Economics Bulletin

Volume 37, Issue 3

Stop breaking down: A graphical analysis of proxy variable and instrumental variable solutions to omitted variable problems

Burkhard Raunig
Central Bank of Austria

Abstract

This graphical analysis discusses cases where proxy variable and instrumental variable solutions to omitted variable problems may break down. The analysis shows that the effectiveness of proxies and instruments depends crucially on the precise causal link between the omitted variable and the explanatory variables included in a structural model.

Burkhard Raunig, Central Bank of Austria, Economic Studies Division, Otto-Wagner-Platz 3, 1090 Vienna, Austria, Phone +43 1 404 20 7219, e-mail: burkhard.raunig@oenb.at. The views expressed in this paper do not necessarily reflect those of the Central Bank of Austria or the Eurosystem.

Citation: Burkhard Raunig, (2017) "Stop breaking down: A graphical analysis of proxy variable and instrumental variable solutions to omitted variable problems", *Economics Bulletin*, Volume 37, Issue 3, pages 1996-2003
Contact: Burkhard Raunig - burkhard.raunig@oenb.at.
Submitted: December 30, 2016. **Published:** August 31, 2017.

1 Introduction

Researchers often face the problem that an important control variable cannot be included in an empirical model to be estimated. Simply ignoring the problem and omitting the control variable may lead to biased and inconsistent estimates of the model coefficients of interest. Consistent estimates may, however, be obtained by using a proxy for the omitted variable or by using instruments for explanatory variables that might be correlated with the omitted variable (e.g. Wooldridge 2010, Ch 4 and Ch 5).[1]

This note examines some cases where proxy and instrumental variable strategies for solving omitted variable problems may break down. These cases have to the best knowledge of the author not been discussed elsewhere. One aim of this note is to provide such a discussion. Another aim is to highlight the usefulness of graphical methods for solving identification problems in structural models.[2]

The analysis in this note builds on the graphical framework outlined in Pearl (2009). This framework has two attractive features that greatly simplify the analysis of structural models: Firstly, structural models are mapped into graphs that make statistical assumptions and presumed causal links between variables explicit. Secondly, simple path tracing rules can be applied to such graphs to check for identification of parameters when the model is linear. The analysis draws also heavily on Pearl (2013) who analyzes a number of issues in causal modeling within this graphical framework.

The graphical analysis that follows shows that proxy variable solutions will break down when the omitted variable has a direct causal link with other explanatory variables. Instrumental variable (IV) strategies for estimating the direct causal effect of the instrumented explanatory variable can fail if the explanatory variable causes an omitted mediating variable.

2 Causal graphs and path tracing rules

This section introduces the tools for analyzing the graphs that appear in this note. Pearl (1995), Pearl (2009), and Chen & Pearl (2014) provide extensive outlines of graphical methods.

Figure 1 shows five graphs. Solid nodes represent observed variables, hollow nodes represent unobserved variables, solid arrows indicate causal links, and curved dashed bi-directed arrows

[1] Panel data offer additional options for solving omitted variable problems. This note does not consider panel data.

[2] Graphical methods for analyzing causal relationships are well known in computer science and statistics but largely unknown in economics.

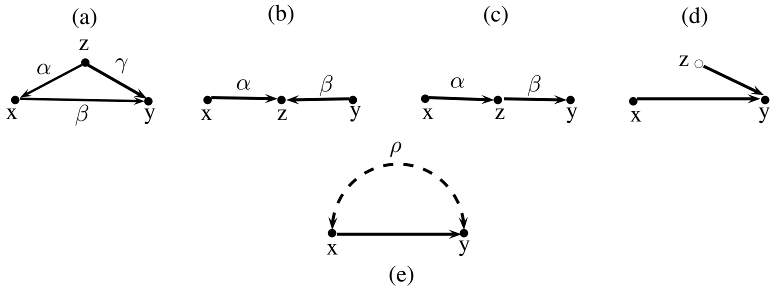


Figure 1: Confounder (a), collider (b), mediator (c), unobserved cause (d), joint unobserved causes (e).

indicate covariances that arise from unspecified causes. Thus, all variables in graphs (a), (b), and (c) are observed, z is unobserved in (d), and x causes y in graph (e) but x and y are also correlated because of neglected causes.

