Causal Identification in TSCS Regression: Overcoming Collider Bias and Harnessing DAGs for Improved Control Variable Selection

Working Paper

June 2023

Please do not cite without authorization.

Manoel Galdino[1]
Davi Moreira[2]
Carolina Dolleans[3]

[1] Assistant professor of Political Science at Universidade de São Paulo. E-mail for correspondence: mgaldino@usp.br
[2] Assistant professor of Political Science at Universidade Federal de Pernambuco
[3] Graduate Student of Political Science at Universidade Federal de Pernambuco

Introduction

The analysis of Time Series Cross Section (TSCS) data plays a crucial role in current research practices within political science, particularly in the fields of International Relations and Comparative Politics. TSCS data offer a powerful combination of spatial and temporal features, making them valuable for empirical analysis.

Recently, a growing body of literature has discussed the challenges and limitations of analyzing TSCS data from a causal inference perspective. These studies have examined assumptions in fixed effect models (Imai and Kim, 2019), the appropriate use of random effect models (Bell and Jones, 2015; Clark and Linzer, 2015), matching techniques for TSCS data (Imai, Kim, and Wang, 2021), and the potential pitfalls of employing two-way fixed effect models with heterogeneous effects in difference-in-differences models (de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2021).

One important aspect that applied researchers need to address in their papers is the selection of control variables when making exogeneity or unconfoundedness assumptions (cf. Imbens, 2004) to identify causal effects. Improper inclusion of controls can introduce biases that prevent the model from being correctly identified and estimating the causal estimand accurately.

By utilizing Directed Acyclic Graphs (DAGs), we can overcome common problems in TSCS regressions related to dynamic panel models and biases resulting from inadvertent inclusion of collider variables as regressors. To the best of our knowledge, this is the first study to focus on how the standard heuristic for selecting control variables can inadvertently introduce collider variables in the context of TSCS data, potentially biasing the results. To fill this gap in the literature, we present an improved heuristic for constructing a DAG that appropriately controls for relevant variables in TSCS data analysis.

The remainder of the paper is organized as follows: First, we review the process by which researchers decide to include variables as controls in regressions. Second, we introduce the fundamental concepts of DAGs and demonstrate their application for correctly identifying causal models in dynamic panels. Third, we discuss the issue of "foreign collider bias" in TSCS data and present strategies to mitigate it. In the final section, we provide our concluding remarks.

Controls for Casual Identification

From a more general point of view for all kinds of observational data, regression with a selection on observables strategy to identify a causal effect must include all relevant variables as controls when attempting to detect a causal effect. Which variables to include as controls depends firstly on the scientific understanding of which variables are crucial to explaining the phenomena of interest.

A second criterion when identifying a causal effect is that the model should be identified, meaning there is no bias in estimating the causal estimand of interest (Cinelli et al., 2021). Thus, it is necessary to assess the identification of the model alongside scientific domain expertise. In practice, most scholars use only some heuristics to decide which controls to include.

A standard heuristic is to review the relevant literature to search for potential causal variables of the outcome of interest and include them in the regression as controls. Here is a typical example of such a heuristic from a paper on foreign aid:

"As the previous literature on aid policy maintains, various other factors shape donor decisions about the allocation of aid resources, including other recipient characteristics and non-developmental donor goals. I include them as control to provide a fully specified model" (Dietrich, 2016, p.81).

We wonder if researchers can easily cite relevant literature backing such a heuristic if hard-pressed. It is possible, however, to point to the fact that this may decrease the unexplained variance in the dependent variable, which improves the precision of the Average Causal Effect (ACE) in finite samples (Hahn, 2004; Pearl, 2013; Cinelli et al., 2021). The problem with this explanation is that more is needed to decide if a variable should be a control. As shown by Cinelli, Forney, & Pearl (2021), it is quite possible to include a variable as a control that may decrease precision and induce bias, even if it is causally related to the outcome of interest.

A similar but slightly better heuristic to identify causal effects is surveying the literature to spot potential confounding variables for which one should control. In econometric parlance, one should avoid omitted variables bias. Here is a typical example of such an approach in a paper on the effect of political regime change on the occurrence of civil war:

“Our model does not attempt to present an inclusive theory of civil war, but level of democracy and political change do not provide a complete explanation. Therefore, we identify a number of control variables – Development, Ethnic Heterogeneity, Proximity of Independence, and International War in Country – whose omission might bias the results for the regime change variable.” (Hegre et. al, 2001, p. 37).

The above example provides a better heuristic for causal inference because it is focused on diminishing the primary source of bias in observational studies, namely, omitted variable bias. In DAG parlance, no omitted variable bias is left if one blocks all backdoors by including all appropriate controls. However, and this is one of the key points we make in the present paper, in the context of TSCS, this heuristic does not assess if they are inadvertently introducing a collider variable that may bias the causal effect of interest, which is a problem, since the introduction of a collider will bias the causal effect.

