Structural Causal Models and the Specification of Time-Series-Cross-Section Models [*]

Adam N. Glynn[†] Kevin M. Quinn[‡]

March 13, 2013

Abstract

The structural causal models (SCM) of Pearl (1995, 2000, 2009) provide a graphical criterion for choosing the "right hand side" variables to include in a model. In this paper, we use SCMs to address the question of whether to include lagged variables in time-series-cross-section (TSCS) models. This question has received a great deal of attention from political methodologists, but unfortunately, the practical advice for applied researchers that comes out of this literature varies considerably from article to article. We attempt to clarify the nature of some of these disagreements and to provide useful tools to reason about the nonparametric identification of causal effects. After clarifying the debate between Beck and Katz (1996, 2011) and Achen (2000) and adding to the discussion by Keele and Kelly (2006), we provide concrete nonparametric identification results for commonly studied TSCS data generating processes. These results are also relevant for the choice of control variables in cross-section (CS) models. We conclude with some general thoughts on how a focus on using the SCM as a tool for proving identification results can help TSCS and CS researchers do better work.

[*] The authors thank Thomas Richardson for introducing them to literature of graphical causal models, and also thank Neal Beck, Andy Eggers, Jim Greiner, Ben Hansen, Gary King, Judea Pearl, Anton Westveld, two anonymous referees, and the participants of 2007 Summer meetings of the Society of Political Methodology for their helpful comments and suggestions. Brandon Stewart provided excellent research assistance. In addition, Quinn thanks the National Science Foundation (grants SES 03-50613 and BCS 05-27513) and the Center for Advanced Study in the Behavioral Sciences for its hospitality and support.

[†] Department of Government and The Institute for Quantitative Social Sciences Harvard University, 1737 Cambridge Street, Cambridge, MA 02138. aglynn@iq.harvard.edu

[‡] UC Berkeley School of Law, 490 Simon Hall, UC Berkeley, Berkeley, CA 94720-7200. kquinn@law.berkeley.edu

1 Introduction

The question of whether to include lagged variables in time-series-cross-section (TSCS) models has received a great deal of attention from political methodologists. Much of this literature examines the question of whether a lagged dependent variable should be included on the right-hand-side of a regression model (Beck and Katz, 1995, 1996, 2011; Achen, 2000; Keele and Kelly, 2006). Less attention has been paid to the question of whether lagged values of independent variables should be included in a regression specification (either with or without lagged values of the dependent variable) although this has been investigated in particular circumstances (Beck and Katz, 1996, 2011).

The practical advice for applied researchers that comes out of this literature varies considerably from article to article. For instance, Beck and Katz (1996, 2011) argue that including a lagged dependent variable is oftentimes an advisable approach to modeling dynamics in TSCS data. On the other hand, Achen (2000) argues that, "...in circumstances often encountered in practice [...] [t]he lagged dependent variable does bias the substantive coefficients toward negligible values and does artificially inflate the effect of the lagged dependent variable" (p. 4). Beck and Katz (2011) respond, making the case that "there is nothing pernicious in using a lagged dependent variable [...] all dynamic models either implicitly or explicitly have such a variable" (p. 331) and

> [T]here is nothing atheoretical about the use of a lagged dependent variable, and there is nothing that should lead anyone to think the use of a lagged dependent variable causes incorrect harm. It may cause "correct" harm, in that it may keep us from incorrectly concluding that x has a big effect when it does not, but that cannot be a bad thing. (p. 336)

Keele and Kelly (2006) provide Monte Carlo evidence regarding the magnitude of biases that exist in certain scenarios. They ultimately take a position somewhere between Beck and Katz (1996, 2011) and Achen (2000), concluding "that while the lagged dependent variable is inappropriate in some circumstances, it remains an appropriate model for the dynamic theories often tested by applied analysts." (Keele and Kelly, 2006, p. 186).

Despite these differences, there do seem to be some important points of commonality among these authors. First, the authors cited above (and nearly all other researchers working in this area) are clearly interested in the estimation of causal effects. Second, the authors above all make the case, to various degrees and in different ways, that a theoretical understanding of the substantive problem, particularly regarding the nature of any temporal dependencies, is an important part of a successful estimation strategy. Third, the authors working in this area make their arguments using the econometric tools of linear (time-series) regression.

We agree that causal effect estimation is an important goal of most TSCS work. We also agree that a theoretical understanding of the problem—particularly regarding the nature of any dependencies—will be an important component of most, if not all, successful estimation strategies. A focus on causal effect estimation and a theoretical understanding of dependencies will be explicit in what follows. However, as we show in this paper, theoretical understanding of the problem can be more easily brought to bear with a different set of tools.

In this paper, rather than relying solely on the familiar econometric tools of linear (time-series) regression, we will make use of tools developed by Pearl (1995, 2000) which are closely related to ideas of Robins (1986); Spirtes et al. (1993). Using the terminology that Pearl (2009) now prefers, we refer to these tools under the heading of the *structural causal model* or SCM for short. We see three major advantages of this addition to the toolkit:

- Once understood, the SCM provides a powerful and easy-to-apply graphical method for researchers to prove that a particular model specification identifies a causal effect of interest.
- Results derived using the SCM do not require the specific distributional or functional form assumptions that are common in the literature on TSCS data (multivariate normality, linear relationships, constant effects, etc.).
- Using the SCM to study identification in TSCS models clarifies the debate between Achen (2000) and Beck and Katz (1996, 2011) and highlights the importance of factors that have not previously received much attention in that debate.

Adopting the SCM allows researchers to receive these benefits without abandoning sound conceptualizations of causal inference they already hold— the SCM is consistent with the Neyman-Rubin causal model (Rubin, 1978; Rosenbaum and Rubin, 1983; Holland, 1986) as well as classic work on linear structural equations and causality, e.g. Haavelmo (1943); Koopmans (1949); Marschak (1950); Simon (1953).

The goal of this paper is to provide a framework based on the SCM for researchers to reason about causal effect identification in the analysis of TSCS data. In doing this, we attempt to clarify the debate between Achen (2000) and Beck and Katz (1996, 2011) regarding the propriety of including lagged dependent variables in TSCS regression specifications. Because the machinery of the SCM is unfamiliar to many political scientists, we spend a considerable portion of this paper presenting the necessary results with unified and consistent notation. While these results are not new, we hope to make the related methods more accessible to TSCS researchers by collecting the necessary results in one place. We also note that this paper is purely about the identification of causal effects. It does not directly address issues of estimation or inference. Finally, it is important to note that SCMs are relevant beyond the TSCS debate discussed in this paper. SCMs may aid in the choice of control variables for even purely cross sectional analyses.

This article is organized as follows. Section 2 addresses the question of what TSCS specifications identify causal effects of interest. Particular attention is paid to the question of when lagged values of dependent and independent variables should be included in TSCS regression specifications. Section 3 describes the SCM in more detail. The goal of this section is to familiarize the reader with the SCM—partly by comparing it to the more familiar (to political scientists) Neyman-Rubin model—so that the reader can use the SCM in practical settings. Section 4 concludes.

2 What Variables Should be Included in Time-Series Cross-Section Models?

The SCM of Pearl (1995, 2000) is a nonparametric generalization of a standard TSCS model, so we can use the graphical identification results from the Pearl (1995, 2000) to clarify the discussion of whether to include lagged dependent variables in the models of (Beck and Katz, 1995, 1996; Achen, 2000; Keele and Kelly, 2006; Beck and Katz, 2011). The basic structure of the problem can be described by adding cross-sectional subscripts to Equations 10-12 of the Keele and Kelly (2006) summary of the (Beck and Katz, 1995, 1996; Achen, 2000) debate:

$$y_{i,t} = \alpha y_{i,t-1} + \beta x_{i,t} + u_{i,t} \tag{1}$$

$$x_{i,t} = \rho x_{i,t-1} + e_{1,i,t} \tag{2}$$

$$u_{i,t} = \phi u_{i,t-1} + e_{2,i,t}; \quad i = 1, ..., N; \quad t = 1, ..., T, \tag{3}$$

where $y_{i,t}$ is the outcome variable of interest, $x_{i,t}$ is the causal variable of interest, $u_{i,t}$ represents unobserved "errors," and all variables have been standardized in order to avoid the need for intercepts. We have added the i subscripts in order to allow for the possibility of $N \geq 1$ units.[1] In practice, $y_{i,t}$ will be regressed on $x_{i,t}$ and perhaps lagged y and x variables in an attempt to estimate the causal parameter β and perhaps other parameters in the model.

We refer to the lagged right-hand-side variables (e.g. $y_{i,t-1}$ and $x_{i,t-1}$) as the *conditioning set*, because we focus on the estimation of β, and these other variables will be included or excluded from the model based solely on whether they improve estimation of β. The analyst can choose between at least the following four conditioning sets: the empty set $\{\}$, a lagged dependent variable $\{y_{i,t-1}\}$, a lagged explanatory variable $\{x_{i,t-1}\}$, and a lagged dependent and explanatory variable $\{y_{i,t-1}, x_{i,t-1}\}$. For example, if the analyst chooses the empty conditioning set, then $y_{i,t}$ will be regressed on only $x_{i,t}$. In practice, the analyst could choose sets that include lags of more than one period, but to further simplify the discussion, we will assume that one period lags are sufficient and that therefore these four sets are the only possibilities. Which of these conditioning sets will be appropriate depends on the parameters in all three equations (1),(2),(3).

2.1 The backdoor criterion for additive linear models

Pearl (1995) showed that general sufficient conditions for causal identification can be derived in terms of causal structure– often represented by a Directed Acyclic Graph (DAG). See Section 3 for

[1] Note that in some applications, the causal variable measured at the same time as the outcome variable may be excluded in favor of a causal variable that is lagged in time (in order to ensure the causal order of the variables). For our purposes, this is consistent with (1) because t need not reflect time exactly as long as it preserves the causal order of the variables.

a formal presentation. For the linear additive models in (1),(2),(3), the causal structure depends on which parameters are possibly non-zero. The discussion of these models often begins with the assumption that $\alpha = 0$ while all other parameters are assumed to be possibly non-zero.

