Journal of Economics, Business & Organization Research

Year: 2025, Vol: 7, Issue: 1, pp. 21-39

Article Type: *Research*

Citation: Erbey, A. (2024). Econometric model specification and D-separation: A Monte Carlo experiment. *Journal of Economics, Business and Organization Research, 7*(1), 21-39. https://doi.org/10.66414/ebor.100346

ECONOMETRIC MODEL SPECIFICATION AND D-SEPARATION: A MONTE CARLO EXPERIMENT

Fernando VERA-SÁNCHEZ[1]

Received / Başvuru: 24.02.2025 **Accepted / Kabul:** 11.09.2025 **Published / Yayın:** 27.12.2025

Abstract

The purpose of this paper is to identify how Directed Acyclic Graphs perform in a commonly occurring setting in economics, namely the use of proxy variables. When estimating the coefficients that describe the relationships between variables in an economic model, theoretical foundations guide the selection of variables to include in the statistical framework. However, some important variables lack measurable scales. Variables such as education, intelligence, or price expectations are often key explanatory factors, but since they are not directly observable, researchers must choose between treating them as omitted variables or using proxy indicators like years of schooling, test scores, or expected future prices. The method of this research aims to compare the traditional use of proxy variables with the methodology based on Directed Acyclic Graphs. To highlight the distinctions between both approaches, using a Monte Carlo simulation. We were able to demonstrate in a Monte Carlo simulation that the proxy variable suggested by existing literature could be worse than the procedure suggested by Directed Acyclic Graphs, these findings are relevant for the specification of economic models. Our conclusions indicate that dealing with an unobservable variable requires careful consideration when introducing a proxy variable. While earlier studies recommend using a proxy to address the issue of missing variables, the Directed Acyclic Graphs method advises a thorough evaluation of the proxy's attributes beyond merely its strong correlation with the unobservable variable.

Keywords: *Proxy Variables, Directed Acyclic Graphs, Monte Carlo Simulation*

[1] Dr., Universidad Anáhuac Puebla, México, fernando.verasa@anahuac.mx, Orcid: 0000-0003-3501-6478

EKONOMETRİK MODEL BELİRLEME VE D-AYRIMI: BİR MONTE CARLO DENEYİ

Öz

Bu çalışmanın amacı, ekonomi literatüründe sıklıkla karşılaşılan bir durum olan vekil (proxy) değişkenlerin kullanımı bağlamında Yönlendirilmiş Çevrimsiz Grafiklerin (Directed Acyclic Graphs, DAG) nasıl bir performans sergilediğini incelemektir. Bir ekonomik modelde değişkenler arasındaki ilişkileri tanımlayan katsayılar tahmin edilirken, istatistiksel çerçeveye dâhil edilecek değişkenlerin seçimi kuramsal temellere dayanmaktadır. Ancak bazı önemli değişkenlerin ölçülebilir ölçekleri bulunmamaktadır. Eğitim, zekâ ya da fiyat beklentileri gibi değişkenler çoğu zaman temel açıklayıcı faktörlerdir; ne var ki doğrudan gözlemlenemedikleri için araştırmacılar bu değişkenleri ya dışlanmış değişkenler olarak ele almak ya da öğrenim süresi, test puanları veya beklenen gelecek fiyatlar gibi vekil göstergeler kullanmak arasında bir tercih yapmak zorunda kalmaktadır. Bu çalışmanın yöntemi, vekil değişkenlerin geleneksel kullanımını Yönlendirilmiş Çevrimsiz Grafiklere dayalı metodoloji ile karşılaştırmayı amaçlamaktadır. Her iki yaklaşım arasındaki farkları ortaya koymak için Monte Carlo simülasyonu kullanılmaktadır. Monte Carlo simülasyonu aracılığıyla, mevcut literatür tarafından önerilen vekil değişkenin, Yönlendirilmiş Çevrimsiz Grafikler tarafından önerilen prosedüre kıyasla daha kötü sonuçlar verebileceği gösterilmiştir. Bu bulgular, ekonomik modellerin belirlenmesi açısından önem taşımaktadır. Sonuçlarımız, gözlemlenemeyen bir değişkenle karşılaşıldığında vekil değişkenin modele dâhil edilmesinin dikkatli bir değerlendirme gerektirdiğine işaret etmektedir. Önceki çalışmalar, eksik değişken sorununu gidermek için vekil değişken kullanımını önermekteyken, Yönlendirilmiş Çevrimsiz Grafikler yaklaşımı, vekil değişkenin yalnızca gözlemlenemeyen değişkenle olan yüksek korelasyonuna değil, onun yapısal özelliklerine de kapsamlı biçimde dikkat edilmesini salık vermektedir.

Anahtar Kelimeler: *Vekil Değişkenler, Yönlendirilmiş Çevrimsiz Grafikler, Monte Carlo Simülasyonu*

1. INTRODUCTION

Directed Acyclic Graph (DAG) methods are techniques that may shed light on problems of model specification in applied econometrics. The basic or fundamental result of these efforts is embedded in the notion of d-separation, which formalizes notions of conditional independence between variables. The objective of this paper is to identify how directed acyclic graphs perform in a commonly occurring setting in economics, namely the use of proxy variables. Factors like education, intelligence, and price expectations frequently play a crucial role in explaining outcomes; however, because they cannot be directly observed, researchers face the decision of either considering them as missing variables or representing them through proxies such as years of education, exam results, or anticipated future prices. In order to provide evidence relative to this area of application, we make use of a Monte Carlo simulation. Monte Carlo simulation is a computational method that employs repeated random sampling to simulate and analyze complex systems or processes. It is used to estimate the probability distributions of possible outcomes when there is uncertainty or variability in the input variables. By running a large number of simulations, this technique helps in understanding the range of potential results, assessing risks, and making informed decisions in situations where deterministic models are difficult or impossible to apply.

