

On a Class of Bias-Amplifying Covariates that Endanger Effect Estimates

Judea Pearl
University of California, Los Angeles
Computer Science Department
Los Angeles, CA, 90095-1596, USA
judea@cs.ucla.edu

November 17, 2009

Abstract

This note deals with a class of covariates that tends to amplify confounding bias in the analysis of causal effects. This class, recently discovered by Wooldridge (2009), includes instrumental variables and variables that have greater influence on treatment selection than on the outcome. We extend the results of Wooldridge by considering non-linear models and show that,

1. the bias-amplifying potential of instrumental variables extends over to non-linear models, though not as sweepingly as in linear models;
2. in non-linear models, conditioning on instrumental variables may introduce new bias where none existed before;
3. in both linear and non-linear models, instrumental variables have no effect on selection-induced bias.

1 Introduction

Current practices of propensity score matching are governed by the belief that adding more covariates to the propensity score can cause no harm (Rosenbaum, 2002, p. 76). For example, a popular tutorial article by D'Agostino, Jr. (1998) states:

“...if one has the ability to measure many of the covariates that are believed to be related to the treatment assignment, then one can be fairly confident that approximately unbiased estimates for the treatment effect can be obtained.” p. 2267.

Rubin (2009) has further reinforced this belief by stating that “To avoid conditioning on some observed covariates... is distinctly frequentist and non-scientific ad hockery.”

This attitude relieves investigators from thinking about cause-effect relationships in the problem but is based, unfortunately, on false premises. Examples abound showing that certain covariates may actually increase bias if included in the propensity score.¹ (Pearl, 1995, 2009a,b; Greenland et al., 1999; Heckman and Navarro-Lozano, 2004; Schisterman et al., 2009). Such covariates include: (1) colliders, (2) intermediate variables on the causal pathways between

¹The same applies to stratifying on these covariates, using them for matching, using them as predictors in regression, or including them in inverse probability weighting.

treatment and outcome, and (3) descendants of such intermediaries (Weinberg 1993; Pearl 2009a, pp. 339–40).

Recently, Wooldridge has identified a new class of covariates that, although not introducing new bias, tend to amplify bias if such exists. Wooldridge has shown that, in linear systems, conditioning on an instrumental variable (that is, a variable that affects treatment and is not associated with other factors that determine outcome) invariably causes an increase in confounding bias if such exists.

This result is far from obvious. An instrumental variable (IV) meets all the requirements that mainstream literature imposes on covariates pending selection:

1. It is a pre-treatment variable, so, it is certainly not affected by the treatment, nor does it interfere with the causal pathways from treatment to outcome.
2. It is related to treatment
3. It is related to outcome
4. It is related to outcome conditioned on treatment.

Thus, an IV seems to behave just like an ordinary confounder that begs to be controlled. Although the analysis of “confounding-equivalence” (Pearl 2009a, pp. 345–6; Pearl and Paz 2009) identifies IV’s as bias modifiers, the idea that the modification always goes in the wrong direction (i.e., increased bias), is rather surprising, and calls for further analysis.

In this note we first re-derive Wooldridge’s results in structural model setting and then extend the analysis in three directions. First, we quantify the condition under which a confounding variable ceases to act as a bias reducer and instead becomes a bias amplifier. Second, we consider non linear systems and show that the bias-amplification potential of instrumental variables extends over to non-linear models, though not as pervasively as in linear models; there are cases where conditioning on an IV reduces bias and, moreover, conditioning on instrumental variables may introduce new bias where none existed before. Finally, we examine the effect of instrumental variables on selection bias created by preferential exclusion of units from the study (as in case-controlled studies,) and show that, in general, conditioning on an IV has no effect on such bias, unless the exclusion depends on factors that cause the treatment

2 Analysis

We will first consider a linear structural model, given in Fig. 1 where X represents treatment, Y is the outcome of interest, U is an unobserved confounder, and Z is an instrumental variable with respect to the causal effect of X on Y .

