DISCUSSION PAPER SERIES

IZA DP No. 10057

Proxy Variables and Nonparametric Identification of Causal Effects

Xavier de Luna
Philip Fowler
Per Johansson

July 2016

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Proxy Variables and Nonparametric Identification of Causal Effects

Xavier de Luna
Umeå University

Philip Fowler
Umeå University

Per Johansson
Uppsala University, IFAU and IZA

Discussion Paper No. 10057
July 2016

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 10057
July 2016

ABSTRACT

Proxy Variables and Nonparametric Identification of Causal Effects

Proxy variables are often used in linear regression models with the aim of removing potential confounding bias. In this paper we formalise proxy variables within the potential outcome framework, giving conditions under which it can be shown that causal effects are nonparametrically identified. We characterise two types of proxy variables and give concrete examples where the proxy conditions introduced may hold by design.

JEL Classification: C14

Keywords: average treatment effect, observational studies, potential outcomes, unobserved confounders

Corresponding author:

Philip Fowler
Department of Statistics, USBE
Umeå University
SE-90187 Umeå
Sweden
E-mail: philip.fowler@umu.se

1 Introduction

Proxy variables are often used in empirical economics and other empirical sciences as substitutes for unobserved confounders when conducting observational studies. However, using substitute variables does not necessarily reduce bias due to confounding to zero and may even increase bias (Frost, 1979). Thus, we call herein proxy variables only such substitute variables which yield identification of a causal effect of interest. Proxy variables have previously been defined in the literature in the context of linear models, using for instance linear projection orthogonality conditions, see Wooldridge (2010, pp. 67-72).

In this note we formalise proxy variables within the potential outcome framework (Imbens and Wooldridge, 2009), giving conditions for which it can be shown that causal effects are nonparametrically identified. This allows us to clarify the use of proxy variables in a general context. Moreover, our approach also allows us to characterise two types of proxy variables, one directly related to the earlier definition mentioned above, and a new type of proxy variable not previously considered in the literature. We also give examples where the proxy conditions introduced may hold by design.

2 Theory on proxy variables

We consider a study with aim to evaluate the effect of a binary treatment T on an outcome Y. Let Y_1, Y_0 be potential outcomes if treated ($T = 1$) or not treated ($T = 0$) respectively, $\boldsymbol{X}$ a set of observed pre-treatment covariates related to T and Y (observed confounders), and $\boldsymbol{U}$ a set of unobserved pre-treatment covariates also related to T and Y (unobserved confounders). We assume that the observed outcome for any given unit is $Y = TY_1 + (1-T)Y_0$, i.e. that consistency and the Stable Unit Treatment Value Assumption, see Rubin (1980), hold. Letting $A \perp\!\!\!\perp B \mid C$ denote that A is conditionally independent of B given C (Dawid, 1979), the following assumptions are used in the sequel.

Assumption 1. [unconfoundedness]

$$\begin{aligned} &i) \quad T \perp\!\!\!\perp Y_0 \mid (\boldsymbol{X}, \boldsymbol{U}), \\ &ii) \quad T \perp\!\!\!\perp Y_1 \mid (\boldsymbol{X}, \boldsymbol{U}). \end{aligned}$$

Assumption 2. [common support]

$$
\begin{aligned}
&i) \quad \Pr(T = 0 \mid \boldsymbol{X}, \boldsymbol{U}) > 0, \\
&ii) \quad \Pr(T = 1 \mid \boldsymbol{X}, \boldsymbol{U}) > 0.
\end{aligned}
$$

If in Assumptions 1 and 2 the set of unobserved covariates $\boldsymbol{U}$ is empty, then the average causal effect $\tau = E(Y_1 - Y_0)$ and the average causal effect on the treated $\tau^t = E(Y_1 - Y_0 \mid T = 1)$ are identified. While if $\boldsymbol{U}$ is empty only for Assumptions 1*i)* and 2*i)* then only τ^t is identified (Imbens and Wooldridge, 2009).