The present study is justified, as considerable work in economics deals with the estimation of relationships between variables where uncertainty exists on proper specification or appropriate measures of variables. Leamer (1978: 4) defines the Axiom of Correct Specification as follows: a) the set of explanatory variables that are thought to determine (linearly) the dependent variable must be: 1) unique, 2) complete, 3) small in number, 4) observable; b) other determinants of the dependent variable must have a probability distribution with at most a few unknown parameters; c) all unknown parameters must be constant. He argues that this axiom is false in its application to a setting of non-experimental inference. Such inference is ubiquitous in applied economics and puts us in the realm of specification search. Directed graphs provide us with one way of conducting that search.

The literature supports using proxy variables when the true variable is unobservable, rather than treating it as missing. However, excluding a relevant variable causes bias and inconsistency (Greene, 2007; Griffiths et al., 1993; Pindyck and Rubinfeld, 1991), while including an irrelevant variable may reduce efficiency despite unbiased coefficients.

Aigner (1974), McCallum (1972), and Wickens (1972) argue that using a poor proxy is preferable to omitting an important variable; Wickens (1972: 760) states that even a weak proxy is better than none. Several studies have examined proxy properties (Gauger, 1989; Greene, 2007; Krasker and Pratt, 1986; Kakimoto and Ohtani, 1985; Murphy and Topel, 1985; Ohtani, 1985; Oxley and McAleer, 1993; Pagan, 1984). More recently, Bollinger and Minier (2015) explored multiple proxies for unobserved variables and established parameter bounds, while de Luna et al. (2017) identified conditions for nonparametric causal inference using proxies. Even when the proxy poorly approximates the unobservable variable and is irrelevant, its negative effect on regression is usually minor (Pindyck and Rubinfeld, 1991).

This paper is organized as follows: In the first section we define the concept of proxy variables, bring forward the concept of DAG and d-separation and suggest why they may be important in economics. Section two develops a model of proxy variables use as a suggested procedure to solve the problem of identification. Also, literature is reviewed in this section. In the third section, we present a simulation design, and Monte Carlo results are provided to demonstrate the usefulness of directed graphs. Results show how recommendations from the literature may be misleading and how directed graphs can improve results. Finally, we draw some general conclusions and offer recommendations.

2. DEFINITIONS

When speaking of the proxy variable, Kennedy (1998) maintains that "often an explanatory variable is unobservable, but a proxy for it can be constructed". We explore the use of proxy variables with the tools of directed acyclic graphs and Monte Carlo analysis. Existing econometrics literature generally recommends the unconditional use of proxy variable when measures of the more theoretically appropriate variable are not available. We demonstrate that this recommendation is not general and, in some cases, described in terms of d-separation, a proxy variable may make matters worse in terms of properties unbiasedness in relation to other coefficient estimates.

When economists want to estimate the coefficients representing the relationships between variables in an economic model, they make use of theory in order to specify the variables to include in the statistical model. However, there are some variables for which scales of measurement do not exist. Variables such as education, intelligence or price expectations could be identified as a major explanatory variable in economic models, but, since these variables are not directly observed, we have to decide between considering this as a missing variable problem or using such proxies as schooling, test scores or future prices instead. The objective is to compare the traditional approach in the use of proxy variables with the approach suggested by directed acyclic graphs. After contrasting these techniques, we make use of a Monte Carlo experiment to illustrate the difference in each technique. We then draw general conclusions and provide recommendations for the use of proxy variables in economics.

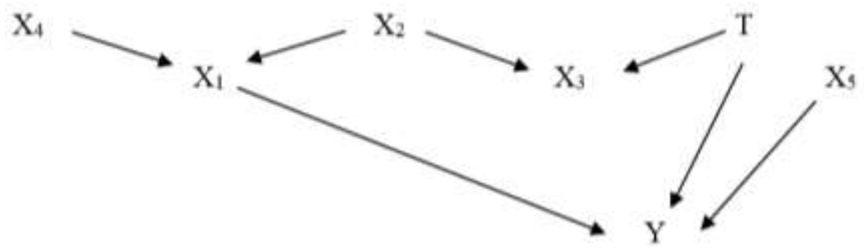
A directed acyclic graph is a picture representing the causal flow of a set of variables. These models represent conditional independence. Pearl (1995) proposed d-separation as a graphical characterization of conditional independence. The main results from d-separation are illustrated by the fact that conditioning of common causes blocks the flow of information between common effects. Spirtes, Glymour, and Scheines (1993) incorporated the notion of d-separation into an algorithm (PC algorithm) for building directed acyclic graphs. The PC algorithm is an ordered set of commands which begins with a general unrestricted set of relationships among variables and proceeds step-wise to remove edges between variables and to direct "causal flow." The algorithm is described in detail in Spirtes, Glymour, and Scheines (1993: 117). PC algorithm and its more refined extensions are marketed as the TETRAD software (Scheines et al., 1994).