With $\alpha = 0$ the causal structure can be represented by the graph in Figure 1 (a). With $\alpha \neq 0$ the causal structure can be represented by the graph in Figure 1 (b). In these graphs the directed edges (i.e., arrows) encode the direction and potential existence of the effects in (1),(2),(3). Closed circles represent observed variables, open circles represent unobserved variables, solid arrows represent effects between observed variables, and dashed arrows represent effects involving an unobserved variable. Note that missing arrows represent the absence of an effect, but existing arrows do not necessarily represent the presence of an effect.

Pearl (1995) demonstrated that when the causal structure was represented by a Directed Acyclic Graph (DAG), such that there are no undirected edges (all edges have an arrow on at least one end) and no cycles (there is no way to follow the direction of arrows and end up at the same point), then whether a conditioning set would identify the causal effect of can be determined from the DAG using the backdoor criterion.

Definition 1 (Back-Door Criterion for Additive Linear Models) *Given an SCM and associated DAG, conditioning on the set of covariates* $\mathbf{Z}$ *identifies the effect of causal variable* X *on outcome* Y *if:*

1. $\mathbf{Z}$ *does not contain any variables on directed paths from* X *to (or through)* Y
2. $\mathbf{Z}$ *blocks all paths between* X *to* Y *with arrows pointing into* X *(i.e., backdoor paths).*

Formal definitions of paths, directed paths and "blocking" are presented in Section 3, but intuitive definitions will suffice for the current discussion. Roughly speaking, directed paths are paths of edges with all arrows pointing in the same direction along the path. For example, in Figure 1 (b) the path $\epsilon_{t-1} \rightarrow \epsilon_t \rightarrow y_t$ is a directed path from ϵ_{t-1} to y_t, and the path $x_t \rightarrow y_t$ is a directed path from x_t to y_t, while the path $x_t \leftarrow x_{t-1} \rightarrow y_{t-1} \rightarrow y_t$ is not a directed path from x_t to y_t.

A path is blocked by a conditioning set $\mathbf{Z}$ when either 1) it contains a chain structure $a \rightarrow b \rightarrow c$ or a fork structure $a \leftarrow b \rightarrow c$ with b in the set Z, or 2) it contains a collider structure $a \rightarrow b \leftarrow c$

where b is not in the set Z nor is any descendent of b in Z (descendent of b would be a variable on a directed path out of b). This blocking condition, known as d-separation, is presented formally in Section 3

It is important to note that "identification" in this criterion corresponds to having the graph associated with a single unit, and having a large number of independent units. Therefore, the criterion will not hold directly for our application unless $T = 2$ and $N \to \infty$. We will address this in the following subsections.

2.2 The Model Without a Causal Effect for the Lagged Dependent Variable

The first model considered in Beck and Katz (1996), Achen (2000), and Keele and Kelly (2006) is the model without a causal effect for the lagged dependent variable (i.e., $\alpha = 0$). With a lengthy derivation, Achen (2000) showed that for this model, the empty conditioning set (i.e., y_t regressed only on x_t) provides a consistent estimator of β, while the conditioning set containing only the lagged dependent variable (i.e., y_t regressed on x_t and y_{t-1}) provides an inconsistent estimator of β. From the standpoint of introductory econometrics, this is a surprising result because $y_{i,t-1}$ appears to be an irrelevant variable if $\alpha = 0$ in (1), and therefore we might expect that including this irrelevant variable would not result in an inconsistent estimator.

Using the backdoor criterion we can arrive at a similar result by simply inspecting Figure 1 (a). With the empty conditioning set ($\{\}$), there is clearly no conditioning variable on a directed path from X, so the first element of the backdoor criterion is satisfied. Furthermore, there is only one backdoor path between x_t and y_t ($x_t \leftarrow x_{t-1} \rightarrow y_{t-1} \leftarrow \epsilon_{t-1} \rightarrow \epsilon_t \rightarrow y_t$), and with the empty conditioning set this path is blocked by the collider structure $\rightarrow y_{t-1} \leftarrow$. Therefore, regressing y_t on x_t identifies β. In contrast, when the conditioning set is the lagged dependent variable ($\{y_{t-1}\}$), the backdoor path is unblocked because we have conditioned on the collider structure $\rightarrow y_{t-1} \leftarrow$ and we have not conditioned on any of the chain or fork structures on this path. Therefore, regressing y_t on x_t and y_{t-1} will not generally identify β. The backdoor criterion also allows us to quickly deduce alternative identifying conditioning sets. In addition to the empty set, the conditioning sets ($\{y_{t-1}, x_{t-1}\}$) and ($\{x_{t-1}\}$) will also identify β because these block the backdoor path at the $\leftarrow x_{t-1} \rightarrow$ fork structure.

It is also important to note that the graphical approach makes explicit the implied causal direction of the equal signs in (1),(2),(3) and therefore the lack of direct effect from $x_{i,t-1}$ to $y_{i,t}$ ($x_{i,t-1}$ only affects $y_{i,t}$ through $x_{i,t}$) and the lack of any effect from $y_{i,t-1}$ to $y_{i,t}$ (there is no directed path from $y_{i,t-1}$ to $y_{i,t}$). If one ignores the implied causal direction from these equations, it is straightforward to manipulate them as in Beck and Katz (1996), pg. 8, Equation 13, and you can obtain the following:

$$y_{i,t} = \phi y_{i,t-1} + x_{i,t}\beta - x_{i,t-1}(\phi\beta) + e_{2,i,t}. \tag{4}$$

Notice that while (4) can be useful for the discussion of estimation (it implies that the $\{y_{t-1}, x_{t-1}\}$ conditioning set is sufficient), it does not explicitly specify that β is meant to represent the causal effect of $x_{i,t}$ on $y_{i,t}$, but that ϕ and $(\phi\beta)$ are not meant to specify causal effects of $y_{i,t-1}$ and $x_{i,t-1}$ on $y_{i,t}$. This lack of clarity as to which parameters correspond to causal effects and which are purely associational parameters can create needless confusion among both methodologists and practitioners—with potentially serious consequences for research.

Indeed, Achen (2000) seems to run afoul of this issue in one of his examples when he argues that because a lagged dependent variable (lagged social welfare expenditures) "has no obvious causal power" it should not be included on the right-hand-side of the regression specification. He goes on to refer to this lagged dependent variable as a "nonsensical variable" (p. 2). As we have shown above, one of the key points to take away from the SCM analysis of this problem is that simply knowing that $y_{i,t-1}$ is not a cause of $y_{i,t}$ does not allow one to conclude that all of the regression specifications that identify the causal effect of $x_{i,t}$ on $y_{i,t}$ exclude $y_{i,t-1}$ from the right-hand-side. In some instances conditioning on values of a lagged dependent variable that exerts no (direct or indirect) causal effect on the outcome variable can be part of a conditioning set that identifies the causal effect of $x_{i,t}$ on $y_{i,t}$. This is important because as shown in (Beck and Katz, 1995, 1996, 2011) there are advantages to including $y_{i,t-1}$ in the specification. A related point that we do not make explicitly with the examples in this section, but that is nonetheless true, is that simply knowing that $y_{i,t-1}$ *is* a cause of $y_{i,t}$ does not allow one to conclude that all of the regression specifications that identify the causal effect of $x_{i,t}$ on $y_{i,t}$ *include* $y_{i,t-1}$ on the right-hand-side.

The appropriateness of the conditioning set that includes a lagged independent variable as well as a lagged dependent variable was discussed in Beck and Katz (1996) (as well as Beck and Katz (2011)) and it is apparent from (4). However, it was also recommended that this solution be used when a Lagrange multiplier test was failed, and the $\{y_{i,t-1}, x_{i,t-1}\}$ set has seen limited use in practice.[2] Furthermore, the $\{y_{i,t-1}, x_{i,t-1}\}$ set was not considered in Achen (2000) or in the Monte Carlo analysis of Keele and Kelly (2006). Further, and to the best of our knowledge, the fact that conditioning on just $x_{i,t-1}$ (along with $x_{i,t}$) identifies the effect of $x_{i,t}$ on $y_{i,t}$ has not been previously recognized in the TSCS literature.

As noted above, the results for the backdoor criterion are derived in terms of independent units ($i = 1, ..., N$), and the implications of these results are not straightforward if one attempts to increase the effective sample size with dependent units such as repeated measurements over time ($t = 1, ..., T$). For example, if one treats each unit (i) as a single observation, perhaps focusing the analysis on the final time period (Y_T) or an average over the final five time periods ($\frac{1}{5}\sum_{t=T-4}^{T} Y_t$), then the graphical identification results are directly applicable. If one attempts to increase the sample size with repeated measurements over time, then as always with dependent observations, some form of additional time-series conditions (e.g., stationarity and ergodicity) will need to be assumed.

In order to explore this, the upper left panels of Figures 2 and 3 provide an extension of the

[2] A search for articles in the *American Political Science Review* from 1995 to 2009 that explicitly discuss lagged dependent or independent variables in a regression specification yielded 88 articles. Of these 88 articles, 49 include $y_{i,t-1}$ on the right-hand-side of at least one regression model, 39 do not include $y_{i,t-1}$ right-hand-side of at least one regression model, and only 11 include $x_{i,t}$, $x_{i,t-1}$ and $y_{i,t-1}$ on the right-hand-side of at least one regression model. Of the articles that include $x_{i,t}$ and $y_{i,t-1}$ but not $x_{i,t-1}$ on the right-hand-side, only one used a Lagrange multiplier test to justify its choice.

Keele and Kelly (2006) Monte Carlo analysis with the $\{x_{i,t-1}\}$ conditioning set (LIV) and the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDVLIV) considered along with the $\{\}$ conditioning set (No Lag), and the $\{y_{i,t-1}\}$ conditioning set (LDV). As in the Keele and Kelly (2006) analysis from Table 1 of that paper, $\alpha = 0$, $N = 1$, $T = 100$, $\beta = .5$, $\phi = .75$, and $\rho = .95$. From the upper left panel of Figure 2 we see that even in the pure time-series analysis ($N = 1$) when $\alpha = 0$ the No Lag, LIV, and the LDVLIV specifications have a bias of zero, while the LDV set has substantial bias. From the upper left panel of Figure 3 we additionally see that the LDVLIV specification provides a slight benefit over the No Lag and LIV specifications in terms of root mean squared error.

2.3 The Model with Serial Correlation and Causal Lagged Dependent Variables

If the lagged dependent variable has a causal effect on the outcome ($\alpha \neq 0$), then the situation is more complicated (Beck and Katz, 1996; Achen, 2000; Beck and Katz, 2011). Figure 1 (b) presents the structural graphical model consistent with one period lags.