In observational studies, it may be the case that, although $\boldsymbol{U}$ is not observed, we have observed variables which may act as proxies for $\boldsymbol{U}$. We now give conditions characterising proxy variables useful for identification of average causal effects. Let $\boldsymbol{P}$ denote a non-empty set of pre-treatment variables not included in the covariate sets defined so far, $\boldsymbol{P} \not\subseteq \{\boldsymbol{X}, \boldsymbol{U}\}$, and let $\boldsymbol{U}$ be non-empty such that $Y_0 \not\perp\!\!\!\perp T \mid \boldsymbol{X}$ and/or $Y_1 \not\perp\!\!\!\perp T \mid \boldsymbol{X}$. A proxy variable will then need to satisfy $Y_0 \perp\!\!\!\perp T \mid (\boldsymbol{X}, \boldsymbol{P})$ (and $Y_1 \perp\!\!\!\perp T \mid (\boldsymbol{X}, \boldsymbol{P})$) in order for τ^t (τ) to be identified. A set of conditions describing useful proxy properties for $\boldsymbol{P}$ are as follows.

Assumption 3. [proxy Type I]

[irrelevance for outcome]		[proxy property]	
i)	$Y_0 \perp\!\!\!\perp (T, \boldsymbol{P}) \mid (\boldsymbol{X}, \boldsymbol{U})$	*iii)*	$T \perp\!\!\!\perp \boldsymbol{U} \mid (\boldsymbol{X}, \boldsymbol{P})$
ii)	$Y_1 \perp\!\!\!\perp (T, \boldsymbol{P}) \mid (\boldsymbol{X}, \boldsymbol{U})$		

This first type of proxy is similar in spirit to Wooldridge's (2010) definition of proxy variables. A proxy variable of Type I is an irrelevant variable for explaining the potential outcomes given the confounders $\boldsymbol{X}, \boldsymbol{U}$ (Assumption 3*i-ii)*). A variable irrelevant for the outcome is useful for identification (see Proposition 1 below) when it makes $\boldsymbol{U}$ irrelevant for the treatment (Assumption 3*iii)*).

We consider further another type of useful proxy variable, which has up to our knowledge not been formalised in the literature.

Assumption 4. [proxy Type II]

[irrelevance for treatment]		[proxy property]	
i)	$T \perp\!\!\!\perp (Y_0, \boldsymbol{P}) \mid (\boldsymbol{X}, \boldsymbol{U})$	*iii)*	$Y_0 \perp\!\!\!\perp \boldsymbol{U} \mid (\boldsymbol{X}, \boldsymbol{P})$
ii)	$T \perp\!\!\!\perp (Y_1, \boldsymbol{P}) \mid (\boldsymbol{X}, \boldsymbol{U})$	*iv)*	$Y_1 \perp\!\!\!\perp \boldsymbol{U} \mid (\boldsymbol{X}, \boldsymbol{P})$

Thus, a proxy variable of Type II is such that it is irrelevant for explaining the treatment assignment given the confounders $(\boldsymbol{X}, \boldsymbol{U})$ (Assumption 4*i-ii)*). A

variable irrelevant for the treatment is useful for identification (see Proposition 2 below) when it makes $\boldsymbol{U}$ irrelevant for the outcome (Assumption 4*iii-iv)*).

We will also need an extension of the common support assumption for identification purposes.

Assumption 5. [support on proxy]

$$\begin{aligned} &i) \quad \Pr(T = 0 \mid \boldsymbol{X}, \boldsymbol{P}) > 0, \\ &ii) \quad \Pr(T = 1 \mid \boldsymbol{X}, \boldsymbol{P}) > 0. \end{aligned}$$

Lemma 1. [Dawid (1979)] For any variables A, B, C and D, it follows that: $A \perp\!\!\!\perp B \mid C$ and $A \perp\!\!\!\perp D \mid (B, C) \iff A \perp\!\!\!\perp (D, B) \mid C$.