A *path* is a sequence of nodes connected by arrows. A path is d-connected if it does not traverse any collider.[3] A variable is a *collider* on a path if two arrows are pointing into it. Thus, the paths $x \to y$ and $x \leftarrow z \to y$ in (a) are d-connected. Furthermore, variables like z and y in (c) are also called descendants of x. The following rules make the concepts of d-separation and d-connection more precise.

d-separation: A path between two variables x and y can be d-separated (or blocked) by a set of nodes Z in two ways. Either (1) the path contains a chain $x \longrightarrow m \longrightarrow y$ or a fork $x \longleftarrow m \longrightarrow y$ such that the middle node m is in the conditioning set Z. Or (2) the path contains a collider $x \longrightarrow m \longleftarrow y$ such that the middle node m (or any of its descendants) is *not* in the conditioning set Z (see Pearl 2009, p16-17).

Thus, the path $x \leftarrow z \to y$ in (a) becomes d-separated once we condition on z (i.e. know the value of z). The path $x \to z \leftarrow y$ in (b) is blocked or d-separated as long as we are *not* conditioning on the collider z.

d-connection: A path between x and y is d-connected conditional on a set of nodes Z if (1) there is a collider-free path between x and y that traverses no member of Z, or (2) a collider (or one of its descendants) is in the conditioning set Z (see Chen & Pearl 2014, p7).

Hence, x and y are d-connected conditional on z in case (a) in Figure 1 whereas x and y become d-connected in case (b) once we condition on z.

Two *path tracing rules* (Wright 1921, 1934) that follow from covariance mathematics yield analytical expressions for covariances between variables in causal graphs (see also Goldberger

[3] The d stands for dependence.

(1972) and Bollen (1989), Ch. 2). Let $\pi_i = c_1 \cdot c_2 \cdot ... \cdot c_n$ be the product of the coefficients along a path i that d-connects two variables x and y. The c_j are either structural coefficients like β in graph (a) or covariances like ρ in graph (e).

The *first rule* states that the covariance between x and y is $\sigma_{xy} = \Sigma_i \pi_i$, i.e. the sum of the π_i over the different d-connected path between x and y. This rule applies when all variables have been standardized (i.e. normalized to have zero mean and unit variance).

The *second rule* states that the product π_i associated with a path between non-standardized variables must be multiplied by the variance of the variable from which the path originates. We will only need the first rule because we will always work with standardized variables to keep the algebra simple.

3 Solutions to the omitted variables problem

Let us now consider a structural model

$$y = \beta x + \gamma q + u \tag{1}$$

where y is determined by the variables x, q, and an error term u. This simple model suffices to outline the basic issues. We are interested in the coefficient β that measures the effect of x on y. As just mentioned, we assume for convenience that all variables have been standardized. Thus, $\sigma_y^2 = \sigma_x^2 = \sigma_q^2 = 1$. Furthermore, the analysis is in terms of population moments.

3.1 Ignoring the omitted variable

Let us start with the case where x and q in equation (1) can be observed (see also Pearl 2013, p158-159). Case (a) in Figure 2 shows the corresponding graph. Two paths d-connect x with y, the direct path $x \rightarrow y$, and the "backdoor" path $x \leftarrow q \rightarrow y$. Conditioning on q blocks this backdoor path. The ordinary least squares (OLS) formula for β in the regression of y on x and q is

$$\beta_{yx.q} = \frac{\sigma_q^2 \sigma_{xy} - \sigma_{xq}\sigma_{qy}}{\sigma_x^2 \sigma_q^2 - (\sigma_{xq})^2}. \tag{2}$$