Even in this more complicated situation the graphical approach clarifies a number of points. It is often stated that OLS cannot be used to consistently estimate the parameters of this model (Achen, 2000; Beck and Katz, 2011). While it is true that an OLS regression of $y_{i,t}$ on $y_{i,t-1}$ and $x_{i,t}$ (the two observed direct causes of $y_{i,t}$) will not consistently estimate either of the two associated causal effects, this does not mean that one cannot identify the causal effect of primary interest (the effect of $x_{i,t}$ on $y_{i,t}$) with a different model specification. Indeed, Figure 1 (b) shows that the effect of $x_{i,t}$ on $y_{i,t}$ identifiable. Under appropriate parametric assumptions, and when each unit is used as a single observation, OLS can provide unbiased estimates of this effect.

To see this, consider Figure 1 (b). Conditioning on $y_{i,t-1}$ (as we would do in a regression of $y_{i,t}$ on $y_{i,t-1}$ and $x_{i,t}$) leaves open a backdoor path from $x_{i,t}$ to $y_{i,t}$ ($x_{i,t} \leftarrow x_{i,t-1} \rightarrow y_{i,t-1} \leftarrow \epsilon_{i,t-1} \rightarrow \epsilon_{i,t} \rightarrow y_{i,t}$) because again, conditioning on the collider at $y_{i,t-1}$ opens up what would otherwise be a closed path. This conditioning set thus does not identify the effect of $x_{i,t}$ on $y_{i,t}$. It is also the case that this conditioning set leaves open a backdoor path from $y_{i,t-1}$ to $y_{i,t}$ ($y_{i,t-1} \leftarrow \epsilon_{i,t-1} \rightarrow \epsilon_{i,t} \rightarrow y_{i,t}$). It follows that this conditioning set does not identify the effect of $y_{i,t-1}$ on $y_{i,t}$. Further, note that there is no feasible conditioning set that will block the backdoor path from $y_{i,t-1}$ to $y_{i,t}$. Consequently, the effect of $y_{i,t-1}$ on $y_{i,t}$ will never be identified via the backdoor criterion and we will not be able to consistently estimate this effect using some form of standard regression adjustment.[3]

However, the conditioning sets $\{x_{i,t-1}\}$ and $\{y_{i,t-1}, x_{i,t-1}\}$ block all backdoor paths from $x_{i,t}$ to $y_{i,t}$. Therefore, β is identified by these conditioning sets. Under the appropriate parametric assumptions and independent observations, OLS can be used to estimate this effect. Put more concretely, an OLS regression of $y_{i,t}$ on $x_{i,t}$ and $x_{i,t-1}$ or on $y_{i,t-1}$, $x_{i,t}$, and $x_{i,t-1}$ will consistently estimate the effect of $x_{i,t}$ on $y_{i,t}$ if the constant linear effects assumption underlying the linear model holds and if we only use each unit as a single observation (e.g., focusing the analysis on the final time period (Y_T) or an average over the final five time periods ($\frac{1}{5}\sum_{t=T-4}^{T} Y_t$)). This is true even though this conditioning set does not allow us to identify the effects of the other right-hand-side variables on $y_{i,t}$. Hence, this provides a counterexample to the standard introductory presentation implying that a lack of identification in one regression parameter implies a lack of identification in all regression parameters.

We further examine the implications of these results for dependent data by again extending the Monte Carlo analysis of Keele and Kelly (2006). Figures 2, 3, 4, and 5 provide results with $\beta = .5$, $\phi = .75$, and $\rho = .95$ that examine the $\{\}$ conditioning set (No Lag), the $\{y_{i,t-1}\}$ conditioning set (LDV), the $\{x_{i,t-1}\}$ conditioning set (LIV), and the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDVLIV) with α set to $0, 0.1, 0.2, 0.5$. Again, Figures 2 and 3 show bias and root mean squared error for a time series analysis with $N = 1$ and $T = 100$, while Figures 4 and 5 show bias and root mean squared error for a TSCS analysis with $N = 15$ and $T = 20$. For both the time-series and time-series-cross-sectional data we see that when $\alpha > 0$ the estimates from both the LIV and LDVLIV specifications are essentially unbiased while the estimates from the No Lag and LDV specifications are noticeably biased. Turning to root mean squared error, we see that the LDVLIV specification performs the

[3] There are instrumental variables approaches to estimating this effect (Beck and Katz, 1996).

best across all values of α.

This analysis demonstrates the efficacy of the backdoor criterion and the graphical approach within additive linear TSCS models. However, the backdoor criterion is also applicable in nonlinear models with heterogenous effects. The next section presents the general result.

3 Frameworks for Reasoning About Causality

In this section we describe key features of the SCM of Pearl (2000). Because this material is likely to be unfamiliar to many readers we begin by sketching the Neyman-Rubin causal model which is likely more familiar to many readers. We then go on to present the SCM, paying particular attention to how aspects of the SCM have direct analogues in the Neyman-Rubin framework. We emphasize at the outset that our use of the SCM is purely to arrive at identification results.

3.1 The Neyman-Rubin Model

In the Neyman-Rubin model, causal effects are defined in terms of potential outcomes: $Y(x, u)$ (i.e. the potential outcome in unit u if X would have been set equal to x) (see Rubin (1978), Rosenbaum and Rubin (1983), Holland (1986)). Here u is thought of as a unit-specific index, and therefore captures any individual-specific effects. It is tempting to think of units as individuals, schools, etc., but in actuality it is more accurate to think of the units as individuals, schools, etc. under a particular set of exogenous background conditions. Thus an an individual at 9:00AM and the same individual at 10:00AM may very well be considered different units. If the value of X received by one unit does not affect the outcomes for other units, then given x and u, $Y(x, u)$ is completely determined. This assumption of non-interference is sometimes called SUTVA (see Angrist et al. (1996)), and we will utilize this assumption throughout this paper.

3.1.1 Unit-Specific Causal Effects

In the Neyman-Rubin model, the potential outcomes are used to define unit-specific causal effects. For simplicity in presentation, we assume that X can only take on the values zero and one.[4] Therefore, the unit-specific causal effect of $X = 1$ on Y relative to the effect of $X = 0$ in unit u is calculated by comparing $Y(1, u)$ to $Y(0, u)$. A common means of comparison is the difference::

$$Y(1, u) - Y(0, u).$$

The key idea is that if it were possible to observe $Y(1, u)$ *and* $Y(0, u)$ for the two levels of the treatment variable (e.g. active treatment and control), then we could observe the unit specific causal effect.

If we assume consistency (Robins, 1986) of the observed outcomes, then we may observe one of these two outcomes for each individual. This assumption requires that the observed outcome for each unit $Y(u)$ matches the potential outcome for unit u for the observed value of X. Formally, this can be written as the following:

$$\mathbf{X}(u) = x \implies \mathbf{Y}(u) = \mathbf{Y}(x, u).$$

and if this holds, then our binary treatment example, the observed Y can be written as the following:

$$Y^{obs}(u) = X(u) \cdot Y(1, u) + (1 - X(u)) \cdot Y(0, u)$$

Unfortunately, consistency does not allow the unit-specific causal effect to be directly observed since u only gets one of either $X = 0$ or $X = 1$ but never both. Holland (1986) calls this the fundamental problem of causal inference.

3.1.2 Population Causal Effects

Given the impossibility of observing individual causal effects, inference is usually confined to the characteristics of populations (sometimes the observed sample of individuals is taken as the entire population). For simplicity, we assume throughout this paper that the parameter of interest is the *average causal effect* from $X = 0$ to $X = 1$. This is defined as

$$ACE \equiv \mathbb{E}[Y(1) - Y(0)],$$

[4] The extension to polytomous or continuous treatment variables will not complicate the discussion of identification in this paper, but will complicate the choice of adjustment strategy.

where the expectation merely represents an average over the units (or the distribution of pertinent background factors) in the population of interest. This parameter has a number of useful properties including the usual decomposition of the expectation of sums which allows us to separately consider the average potential outcomes under treatment and control.

$$ACE \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

Unfortunately, these averages are not observed in general. Instead we observe averages of potential outcomes over the subpopulations that actually received treatment and control. Hence we can identify the potentially similar parameter that Holland (1986) calls the prima facie average causal effect (the second line is due to consistency):

$$\begin{aligned} ACE^{pf} &\equiv \mathbb{E}[Y(1)|X=1] - \mathbb{E}[Y(0)|X=0] \\ &= \mathbb{E}[Y|X=1] - \mathbb{E}[Y|X=0] \end{aligned}$$

3.1.3 Ignorability and the Identification of Population Causal Effects

The Neyman-Rubin model makes clear that the following will not hold in general,

$$\mathbb{E}[Y(0)] = \mathbb{E}[Y(0)|X=0] \tag{5}$$

$$\mathbb{E}[Y(1)] = \mathbb{E}[Y(1)|X=1], \tag{6}$$

because averages over subpopulations need not match averages over the population. However, it is sufficient to assume the equalities in (5) and (6) in order to identify the ACE. This assumption, sometimes known as mean ignorability[5], is usually hard to justify because the subpopulation that receives treatment is often quite different from the subpopulation that receives control. Random treatment assignment for a large population is an example where the subpopulations will be similar.

It is often possible to "weaken" ignorability assumptions by conditioning on a set of background

[5] Although we focus on identification in this paper, there are other inferential goals, and hence it is often necessary to make stronger ignorability assumptions. Rosenbaum and Rubin (1983) describes sufficient ignorability assumptions for a variety of inferential tasks.

variables which we will denote as $\mathbf{Z}$. Hence, even if (5) and (6) do not hold, we may believe that,

$$\mathbb{E}[Y(0)|\mathbf{Z}] \;=\; \mathbb{E}[Y(0)|X=0,\mathbf{Z}] \tag{7}$$

$$\mathbb{E}[Y(1)|\mathbf{Z}] \;=\; \mathbb{E}[Y(1)|X=1,\mathbf{Z}], \tag{8}$$

hold for some set (or sets) of $\mathbf{Z}$. The equalities in (5) and (6) allow the identification of average causal effects within the strata defined by $\mathbf{Z}$, and these can then be combined through a weighted average to identify the overall ACE. When $\mathbf{Z}$ lives in a high dimensional space, this averaging can present considerable practical difficulty, so in order to confine the discussion to the issues considered in this paper, we assume throughout that $\mathbf{Z}$ is discrete and has low dimension or that the joint distribution of all variables has a simple parametric form.