There are studies which identified conditions for the total effect between variables using directed acyclic graphs, see Kuroki and Pearl (2014), Pearl 2009, and Stanghellini (2004). In the last case, the author uses instrumental variables and DAG to identify relevant parameters with one explanatory variable. Also, Bellemare et al. (2017) specified the conditions under which lagged explanatory variables are appropriate responses to endogeneity concerns using DAG. Sperrin and Martin (2020) conducted a simulation study to determine when the use of a missing indicator, combined with multiple imputation, would reduce bias for causal effect estimation. Finally, Shingaki et al. (2021) proposed an integrated generalized instrumental variable estimator, based on DAG, for severable variables.

3. SUGGESTED PROCEDURE

When theory suggests that it is relevant to include a variable in an economic model but it is unobservable, we may be tempted to use another variable as a proxy. The case we focus on is represented in Figure 1, using a graph as given in Spirtes, et al (1995).

Figure 1. DAG on Six Observable Variables and One Unobservable Variable



In this model we see that variable Y is the result of the influence of three variables X_1, X_5 and T; however, T is not observable. Since X_3 is a child of T, we may be tempted to use X_3 as a proxy variable for T.

The temptation to use a proxy variable is supported by the literature. If we know that we are unable to include the true variable, we can ignore this influence and consider the case of missing variables. However, as many textbooks show the cost of excluding a variable that should appear in the model is, on the one hand, bias and inconsistency, see Greene (2007), Griffiths, Hill and Judge (1993), Pindyck and Rubinfeld (1991). On the other hand, the cost of adding an irrelevant variable may lead to the loss of efficiency even when the coefficient is unbiased.

Aigner (1974), McCallum (1972) and Wickens (1972) conclude that it is better to use a poor proxy than ignore the presence of a relevant variable in the regression equation. Wickens (1972: 760) writes: "*Thus it is better to use even a poor proxy than to use none at all and omit the unobservable variable.*" Other studies have estimated the characteristics of this proxy variable, see in particular, Gauger (1989), Greene (2007), and Krasker and Pratt (1986), Kakimoto and Ohtani (1985), Murphy and Topel (1985), Ohtani (1985), Oxley and McAleer (1993), Pagan (1984). Recently, Bollinger and Minier (2015) considered the use of multiple proxy measures for an unobserved variable. They derived a set of bounds for all parameters in the model. de Luna et al., (2017) considered conditions to estimate proxy variables where causal effects are nonparametrically identified.

In regards to proxy and instrumental variables, Hu and Schennach (2008) proposed an approach that relies on the eigenvalue-eigen function decomposition of an integral operator associated with specific joint probability densities, in order to deal with instrumental variables. Fé (2012) considered heteroskedasticity adaptive error component models, while estimating instrumental variables. Galvao et al. (2017) proposed an alternative solution to the endogeneity problem by explicitly modeling the joint interaction of the endogenous variable, whereas Hu and Sasaki (2018) derived a closed-form identification of the model structure by explicitly solving integral equations. Namba and Xu (2018) considered the condition when some relevant regressors are unobservable: they derive the explicit formula of predictive mean squared error of a general family of shrinkage estimators. Sajons (2020) proposed the experimentally randomized instrumental variable to adjust the endogeneity bias. Rothenhäusler et al. (2021) uses anchor regression to predict a response. Papadopoulos (2022) proposes likelihood inference and Gaussian Copula, in order to account for endogeneity in linear regression models without the need for instrumental variables. Ghosh et al. (2021) consider the case of the heterogeneous endogeneity to increase the precision over the traditional instrumental variable, while Kim and Petrin (2022) propose the conditional moment restrictions with control function estimators to deal with the problem of endogeneity in non-parametrically instrumental variables.

Even in the case when the proxy variable is a poor approximation of the unobservable variable that becomes irrelevant in the model, the negative effect in the regression will not be severe. Pindyck and Rubinfeld (1991: 165) describe this situation as follows:

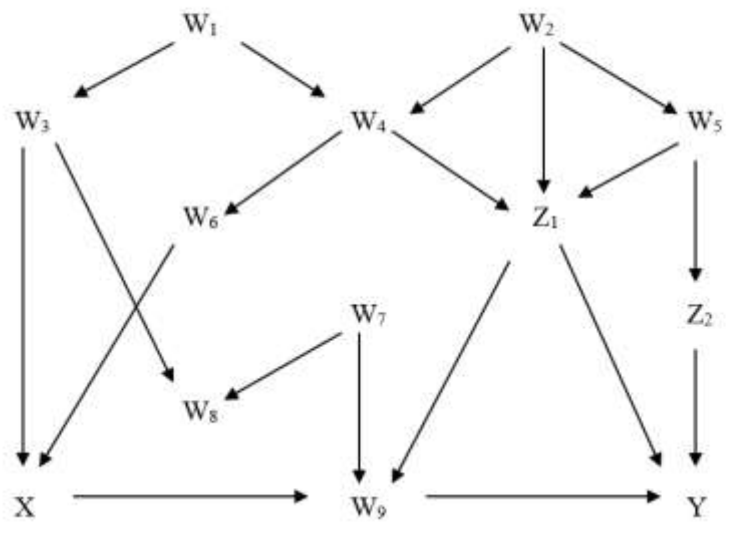
"*The inclusion of an irrelevant variable does not bias the slope parameter estimates of any of the slope variables which appear in the true model. It is not difficult to show that the intercept of the equation is unbiased as well and that the estimate of the coefficient of [the irrelevant variable] will have an expected value of 0.*"