Path tracing yields $\sigma_{xy} = \beta + \alpha\gamma$, $\sigma_{xq} = \alpha$, and $\sigma_{qy} = \gamma + \alpha\beta$. Plugging into (2) verifies that

$$\beta_{yx.q} = \frac{\beta + \alpha\gamma - \alpha(\gamma + \alpha\beta)}{1 - \alpha^2} = \frac{\beta(1 - \alpha^2)}{1 - \alpha^2} = \beta. \tag{3}$$

Note that here the direction of the causal link between x and q does not matter. Reversing the causal link between x and q does not change the result.

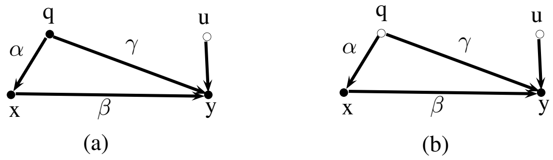


Figure 2: Causal graphs for equation (4) when q is (a) observed and (b) unobserved.

In graph (b) the variable q is unobserved and the backdoor path $x \leftarrow q \rightarrow y$ is now unblocked. The OLS formula for β in the regression of y on x alone is

$$\beta_{yx} = \frac{\sigma_{xy}}{\sigma_x^2}. \tag{4}$$

Substituting for σ_{xy} shows that $\beta_{yx} = \beta + \alpha\gamma$. Thus the regression yields inconsistent estimates for β in large samples when q is omitted unless q is (asymptotically) uncorrelated with x or q is irrelevant for explaining y.

3.2 Proxies

Let us now consider a proxy p of the form

$$q = \delta p + v \tag{5}$$

where the error v is uncorrelated with p. Substituting for the unobserved variable q in (1) yields the model

$$y = \beta x + \gamma\delta p + (\gamma v + u). \tag{6}$$

Estimating (6) with OLS is appropriate when the proxy p fulfills two statistical requirements (Wooldridge 2010, Ch 4): First, p must be redundant. Thus, the expectation of y conditional on x, q and p must not depend on p, i.e. $E(y|x, q, p) = E(y|x, q)$. Second, the omitted variable q must be uncorrelated with x conditional on p, i.e. $E(q|p, x) = E(q|p)$. A variable that fulfills both requirements is sometimes called a "perfect" proxy.

Graph (a) in Figure 3 provides an example where p is a perfect proxy. Conditioning on p blocks the backdoor path $x \dashleftarrow\dashrightarrow p \rightarrow q \rightarrow y$. Path tracing yields $\sigma_{xy} = \beta + \rho\delta\gamma$, $\sigma_{xp} = \rho$, and $\sigma_{py} = \delta\gamma + \rho\beta$. Plugging into the OLS formula for β in the regression of y on x and p

$$\beta_{yx.p} = \frac{\sigma_{xy} - \sigma_{xp}\sigma_{py}}{\sigma_x^2\sigma_p^2 - (\sigma_{xp})^2} \tag{7}$$

shows that $\beta_{yx.p} = \beta$.

In graphs (b) and (c) the variable p violates the second requirement for a perfect proxy. In (b) the variable x affects y directly and indirectly via q. Conditioning on p does not block the

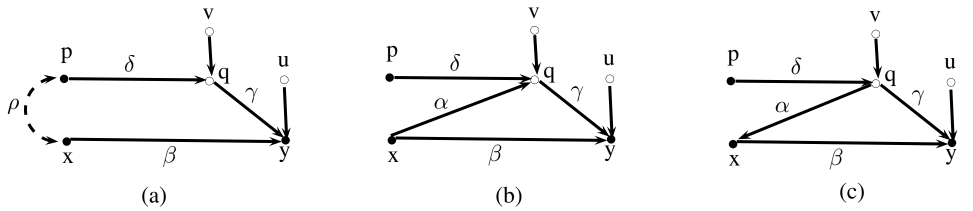


Figure 3: Proxy variable solution when x and q are (a) indirectly related, (b) when x causes q, and (c) when q causes x.