Proposition 1. If Assumptions 3*i),iii)* and 5*i)* hold, then τ^t is identified. Moreover, if Assumptions 3 and 5 hold, then both τ and τ^t are identified.

Proof. By Lemma 1 we have that

$$T \perp\!\!\!\perp \boldsymbol{U} \mid (\boldsymbol{X}, \boldsymbol{P}) \text{ and } T \perp\!\!\!\perp Y_0 \mid (\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{P}) \iff T \perp\!\!\!\perp (Y_0, \boldsymbol{U}) \mid (\boldsymbol{X}, \boldsymbol{P}). \quad (1)$$

The first part of the left-hand side of (1) holds by Assumption 3*iii)*. The second part of the left-hand side of (1) holds by Assumption 3*i)*, using Lemma 1 to note that $Y_0 \perp\!\!\!\perp (T, \boldsymbol{P}) \mid (\boldsymbol{X}, \boldsymbol{U}) \Rightarrow Y_0 \perp\!\!\!\perp T \mid (\boldsymbol{U}, \boldsymbol{X}, \boldsymbol{P})$. Since the left-hand side of (1) holds, it follows that $T \perp\!\!\!\perp (Y_0, \boldsymbol{U}) \mid (\boldsymbol{X}, \boldsymbol{P})$, which by Lemma 1 implies that $T \perp\!\!\!\perp Y_0 \mid (\boldsymbol{X}, \boldsymbol{P})$. Thus, assuming further Assumption 5*i)* yields identification of τ^t. Similarly, if Assumption 3*ii)* holds, then $T \perp\!\!\!\perp Y_1 \mid (\boldsymbol{X}, \boldsymbol{P})$. Finally, if Assumptions 3 and 5 hold, then τ is identified. □

Proposition 2. If Assumptions 4*i),iii)* and 5*i)* hold, then τ^t is identified. Moreover, if Assumptions 4 and 5 hold, then both τ and τ^t are identified.

Proof. The proof is similar to the proof of Proposition 1 and thus omitted. □

3 Proxy variables by design

Proxy variables may be obtained by design and here we give some examples. For the sake of simplicity, we focus on univariate proxy variables P in the sequel.

3.1 Proxy Type I: outcome prediction

We characterise here a natural situation where a proxy of Type I arises. Let

$$Y_0 = h(\boldsymbol{X}, \boldsymbol{U}) + \varepsilon_Y, \tag{2}$$

where ε_Y is exogenous and $h(\boldsymbol{X}, \boldsymbol{U}) = E(Y_0 \mid \boldsymbol{X}, \boldsymbol{U})$. Assume that a prediction P of Y_0, made before the treatment assignment, is available such that

$$P = h(\boldsymbol{X}, \boldsymbol{U}) + \varepsilon_P, \tag{3}$$

where $\varepsilon_P \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{U}, Y_0)$ and $E(\varepsilon_P) = 0$, i.e. the prediction is unbiased. Consider further a study design where the treatment assignment is a function of P and $\boldsymbol{X}$ as follows:

$$T^* = k\left(P, \boldsymbol{X}\right) + \varepsilon_T, \tag{4}$$

for some function $k(\cdot)$, with ε_T exogenous and where $Var(\varepsilon_T) > 0$. Let the treatment assignment be such that $T = 1$ if $T^* > 0$ and $T = 0$ otherwise. By exogeneity of ε_Y, we have that $Y_0 \perp\!\!\!\perp (T, P) \mid (\boldsymbol{X}, \boldsymbol{U})$, i.e. Assumption 3*i)* holds. Also, $T \perp\!\!\!\perp \boldsymbol{U} \mid (\boldsymbol{X}, P)$ by design, i.e., Assumption 3*iii)* is fulfilled. Suppose further that $k(\cdot)$ and ε_T are chosen in such a way that Assumption 5*i)* is fulfilled. Note that the design error ε_T is necessary in order for $Pr(T = 0 \mid \boldsymbol{X}, P) > 0$. Then τ^t is identified by Proposition 1.