As far as we know, those are the two most important heuristics researchers use to decide which variables to include as controls. What about formal approaches that go beyond Heuristics and formalize criteria for including controls? Heckman (2008) argues that the researcher needs to explicitly model the selection into treatment by agents, which imply to model the controls in a selection on observables strategy. Formally modeling agents' choices is a way to deal with this problem because one can formally assess if the causal effect is identified. However, this may be impossible in practice due to tractability constraints or a lack of precise scientific knowledge on the matter.

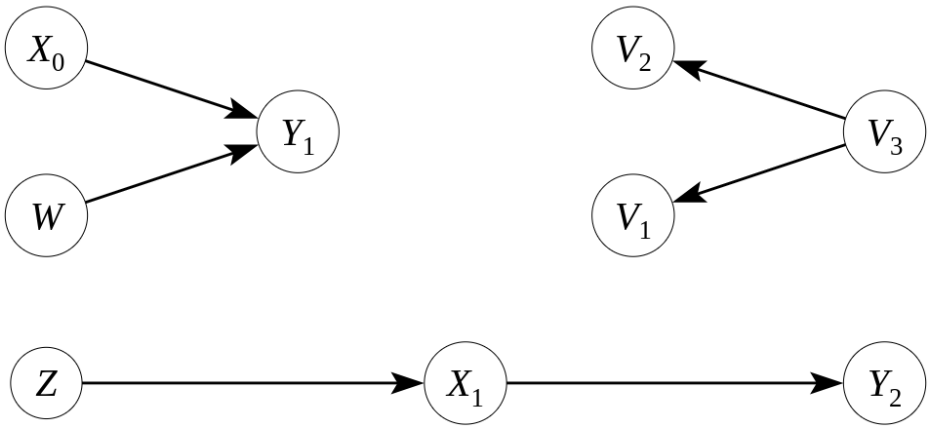
Another possible approach is causal discovery (Spirtes & Zhang, 2016; Glymour et al., 2019; Duart et al., 2021). However, it is based on independence relations found in the data. It thus presupposes that all relevant variables have been included, which is precisely the problem we need to solve in the first place.

A third approach is the one Pearl advocated and used in the present paper. The researcher should draw a Directed Acyclic Graph to determine whether the causal impact is identified, notably whether any controls are missing or whether an erroneously included "bad control" actually biases the causal effect. Adopting this approach, she can even think about possible unobservable variables that will be impossible to control in the study, fairly setting the study's limits. To see how it can help, we will introduce the reader to DAGs and provide an example of an application in the context of TSCS.

Preliminaries and Basic Terminology for DAGs

Directed Acyclic Graphs (DAGs) are part of an approach to causal inference that was first developed at the beginning of the 21st century by Philip Wright (1928) and Sewall Wright (1934). It has become a progressive research program with the work of Judea Pearl and his collaborators (Pearl & Mackenzie, 2018), after which it has been applied in several domains, like epidemiology and computer science and, more recently in political science (Pearl & Mackenzie, 2018; Imbens, 2020; Yao et al., 2021).

All DAGs comprise three primary structures: chains, forks, and colliders (or inverted forks). The graphic below presents the three structures:



The top left DAG is an inverted fork. It contains a collider $Y_{1,,}$, representing common effects, i.e., two (or more) variables causing a third variable. The top right DAG is a fork. It has a variable that is a common cause of other variables, such as V_3. Lastly, the bottom DAG is a chain, and it contains a mediator, X_1. A collider, contrary to chains and forks, does not induce association among variables. However, conditioning on it (or on its descendants) does induce an association among its causes.

This means that if you have a fork, you should control for the common variable to avoid omitted variable bias. In that case, we say that we blocked a backdoor path. If we block all backdoor paths, then there is no omitted variable bias. Caution is needed when there is a mediator. If one is interested in the total effect of a variable (say, Z, above) on another (Y_2), then controlling for the mediator (X_1) is wrong since it will block that specific causal path. If,

on the other hand, one is interested only in the direct (unmediated) effect, then one should control for the mediator.
Last but not least, if one is interested in, say, the effect of X_o on W (or of W on X_o), you should never control for a collider such as Y_1 in the top left DAG above since it will create a spurious association between variables.

We can then interpret the heuristic behind including controls to deconfound the estimate. It is trying to block all backdoor paths. However, it says nothing about what to do if the control is a mediator or a collider.

