Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools

Joseph G. Altonji

Yale University and National Bureau of Economic Research

Todd E. Elder

University of Illinois at Urbana-Champaign

Christopher R. Taber

Northwestern University and National Bureau of Economic Research

In this paper we measure the effect of Catholic high school attendance on educational attainment and test scores. Because we do not have a good instrumental variable for Catholic school attendance, we develop new estimation methods based on the idea that the amount of selection on the observed explanatory variables in a model provides a guide to the amount of selection on the unobservables. We also propose an informal way to assess selectivity bias based on measuring the ratio of

We thank Timothy Donohue and Emiko Usui for excellent research assistance. We received helpful comments from Glen Cain, Tim Conley, Thomas DeLeire, Steve Levitt, Lars Hansen, James Heckman, Robert LaLonde, Jean-Marc Robin, George Jakubson, and especially Derek Neal and the anonymous referees, as well as participants in seminars at the American Economic Association winter meetings (January 2000), Boston College, Cornell University, CREST-INSEE, Duke University, Institute for the Study of Labor (Bonn), Johns Hopkins University, Harvard University, Massachusetts Institute of Technology, Northwestern University, Princeton University, University of Chicago, University College London, University of Florida, University of Maryland, University of Missouri, University of Rochester, and University of Wisconsin–Madison. We are grateful for support from the National Science Foundation grant 9512009 (Altonji), the National Institute of Child Health and Development grant R01 HD36480-03 (Altonji and Taber), and the Institute for Policy Research, Northwestern University.

[*Journal of Political Economy*, 2005, vol. 113, no. 1]
 0022-3808/2005/11301-0001$10.00

- Pioneer the method of bounding $Corr(d, \varepsilon)$ by looking at $Corr(d, X\beta)$

- Pioneer the method of bounding $Corr(d, \varepsilon)$ by looking at $Corr(d, X\beta)$
- Setting: estimating causal effect of attending Catholic high school

- Pioneer the method of bounding $Corr(d, \varepsilon)$ by looking at $Corr(d, X\beta)$
- Setting: estimating causal effect of attending Catholic high school
- Lots of selection, no random variation available

- Pioneer the method of bounding $Corr(d, \varepsilon)$ by looking at $Corr(d, X\beta)$
- Setting: estimating causal effect of attending Catholic high school
- Lots of selection, no random variation available
- Argue that $Corr(d, \varepsilon) = Corr(d, X\beta)$ is a good bound

Why assume $Corr(d, \varepsilon) = Corr(d, X\beta)$?

Why assume $Corr(d, \varepsilon) = Corr(d, X\beta)$?

- X is only a subset of everything that affects y

Why assume $Corr(d, \varepsilon) = Corr(d, X\beta)$?

- X is only a subset of everything that affects y
- Data collected for multiple purposes

Why assume $Corr(d, \varepsilon) = Corr(d, X\beta)$?

- X is only a subset of everything that affects y
- Data collected for multiple purposes
- Data costly to collect, some variables impossible to measure

Why assume $Corr(d, \varepsilon) = Corr(d, X\beta)$?

- X is only a subset of everything that affects y
- Data collected for multiple purposes
- Data costly to collect, some variables impossible to measure
- Thus X is probably a random subset of everything affecting y

$$Corr(d, \varepsilon) \overset{?}{\lessgtr} Corr(d, X\beta)$$

$$Corr(d, \varepsilon) \overset{?}{\lessgtr} Corr(d, X\beta)$$

- Application-specific question requiring careful thinking

$$Corr(d, \varepsilon) \overset{?}{\lessgtr} Corr(d, X\beta)$$

- Application-specific question requiring careful thinking
- What are sources of selection bias? Why is R^2 low?

$$Corr(d, \varepsilon) \overset{?}{\lessgtr} Corr(d, X\beta)$$

- Application-specific question requiring careful thinking
- What are sources of selection bias? Why is R^2 low?
- Selection bias? Measurement error? Irreducible uncertainty?