We need to compare three quantities:

$$A_1 = \frac{\partial}{\partial x} E(Y|do(x))$$

$$A_2 = \frac{\partial}{\partial x} E(Y|x)$$

$$A_3 = \frac{\partial}{\partial x} E(Y|x, z)$$

A_1 is the incremental causal effect² of X on Y , A_2 is the (unconditional) incremental dependence of Y on X , given by the regression coefficient of Y on X , and A_3 is the incremental conditional dependence of Y on X , given by the coefficient of X in the regression of Y on X and Z .

The unadjusted bias is given by the difference

$$B_0 = A_2 - A_1$$

while the bias after conditioning on $Z = z$ is given by

$$B_z = A_3 - A_1$$

Our task is to compare the magnitudes of B_0 and B_z under various assumptions about the data-generating model.

For the model of Fig. 1, we have:

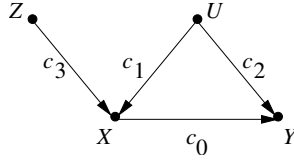


Figure 1: Linear model with instrumental variable Z and confounder U .

$$A_1 = c_0 \tag{1}$$

$$A_2 = c_0 + c_1 c_2 \tag{2}$$

$$\begin{aligned} A_3 &= \frac{\partial}{\partial x} E(Y|x, z) = \frac{\partial}{\partial x} \sum_u E(Y|x, z, u) P(u|x, z) \\ &= \frac{\partial}{\partial x} \sum_u E(Y|x, u) P(u|x, z) \\ &= \frac{\partial}{\partial x} \sum_u (c_0 x + c_2 u) P(u|x, z) \\ &= c_0 + c_2 \frac{\partial}{\partial x} E(U|x, z) \end{aligned} \tag{3}$$

Eq. (1) follows from the definition of A_1 , Eq. (2) follows from Wright's rule of path analysis (assuming zero-mean, standardized variables), and Eq. (3) was derived above using the structural assumption $E(Y|x, z, u) = c_0 x + c_2 u$.

Our next step is to evaluate the right hand side of Eq. (3) in terms of the structural coefficients c_3 and c_1 . Starting with the regression of U on X and Z :

$$u = \beta x + \alpha z + \epsilon \quad \epsilon \perp\!\!\!\perp x, z \tag{4}$$

²Readers versed in potential-outcome notation can identify $E(Y|do(x))$ with the counterfactual expression $E(Y_x)$; both are given by the conditional expectation $E(Y|x)$ in a modified structural model, with c_1 and c_3 set to zero (see (Pearl, 2009a)).

we note that $E(U|x, z) = \beta x + \alpha z$ and, so, to evaluate (3) we need to express β and α in terms of c_3 and c_1 . Multiplying by Z and X , taking expectations, and invoking the instrumental assumption $cov(Z, U) = 0$, we obtain:

$$\begin{aligned} E(UZ) &= 0 = \beta E(XZ) + \alpha E(Z^2) = \beta c_3 + \alpha \\ E(UX) &= c_1 = \beta E(X^2) + \alpha E(XZ) = \beta + \alpha c_3 \end{aligned}$$

yielding:

$$\beta = \frac{c_1}{1 - c_3^2} \quad \alpha = -\frac{c_1 c_3}{1 - c_3^2} \quad (5)$$

Substituting (5) in (3) enables us to evaluate A_3 , giving:

$$A_3 = c_0 + c_2 \frac{\partial}{\partial x} (\beta x + \alpha z) = c_0 + \frac{c_2 c_1}{1 - c_3^2} \quad (6)$$

We now have A_1, A_2 , and A_3 evaluated, from which we can compute the biases B_0 and B_z , giving

$$B_z = \frac{c_2 c_1}{1 - c_3^2}, \quad B_0 = c_1 c_2, \quad B_z = \frac{B_0}{1 - c_3^2} \quad (7)$$

Clearly, $|B_z| \geq |B_0|$ regardless of the signs of c_1 and c_2 with strict inequality holding whenever $|B_0| > 0$. The same result holds when U is a vector of confounding variables. Thus, conditioning on Z amplifies the unconditioned bias by a factor $\frac{1}{1 - c_3^2}$.