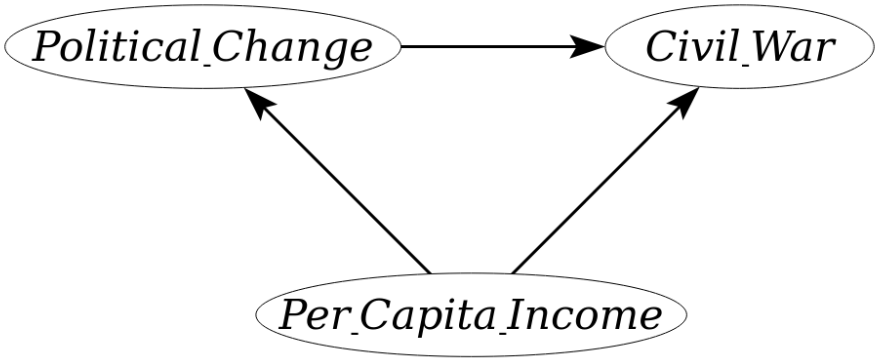
One of the focuses of this paper is the case of inadvertently including a collider. Nevertheless, before looking into this case, it is easier to see how a DAG can help us avoid omitted variable bias in a common TSCS application: the so-called dynamic panel model..

A Simple DAG Application in the Context of TSCS

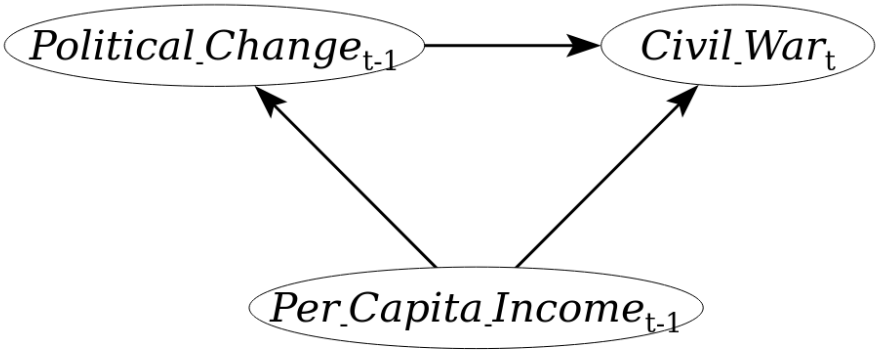
To motivate the reader on the usefulness of using a DAG, consider a typical setup in comparative politics in a very simplified setting.

Suppose the researcher is interested in knowing the causal effect of political regime change on civil war[4], like in Hegre et al. (2001) that we quoted when presenting a heuristic for including controls. Suppose also that the only possible (measured) confounder is per capita income (it can cause civil war and democratization[5]). Thus, to avoid omitted variable bias, the researcher decides to control for per capita income in the regression. Here is the DAG that represents this situation.

[4] There are many papers like this. See, for example, Hegre et. al. (2001), for an applied paper, with almost three thousand citations in google scholar, or Blattman & Miguel (2010) for a review of the literature.
[5] There are, of course, many more potential confounding, such as economic growth (Blattman & Miguel, 2010).



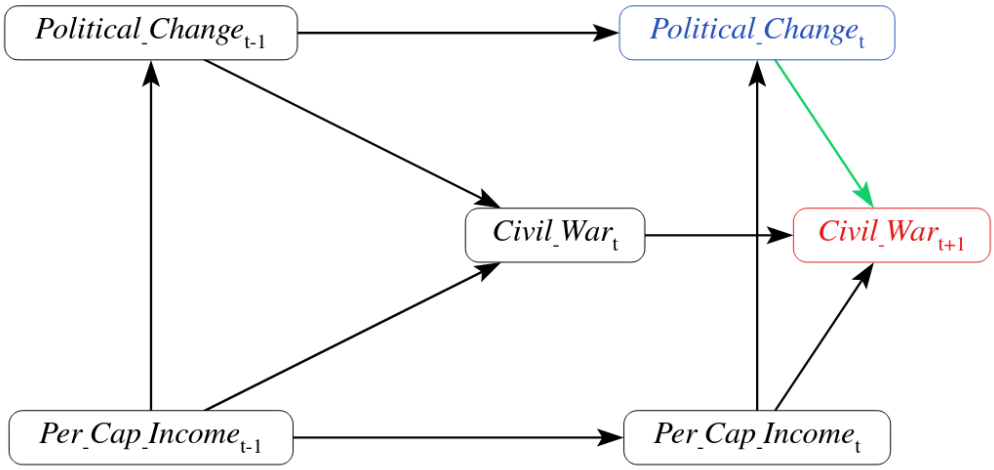
In the context of TSCS data, we need to index the above variables by t. Suppose there is also concern about reverse causality or that it is hard to measure when a civil war starts. The researcher decides to lag both per capita income and political change in the regression to circumvent such possibilities and make the results more robust.



With such a model in mind and a dataset in the long format, in software like R, the researcher would run something like lm(civil_war ~ poli_change_lag + per_cap_income_lag, data=df) for a linear probability model and check if the effect is significant.

The software code and the above DAG are deceptively simple. In fact, they obscure that a dynamic process evolves over time.