backdoor path $x \to q \to y$ and path tracing shows that $\beta_{yx.p} = \beta + \alpha\gamma$ captures the total (i.e. direct + indirect) effect of x on y rather than the direct effect β. In (c) the variable q is a confounder. Conditioning on p is ineffective because the path $x \leftarrow q \to y$ remains unblocked. Path tracing demonstrates that OLS yields $\beta_{yx.p} = \beta + [\alpha\gamma(1-\delta^2)/(1-\alpha^2\delta^2)]$ in this case.

Wooldridge (2010) states (p 68) that a perfect proxy variable p must be "closely enough related to the omitted variable" so that the other explanatory variables are partially uncorrelated with the omitted variable once p is included in the equation to be estimated.

The graphical analysis makes this statement more transparent: Proxy variables only work perfectly when they, as in case (a), block all paths between the explanatory variables and the omitted variable. In case (b) the proxy p would have to be an intermediate cause of q to work. In case (c) the proxy as given by equation (5) never works because the proxy cannot block the direct causal link between the omitted variable and the explanatory variable.

3.3 Instruments

Another way to solve the omitted variable problem is to let the omitted variable q be part of the error term and to use instruments for explanatory variables that might be correlated with q. Model (1) becomes

$$y = \beta x + e \tag{8}$$

where the error $e = (\gamma q + u)$. Let us assume that an instrumental variable z that is correlated with x but uncorrelated with the error term e is available. The IV formula for β is

$$\beta_{yx}^{IV} = \frac{\sigma_{zy}}{\sigma_{zx}}. \tag{9}$$

Figure 4 shows two cases where an instrument is used to solve the omitted variable problem. The IV strategy works in case (a) where q is a confounder. Here x becomes a collider that blocks the path $z \to x \leftarrow q \to y$. Path tracing yields $\sigma_{zy} = \delta\beta$ and $\sigma_{zx} = \delta$. Plugging into (9) gives $\beta_{yx}^{IV} = \beta$.

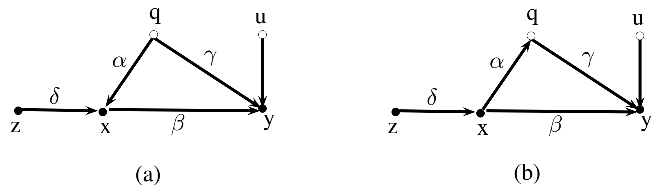


Figure 4: Instrument for x when (a) q causes x, and (b) when x causes q.

Consider now case (b). Here x also affects y indirectly via q. No instrument is required if one is interested in the total effect (i.e. the direct and indirect effect) of x on y. OLS will work. In certain cases, however, one may be interested in the direct effect of x on y.

Consider an example, where x is years of schooling, y is wages, and q is work experience which is unobserved (see Morgan & Winship 2015, Ch 10.1). Schooling may have a direct positive effect on wages and an indirect negative effect via reduced work experience. The goal may be to quantify the direct positive effect of schooling on wages.

Instrumenting x does not work in case (b). The path $z \to x \to q \to y$ is unblocked. Now $\sigma_{zy} = \delta\beta + \delta\alpha\gamma$ and $\beta_{yx}^{IV} = \beta + \alpha\gamma$ yields the total effect of x on y and not the direct effect β that we want to estimate. To obtain the direct effect β one needs either a "perfect" proxy for q (i.e. a variable p such that $x \to p \to q$) or an indicator of q to which an instrument can be applied. Brito & Pearl (2002) provide graphical rules and further results for IV identification.

