outcome (even if only indirectly) because they are the ones that lie on a potential *backdoor* and mediating path in the causal diagram. The sources used in the longlisting of variables should be documented using the same standards that are currently used for meta-analyses and systematic reviews. Of course, potential variables with causal relationships can be longlisted even if they were not documented in prior studies if researchers have a clear justification for doing so.

After longlisting the variables, the variables should be shortlisted to focus on the most relevant variables. These inclusion decisions should focus on two aspects: 1) how strong is the presumed causal relationship and how much evidence there is to support its existence and 2) are any of the causes correlated with the sample selection criteria. These two criteria mirror the use of causal graphs for addressing spurious associations and endogenous selection (Elwert, 2013). After compiling the shortlist, the next step is to compile the causal diagram, which is a simple drawing exercise where the content of the shortlist is presented in a graphical format.

How to apply the causal diagram to choose the controls

Once the causal diagram has been constructed, the choice of controls involves just the application of causal identification rules based on the diagram, which is a mechanical exercise (Elwert, 2013; Pearl & Mackenzie, 2018). In the full paper, we explain this process with an empirical example. In short, the application of the causal diagram involves the use of the taxonomy shown in Table 1 and classifying each potential control into the categories of good, bad, tricky, and unnecessary. Good controls should always be included, and bad controls left out. Tricky cases call for researcher judgment and sensitivity analysis. Unnecessary controls should generally be left out (unless there are strong reasons to believe that the situation corresponds to model 8, in which case the control can be included to improve estimation precision but is not a requirement for causal identification of the treatment).

10.5465/AMBPP.2022.294

Sensitivity analysis

If unobserved confounding cannot be ruled out (as in model 6 of Table 2), sensitivity analysis should be carried out to quantify the robustness of the results. This can be done in multiple ways. The first is by using the *impact* of the unobserved confounder Z, which Frank (2000) defines as the product of the two path coefficients $R_{Y\sim Z|X}R_{T\sim Z|X}$. Based on this definition, we can calculate a threshold (the so-called *impact threshold for a confounding variable*, ITCV) at which the impact of Z would be large enough to turn the focal estimate statistically indistinguishable from zero. Computing the impact threshold requires information on the estimated coefficient, its standard error, and the degrees of freedom, which all can be obtained from published studies (Rosenberg, Xu, Frank, 2018; Busenbark, Yoon, Gamache, and Withers, 2022). The second approach is Oster (2019), which extends Frank's work and provides a formal identification result that allows assessing the magnitude of the omitted variable bias under realistic scenarios. This approach involves benchmarking the strength of Z with that of the observed covariates $\mathbf{X}$, which serves as a useful basis for comparison. The third approach is Cinelli and Hazlett (2020), which improves on both previous procedures. Compared to Frank (2000), their proposed method allows for testing robustness to arbitrary non-zero null hypotheses (not just $H_0: \tau = 0$). Compared to Oster (2019), their approach produces results that have a more straightforward interpretation, making it easy for applied researchers to use their substance knowledge when judging the robustness of their causal conclusions. The full paper provides an empirical demonstration of these approaches.

CONCLUSION

Our framework makes it clear that the debate on whether to include fewer or more variables is not a productive one. Causal diagrams shift the discussion towards a transparent framework that arrives at an efficient and even potentially automatable solution for the causal identification problem (Textor and Liśkiewicz, 2011). Not only do causal diagrams provide a framework for finding a suitable set of controls that will achieve causal identification for the causal effect of interest, but they also point out whether an additional control (if for instance it was used in previous literature or recommended by a reviewer in the peer-review process) should be included. If the control variable would open a backdoor path, the control variable should be avoided, and if it does not open a backdoor path and causal identification is maintained, the control variable would be irrelevant.

If all control variables are observed, then causal identification is possible through selection-on-observables. If there is no suitable data in which all control variables are observed, alternative solutions would be to employ instrumental variables, fixed-effects models (if unobserved confounders are time-constant), or to resort to experimental or quasi-experimental methods.