Example 1 (Outcome prediction proxy by design). Consider the situation where a treatment T is a social program for the unemployed, whose effect on duration to employment, Y, we want to evaluate. Suppose treatment is assigned by case workers after interviews with eligible individuals. A set of individual and labor market characteristics $\boldsymbol{X}$ are recorded at the time of the interview. At that time, the case worker also makes a prediction P of unemployment duration, would the individual not be assigned to treatment (prediction of Y_0). Then, arguably the case workers will provide an unbiased prediction of Y_0, based on $\boldsymbol{X}$ and other unobserved information $\boldsymbol{U}$ obtained at interview, i.e. such that (2-3) hold. Furthermore, if we believe that P summarises all information in $\boldsymbol{U}$ necessary to make the treatment assignment decision, such that (4) holds, then P is a proxy of Type I. In practice, the latter statement may be difficult to ensure by design and a sensitivity analysis to Assumption 3*iii)* may be useful.

3.2 Proxy Type II: lagged outcome

A Type II proxy variable may be available in a follow up setting with three time periods, $t = 0, 1, 2$. Assume that the outcome Y is observed at time $t = 2$. Further, let $\boldsymbol{X}$ and $\boldsymbol{U}$ be defined at baseline ($t = 0$), with $\boldsymbol{X}$ potentially including the outcome measured at $t = 0$. We also observe the outcome at $t = 1$, denoted Y^l, simultaneously as treatment T is assigned. Then, with such a design it may be realistic to assume that

$$Y^l = l(\boldsymbol{X}, \boldsymbol{U}) + \varepsilon_L, \qquad T^* = m(\boldsymbol{X}, \boldsymbol{U}) + \varepsilon_T,$$
$$T = 1 \text{ if } T^* > 0 \text{ and } T = 0 \text{ otherwise,}$$

for some functions $l(\cdot)$ and $m(\cdot)$ and where ε_L and ε_T are exogenous error terms. Furthermore, if we have

$$Y_0 = q(\boldsymbol{X}, Y^l) + \varepsilon_Y, \tag{5}$$

for some function $q(\cdot)$ and where the error term ε_Y is exogenous, then $T \perp\!\!\!\perp (Y^l, Y_0) \mid (\boldsymbol{X}, \boldsymbol{U})$. Thus, by design $P = Y^l$ fulfills Assumption 4*i)*, i.e. Y_l is irrelevant for the treatment assignment T. Moreover, $Y_0 \perp\!\!\!\perp \boldsymbol{U} \mid (\boldsymbol{X}, Y^l)$, i.e. Assumption 4*iii)* also holds. The validity of (5) should be investigated through a sensitivity analysis. Finally, to guarantee that 5*i)* holds here, a sufficient condition is that Assumption 2 holds together with $Pr(\boldsymbol{U} \mid \boldsymbol{X}, Y^l) > 0$.

Example 2 (Lagged outcome proxy design). An example of a lagged outcome proxy design is given in Wooldridge (2010, Example 4.4), where data on Michigan manufacturing firms is discussed with the purpose to estimate the effect of job training grants (T) on firms' productivity. A factor giving a measure of the latter is log scrap rate (number of items out of 100 that must be scrapped) $-$ Y here. Wooldridge used years 1988 and 1987 for the purpose of illustration, that is where T and outcome Y are measured in 1988, and argued that Y_{87} (log scrap rate in 1987) is a proxy of Type I, i.e. in our framework such that $T \perp\!\!\!\perp U \mid Y_{87}$, where U represents unobserved productivity factors. However, one may arguably think that it is more realistic to see Y_{87} as a proxy of Type II, i.e. such that $Y \perp\!\!\!\perp U \mid Y_{87}$.