3.2 The Structural Causal Model of Pearl

The structural causal model (Pearl, 1995, 2000) and its close relatives (Spirtes et al., 1993; Robins, 1986) provide additional structure to the Neyman-Rubin model. In what follows, we adapt Pearl's presentation to the case considered in this paper (a single intervention variable, a single outcome variable, and no interference).

3.2.1 A Unit-Level Causal Model

Adapting the definition from Pearl (2000), we define a unit-specific causal model for a single outcome and intervention variable to be the triple:

Definition 2 (Unit-Level Causal Model)

$$M = \langle \mathbf{U}, \mathbf{V} \equiv \{Y, X, \mathbf{Z}\}, \mathbf{h} \rangle$$

where:

1. $\mathbf{U}$ *is a set of* exogenous *background variables.*
2. $\mathbf{V} \equiv \{Y, X, \mathbf{Z}\}$ *is a set of* endogenous *variables.* Y *is the outcome variable,* X *is the intervention variable, and* $\mathbf{Z}$ *is a set of potential control variables (some possibly unobserved),*
3. $\mathbf{h}$ *is a set of functions that defines the endogenous variables (one for each endogenous variable).*

Furthermore, we will utilize the following conventions/assumptions/clarifications (some portions of these statements are redundant, but we include them all in order to provide intuition):

1. The set of exogenous variables is rich enough to define a unit. Therefore, we can think of the unit index u from Section 3.1 as a function of the realized vector of exogenous variables $\mathbf{U} = \mathbf{u}$.
2. We assume a causal order to all endogenous variables such that an endogenous variable may not be an input to the functions of any of the endogenous inputs to its function. This type of causal model is sometimes known as recursive.
3. We assume that given the values of the exogenous variables ($\mathbf{U} = \mathbf{u}$), the endogenous variables are uniquely determined by the functions $\mathbf{h}$.

Given the definition and assumptions, consider the following simple example:

$$Z \leftarrow h_Z(u_1)$$

$$X \leftarrow h_X(z, u_2)$$

$$Y \leftarrow h_Y(x, z, u_3)$$

Z is a *deterministic* function of u_1, X is *deterministic* function of z and u_2, and Y is a *deterministic* function of x, z, and u_3. The assignment notation ($\leftarrow$) above is to make clear that the assignment in these functions is asymmetric, and we label the entire model M.

The model M is non-parametric in that no assumptions are made about h_Z, h_X, h_Y, U_1, U_2 and U_3. Given the causal order assumption, the endogenous variables can be written as a function of the exogenous variables. For example, in our simple model, y can be written as a function of all other variables and functions:

$$Y \leftarrow h_Y(h_X(h_Z(u_1), u_2), h_Z(u_1), u_3).$$

Therefore, $Y(\mathbf{u})$ denotes the unique value of Y generated by model M given $\mathbf{U} = \mathbf{u}$, and is analogous to $Y^{obs}(u)$ from the Neyman-Rubin model.

Now consider intervening in the system to set variable X equal to a particular value x without *directly* disturbing any of the other variables in the system. This involves creating a submodel in which the function for X is removed and X becomes an exogenous variable.

Definition 3 (Unit-Level Submodel) *Let M be a unit-level causal model, X be the intervention variable in the set of endogenous variables $\mathbf{V}$ and x be a particular realization of $\mathbf{X}$. A* unit-level submodel *M_x of M is the unit-level causal model*

$$M_x = \langle \mathbf{U}, \mathbf{V}, \mathbf{h}_x \rangle$$

where

$$\mathbf{h}_x = \mathbf{h}/h_X$$

For example, consider the simple example again with X set to zero.

$$Z \leftarrow h_Z(u_1)$$

$$X \leftarrow 0$$

$$Y \leftarrow h_Y(0, z, u_3)$$

We call this system of equations submodel M_x.

Within a submodel, we can define potential outcomes that are analogous to the potential outcomes from the Neyman-Rubin model by solving for the functional output of Y under the submodel.

Definition 4 (Potential Outcome) *Let M denote a unit-level causal model, and set the intervention variable X equal to value x, then the* potential outcome $Y(x, \mathbf{u})$ *denotes the unique solution for Y as determined by x, the exogenous variables in the model and the functions $\mathbf{h}_x$.*

If u were observed, the model M and submodel M_x would define unit-specific causal effects analogous to unit-specific causal effects in the Neyman-Rubin model. Consider the following simple non-parametric example:

Suppose that a "get out the vote" (GOTV) study randomly assigns a small and geographically separated group of registered voters[6] to either treatment (phone call) or control (no phone call), and after the election voters are observed to have voted or not voted. Unbeknownst to the study designers, some of the registered voters will be so put off by a phone call that they will not vote (even if they would have voted without the phone call). If we let $X = \{0 \text{ (no call)}, 1 \text{ (call)}\}$ be treatment assignment, $Y = \{0 \text{ (no vote)}, 1 \text{ (vote)}\}$ be voting status, U_1 be the exogenous treatment randomization mechanism, and U_2 be the exogenous variable that describes each registered voter's potential response to treatment, then this scenario can be conceptualized within the SCM

[6] The registered voters are selected for the study so that the assumption of no interference is plausible.

framework with the following set of functional assignments:

$$x \leftarrow h_X(u_1)$$

$$y \leftarrow h_Y(x, u_2)$$

In the parlance of the Neyman-Rubin model, h_X represents the treatment assignment mechanism. Furthermore, with binary treatment and binary outcome, the domain of U_2 (and hence registered voters) can be partitioned into four potential outcome equivalence classes: those who would vote regardless of whether they received treatment or control (Always Vote: $Y(0, \mathbf{u}) = 1, Y(1, \mathbf{u}) = 1$), those who would not vote regardless of whether they received treatment or control (Never Vote: $Y(0, \mathbf{u}) = 0, Y(1, \mathbf{u}) = 0$), those who are treatable and would vote with a phone call and would not vote without (Encouraged by Call): $Y(0, \mathbf{u}) = 0, Y(1, \mathbf{u}) = 1$), and those with the potential to be put off by the phone call so that they would not vote with the call, and would vote without the call (Discouraged by Call): $Y(0, \mathbf{u}) = 1, Y(1, \mathbf{u}) = 0$). Therefore, given the values of the exogenous variables, the observed variables and potential outcomes are determined. Furthermore, unit specific causal effects are defined by the potential outcome equivalence classes. The unit-specific effect is zero for the "Always Vote" and "Never Vote" units, one for the "Encouraged" units, and negative one for the "Discouraged units". This example demonstrates the fully nonparametric nature of the model (no additivity, monotonicity, or functional form assumptions).

In this example as in most cases, $\mathbf{u}$ is not observed, and hence we do not observe the unit-specific causal effects. The solution, as in the Neyman-Rubin framework, is to shift inferential focus to population causal effects.

3.2.2 A Population Causal Model

We can create a population causal model from the deterministic causal model of the previous subsection by assuming a distribution over U.

Definition 5 (A Population Level Causal Model) *A population-level model is a pair*

$$\langle M, F_{\mathbf{U}}(\mathbf{u}) \rangle$$

where M is the unit-level causal model and $F_{\mathbf{U}}(\mathbf{u})$ is a cumulative distribution function defined over the domain of U.

Again, we note the following conventions/assumptions/clarifications:

1. We configure the exogenous variables to be independent of each other, so the distribution $F_{\mathbf{U}}(\mathbf{u})$ factors accordingly. This can be accomplished by combining dependent exogenous variables into a single exogenous variable.
2. Some authors include as endogenous all variables that are inputs to more than one function (i.e. common cause variables). We do not use this convention.
3. The distribution over the endogenous variables is uniquely determined by $F_{\mathbf{U}}(\mathbf{u})$ and $\mathbf{h}$.

The model M along with an assumption as to the distribution of $\mathbf{U}$ generates the "pre-intervention" distribution $F_{\mathbf{U},\mathbf{V}}(\mathbf{U}, \mathbf{V})$, and given the assumption of a recursive causal model, this joint distribution is uniquely defined. *One can estimate the marginal distribution of the observed variables in* $\mathbf{V}$ *directly from observational data without making untestable assumptions.* For example, the "pre-intervention" outcome distribution can be derived by integrating over $F_{\mathbf{U},\mathbf{V}}(\mathbf{U}, \mathbf{V})$ and is written in the standard notation $F_Y(y)$.

Continuing the GOTV example from the previous section, the finite number of treatment assignments and potential outcome equivalence classes allows the interpretation of $F_{\mathbf{U}}(\mathbf{u})$ in terms of the population proportions of individuals. Hence, $F_{\mathbf{U}}(\mathbf{u})$ describes the proportions of (Always Vote, Never Vote, Encouraged by Call, Discouraged by Call) individuals in the population, and the proportions of (Treatment, Control) individuals in the population. Furthermore, these distributions in combination with the causal model define the proportions of (Vote, No Vote) individuals in the population.[7]

[7] Note that given our conventions, any dependencies between the exogenous $\mathbf{u}$ variables need to be modeled explicitly through the redefinition of the $\mathbf{U}$ variables. In this example, if we didn't have random treatment assignment, then we might believe that experienced GOTV workers could target their efforts at those "Encouraged" registered voters. Therefore, dependence between U_1 and U_2 could be accommodated by combining them into a single exogenous

Because $F_{\mathbf{U}}(\mathbf{u})$ remains unchanged for the submodel M_x, we can ask what the probability distribution of $Y(x, \mathbf{u})$ is for a $\mathbf{u}$ randomly drawn from the population distribution of $\mathbf{U}$. In other words, what is the distribution of Y in the population after the intervention on X. This quantity is denoted $F_{Y(x)}(y)$ and is called the post-intervention distribution of Y. *Post-intervention distributions are not directly estimable from observational data without untestable causal assumptions.* With probability distributions defined over post-intervention distributions, the expectations and average causal effects are written as

$$\mathbb{E}[Y(x)] \equiv \int y dF_{Y(x)},$$

and obviously

$$ACE \equiv \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$$

As in the Neyman-Rubin model, we would like to establish situations in which the observable pre-intervention distribution identifies averages over the unobserved post-intervention distribution. This task is simplified by representing SCMs as directed acyclic graphs.