$$Corr(d, \varepsilon) \overset{?}{\lessgtr} Corr(d, X\beta)$$

- Application-specific question requiring careful thinking
- What are sources of selection bias? Why is R^2 low?
- Selection bias? Measurement error? Irreducible uncertainty?
- If lots of irreducible uncertainty: $Corr(d, \varepsilon) \ll Corr(d, X\beta)$

$$Corr(d, \varepsilon) \overset{?}{\lessgtr} Corr(d, X\beta)$$

- Application-specific question requiring careful thinking
- What are sources of selection bias? Why is R^2 low?
- Selection bias? Measurement error? Irreducible uncertainty?
- If lots of irreducible uncertainty: $Corr(d, \varepsilon) \ll Corr(d, X\beta)$
- Opposite true if not much irreducible uncertainty

$$Corr(d, \varepsilon) \overset{?}{\lessgtr} Corr(d, X\beta)$$

- Application-specific question requiring careful thinking
- What are sources of selection bias? Why is R^2 low?
- Selection bias? Measurement error? Irreducible uncertainty?
- If lots of irreducible uncertainty: $Corr(d, \varepsilon) \ll Corr(d, X\beta)$
- Opposite true if not much irreducible uncertainty
- Typically assume $Corr(d, \varepsilon)$ and $Corr(d, X\beta)$ have same sign

Research Article

Brian Krauth*

Bounding a Linear Causal Effect Using Relative Correlation Restrictions

DOI 10.1515/jem-2013-0013
Previously published online June 4, 2015

Abstract: This paper describes and implements a simple partial solution to the most common problem in applied microeconometrics: estimating a linear causal effect with a potentially endogenous explanatory variable and no suitable instrumental variables. Empirical researchers faced with this situation can either assume away the endogeneity or accept that the effect of interest is not identified. This paper describes a middle ground in which the researcher assumes plausible but nontrivial restrictions on the correlation between the variable of interest and relevant unobserved variables relative to the correlation between the variable of interest and observed control variables. Given such relative correlation restrictions, the researcher can then estimate informative bounds on the effect and assess the sensitivity of conventional estimates to plausible deviations from exogeneity. Two empirical applications demonstrate the potential usefulness of this method for both experimental and observational data.

Keywords: endogeneity; partial identification; sensitivity analysis.

1 Introduction

Applied researchers often find themselves attempting to measure the effect of a variable of interest on an outcome where the best available research design is a linear regression of the outcome on the variable of interest and some control variables. This standard research design requires the researcher to assume that the variable of interest is exogenous, or uncorrelated with the unobserved term in the regression. Exogeneity is a strong and potentially incorrect assumption whose failure will produce biased estimates, yet researchers and policymakers will often prefer a potentially biased estimate to no estimate at all, or to waiting for a better research design to appear.

This paper develops a middle ground between assuming exogeneity and giving up. It works by defining deviations from exogeneity in terms of a sensitivity parameter that describes the (unobserved) correlation between the variable of interest and the regression "error" term relative to the (observed) correlation between the variable of interest and the control variables. The strong assumption of exogeneity in the conventional regression analysis can then be replaced with a weaker assumption that this sensitivity parameter falls in some known range. Provided that this range is not too wide, the effect is partially identified – the researcher can place upper and lower bounds on its value – and can be subjected to hypothesis tests and confidence intervals with the usual interpretation.

The general idea of obtaining partial identification of a causal or structural parameter by imposing restrictions on some relative correlation parameter has been used in a number of previous papers. Settings considered in this literature include nonparametric treatment effects (Manski 1994, 2003; Rosenbaum 2002), parametric treatment selection (Imbens 2003; Altonji, Elder, and Taber 2005b) peer effects (Krauth 2007),

*Corresponding author: Brian Krauth, Department of Economics, Simon Fraser University, 8888 University Dr Burnaby BC V5A 1S6, Canada, E-mail: bkrauth@sfu.ca

- Generalizes Altonji, Elder and Taber (2005)

- Generalizes Altonji, Elder and Taber (2005)
- Allows for relative correlation restriction (RCR):

$$Corr(d, \varepsilon) = \lambda Corr(d, X\beta)$$

Two uses of λ:

Two uses of λ:

1. Assume $\lambda \in [\lambda_L, \lambda_H]$ and estimate corresponding α's in interval $[\alpha_L, \alpha_H]$

Two uses of λ:

1. Assume $\lambda \in [\lambda_L, \lambda_H]$ and estimate corresponding α's in interval $[\alpha_L, \alpha_H]$

2. Estimate α by OLS, then find smallest (absolute) value of λ such that OLS estimate is statistically zero

Unobservable Selection and Coefficient Stability: Theory and Evidence

Emily OSTER
Department of Economics, Brown University, Providence, Rhode Island, and NBER (*emily_oster@brown.edu*)

A common approach to evaluating robustness to omitted variable bias is to observe coefficient movements after inclusion of controls. This is informative only if selection on observables is informative about selection on unobservables. Although this link is known in theory in existing literature, very few empirical articles approach this formally. I develop an extension of the theory that connects bias explicitly to coefficient stability. I show that it is necessary to take into account coefficient and R-squared movements. I develop a formal bounding argument. I show two validation exercises and discuss application to the economics literature. Supplementary materials for this article are available online.