In the specific case of Figure 1, the existing literature would suggest the use of X_3 as a proxy for the unobservable variable T. The result of this decision would appear as follows: consider that one is interested in unbiased ordinary least squares (OLS) estimates of coefficients (α_i) associated with a causal model: $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_5 + \alpha_3 T$. If Y, X_1 and X_5 are observable, while T is not, but $X_3 = g(T)$ is observable, then X_3 can be used as a proxy variable for T. Since X_3 is related with T, and T is a cause of X_3, we could assume that the estimated model will result in unbiased OLS estimates of the coefficients associated with X_1 and X_5, if X_3 is d-separated from X_1 and X_5. If X_3 is not d-separated from X_1 or X_5, the ordinary least squares estimates will be biased. This assumption is related to the analysis that we will present with a simulation in the next section.

Pearl (1996) defines the decision of whether to include a variable in an estimation as a statistical adjustment problem. The following example (taken from Pearl, 2009) will clarify the directed graph approach in this case. Consider an observational study where we wish to find the effect of X on Y, for example, treatment on response. We can think of many factors that are relevant to the problem: some are affected by the treatment, while some are affecting the treatment and some are affecting both treatment and response. Some of these factors may be unmeasurable, such as genetics, traits or lifestyle, while others are measurable, such as gender, age, and salary level. Our challenge is to select a subset of these factors for measurement and adjustment, namely, if we compare subjects under the same value of those measurements and average, we get the correct result. From the graph depicted in Figure 2, the procedure suggested to accomplish this task shows the following:

Test if Z_1 and Z_2 are sufficient measurements,
to find the effect of X on Y.
Step 1: Z_1 and Z_2 should not be descendants of X and Y.
Step 2: Delete all non-ancestors of {X, Y, Z}
Step 3: Delete all arc emanating from X.
Step 4: Connect any two parents sharing a common child.
Step 5: Strip arrow-heads from all edges.
Step 6: Delete Z_1 and Z_2.
Test: If X is disconnected from Y in the remaining graph,
then Z_1 and Z_2 are appropriate measurements.

Figure 2. DAG of Pearl to Illustrate Adjustment Problem, Pearl (2009)



4. SIMULATION DESIGN AND RESULTS

Consider the example given in Figure 1 applying the d-separation procedure. Should we include X_3 in the regression of X_1 on Y? We conclude that X_3 is not an appropriate measure in the estimation of X_1 on Y, meaning that we should run a regression with X_1 and X_5 alone. See the suggestion from the existing literature that even a poor proxy like X_3 is better than ignoring a relevant variable T.

Since literature and directed graphs suggest different approaches in the use of proxy variables in this case, we generated a Monte Carlo simulation model, in order to identify which procedure is better.

Assume the following model:

$$X_4 = \varepsilon_4$$
$$X_2 = \varepsilon_2$$
$$X_5 = \varepsilon_5$$
$$T = \varepsilon_T$$
$$Y = a_1 X_1 + a_2 X_5 + a_3 T + \varepsilon_y$$
$$X_1 = a_4 X_2 + a_5 X_4 + \varepsilon_1$$
$$X_3 = a_6 X_2 + a_7 T + \varepsilon_3$$

Assume that all exogenous variables, including the error terms, are independent and distributed normally with zero mean and unit variance, and that none of the a_i vanish. We generate each one of these variables with 1,000 data points using Excel random generation number and specified parameters, like number 4 for the relationship for coefficient (a_3). The causal graphical representation of this model is described in Figure 1.

From here we run 16 different regressions from the combination of Y on X_1 and the inclusion or exclusion of all the other observed variables. We run each regression 100 times using EViews and then we calculate the estimated coefficient and the standard error at its means. We focus our attention on the following two questions. Which regression gives an unbiased consistent estimator of the coefficient of X_1 on Y? Also, which regressions give a significantly different from zero coefficient of X_3 on Y?

The results of these regressions are shown in Table 1. For each one of the 16 models we include the dependent variable Y and, as explanatory variables, a constant, X1 and a combination of variables X2 to X5. The last two columns in the table answer the questions above. Table 1 presents the comparative results obtained from each of the models. The statistical significance of the coefficients is evaluated using the p-value, with a threshold set at the 0.05 significance level. This means that if the p-value associated with a given coefficient is less than 0.05, we can conclude that the coefficient is statistically significant, implying that there is strong evidence against the null hypothesis. Additionally, we assess whether the estimators are biased or unbiased by analyzing the estimated value of parameter number 4, as defined in the Monte Carlo simulation framework. This parameter serves as a benchmark to determine the accuracy and reliability of the estimations provided by each model.