3.4 Indicators

Let us now assume that two indicators $i = \delta q + w$ and $z = \lambda q + r$ for q are available. The errors w and r are assumed to be uncorrelated. Rearranging yields $q = (1/\delta)i - (1/\delta)w$ and substituting for q in (1) gives

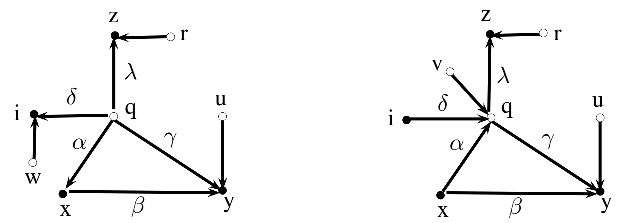
$$y = \beta x + (\gamma/\delta)i + (u - (\gamma/\delta)w). \tag{10}$$

The other indicator z can now serve as an instrument for i. One could of course also use any other valid instrument for the indicator i.

Figure 5 (a) shows the graph for the multiple indicator strategy when the omitted variable q is a confounder. The IV formula for β in the regression of y on x and i is

$$\beta_{yx.i}^{IV} = \frac{\sigma_{zi}\sigma_{xy} - \sigma_{xi}\sigma_{zy}}{\sigma_x^2\sigma_{zi} - \sigma_{xi}\sigma_{zx}}. \tag{11}$$

Path tracing gives $\sigma_{zi} = \lambda\delta$, $\sigma_{xy} = \beta + \alpha\gamma$, $\sigma_{xi} = \alpha\delta$, $\sigma_{zy} = \lambda\gamma + \lambda\alpha\beta$, and $\sigma_{zx} = \lambda\alpha$. Plugging into (11) shows that $\beta_{yx.i}^{IV} = \beta$.



(a) (b)

Figure 5: Multiple indicator solution when q causes i and z as in (a), and (b) when i causes q.

Note that this IV strategy does, unlike instrumenting x, not depend on the direction of the causal link between x and q as long as i depends on q. To see this just replace $q \rightarrow x$ with $x \rightarrow q$ in case (a) in Figure 5.

However, when i causes q instead, i.e. $q = \delta i + v$, then instrumenting i with z yields β only when q is a confounder. The strategy breaks down when x causes q as in case (b). Then one obtains the total effect $\beta + \alpha\gamma$. In the later case the strategy fails because z instruments the ineffective conditioning variable i.

4 Conclusion

The graphical analysis presented in this note discussed some simple examples where proxy variable and instrumental variable strategies to solve an omitted variable problem fail. The examples demonstrate that the effectiveness of proxy and instrumental variable strategies depends crucially on the causal links between explanatory variables and omitted variables. These links must be properly taken into account.

References

Bollen, K. A. (1989), *Structural Equations with latent Variables*, Wiley.

Brito, C. & Pearl, J. (2002), 'A graphical criterion for the identification of causal effects in linear models', Proceedings of the Eighteenth National Conference on Artificial Intelligence, AAAI Press/The MIT Press: Menlo Park, CA. 533–538.

Chen, B. & Pearl, J. (2014), Graphical tools for linear structural equation modeling, Technical report, r-432, University of California.

Goldberger, A. S. (1972), 'Structural equation methods in the social sciences', *Econometrica* **40**(6), 979–1001.

Morgan, S. L. & Winship, C. (2015), *Counterfactuals and causal inference: methods and principles for social research, 2nd Edition*, Cambridge University Press.

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

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

Pearl, J. (2013), 'Linear models: A useful "microscope" for causal analysis', *Journal of Causal Inference* **1**(1), 155–170.

Wooldridge, J. M. (2010), *Econometric Analysis of Cross Section and Panel Data, second edition*, The MIT Press.

Wright, S. (1921), 'Correlation and causation', *Journal of Agricultural Research* **20**, 557–585.

Wright, S. (1934), 'The method of path coefficients', *The Annals of Mathematical Statistics* **5**(3), 161–215.