Let us unpack what is behind such a dynamic process and how it relates to a research question by considering a DAG over two periods for the outcome variable.



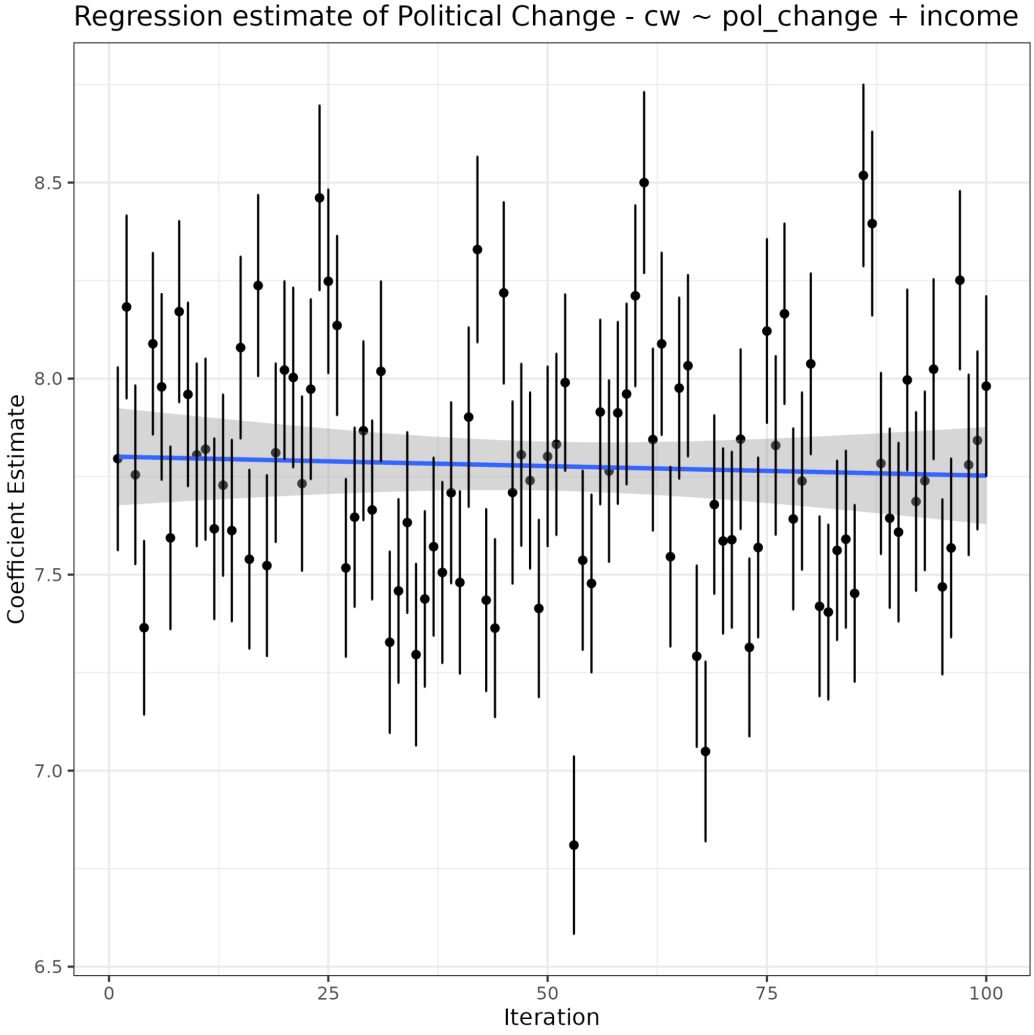
Firstly, in the DAG above, we included arrows from past levels of variables to current ones. Such a setting is likely in the social sciences, where persistence and inertia are the norm. In particular, we know that: per capita income at time t is correlated with its value at $t + 1$, if there was political change in time t it will change the likelihood of another political change in time $t + 1$. So, a more believable DAG should include such arrows.

Secondly, it is not easy to say which variable is the treatment or outcome. In a DAG, both the treatment and the outcome are represented by a single node. In contrast, we have two nodes that are the treatments: political change at time $t - 1$ and at time t and three nodes that are the outcomes: civil war at $t - 1$, t and $t + 1$.