Table 1. Summary Results in Monte Carlo Experiments

Model	X_1	X_2	X_3	X_4	X_5	X_1 on Y	X_3 on Y
1	✓					unbias	
2	✓	✓				unbias	
3	✓		✓			**bias**	Significant
4	✓			✓		unbias	
5	✓				✓	unbias	
6	✓	✓	✓			unbias	Significant
7	✓	✓		✓		unbias	
8	✓	✓			✓	unbias	
9	✓		✓	✓		**bias**	Significant
10	✓		✓		✓	**bias**	Significant
11	✓			✓	✓	unbias	
12	✓	✓	✓	✓		unbias	Significant
13	✓	✓	✓		✓	unbias	Significant
14	✓	✓		✓	✓	unbias	
15	✓		✓	✓	✓	**bias**	Significant
16	✓	✓	✓	✓	✓	unbias	Significant

We can see that there are four cases where the coefficient of X_1 on Y is biased. These cases coincide with the inclusion of X_3. In each case where X_3 is included, the OLS gives a significantly different from zero coefficient to this variable, even when we know that is not the case in the true model.

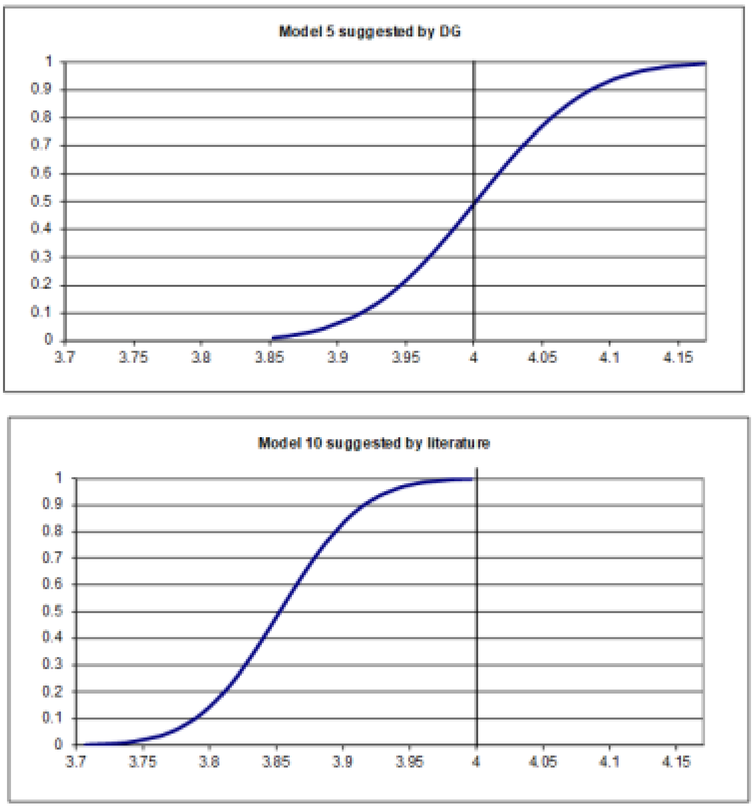
From literature we could suggest a model like the one described as number 10, which includes X_1, X_5 and X_3 as a proxy of the unobservable variable T. However, this model gives a biased coefficient of X_1 on Y and a significant coefficient of X_3. The d-separation approach will suggest not to include X_3 on the regression and use a model such as number 5, which gives an unbiased estimator of X_1 on Y.

Another question arises from this table. We have four cases where the inclusion of X_3 will not cause any harm, i.e. even when X_3 is significant, the coefficient of X_1 on Y is unbiased.

The answer is again due to the concept of d-separation. In these four cases we are including the variable X_2, which, as we have seen before, d-separates the effects of X_1 and T, breaking the correlation between these two variables.

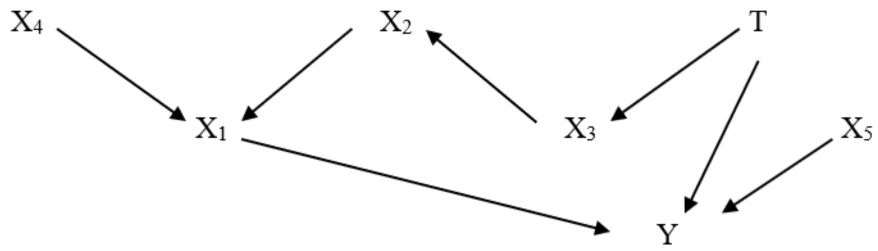
The problem with the traditional approach is that it assumes that all regressors are correlated, and it worries about correlation, but we can see that it is the conditional correlation that is significant, a concept explored in directed graphs. Some may argue that bias is not the best criteria to test the different approaches studied here. In Figure 3 we estimate the cumulative distribution function in order to compare bias and consistency. The simulation parameter is 4, in model 5 is consistent, in model 10 is biased.

Figure 3. Cumulative Distributions on Realizations of Parameters Estimates for Model 5 and 10; True Value of B is 4.0 in both Simulations



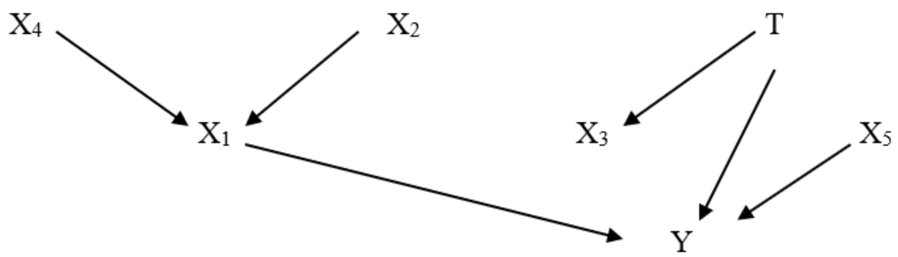
Other interesting results from this simulation are found when we change the causal relationship of X_3 with X_2 and X_1. If we turn to the model described in Figure 4, where X_3 is now a cause of X_2, we find that, in this case, there are four cases out of the 16 regressions where the coefficient of X_1 on Y is biased. These cases do not include X_2 or X_3, which could be used as a proxy variable for T, suggesting that ignoring the presence of the relevant variable T will generate biased estimates, as suggested in literature.