3 Intuition

For intuitive understanding of this phenomenon, consider the transition from $X = 0$ to $X = 1$, assuming a simplified equation $X = U + cZ$. By conditioning on $Z = z$, we are comparing units for which $U + cz = 0$ with those for which $U + cz = 1$. The mean difference in U is of course unity

$$E(U|X = 1, Z = z) - E(U|X = 0, Z = z) = 1$$

and this difference will be transmitted to Y and translate into the bias

$$E(Y|X = 1, Z = z) - E(Y|X = 0, Z = z) - c_0 = c_2$$

If, on the other hand, we do not condition on Z but let it vary freely, the variation in Z will “absorb,” or “account for” part of the change in X , and only part of that change will be transmitted through $E(U)$ onto $E(Y)$. To see it formally, consider again the mean difference (in U) between units for which $U + cZ = 0$ and those for which $U + cZ = 1$, but now both U and Z are variables.

Writing

$$\begin{aligned} E(U|x) &= \sum_z E(U|x, z) P(z|x) \\ &= \sum_z (x - cz) P(z|x) \\ &= x - c E(Z|x) \end{aligned}$$

we have

$$E(U|X = 1) - E(U|X = 0) = 1 - c(E(Z|X = 1) - E(Z|X = 0)).$$

The second term on the right is always positive, regardless of whether X and Z are positively or negatively correlated (respectively, $c > 0$ or $c < 0$).³ We conclude therefore that mean difference in U diminishes when we refrain from conditioning on Z :

$$E(U|X = 1) - E(U|X = 0) < E(U|X = 1, Z = z) - E(U|X = 0, Z = z) = 1$$

and the bias transmitted onto Y will likewise be reduced:

$$E(Y|X = 1) - E(Y|X = 0) < E(Y|X = 1, Z = z) - E(Y|X = 0, Z = z).$$

4 Extensions

4.1 The line between instruments and confounders

We now extend this result in three directions. First we relax the assumption that Z is a “perfect” instrumental variable by allowing it to influence Y directly as shown in Fig. 2. We ask for the relative values of c_3 and c_4 that would turn Z from a bias-amplifier to a bias-reducer. Repeating

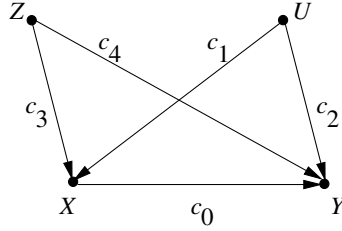


Figure 2: Model in which Z is both a confounder and an (imperfect) instrumental variable relative to X, Y .

the derivation of (6) with $E(Y|x, z, u) = c_0x + c_4z + c_2u$ leaves A_3 and B_z the same, but changes A_2 to read

$$A_2 = \frac{\partial}{\partial x} E(Y|x) = c_0 + c_2c_1 + c_3c_4 \quad (8)$$

We see that Z becomes a bias-reducer when

$$B_z \leq B_0 = c_2c_1 + c_3c_4$$

or

$$\frac{c_4}{c_3} \geq \frac{c_2c_1}{1 - c_3^2} \quad (9)$$

Thus, for Z to become a bias-reducer, the effect of Z on Y must exceed its effect on X by a factor $c_2c_1/1 - c_3^2$. This may be a tall order to meet when c_3 is close to unity. Ironically, this means that the best predictors (of X) are also the most dangerous bias amplifiers.

³More formally, we have:

$$\begin{aligned} E(U|X = x)/x &= Cov(XU)/Var(X) \\ &= Cov[U(U + cZ)]/Var[(U + cZ)^2] \\ &= Var(U)/(Var(U) + c^2Var(Z)) < 1 \end{aligned}$$

Thus, only a fraction of the unit change in X will be transported over to U , and then to Y .

This finding should be of concern to program evaluation researchers and propensity score analysts. Pre-treatment covariates are chosen for control or for propensity score matching, not because they are good predictors of treatment or outcome, but because they are deemed likely to reduce bias, when such is suspected. The analysis of this section shows that being a good predictor of treatment assignment compromises the bias-reducing potential of a covariate, for it tends to amplify bias due to other, uncontrolled confounders. One would do better therefore to rank order covariates based on their importance with respect to the outcome variable, a strategy advocated by Brookhart et al. (2006) and Hill (2008) which, in light of Eq. (9), deserves a general endorsement.