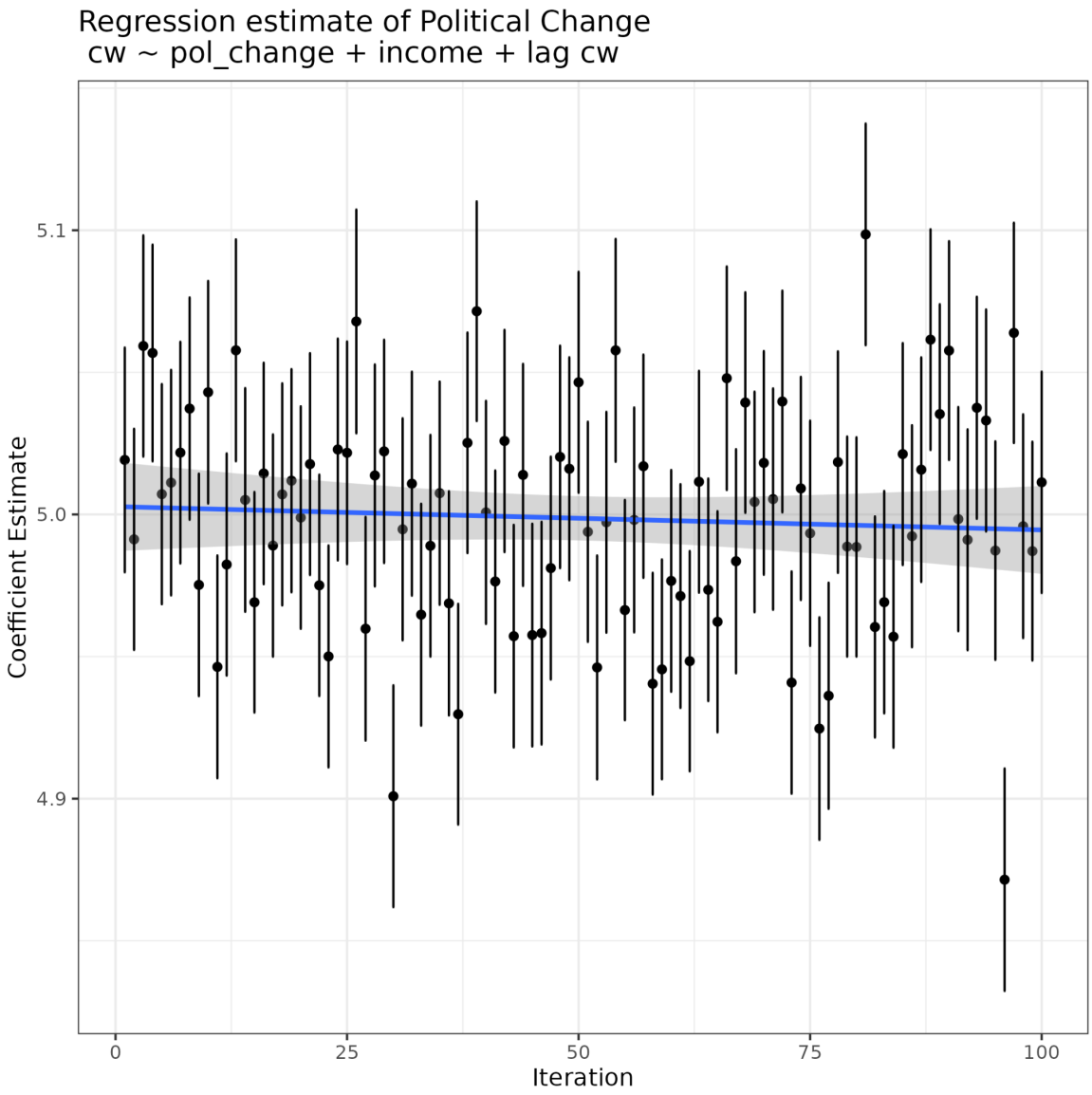
To circumvent this in the DAG above, consider only period two. Then, the causal effect of interest is political change at t on civil war at $t + 1$. In this case, it is easy to see from the DAG above that two backdoors opened, both going through civil war lagged. Only a dynamic panel model can recover the actual causal effect.

In this case, whenever there is variation in the treatment effect over time, the causal effect will be estimated with bias, which will depend on the distribution of variations of treatment allocation across countries and time.

We ran a simple simulation[6] based on the above DAG and assumed that civil war is a normal variable to make things easier to interpret. In every iteration of the simulation, we kept everything constant but the error terms and, as a result, who would experience civil war or not in the end. Other than that, every iteration was the same. There are no heterogeneous effects over time. The first chart is based on a model in which only lagged predictors for political change and per capita income were included. The second chart added a lagged dependent variable to the model. The true effect of political change on civil war is 5.



[6] The simulation code is available in the Github repository of one of the authors: https://github.com/mgaldino/beware_collider/blob/main/scripts/sim_rq.R



The first model needs to be corrected, but the dynamic panel model is, on average, correct. Based on the DAG, we already knew that that was the case. Thus, a DAG can save us much effort and help to specify a model that can identify a causal effect correctly.