Figure 4. DAG with Connection between X_1 and X_3



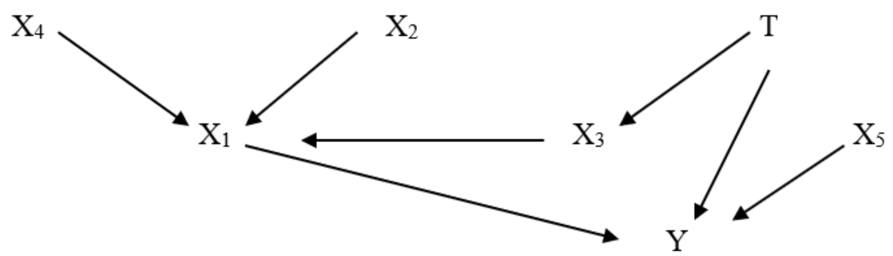
We could also have the case where there is no relationship between X_3 and X_2, as described in Figure 5. As the existing literature suggests, this is the only case where we can ignore the presence of T, since X_1 and T are not correlated. Each one of the 16 regressions gives unbiased estimates of X_1 on Y.

Figure 5. DAG without Connection between X_1 and X_3



The last case studied in this section is the one described in Figure 6, where X_3 is a direct cause of X_1. Even when the inclusion of X_1 and X_3 in the regression of Y generates multicollinearity, we can say that X_3 d-separates the effect of X_1 and T. The results from the simulation show that every time we include X_3 as a proxy of T, these models give an unbiased coefficient, and the models that include X_3 are the only unbiased models.

Figure 6. DAG with X_1 Dependent on X_3



However, we may need to explain why, if the true model is described in the following equation: $Y = X_1\beta_1 + X_5\beta_5 + T\beta_T + \varepsilon_1$, it is possible to find unbiased estimates of β_1 using the following model, which ignores the presence of two relevant variables: $Y = X_1\beta_1 + \varepsilon_2$. In order to answer this question, we will make use of the OLS estimate of β_1. By definition, the estimate b_1 is:

$$
\begin{aligned}
b_1 &= \left(X_1^{'}X_1\right)^{-1} X_1^{'}Y \\
&= \left(X_1^{'}X_1\right)^{-1} X_1^{'}\left(X_1\beta_1 + X_5\beta_5 + T\beta_T + \varepsilon_1\right) \\
&= \left(X_1^{'}X_1\right)^{-1} X_1^{'}X_1\beta_1 + \left(X_1^{'}X_1\right)^{-1} X_1^{'}X_5\beta_5 + \left(X_1^{'}X_1\right)^{-1} X_1^{'}T\beta_T + \left(X_1^{'}X_1\right)^{-1} X_1^{'}\varepsilon_1
\end{aligned}
$$

Taking expectations, it simplifies to

$$E[b_1] = \beta_1 + \left(X_1^{'}X_1\right)^{-1} X_1^{'}X_5\beta_5 + \left(X_1^{'}X_1\right)^{-1} X_1^{'}T\beta_T$$

since we know that $\beta_5 \neq 0$ and $\beta_T \neq 0$. The only case where $E[b_1]$ is unbiased is the case where $\left(X_1^{'}X_5\right)$ and $\left(X_1^{'}T\right)$ are zero. The case where $\left(X_1^{'}X_5\right)$ goes to zero can be clearly identified in Figure 1, where there is no relationship between these two variables. Both are parents of variable Y, but they are d-separated by this variable, while X_1 and X_5 are independent of each other.

The case of X_1 and T is more complicated. We could have different cases that explain the correlation between X_1 and T, as we saw in Figures 1, 4, 5 and 6. The correlation between X_1 and T for Figures 1 and 5 is –0.0029 and 0.0035 respectively, which explains the unbiased estimate of β_1 by OLS. Figure 5 shows no connection between X_1 and T, except that both are parents of variable Y. Figure 1 shows a connection between X_1 and T through X_2 and X_3. However, X_3 is breaking the correlation between X_1 and T (this is what d-separation does), and this is also why conditional correlation becomes relevant. X_1 and T are d-separated. Given X_3, the correlation goes to zero and the OLS estimate of β_1 is unbiased.

In the case of Figures 4 and 6, the correlation between X_1 and T is 0.4512 and 0.4435. There is no d-separation between these variables; they are d-connected; the correlation is high, and the estimates of β_1 by OLS ignoring a proxy variable are biased. The only difference between Figure 1 and 4 is the direction of the arrow from variable X_3 to X_2. It seems like a small detail since X_3 is still in a direct relationship with T and could still be considered a good proxy variable. However, the change in direction changes the correlation between X_1 and T and the biasedness of the OLS coefficient (b_1). In Figure 1, X_1 is d-separated from T. In Figure 4, X_1 is d-connected with T.