4.2 A glimpse at non-linear systems

The second extension deals with the question of whether the amplification phenomenon described in (7), which prevails over all linear models regardless of parameter values extends over to non-linear models as well. We will show that, although the phenomenon persists in non-linear model, it is not as pervasive – there are non-linear models for which $|B_0| > |B_2|$.

Consider the model of Fig. 1, in which the equation determining X remains the same, but the one for Y becomes non-linear in X :

$$Y = f(x) + ug(x) + \epsilon'$$

This yields

$$\begin{aligned} E(Y|x, z) &= f(x) + g(x)E(U|x, z) \\ &= f(x) + g(x)(\beta x + \alpha z) \\ &= f(x) + \beta g(x)(x - c_3 z) \end{aligned}$$

with β and α given in (5); and

$$\begin{aligned} E(Y|x) &= f(x) + g(x)E(U|x) \\ &= f(x) + g(x)c_1 x \end{aligned}$$

Consequently, A_1 , A_2 , and A_3 evaluate to

$$\begin{aligned} A_1 &= \frac{\partial}{\partial x} E(Y|do(x)) = \frac{\partial}{\partial x} E(f(x) + ug(x)) = f'(x) \\ A_2 &= \frac{\partial}{\partial x} E(Y|x) = f'(x) + c_1(xg'(x) + g(x)) \\ A_3 &= \frac{\partial}{\partial x} E(Y|x, z) = f'(x) + \beta(xg'(x) + g(x) - c_3g'(x)z) \end{aligned} \tag{10}$$

and the two bias measures become

$$\begin{aligned} B_0 &= c_1(xg'(x) + g(x)) \\ B_z &= \frac{1}{1 - c_3^2} [c_1(xg'(x) + g(x) - c_3g'(x)z)] \\ &= \frac{1}{1 - c_3^2} (B_0 - c_1c_3g'(x)z) \end{aligned} \tag{11}$$

Clearly, if $B_0 \geq 0$ and $c_1 c_3 g'(x) z > 0$, we can get $|B_z| < |B_0|$. This means that conditioning on Z may reduce confounding bias, even though Z is a perfect instrument and both Y and X are linear in U . Note that, owed to the non-linearity of $Y(x, u)$, the conditional bias depends on the value of Z and, moreover, for $Z = 0$ we obtain the same bias amplification as in the linear case (Eq. (7)).

Equation (11) also shows that conditioning on Z can introduce bias where none exists. This occurs when $c_1 > 0$ and $g(x) = A/x$, a condition that yields $B_0 = 0$ and $B_z > 0$. This potential of instrumental variables to produce new bias is suppressed in linear systems, as seen in Eq. (7), but is unleashed in non-linear systems. Still, this can only occur when Z and Y are dependent given X (see Pearl and Paz (2009)); it will not occur therefore in situations where the zero bias condition $B_0 = 0$ is *structural*, that is, where one of the structural equations $x = h(z, u)$ or $y = h'(x, u)$ is trivial in its u argument.⁴

4.3 The resilience of Selection Bias

Our final extension concerns the effect of instrumental variables on selection bias, that is, bias induced by preferential selection of units for data analysis which is often governed by unknown factors including treatment, outcome and their consequences. Case control studies are particularly susceptible to such bias, e.g., when the outcome is a disease or complication that warrants reporting (see (Glymour and Greenland 2008, pp. 111–37; Robins 2001; Hernán et al. 2004)).

To illuminate the nature of this bias, consider the linear model of Fig. 3 in which S is a variable affected by both X and Y , indicating entry into the data pool. Such preferential selection

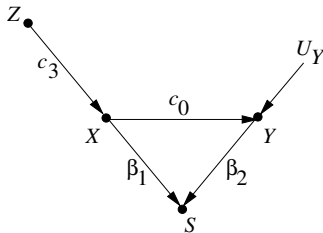


Figure 3: A model illustrating selection bias; conditioning on S induces spurious associations between X and Y .

to the data pool amounts to conditioning on S , and creates spurious association between X and Y through two mechanisms. First, conditioning on the collider S induces spurious association between its parents, X and Y . Second, S is also a descendant of a “virtual collider” Y , whose parents are X and the error term U_Y (also called “omitted factors”) which is always present, though often not shown.⁵ The first mechanism is suppressed when β_1 is zero, while the latter is suppressed when c_0 is zero; both are suppressed when β_2 is zero.