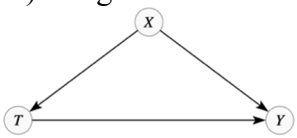
REFERENCES AVAILABLE FROM THE AUTHORS

10.5465/AMBPP.2022.294

Table 1. Taxonomy of controls

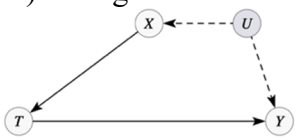
GOOD CONTROLS

1.) Exogenous



- X create a spurious correlation on the path $T \leftarrow X \rightarrow Y$, which can be blocked by controlling for X.
- X is not causally affected by any other variable and thus exogenous.

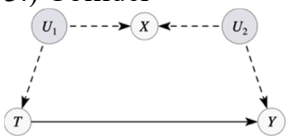
2.) Endogenous



- X is not a causal determinant of Y anymore, but only correlated with it due to the unobserved confounder U.
- The backdoor path $T \leftarrow X \leftarrow U \rightarrow Y$ leads to a spurious correlation, which can be blocked by controlling for X.

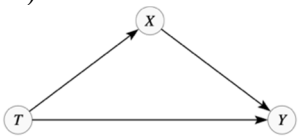
BAD CONTROLS

3.) Collider



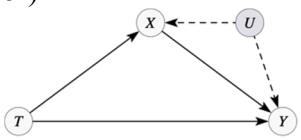
- X correlates with T and Y due to unobserved confounders.
- There is no spurious correlation because the backdoor path $T \leftarrow U_1 \rightarrow X \leftarrow U_2 \rightarrow Y$ is blocked by the collider X.
- Controlling for X unblocks the path and leads to collider bias.

4.) Mediator



- X causally affected by T.
- Controlling for X blocks the path $T \rightarrow X \rightarrow Y$.
- In linear models with constant effects, this allows identifying the direct effect of $T \rightarrow Y$ from the mediation effect.

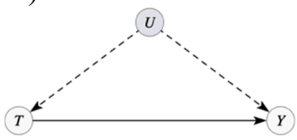
5.) Mediator as collider



- X is a collider on the path $T \rightarrow X \leftarrow U \rightarrow Y$.
- Controlling for X introduces collider bias.
- The direct effect of $T \rightarrow Y$ is not identifiable and disentangling causal mechanisms fail.

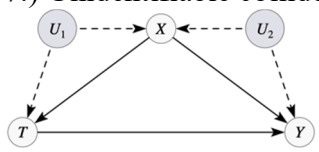
TRICKY CASES

6.) Unobservable confound



- U is an unobservable confounder that joint affects T and Y.
- The backdoor path $X \leftarrow U \rightarrow Y$ cannot be blocked.
- The causal effect of T on Y is not identifiable via covariate adjustment.

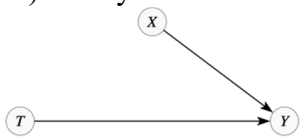
7.) Unidentifiable collider



- X is both a collider on the path $T \leftarrow U_1 \rightarrow X \leftarrow U_2 \rightarrow Y$ as well as a confounder on the path $T \leftarrow X \rightarrow Y$.
- Controlling for X reduces confounding but creates collider bias at the same time so the causal effect of T on Y remains unidentifiable.

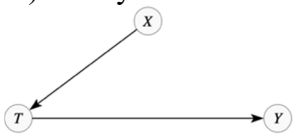
UNNECESSARY CONTROLS

8.) X only affects outcome



- X affects only outcome Y.
- Controlling for X is not necessary in a regression of Y on T.
- Controlling for X might reduce estimation error and result in higher precision.

9.) X only affects treatment



- X affects only treatment T.
- Controlling for X is therefore not necessary in a regression of Y on T.
- Controlling for X might increase estimation error and result in lower precision.