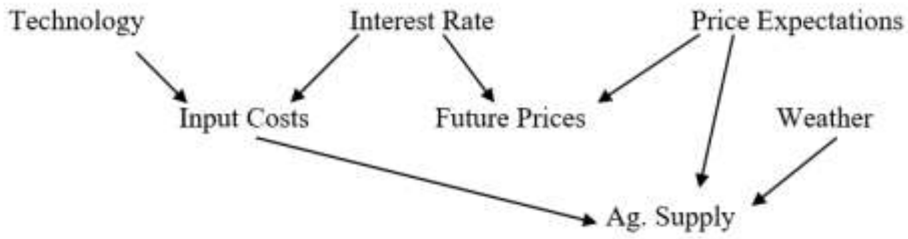
5. CONCLUSIONS

When estimating relationships in economic models, theory guides variable selection. However, some key factors—like education, intelligence, or price expectations—cannot be directly measured. As a result, researchers must either treat them as omitted variables or use proxies such as years of schooling, test scores, or expected prices.

Our results suggest that when we face the problem of an unobservable variable, caution must be exercised in the introduction of a proxy variable. Previous literature suggests the inclusion of a proxy in order to avoid the problem of missing variables. However, the d-separation approach emphasizes the importance of thoroughly evaluating the characteristics of a proxy variable beyond merely its strong correlation with the unobservable variable it aims to represent. Simply relying on high correlation can be misleading and may overlook other critical aspects that affect the quality of the estimation. To illustrate this point, we conducted a detailed Monte Carlo simulation based on the example presented in Figure 1. In this simulation, we generated a dataset consisting of 1,000 observations for five variables and estimated 16 different regression models to analyze the effects of including the proxy variable. The results revealed that, contrary to common intuition, incorporating the proxy variable does not always improve the model. In fact, it can sometimes introduce bias or distort the estimation of the specific coefficient of interest, ultimately doing more harm than good. This finding highlights the necessity of carefully scrutinizing proxy variables before their inclusion in empirical models, as improper use may compromise the validity of the conclusions drawn.

As an example of political implications, this problem can be adapted to an economic model like the one described in Figure 7. In this case, we can assume that agricultural supply depends on variables such as input costs, weather and price expectations. However, we cannot generate a measurement of expectations. A good proxy could be future prices, since we assume that this variable reflects the behavior of price expectations. However, we also need to remember that future prices could be the reflection of other variables such as interest rates.

Figure 7. A Hypothetical Example from Agriculture



The recommendation generated by directed graphs on dealing with proxy variables is the following. If we want to estimate a specific coefficient representing the relationship between variables, we could graph the causal flow of all these variables making use of the Tetrad software. From here we can apply the d-separation technique to identify which variables are relevant to include in a regression.

A simple rule suggested is as follows: include the proxy variable if it is d-separated from the observed relevant variable, but do not include the proxy variable if is not d-separated from the observed relevant variable.

The present study is not without certain limitations; therefore, several aspects should be addressed in future research. First, although we have shed light on the use of proxy variables in economics, the broader application and generalization of the Directed Acyclic Graph (DAG) technique may require more time and further development. Second, it would be valuable to compare the results of proxy variable usage across different studies, especially within the context of public policy research, to better understand their practical implications and effectiveness. Future work could also explore alternative methodologies for handling unobservable variables and investigate how these approaches perform in various empirical settings.

References

Aigner, D. J. (1974) MSE dominance of least squares with errors of observation. *Journal of Econometrics, 2*(4), 365-372. https://doi.org/10.1016/0304-4076(74)90020-7

Bellemare, M. F., Masaki, T., & Pepinsky, T. B. (2017). Lagged explanatory variables and the estimation of causal effect. *Journal of Politics, 79*(3), 949-963. https://doi.org/10.1086/690946

Bollinger, C. R., & Minier J. (2015). On the robustness of coefficient estimates to the inclusion of proxy variables. *Journal of Econometric Methods, 4*(1), 101-122. https://doi.org/10.1515/jem-2012-0008

de Luna, X., Fowler, P., & Johansson, P. (2017). Proxy variables and nonparametric identification of causal effects. *Economics Letters, 150*, 152-154. https://doi.org/10.1016/j.econlet.2016.11.018

Fé, E. (2012). Instrumental variable estimation of heteroskedasticity adaptive error component models. *Statistical Papers, 53*(3), 577-615. https://doi.org/10.1007/s00362-011-0366-5

Galvao, A. F., Montes-Rojas, G., & Song, S. (2017). Endogeneity bias modeling using observables. *Economics Letters, 152*, 41-45. https://doi.org/10.1016/j.econlet.2016.12.021

Gauger, J. (1989). The generated regressor correction: Impacts upon inferences in hypothesis testing. *Journal of Macroeconomics, 11*(3), 383-95. https://doi.org/10.1016/0164-0704(89)90065-7

Ghosh, P., Grier, K., & Kim, J. (2021). Heterogeneous endogeneity. *Statistical Papers, 62*(2), 847-886. https://doi.org/10.1007/s00362-019-01116-9

Greene W. H. (2007). *Econometric analysis* (6th edition). Prentice Hall.

Griffiths, W. E., Hill, R. C., & Judge, G. G. (1993). *Learning and practicing econometrics*. John Wiley & Sons.