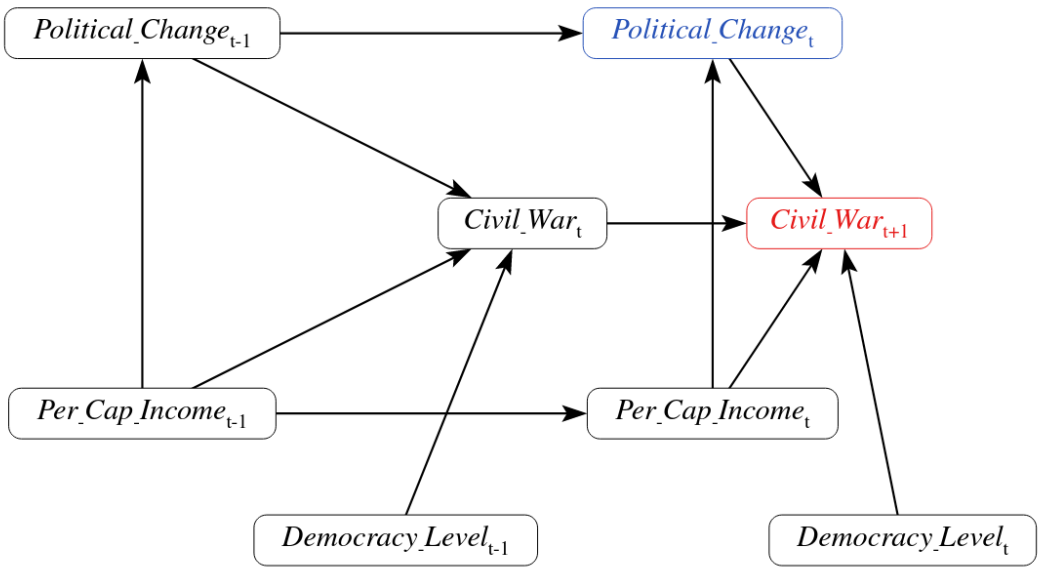
Collider Bias

As mentioned, the problem with including as many controls as possible is that one may inadvertently include what is called in the Potential Outcomes framework a “bad control” (Angrist & Pischke, 2009). A bad control is either a mediator or a collider. The Potential Outcomes framework provides very generic advice on what is a bad control. When it is a mediator, the lack of precise advice is not a problem because knowing that a mediator cannot

be a control variable if one is interested in the total effect is pretty intuitive. However, in the case of colliders, the intuition should be more straightforward, and we need a more formal approach, namely, using DAGs.

It is important to note that a collider is only a problem if it is in the path between the treatment (our x variable) and the outcome (the y variable). In general, any DAG in a social science context will have plenty of colliders that create spurious associations between variables. However, that is not a problem because those associations are not of interest in a given research. Thus, the DAG is built with a clear goal: to allow one to check if the causal effect of the treatment on the outcome is identified. All other causal relations are only important insofar as they help to identify the causal effect of interest. They are not of interest per se.

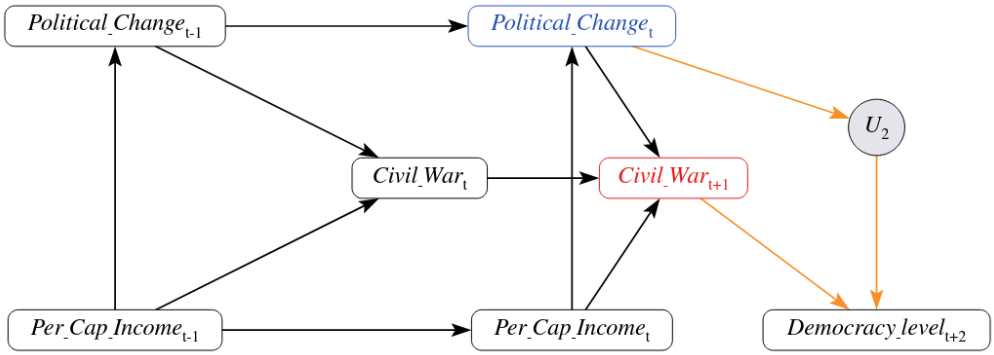
Let us return to our previous example of political change and civil war. Suppose, like Hegre et al., one wants to add other controls to a regression, such as democracy level. The DAG now is as follows:



No harm is apparently done, and the causal effect of interests is identified as before. However, it is plausible that after a civil war onset, the democracy level of the country will

drop. In our DAG, this means there is a missing arrow from current civil war nodes to current democracy levels nodes.

However, suppose one looks at the literature on the causes of democracy levels. In that case, one will find several variables that may cause it and are not included in the graph because they are not causally related to civil war. The current heuristics for including controls need to tell us to review the literature on potential causes of the controls. Why would it? If, in this "foreign" literature, there are causal paths from one of your predictors to another, not included in the DAG, that also causes the control, then we have a collider bias problem. We call it a foreign collider bias.