3.2.3 A Graphical Criterion for Causal Identification

We begin with some basic terminology. A directed graph is defined as the following:

Definition 6 (Directed Graph) *A* directed graph *$\mathcal{G}$ is a pair $\langle \mathcal{V}, \mathcal{E} \rangle$ where $\mathcal{V}$ is a finite set of* vertices *(a.k.a. nodes) and $\mathcal{E}$ is the set of* directed edges *(a.k.a. directed arcs or directed links). Each directed edge in $\mathcal{E}$ is an ordered pair of distinct vertices from $\mathcal{V} \times \mathcal{V}$. A directed edge $(V_i, V_j) \in \mathcal{E}$ is also denoted $V_i \rightarrow V_j$.*

In this paper, we think of each $V \in \mathcal{V}$ as being a (possibly non-scalar) random variable and each directed edge $(V_i, V_j) \in \mathcal{E}$ as a causal relationship between V_i and V_j (to be made explicit later). Within this framework, a number of additional definitions will be useful.

variable:

$$\begin{aligned} \mathbf{U} &\equiv \{U_1, U_2\} \\ x &\leftarrow h_X(\mathbf{u}) \\ y &\leftarrow h_Y(x, \mathbf{u}). \end{aligned}$$

Definition 7 (Path) *A* path *from V_i to V_j in a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a sequence of distinct nodes $V_i = X_0, \ldots, X_n = V_j$ such that $(X_{k-1}, X_k) \in \mathcal{E}$ or $(X_k, X_{k-1}) \in \mathcal{E}$ for each $k = 1, \ldots, n$.*

Note that a path cannot visit the same node twice, and the direction of the edges does not matter. For instance, a path from V_1 to V_3 exists in each of the following four graphs.

$$V_1 \rightarrow \quad V_2 \quad \rightarrow V_3 \tag{9}$$

$$V_1 \leftarrow \quad V_2 \quad \rightarrow V_3 \tag{10}$$

$$V_1 \rightarrow \quad V_2 \quad \leftarrow V_3 \tag{11}$$

$$V_1 \leftarrow \quad V_2 \quad \leftarrow V_3 \tag{12}$$

While all of these relationships represent paths from V_1 to V_3, it is useful to make some distinctions between these types of paths and vertices.

Definition 8 (Directed Path) *A* directed path *from V_i to V_j in a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a sequence of distinct nodes $V_i = X_0, \ldots, X_n = V_j$ such that $(X_{k-1}, X_k) \in \mathcal{E}$ and $(X_{k-1}, X_k) \notin \mathcal{E}$ for each $k = 1, \ldots, n$. We write $V_i \rightsquigarrow V_j$ to denote a directed path from V_i to V_j.*

Informally, a *directed path* from V_1 to V_k requires that all edges point toward V_k along the path. For instance, (9) depicts a directed path from V_1 to V_3. We can distinguish further between the two-edge paths depicted in (9),(10),(11), and (12). In particular, we will call (9) and (12) chain structures, we will call (10) a fork structure, and we will call (11) a collider structure. Notice that this terminology will extend to situations where these structures appear as subpaths in longer paths. We will also find it useful to define some familial relations within directed graphs. The notions of children, and descendents can be defined as the following:

Definition 9 (Children) *In a directed graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ the set of* children *of a node $V \in \mathcal{V}$ is defined to be:*

$$ch(V) = \{Z \in \mathcal{V} : (V, Z) \in \mathcal{E}\}$$

Definition 10 (Descendents) *In a directed graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ the set of* descendents *of a node $V \in \mathcal{V}$ is defined to be:*

$$de(V) = \{Z \in \mathcal{V} : V \rightsquigarrow Z\}$$

Put more informally, the children of V are all nodes to which there is a directed edge from V, and the set of descendents of V consists of the vertices to which there exists a directed path from V. Analogous definitions can be constructed for parents and ancestors. Using these familial notions, we can now define the class of graphs that we will utilize throughout the rest of this paper. A directed graph that does not have cycles (i.e., no vertex in the graph is a descendent of itself) is said to be a *directed acyclic graph (DAG).*

DAGs are useful in causal modeling, because they form a compact representation of the assumptions implicit in recursive SCMs. Furthermore, a causal markov condition connects SCMs to graphical rules for deriving conditional independence relations between the pre-intervention and *post-intervention* random variables. Therefore, graphs provide a shorthand for deriving conditional ignorability conditions.

We use the edges in a DAG to represent the inputs to the functions of a corresponding SCM. The rules for forming a DAG G_M from a SCM M are the following:

1. Represent each unobserved variable with an open vertex.
2. Represent each observed variable with a closed vertex.
3. For each assignment operation in M draw an edge from each variable on the right-hand-side of the $\leftarrow$ operator to the variable on the left-hand-side using a solid line when both vertices are observed and a dashed line when one vertex is unobserved.

In the interest of graphical simplicity, many authors will delete nodes and edges that do not affect the results of graphical tests of interest. For example, exogenous variables that point into a single endogenous variable can often be removed.[8] However, the removal of exogenous variables from the graph may obscure the fact that the SCM defines individual causal effects. To avoid such confusion, in our presentation all endogenous variables have at least one exogenous variable pointing into them (with the exception of the illustrative example in Figure 6(b)).

[8] In some treatments of this material, exogenous variables always point into a single endogenous variable and dependencies among the variables are represented by dashed arcs (Pearl, 2000, Ch. 3).

A simple example will make this process more clear. Consider again the following structural model M:

$$z_1 \leftarrow h_{Z_1}(u_1, u_2) \tag{13}$$

$$x \leftarrow h_X(z_1, u_3) \tag{14}$$

$$z_2 \leftarrow h_{Z_2}(x, u_1, u_4) \tag{15}$$

$$y \leftarrow h_Y(x, z_1, z_2, u_5) \tag{16}$$

In Figure 6 (a), G_M is constructed using rules 1-3. A pruned version of G_M is drawn in Figure 6 (b) in which vertices and edges that are unnecessary for identifying the effect of X on Y have been dropped. The exact interpretation of the graphs in Figure 6 will wait until we discuss d-separation in the next section.

Given an SCM model M and an associated causal DAG, we can read conditional independence relations from such a model with the concept of *d-separation* (Geiger et al., 1990).

Definition 11 (d-Separation) *Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ be a DAG and X, Y, and Z be disjoint subsets of $\mathcal{V}$. X is said to be d-*separated *from Y by Z in $\mathcal{G}$ if and only if Z* blocks *every path from a vertex in X to a vertex in Y.*

A path p is said to be blocked *by a set of vertices Z if and only if at least one of the following conditions hold:*

1. *p contains a chain structure $a \rightarrow b \rightarrow c$ or a fork structure $a \leftarrow b \rightarrow c$ where the node b is in Z*
2. *p contains a collider structure $a \rightarrow b \leftarrow c$ where b is* not *in Z and no descendent of b is in Z*

If X is not *d-separated from Y by Z we say that X is d-*connected *to Y by Z.*

The d-separation criterion is incredibly powerful in the SCM framework, because of the following theorem (Geiger et al., 1990).

Theorem 1 (Probabilistic Implications of d-Separation) *Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ be a DAG and X, Y, and Z be disjoint subsets of $\mathcal{V}$.*

If Z d-separates X from Y in $\mathcal{G}$, then X is conditionally independent of Y given Z in every *distribution compatible with $\mathcal{G}$.*

For an SCM M, the joint distribution of the exogenous and endogenous variables is compatible with a graph G_M that is drawn using the rules from the previous subsection. Therefore, we can read conditional independence[9] relations (and hence ignorability conditions) from the graph, and these form the basis of the causal identification criterion of the next subsection.

As noted in the previous subsection, there is a simple graphical criterion (Pearl, 2000) that can be checked to see if a given set $\mathbf{Z}$ is sufficient to control confounding bias. This criterion can be stated as follows.

Definition 12 (Back-Door Criterion) *Given a causal model M and associated causal graph G_M, A set of covariates $\mathbf{Z}$ satisfies the back-door criterion for a causal variable X and outcome Y if:*

1. *$\mathbf{Z}$ does not block any directed paths from X to (or through) Y*
2. *$\mathbf{Z}$ blocks all paths from X to Y that are not directed paths*

where "blocking" is defined as in Definition 11 (d-Separation).

If $\mathbf{Z}$ satisfies the back-door criterion then an ignorability condition ($Y(x) \perp\!\!\!\perp X | \mathbf{Z}$) holds (Pearl, 2000), and the potential outcome distribution can be calculated using the standard stratification adjustment (Cochran, 1968; Rubin, 1977):

$$f_{Y(x)}(y) = \int_{\mathbf{z}} f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}) f_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z}$$

or

$$f_{Y(x)}(y) = \sum_{\mathbf{z}} f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}) f_{\mathbf{Z}}(\mathbf{z})$$

depending on whether $\mathbf{Z}$ is continuous or discrete and where $\mathbf{Z}$ may be multivariate. Pearl refers to this as the *back-door* adjustment.[10] Since if $\mathbf{Z}$ satisfies the back-door criterion the standard

[9] Careful readers will note that the implication in Theorem 1 does not go both ways. In particular, there may be conditional independence relations in a joint distribution that are not represented by d-separation in the associated graph. However, these situations usually depend on the rare circumstances of zero effects, exact cancellation of effects, or endogenous variables that are a deterministic function of only endogenous variables. Therefore, this important caveat will not affect the results of this paper.

[10] Note that $f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z})$ must exist—and thus $f_{X,\mathbf{Z}}(x, \mathbf{z})$ must be non-zero for all x and $\mathbf{z}$—in order for the back-door adjustment to be valid. Put slightly differently, this method of adjustment requires the distributions of the measured confounders to have the same support in the treated and control groups if an average causal effect is to be estimated. This is something that is well understood by political scientists who employ matching estimators of causal effects (Ho et al. (2007), see also King and Zeng (2006)).

stratification adjustment is appropriate, it follows that matching or stratifying on $\Pr(x|\mathbf{z})$ (the propensity score given a realized value $\mathbf{z}$ of $\mathbf{Z}$), along with related adjustments that make use of conditional ignorability, will also be appropriate (Rosenbaum and Rubin, 1983, 1984). As we will see below, this is true regardless of whether all (or even any) of the variables that affect treatment assignment are in $\mathbf{Z}$– all that is required is that conditional ignorability hold given $\mathbf{Z}$.[11]

Again, the major advantage of this graphical approach to the identification of causal effects is that it is framed in terms of a series of local assumptions about causal mechanisms. These local assumptions are often easier to consider, debate, and possibly reject as unbelievable than the single global assumption of conditional ignorability.