⁴The condition $g(x) = A/x$ that gave rise to $B_0 = 0$ can be thought of as “unstable” (Pearl, 2009a, Ch. 2) for it depends critically on the exact form of the function $y = h'(x, u)$.

⁵See Pearl (2009a, pp. 339–41) for further explanation of this bias mechanism, which seems to have escaped the taxonomies in Hernán et al. (2004) and Schisterman et al. (2009).

Writing

$$A_1 = \frac{\partial}{\partial x} E(Y|do(x)) = c_0 \quad (12)$$

$$A_2 = \frac{\partial}{\partial x} E(Y|x, s) \quad (13)$$

$$A_3 = \frac{\partial}{\partial x} E(Y|x, z, s) \quad (14)$$

the bias due to conditioning on S , $A_2 - A_1$, can be calculated through the usual method of expectations (as in Eqs. (3)-(6)) and yields:

$$B_0 = A_2 - A_1 = \frac{-\beta_2(1 - c_0^2)(\beta_1 + c_0\beta_2)}{1 - (\beta_1 + c_0\beta_2)^2} \quad (15)$$

We see that B_0 can be substantial and it vanishes if and only if one of the following conditions holds:

$$\beta_2 = 0, \quad c_0^2 = 1 \text{ or } \beta_1 = -c_0\beta_2.$$

In view of the amplification effect of IV's on confounding bias, one may be tempted to surmise that a similar effect can be expected vis-à-vis selection bias. This however is not the case. Conditioning on Z has no effect whatsoever on selection-induced bias, formally,

$$A_3 = \frac{\partial}{\partial x} E(Y|x, s, z) = \frac{\partial}{\partial x} E(Y|x, s) = A_2 \quad (16)$$

This equality can be derived, of course, from the parametric model of Fig. 3, going through the necessary (yet painstaking) steps of algebraic manipulations. However, the validity of this equality is much broader, for it holds in non-linear systems as well. This can be seen immediately from the structure of the diagram of Fig. 3, which asserts that, regardless of the functional relationships between the variables in the diagram, Y and Z are independent given X and Z ,⁶ which entails Eq. (16).

We thus conclude that selection bias differs fundamentally from confounding bias in that the former, as distinct from the latter, is insensitive to conditioning on an IV. This distinction can be used in practice to detect the presence of confounding bias. If one has a solid theoretical basis to believe that a variable Z is a valid instrument relative to the effect of X on Y , and if data shows that the association between X and Y changes across strata of Z , chances are the study is marred by confounding bias, and remedial steps are necessary, possibly through covariate control. Conversely, if no such changes can be detected in the data, chances are no confounding bias exists, though the study can still be contaminated with selection-induced bias, resistant to covariate control.⁷

It is important to keep in mind though that the selection bias analyzed above was “pure,” in the sense of inducing no confounding component. In general, if the reasons for excluding units from the study data involves ancestors of X , confounding bias may also be induced, as shown in Fig. 4. S_3 induces “pure” confounding bias that would be amplified by conditioning on Z , while S_2 induces a “pure” selection bias that will be impervious to conditioning on Z . S_1 induces both

⁶This is verified through the d -separation rule (Pearl, 2009a, pp. 335–7), which identifies the conditional independencies implied by a system of non-linear structural equations.

⁷I use the cautionary term “chances are” to allow for the rare possibilities that Z introduces its own bias (see footnote 4) or that the bias will not be changed by Z . These possibilities are not structural, in the sense that they require fine tuning of the functional relationships in the model.

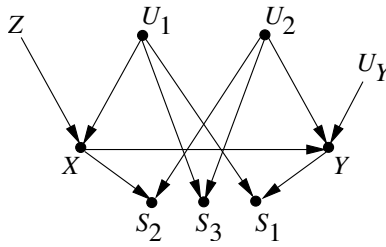


Figure 4: Model in which S_2 induces selection bias, S_3 induces confounding bias, and S_1 induces both.

selection and confounding bias through to the path $X - U_1 - S_1 - Y$; the latter will be sensitive to Z . Conditioning on U_2 will eliminate the entire bias induced by S_2 , while conditioning on U_1 will eliminate only the confounding part of the bias induced by S_1 ; the selection-induced part, due to the association created between X and U_Y , cannot be eliminated by any method. S_3 induces purely confounding bias, which can be eliminated by conditioning on either U_1 or U_2 .