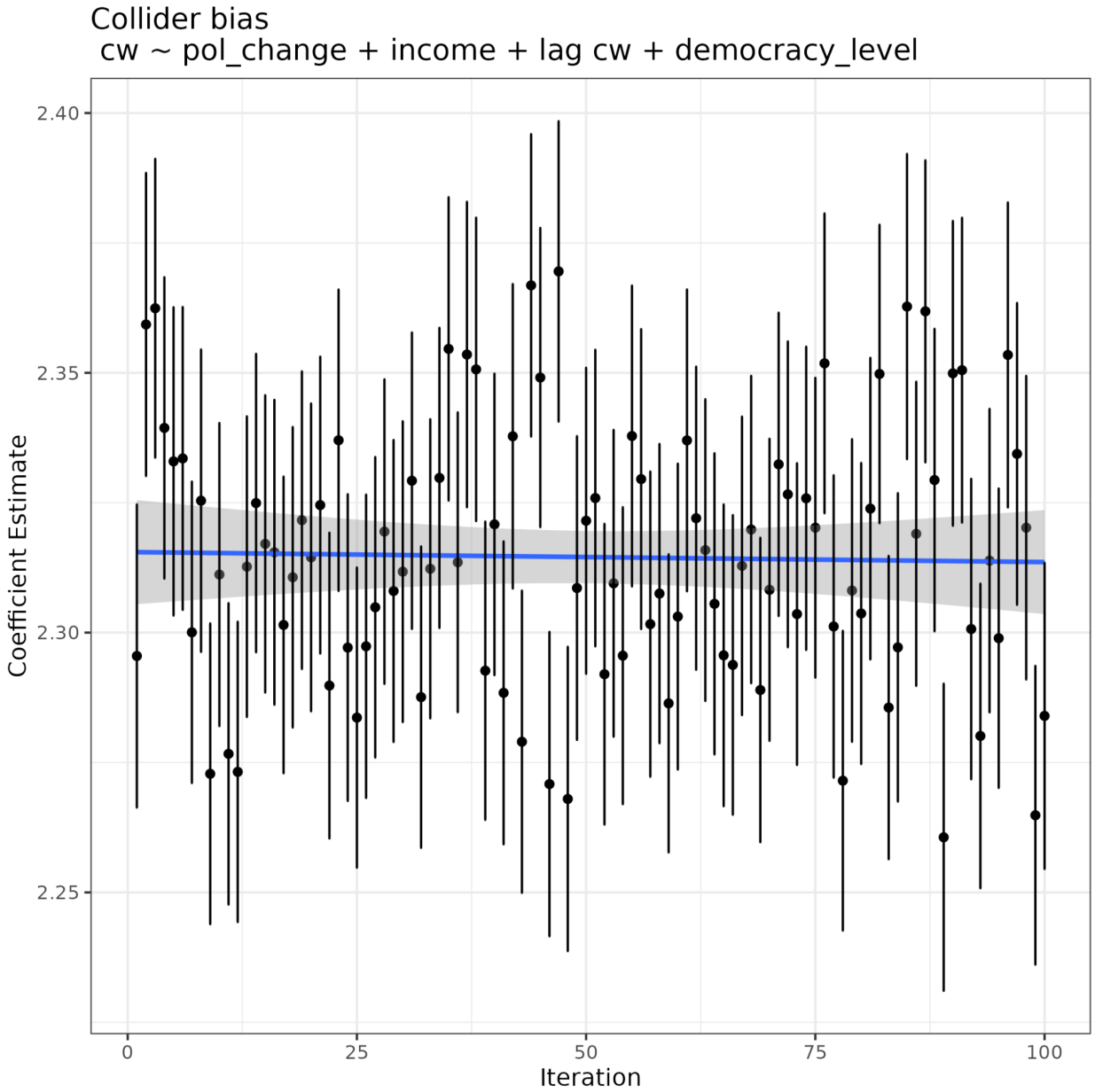
The DAG above may be a perfect example of the saying that a picture is worth a thousand words. In the above DAG, we excluded the past values of democracy level, causing civil war for simplicity. The variable U_2 is not included in the regression because no one is concerned about it. Our standard heuristics suggest including a variable in a regression as a predictor if i) it can cause our dependent variable or ii) if it is a potential source of confounding. On the other hand, the same heuristics suggest that we should include democracy level because it can cause civil war (or perhaps is a potential confounding of political change). However, introducing it as a regressor biases the regression.

To see why it is a collider, notice that democracy level at time $t + 1$ is a common effect of U_2 and Political Change at time t. By conditioning on democracy level (including it as a control), we open a backdoor path connecting $U_{2...}$ to Civil War.

This problem is specific to TSCS data because of the usage of lagged variables and the possibility of reverse causality. In the example at hand, the democracy level variable enters

the regression lagged to avoid reverse causality. However, the reverse causality means there is an arrow vom the dependent variable to current values of democracy level. If there are any other variables, not observed, caused by police change that causes current values of democracy level, there is collider bias.