4 Conclusion

In this article we have used the SCM to study the identification of causal effects in the context of TSCS data and to clarify aspects of the debate between Achen (2000) and Beck and Katz (1996, 2011). A key takeaway point that emerges from this analysis is that, for data generating processes consistent with either Figure 1 (a) or Figure 1 (b) above, one can identify the causal effect of $x_{i,t}$ on $y_{i,t}$ by conditioning on either $x_{i,t-1}$ and $y_{i,t-1}$ or just on $x_{i,t-1}$ in addition to $x_{i,t}$. This is a very general result and does not depend on specific distributional assumptions or assumptions regarding constant, linear effects.

One's choice of tools affects how one works and even how one thinks about that work. Using the SCM within the context of TSCS data is no exception. Perhaps the biggest contribution of this article is to demonstrate that use of the SCM brings attention to a number of previously neglected facts that are relevant for practical TSCS research. These include the following:

- Simply knowing that $y_{i,t-1}$ is or is not a cause of $y_{i,t}$ does not provide enough information to know whether $y_{i,t-1}$ should or should not be included in a conditioning set. The same is true for $x_{i,t-1}$ and other potential conditioning set variables.

[11] We note in passing the obvious point that the results of Rosenbaum and Rubin (1983) show that if conditional ignorability holds given $\mathbf{Z}$ then using $\Pr(x|\mathbf{z})$ or any other balancing score as an adjustment covariate is appropriate. They do not show that subclassifying or matching on $\Pr(x|\mathbf{z})$ or any other balancing score for arbitrary $\mathbf{Z}$ produces conditional ignorability of treatment assignment.

- It is not necessary to consistently estimate the effect of each conditioning set variable on $y_{i,t}$ in order to use that conditioning set to identify the effect of $x_{i,t}$ on $y_{i,t}$.
- Given a particular data generating process, there may be multiple, distinct conditioning sets that identify a causal effect of interest.

The SCM also offers a number of more general benefits for researchers interested in reasoning about the identifiability of causal effects in complicated settings. First, once it is understood, the SCM provides a powerful and easy-to-apply method for researchers to prove that a particular conditioning set identifies a causal effect of interest. The SCM forces a researcher to make his or her assumptions about causal dependencies explicit. Once this is done, it is a relatively straightforward task to check to see if a particular conditioning set satisfies the backdoor criterion and consequently identifies the causal effect of interest. No linear algebra, calculus, or probability theory is required.

Second, identification results derived using the SCM do not require specific distributional or functional form assumptions (multivariate normality, linear relationships, constant effects, etc.). The results are nonparametric and are thus consistent with any population distribution and a wide variety of estimation strategies. For instance, the nonparametric identification results for TSCS data in this paper are equally valid for Gaussian outcomes, binary outcomes, negative-binomial outcomes, etc. Further, while we have, at some points, assumed constant linear effects in order to be consistent with Beck and Katz (1995, 1996, 2011); Achen (2000), and Keele and Kelly (2006), this was not necessary. We could have easily used much weaker functional form assumptions to motivate estimators other than linear regression (such as matching, weighting, or even simple stratification estimators).[12]

However, the SCM results are built around an assumption of independence between units so attempts to use repeated measurements over time to increase the effective sample size will as always require additional assumptions. In this paper we have shown that when the usual linear model time-

[12] By their nature, identification results are large sample results. Thus while the identification results in this paper will continue to hold under other distributional assumptions and functional form assumptions there is no guarantee that the finite sample Monte Carlo results in this paper will continue to provide the correct guidance for data generating processes that are not at least approximately the same as those considered in the Monte Carlo experiments. In particular, there is some reason to believe that data generating processes with highly nonlinear conditional mean functions may require much larger sample sizes in order to reliably estimate the effects of interest.

series assumptions are appropriate, the insights derived from SCMs provide benefits in terms of bias and root mean square error. Future work should address other specifications.

Finally, while the analysis in this paper has focused on what might be termed "short run" effects (e.g. the effect of x_t on y_t). It is straightforward within the SCM framework to address "long run" questions (e.g., the effect of x_t on y_{t+20}). All of the graphical identification rules extend to the these and other alternative effects. Again, the only difficulty comes when dependent observations are used to increase the sample size. It is also possible to use this framework to address the effects of more complicated static and dynamic treatment regimes (Robins, 1986; Blackwell, 2013).

We conclude with some general suggestions for applied TSCS researchers. Many of these points will be obvious to seasoned researchers. Nonetheless, the fact that few researchers appear to follow these recommendations suggests that stating what is obvious to some may be of use to many TSCS researchers.

First, we recommend that researchers clearly define the causal effect they are interested in estimating. Every researcher should be able to clearly state what their estimand is and all efforts should be devoted to identifying and estimating that causal effect. Note that different estimands will almost always require different identification and estimation strategies. A regression specification that identifies one causal effect of interest will not necessarily identify another causal effect of interest.

Once the causal effect of interest is clearly defined, the TSCS researcher should work out as many plausible models of the causal dependencies as possible and operationalize these models as SCMs. Minimally, the TSCS researcher should work out a substantively defensible SCM for the problem at hand.

Once the SCM (or more preferably a variety of plausible SCMs) are worked out, the researcher should use the backdoor criterion to determine which conditioning sets, if any, identify the causal effect of interest under the assumptions embodied in the particular SCM of interest. The researcher can then choose an estimation method that makes sense given the nature of their data and their identification results.

The researcher can then estimate the effect of interest. In situations where a single SCM yields multiple conditioning sets that each identify the same causal effect, the researcher can estimate the effect using each of the conditioning sets and check to see if the results are substantially similar. One could construct formal hypothesis tests to test the null hypothesis that each of these estimators are in fact estimating the same thing (i.e., tests of overidentifying restrictions). In situations where multiple causal structures (SCMs) are plausible then the researcher can estimate the causal effect of interest under each plausible set of assumptions. If the estimates are qualitatively similar the researcher might be willing to conclude with some confidence that they have a good understanding of the approximate size of the effect of interest. If, on the other hand, the estimates differ considerably across plausible SCMs then the researcher is not in a position to make strong inferences about the causal effect of interest (Pearl, 2004).

References

Achen, Christopher H. 2000. "Why Lagged Dependent Variables Can Suppress the Explanatory Power of Other Independent Variables." Paper prepared for the 2000 Annual Meeting of the Society for Political Methodology.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91:444–455.

Beck, N., and J.N. Katz. 1995. "What to do (and not to do) with time-series cross-section data." *American Political Science Review* pp. 634–647.

Beck, N., and J.N. Katz. 1996. "Nuisance vs. substance: specifying and estimating time-series-cross-section models." *Political analysis* 6(1):1.

Beck, Nathaniel, and Jonathan N. Katz. 2011. "Modeling Dynamics in Time-Series-Cross-Section Political Economy Data." *Annual Review of Political Science* 14:331–352.

Blackwell, Matthew. 2013. "A Framework for Dynamic Causal Inference in Political Science." *American Journal of Political Science* forthcoming.

Cochran, William G. 1968. "The Effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies." *Biometrics* 24(2):295–313.

Geiger, D., T. Verma, and J. Pearl. 1990. "Identifying independence in Bayesian networks." *NETWORKS.* 20(5):507–534.

Haavelmo, Trygve. 1943. "The Statistical Implications of a System of Simultaneous Equations." *Econometrica* 11:1–12.

Ho, Daniel, Kosuke Imai, Gary King, and Elizabeth Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15:199–236.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81:945–960.

Keele, Luke, and Nathan J. Kelly. 2006. "Dynamic Models for Dynamic Theories: The Ins and Outs of Lagged Dependent Variables." *Political Analysis* 14:186–205.

King, Gary, and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159.

Koopmans, Tjalling C. 1949. "Identification Problems in Economic Model Construction." *Econometrica* 17:125–144.

Marschak, Jacob. 1950. "Statistical Inference in Economics: An Introduction." In *Statistical Inference in Dynamic Economic Models* (Cowles Commission Monograph no. 10) (Tjalling Koopmans, editor), New York: John Wiley & Sons.

Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669–710.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.

Pearl, Judea. 2004. "Robustness of Causal Claims." Technical Report R-320, University of California, Los Angeles.

Pearl, Judea. 2009. “Causal Inference in Statistics: An Overview.” *Statistics Surveys* 3:96–146.

Robins, J.M. 1986. “A new aproach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect.” *Mathematical Modeling* 7:1393–1512.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika* 70:41–55.

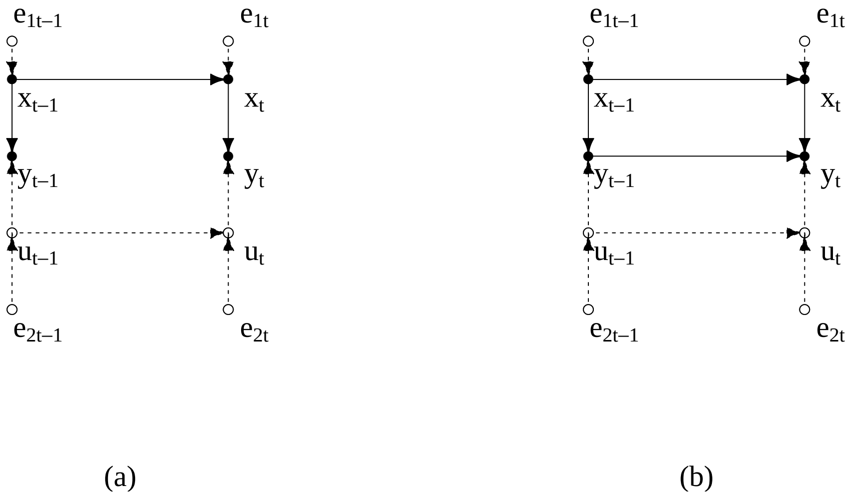
Rosenbaum, Paul R., and Donald B. Rubin. 1984. “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.” *Journal of the American Statistical Association* 79:516–524.