4 Parametric modelling

We now turn our attention to a linear models where a variable P is a proxy variable of Type I. Suppose that we have potential outcomes such that:

$$Y_0 = \alpha_0 + \boldsymbol{\beta}_0' \boldsymbol{X} + \gamma U + \varepsilon_0, \tag{6}$$

$$Y_1 = \alpha_1 + \boldsymbol{\beta}_1' \boldsymbol{X} + \gamma U + \varepsilon_1, \tag{7}$$

where ε_j, $j = 0, 1$, are exogenous variables with mean zero and independent of each other. Let P be such that (3) holds. Then $Y_j \perp\!\!\!\perp (P, T) \mid (\boldsymbol{X}, U)$, $j = 0, 1$, and Assumptions 3*i-ii)* are fulfilled.

By Lemma 1 it follows from Assumption 3 that $Y_j \perp\!\!\!\perp P \mid (\boldsymbol{X}, U, T)$, $j = 0, 1$. By consistency it follows that $Y \perp\!\!\!\perp P \mid (\boldsymbol{X}, U, T)$. This implies that $E(Y \mid T, \boldsymbol{X}, U, P) = E(Y \mid T, \boldsymbol{X}, U)$, which is in analogy with the redundancy condition in Wooldridge (2010, page 68). Furthermore, write

$$U = E(U \mid \boldsymbol{X}, P) + r, \tag{8}$$

where $E(U \mid \boldsymbol{X}, P) = \theta_0 + \boldsymbol{\theta}' \boldsymbol{X} + \phi P$ and assume that $r \perp\!\!\!\perp T \mid (\boldsymbol{X}, P)$. Then, $U \perp\!\!\!\perp T \mid (\boldsymbol{X}, P)$, i.e. P fulfills Assumption 3*iii)*. Given (8) it also follows that $L(U \mid 1, \boldsymbol{X}, P, T) = L(U \mid 1, \boldsymbol{X}, P)$, where $L(A \mid B)$ is the linear projection of A on B. This corresponds to condition (4.26) in Wooldridge (2010, page 68). In summary, in this situation, P is a proxy of Type I and a proxy as defined by Wooldridge (2010). If Assumption 5 holds, then, by Proposition 1, τ is identified. Note however that if γ in (6) and (7) instead is γ_0 and γ_1 respectively, then identification is not achieved through a linear model.

5 Discussion

Proxies are often used in empirical economics in order to block unobserved confounding in observational studies. In this paper we have given formal conditions under which proxies yield nonparametric identification of average causal effects.

In many applications, an unobserved characteristic is replaced by an observed variable believed to be a function of the former, in the spirit of (3). For example, in Wooldridge (2010, Example 4.3), ability is replaced by IQ. The key issue is whether such a variable is a proxy as defined in this article, and in particular whether Assumption 3*iii)* holds or not. In the ability-IQ situation, it seems reasonable to believe that $IQ = fct(\text{Ability}) + \varepsilon_{IQ}$. However, assuming that $T^* = fct(IQ) + \varepsilon_T$ (in the sense of (4)) is not realistic since one expects instead $T^* = fct(\text{Ability}) + \varepsilon_T$ to hold. Thus, IQ is not a proxy as defined herein, but rather a measure of ability with error. Conditioning on the latter

may yield bias; see Pearl (2010).

Acknowledgments

We are grateful to Ingeborg Waernbaum and Inga Laukaityte for helpful comments that have improved the paper. Financial support from the Swedish Council for Working Life and Social Research (DNR 2009-0826) is gratefully acknowledged.

References

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31.

Frost, P. A. (1979). Proxy variables and specification bias. *The Review of Economics and Statistics*, 61(2):323–325.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.

Pearl, J. (2010). On measurement bias in causal inference. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pages 425–432. AUAI Press.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, Massachusetts, 2nd edition.