We simulated to show that the previous specification, which did recover the actual causal effect, fails to do so with the inclusion of a control variable in the new DAG – we omitted the arrows from lagged democracy level to civil war to make things simple.



In the chart above, we ran a regression including lagged democracy level as a control. In the true Data Generating Process (DGP), it did not cause civil war. However, civil war caused the current values of democracy level, and an unobserved variable, U, caused democracy level and was caused by political changThethe DGP was the same in the next chart. However, we excluded democracy level from the regression. As we can see, the true causal effect of 5 is recovered on average.

Concluding Remarks

The existing literature has drawn the attention of applied researchers to challenges in the causal identification of standard Time Series Cross-Sectional (TSCS) regression models, particularly those related to heterogeneous treatment effects over time. In this article, we contribute to this literature by highlighting another prevalent issue: introducing a specific form of collider bias, which we refer to as "foreign collider bias." Through the application of Directed Acyclic Graphs (DAGs) in TSCS regression, we can determine when to include a variable in a regression model.

The conventional heuristics employed by researchers to decide which variables to control for are insufficient in avoiding the introduction of collider bias in TSCS data. The temporal dimension of the data offers the possibility of utilizing lagged values of variables to mitigate the problem of reverse causality. However, this seemingly advantageous feature opens up the potential for introducing a collider into the regression. Consequently, researchers must construct their DAGs while considering this possibility. When reviewing the literature to determine the controls to include, any variable susceptible to reverse causality should be treated as a potential collider. The review should encompass a discussion of the potential causes of these controls to ascertain if the treatment itself is also causing them.

Another important characteristic of TSCS data is that lagged variables can influence the current values of variables due to inertia or persistence effects over time. In such cases, it becomes evident from the DAG that excluding a lagged dependent variable from the model biases the results of the causal effect of the treatment of interest.

As researchers increasingly recognize the numerous advantages of utilizing DAGs in TSCS regression, we anticipate that it will become an indispensable tool for mitigating foreign collider issues and accurately specifying models in the presence of carryover effects.

References

Angrist, J. D., & Pischke, J. S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton university press.

Blattman, C., & Miguel, E. (2010). Civil war. Journal of Economic Literature, 48(1), 3-57.

Callaway, B. and Sant'Anna, P. H. (2021), 'Difference-in-differences with multiple time periods,' Journal of Econometrics 225(2), 200–230.

de Chaisemartin, C. and D'Haultfœuille, X. (2020), 'Two-way fixed effects estimators with heterogeneous treatment effects', American Economic Review 110(9), 2964–2996.

Franzese, R. J., & Hays, J. C. (2007). Spatial econometric models of cross-sectional interdependence in political science panel and time-series-cross-section data. Political analysis, 15(2), 140-164.

Cinelli, C., Forney, A., & Pearl, J. (2020). A crash course in good and bad controls. Available at SSRN 3689437.

Imai, K., & Kim, I. S. (2019). When should we use unit fixed effects regression models for causal inference with longitudinal data? American Journal of Political Science, 63(2), 467-490.

Heckman, J. J. (2008). Econometric causality. International statistical review, 76(1), 1-27.

Duarte, G., Finkelstein, N., Knox, D., Mummolo, J., & Shpitser, I. (2021). An automated approach to causal inference in discrete settings. arXiv preprint arXiv:2109.13471.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of causal discovery methods based on graphical models. Frontiers in Genetics, 10:524

Hegre, H. (2001). Toward a democratic civil peace? Democracy, political change, and civil war, 1816–1992. American political science review, 95(1), 33-48.

Philip G Wright. Tariff on animal and vegetable oils. Macmillan Company, New York, 1928.
Sewall Wright. The method of path coefficients. The annals of mathematical statistics, 5(3): 161–215, 1934.

Spirtes, P., & Zhang, K. (2016, December). Causal discovery and inference: concepts and recent methodological advances. In Applied informatics (Vol. 3, No. 1, pp. 1-28). SpringerOpen.

Newman, B. J., & Hartman, T. K. (2019). Mass shootings and public support for gun control. British Journal of Political Science, 49(4), 1527-1553.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. Review of Economics and statistics, 86(1), 4-29.

Dietrich, S. (2016). Donor political economies and the pursuit of aid effectiveness. International Organization, 70(1), 65-102.

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

Imbens, G. W. (2020). Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. Journal of Economic Literature, 58(4), 1129-1179.

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

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

Cinelli, C., Forney, A., & Pearl, J. (2021). A crash course in good and bad controls. Sociological Methods & Research, 00491241221099552.

Newman, B. J., & Hartman, T. K. (2019). Mass shootings and public support for gun control. British Journal of Political Science, 49(4), 1527-1553.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., & Zhang, A. (2021). A survey on causal inference. ACM Transactions on Knowledge Discovery from Data (TKDD), 15(5), 1-46.