Rubin, Donald B. 1977. “Assignment to Treatment Group on the Basis of a Covariate.” *Journal of Educational Statistics* 2(1):1–26.

Rubin, Donald B. 1978. “Bayesian Inference for Causal Effects: The Role of Randomization.” *The Annals of Statistics* 6(1):34–58.

Simon, Herbert A. 1953. “Causal Ordering and Identifiability.” In *Studies in Econometric Method* (W.C. Hood, and T.C Koopmans, editors), New York: John Wiley & Sons.

Spirtes, P., C. Glymour, and R. Scheines. 1993. *Causation, Prediction, and Search*. New York: Springer.



(a)

(b)

Figure 1: *Causal graphs consistent with the models considered in Beck and Katz (1996), Achen (2000), and Keele and Kelly (2006).* Unit subscripts (i) have been removed from the graph for clarity. In **Panel (a)** the empty conditioning set $\{\}$ will identify the average effect of $x_{i,t}$ on $y_{i,t}$. However, the conditioning set that includes only the lagged dependent variable $\{y_{i,t-1}\}$ will induce bias because conditioning on the collider $y_{i,t-1}$ unblocks the backdoor path from $y_{i,t}$ to $x_{i,t}$. If a right-hand-side regression variable is defined to be irrelevant when it does not have a causal effect on the outcome, then this model invalidates the standard econometric result that the inclusion of irrelevant variables cannot bias a regression result. However, if one conditions on $x_{i,t-1}$ in addition to $y_{i,t-1}$, then β will be identified: the backdoor path from $y_{i,t}$ on $x_{i,t}$ is blocked by the conditioned fork structure at $x_{i,t-1}$ for the conditioning set $\{y_{i,t-1}, x_{i,t-1}\}$. It is also the case that conditioning on just $x_{i,t-1}$ (in addition to $x_{i,t}$) will identify β. In **Panel (b)** there are no conditioning sets using the observed variables that identify the average effect of $y_{i,t-1}$ on $y_{i,t}$. However, the conditioning sets $\{x_{i,t-1}\}$ and $\{y_{i,t-1}, x_{i,t-1}\}$ identifies the average effect of $x_{i,t}$ on $y_{i,t}$. The fact that the conditioning set $\{y_{i,t-1}, x_{i,t-1}\}$ identifies β provides a counterexample to the idea that bias in one regression parameter implies bias in all regression parameters. Identification in all of these scenarios is derived in terms of independent units. The Monte Carlo analyses in Figures 2, 3, 4, and 5 demonstrate that these identification results have implications for dependent observations.

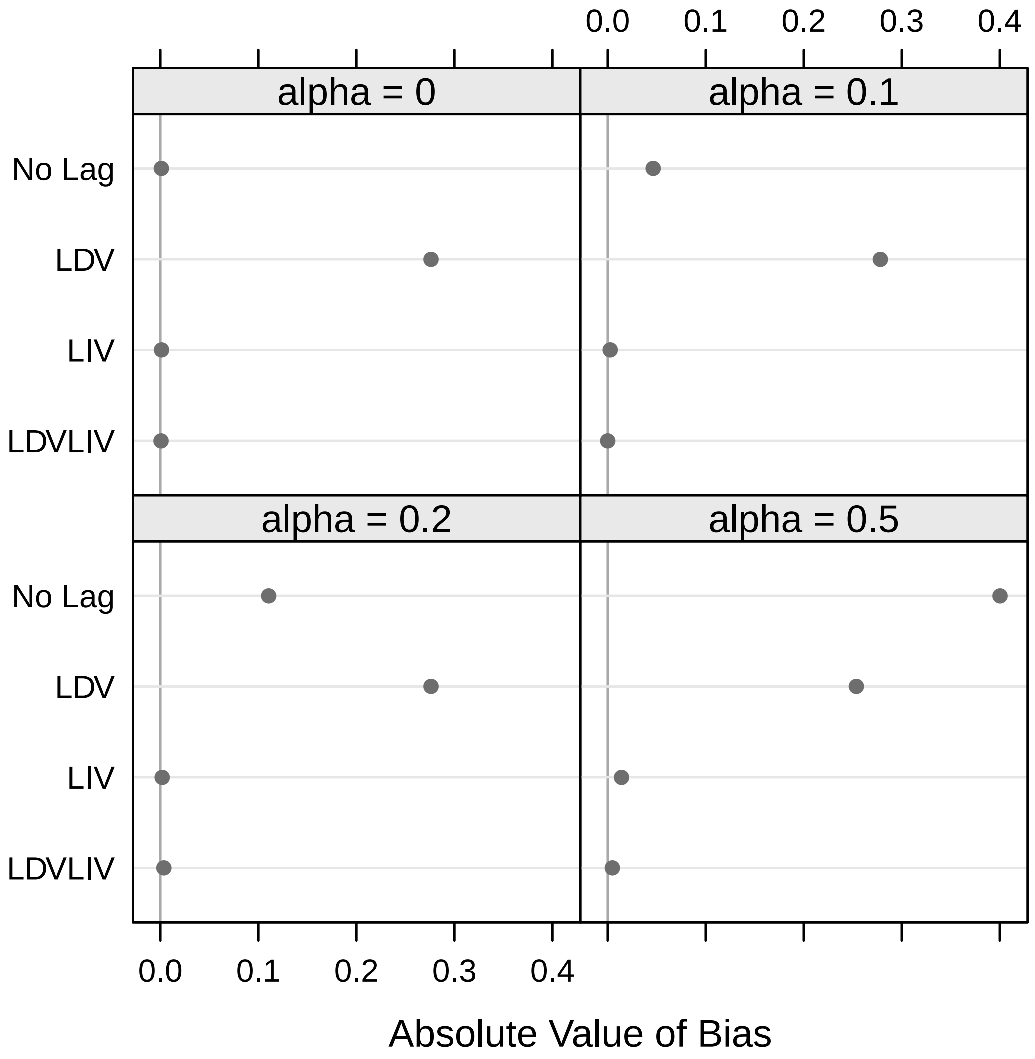


Figure 2: *Extension of the Keele and Kelly (2006) Time Series Monte Carlo analysis of bias consistent with the causal graphs in Figure 2 (a) (upper left panel) and Figure 2 (b) (other three panels).* Monte Carlo analysis of bias with the {} conditioning set (No Lag), the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDV), and the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDVLIV). $N = 1$, $T = 100$, $\beta = .5$, $\phi = .75$, and $\rho = .95$. As noted in Keele and Kelly (2006), some of these specifications violate stationarity.

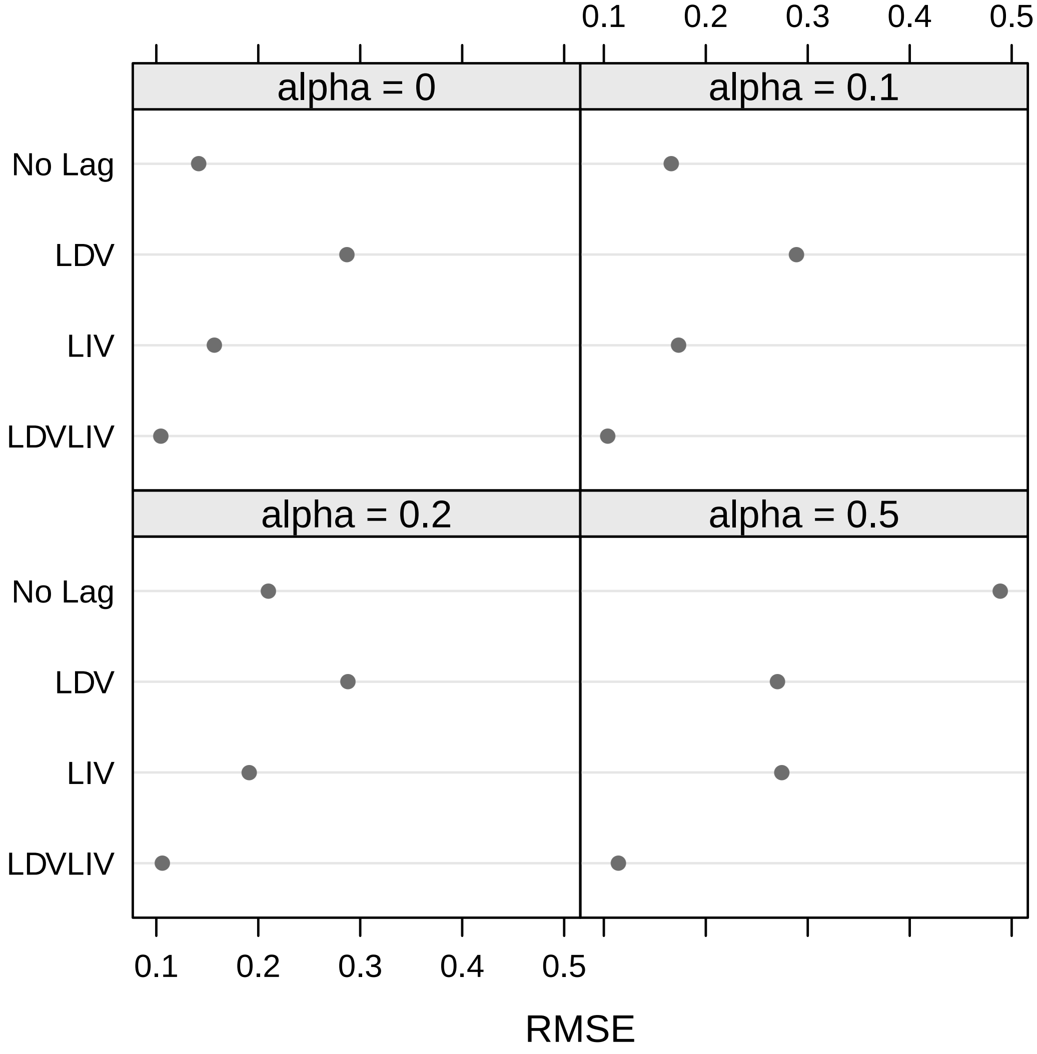


Figure 3: *Time Series Monte Carlo analysis of root mean squared error consistent the Keele and Kelly (2006) Monte Carlo analysis and with the causal graphs in Figure 2 (a) (upper left panel) and Figure 2 (b) (other three panels).* Monte Carlo analysis of bias with the $\{\}$ conditioning set (No Lag), the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDV), and the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDVLIV). $N = 1$, $T = 100$, $\beta = .5$, $\phi = .75$, and $\rho = .95$. As noted in Keele and Kelly (2006), some of these specifications violate stationarity.