KEY WORDS: Coefficient stability; Selection; Omitted variable bias.

1. INTRODUCTION

Concerns about omitted variable bias are common to most or all nonexperimental work in economics. (Despite recent trends, this still makes up the vast majority of results within economics: in 2012 the combination of the *American Economic Review*, the *Quarterly Journal of Economics*, and the *Journal of Political Economy* published 69 empirical, nonstructural articles, only 11 of which were randomized.) The most straightforward approach to such concerns is to include controls that can be observed. Angrist and Pischke (2010) argued that among the major advances in empirical economics in the past two decades is greater effort to identify the most important threats to validity, and to address them with appropriate selection of controls.

In some cases it is possible to argue that a control (or set of controls) fully captures a particular omitted variable. However, in many cases observed controls are an incomplete proxy for the true omitted variable or variables. For example, it is common in many applications to worry about confounding from socioeconomic status. Researchers often include controls to capture this, but typically with the acknowledgment that the controls observed in a typical dataset (e.g., education and income categories, race groups) do not perfectly capture overall socioeconomic status. Similarly, in many cross-country or cross-regional analyses authors seek to control for geographic differences across areas, but observed controls (extent of mountains, access to water) are incomplete proxies for the true omitted factor.

A common approach in these situations is to explore the *sensitivity* of treatment effects to the inclusion of observed controls. If a coefficient is stable after inclusion of the observed controls, this is taken as a sign that omitted variable bias is limited. (The next section and the final section of this article will discuss more explicitly the use of this approach within economics, but it is worth noting that it is the link between coefficient stability and omitted variable bias is often quite direct. For example, Chiappori, Oreffice, and Quintana-Domeque (2012) stated: "It is reassuring that the estimates are very similar in the standard and the augmented specifications, indicating that our results are unlikely to be driven by omitted variables bias." Similarly, Lacetera, Pope, and Sydnor (2012) stated: "These controls do not change the coefficient estimates meaningfully, and the stability of the estimates from columns 4 through 7 suggests that controlling for the model and age of the car accounts for most of the relevant selection.") The intuitive appeal of this approach lies in the idea that the bias arising from the observed (imperfect) controls is informative about the bias that arises from the full set including the unobserved components. This is not, however, implied by the baseline assumptions underlying the linear model.

Formally, using the observables to identify the bias from the unobservables requires making further assumptions about the covariance properties of the two sets. In particular, the case in which the omitted variable bias is fully identified by the observed controls corresponds to the extreme assumption that the relationship between treatment and unobservables can be fully recovered from the relationship between treatment and observables (Murphy and Topel 1990; Altonji, Elder, and Taber 2005a; Altonji et al. 2011).

Even under this most optimistic assumption, however, coefficient movements alone are not a sufficient statistic to calculate bias. To illustrate why, consider the case of a researcher estimating wage returns to education with individual ability as the only confounder, and where there are two orthogonal components of ability, one of which has a higher variance than the other. Assume wages would be fully explained if both ability components were observed but, in practice, the researcher sees only one of the two. The coefficient will appear much more stable if the observed ability control is the lower variance one, but this is not because the bias is smaller but simply because less of the wage outcome is explained by the controls.

Journal of Business & Economic Statistics
April 2019, Vol. 37, No. 2
DOI: 10.1080/07350015.2016.1227711
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jbes.

- Also generalizes Altonji et al. (2005)

- Also generalizes Altonji et al. (2005)
- Focuses on comparing movements in α with movements in R^2

- Also generalizes Altonji et al. (2005)
- Focuses on comparing movements in α with movements in R^2
- Intuition: if we could observe all unobservables, then $R^2 = 1$

- Also generalizes Altonji et al. (2005)
- Focuses on comparing movements in α with movements in R^2
- Intuition: if we could observe all unobservables, then $R^2 = 1$
- Value of α when $R^2 = 1$ represents true causal value

- Also generalizes Altonji et al. (2005)
- Focuses on comparing movements in α with movements in R^2
- Intuition: if we could observe all unobservables, then $R^2 = 1$
- Value of α when $R^2 = 1$ represents true causal value
- If measurement error in y, consider $R_{\text{max}} < 1$ instead

- Also generalizes Altonji et al. (2005)
- Focuses on comparing movements in α with movements in R^2
- Intuition: if we could observe all unobservables, then $R^2 = 1$
- Value of α when $R^2 = 1$ represents true causal value
- If measurement error in y, consider $R_{\text{max}} < 1$ instead
- Implementation closely similar to Krauth (2016)