Formally, the distinction between confounding and selection bias can be articulated thus: Confounding bias is any $X - Y$ association that is attributable to paths traversing ancestors of X (i.e., factors affecting treatment.) Operationally, the distinction may refer to experimental paradigm: Confounding bias is any $X - Y$ association that can be eliminated by randomization. The theory of graphical models establishes the equivalence of the two criteria.

5 Conclusions

We have examined the effect of instrumental variables on various types of bias. We first showed that, while in linear systems conditioning on an IV always amplifies confounding bias (if such exists), bias in non-linear systems may be amplified as well as attenuated. In some cases an IV may introduce new bias where none exists. We further examined the effect of IV's on selection-induced bias and showed that no such effect exists as long as the bias contains no confounding component. A formal criterion for distinguishing the two types of bias sources is introduced, and the possibility of using IV-sensitivity as a diagnostic tool for bias detection is suggested. From a practical viewpoint, the immediate implication of this analysis is that covariates should be chosen based on their importance with respect to the outcome, rather than the treatment.

Acknowledgments

This note has benefited from discussions with Jeffrey Wooldridge and was supported in parts by grants from NIH #1R01 LM009961-01, NSF #IIS-0914211, and ONR #N000-14-09-1-0665.

References

- BROOKHART, M., SCHNEEWEISS, S., ROTHMAN, K., GLYNN, R., AVORN, J. and STÜRMER, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology* **163** 1149–1156.
- D'AGOSTINO, JR., R. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine* **17** 2265–2281.

- GLYMOUR, M. and GREENLAND, S. (2008). Causal diagrams. In *Modern Epidemiology* (K. Rothman, S. Greenland and T. Lash, eds.), 3rd ed. Lippincott Williams & Wilkins, Philadelphia, PA, 183–209.
- GREENLAND, S., PEARL, J. and ROBINS, J. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10** 37–48.
- HECKMAN, J. and NAVARRO-LOZANO, S. (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics* **86** 30–57.
- HERNÁN, M., HERNÁNDEZ-DÍAZ, S. and ROBINS, J. (2004). A structural approach to selection bias. *Epidemiology* **15** 615–625.
- HILL, J. (2008). Discussion of research using propensity-score matching: Comments on ‘A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003’ by Peter Austin, *statistics in medicine*. *Statistics in Medicine* **27** 2055–2061.
- PEARL, J. (1995). Causal diagrams for empirical research. *Biometrika* **82** 669–710.
- PEARL, J. (2009a). *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge University Press, New York.
- PEARL, J. (2009b). Letter to the editor: Remarks on the method of propensity scores. *Statistics in Medicine* **28** 1415–1416. <http://ftp.cs.ucla.edu/pub/stat_ser/r345-sim.pdf>.
- PEARL, J. and PAZ, A. (2009). Confounding equivalence in observational studies. Tech. Rep. TR-343, University of California, Los Angeles, CA. <http://ftp.cs.ucla.edu/pub/stat_ser/r343.pdf>.
- ROBINS, J. (2001). Data, design, and background knowledge in etiologic inference. *Epidemiology* **12** 313–320.
- ROSENBAUM, P. (2002). *Observational Studies*. 2nd ed. Springer-Verlag, New York.
- RUBIN, D. (2009). Author’s reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment group? *Statistics in Medicine* **28** 1420–1423.
- SCHISTERMAN, E., COLE, S. and PLATT, R. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* **20** 488–495.
- WEINBERG, C. (1993). Toward a clearer definition of confounding. *American Journal of Epidemiology* **137** 1–8.
- WOOLDRIDGE, J. (2009). Should instrumental variables be used as matching variables? Tech. Rep. <<https://www.msu.edu/~ec/faculty/wooldridge/current%20research/treat1r6.pdf>>, Michigan State University, MI.