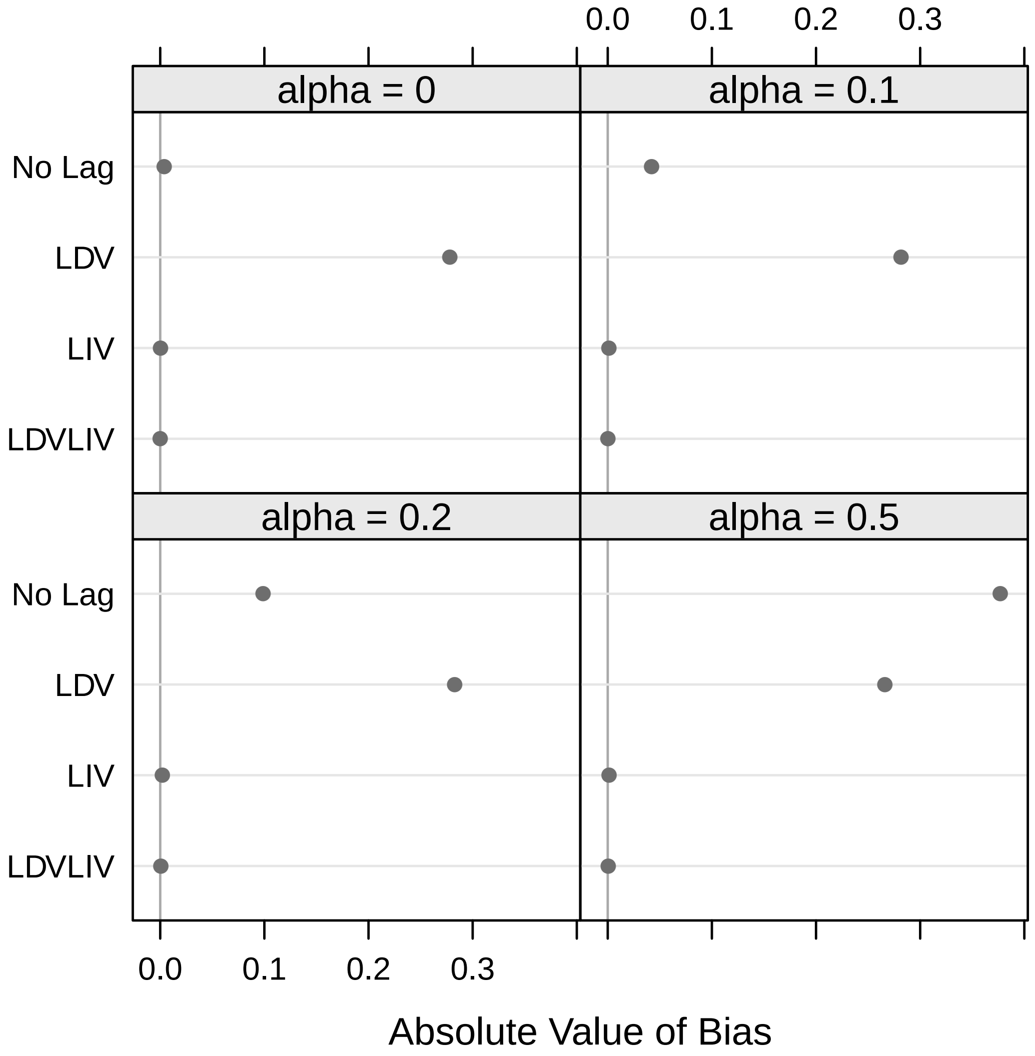


Figure 4: *TSCS Monte Carlo analysis of bias with $N = 15$, $T = 20$, consistent with the causal graphs in Figure 2 (a) (upper left panel) and Figure 2 (b) (other three panels).* Monte Carlo analysis of bias with the {} conditioning set (No Lag), the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDV), and the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDVLIV). $\beta = .5$, $\phi = .75$, and $\rho = .95$. As noted in Keele and Kelly (2006), some of these specifications violate stationarity.

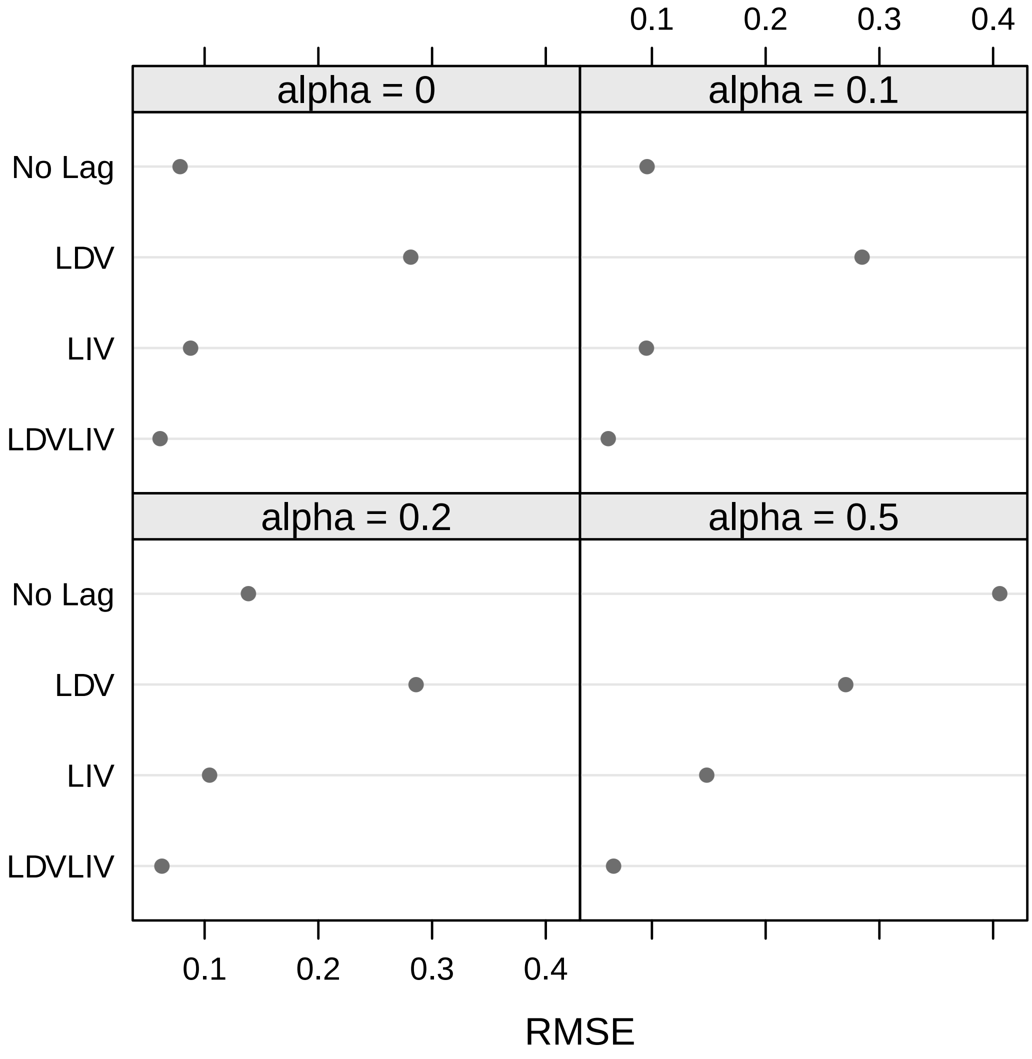


Figure 5: *TSCS Monte Carlo analysis of root mean squared error with* $N = 15$*,* $T = 20$*, consistent with the causal graphs in Figure 2 (a) (upper left panel) and Figure 2 (b) (other three panels).* Monte Carlo analysis of bias with the $\{\}$ conditioning set (No Lag), the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDV), and the $\{y_{i,t-1}, x_{i,t-1}\}$ conditioning set (LDVLIV). $\beta = .5$, $\phi = .75$, and $\rho = .95$. As noted in Keele and Kelly (2006), some of these specifications violate stationarity.

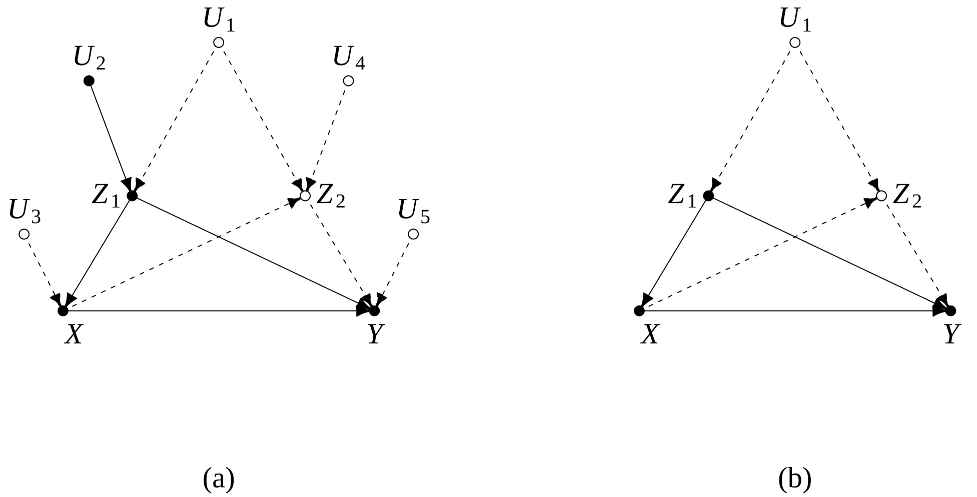


Figure 6: *Graphical Model Consistent with Structural Equations 13 - 16.* Panel (a) shows all exogenous variables and their associated edges. Panel (b) removes superfluous exogenous variables and their edges. Note that observability is neither necessary nor sufficient for a variable to be exogenous. Here $U_1, \ldots, U_5$ are the exogenous variables. U_2 is observed but the other U variables are not. Further, one of the endogenous variables (Z_2) is unobserved while the other endogenous variables are observed.