Hu, Y., & Sasaki, Y. (2018). Closed-form identification of dynamic discrete choice models with proxies for unobserved state variables. *Econometric Theory, 34*(1), 166-185. https://doi.org/10.1017/S0266466617000081

Hu, Y., & Schennach, S. M. (2008). Instrumental variable treatment of nonclassical measurement error models. *Econometrica, 76*(1), 195-216. https://doi.org/10.1111/j.0012-9682.2008.00823.x

Kakimoto, S., & K. Ohtani. (1985). On the use of proxy variable in the test of homoscedasticity. *Economic Letters, 18*(2-3), 153-156. https://doi.org/10.1016/0165-1765(85)90171-5

Kennedy, P. (1998). *A guide to econometrics* (4th edition). MIT Press.

Kim, K. I., & Petrin, A. (2022). A generalized non-parametric instrumental variable-control function approach to estimation in nonlinear settings. *Journal of Econometric Methods, 11*(1), 91-125. https://doi.org/10.1515/jem-2021-0038

Krasker, W. S., & Pratt, J. W., (1986). Bounding the effects of proxy variables on regression coefficients. *Econometrica, 54*(3), 641-655. https://doi.org/10.2307/1911312

Kuroki M., & Pearl J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika, 101*(2), 423-437. https://doi.org/10.1093/biomet/ast066

Leamer, E. E. (1978). *Specification searches ad hoc inference with nonexperimental data*. John Wiley & Sons.

McCallum, B. T. (1972). Relative asymptotic bias from errors of omission and measurement. *Econometrica, 40*(4), 757-758. https://doi.org/10.2307/1912970

Murphy, K., & Topel, R. (1985). Estimation and inference in two step econometric models. *Journal of Business & Economic Statistics, 3*(4), 370-379. https://doi.org/10.1080/07350015.1985.10509471

Namba, A., & Xu, H. (2018). PMSE dominance of the positive-part shrinkage estimator in a regression model with proxy variables. *Journal of Statistical Computation & Simulation, 88*(15), 2893-2908. https://doi.org/10.1080/00949655.2018.1491576

Ohtani, K. (1985). A note on the use of a proxy variable in testing hypotheses. *Economics Letters, 17*(1-2), 107-110. https://doi.org/10.1016/0165-1765(85)90137-5

Oxley, L., & McAleer, M. (1993). Econometric issues in macroeconomic models with generated regressors. *Journal of Economic Surveys, 7*(1), 1-40. https://doi.org/10.1111/j.1467-6419.1993.tb00158.x

Pagan, A. R. (1984). Econometric issues in the analysis of regressions with generated regressors. *International Economic Review, 25*(1), 221-247. https://doi.org/10.2307/2648877

Papadopoulos, A. (2022). Accounting for endogeneity in regression models using copulas: A step-by-step guide for empirical studies. *Journal of Econometric Methods, 11*(1), 127-154. https://doi.org/10.1515/jem-2020-0007

Pearl J. (2009). *Causality: Models, reasoning and inference* (2nd edition). Cambridge University Press.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika, 82*(4), 669-688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (1996). *The art and science of cause and effect* (Technical Report No. R-248). Cognitive Science Laboratory, University of California, Los Angeles.

Pindyck, R. S., & Rubinfeld, D. L. (1991). *Econometric models and economic forecasts*. McGraw Hill.

Rothenhäusler, D., Meinshausen, N., Bühlmann, P., & Peters, J. (2021). Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 83*(2), 215-246. https://doi.org/10.1111/rssb.12398

Sajons, G. B. (2020). Estimating the causal effect of measured endogenous variables: A tutorial on experimentally randomized instrumental variables. *Leadership Quarterly, 31*(5), 101348. https://doi.org/10.1016/j.leaqua.2019.101348

Scheines, P., Spirtes, P., Glymour, C., & Meek. C. (1994). *Tetrad II user's manual and software*. Lawrence Erlbaum Associates, Inc.

Shingaki, R., Kanda, H., & Kuroki, M. (2021). Selection and integration of generalized instrumental variables for estimating total effects. *Statistical Papers, 62*(5), 2355-2381. https://doi.org/10.1007/s00362-020-01190-4

Sperrin, M., & Martin, G. P. (2020). Multiple imputation with missing indicators as proxies for unmeasured variables: simulation study. *BMC Medical Research Methodology, 20*(1), 185. https://doi.org/10.1186/s12874-020-01068-x

Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. In P. Besnard & S. Hanks (Eds.), *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (pp. 491-498). Morgan Kaufmann Publishers.

Spirtes, P., Glymour, C., & Scheines. R. (1993). *Causation, prediction, and search*. Springer. https://doi.org/10.1007/978-1-4612-2748-9

Stanghellini E. (2004). Instrumental variables in Gaussian directed acyclic graph models with an unobserved confounder. *Environmetrics, 15*(5), 463-469. https://doi.org/10.1002/env.676

Wickens, M. R. (1972). A note on the use of proxy variables. *Econometrica, 40*(4), 759-761. https://doi.org/10.2307/1912971

Declaration of Contribution Rate: The entire study has been prepared only by the responsible author.
Declaration of Support and Appreciation: The research did not receive any support from any institution or organisation.
Declaration of Conflict: The author declares that there is no conflict of interest.
In this study, the rules stated in the **"Higher Education Institutions Scientific Research and Publication Ethics Directive"** were followed.
This article has been screened with **similarity** detection software.