Bounding a linear causal effect using relative correlation restrictions [*]

Brian Krauth
Department of Economics
Simon Fraser University

January 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: sensitivity analysis, partial identification, endogeneity

[*] This research has benefited from comments by seminar audiences at Duke, Guelph, McMaster, Toronto, Virginia, Waterloo, York, the Federal Trade Commission, the 2008 CEA meetings, and the 2006 Joint Statistical Meetings. All errors are mine. Author contact information: email bkrauth@sfu.ca. The author acknowledges support from the Social Sciences and Humanities Research Council of Canada. Stata code is available at http://www.sfu.ca/b̃krauth/code.

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), linear regression (Altonji, Elder, and Taber, 2005a), instrumental variables (Altonji et al., 2005a, Conley, Hansen, and Rossi, 2012, Kraay, 2012, Nevo and Rosen, 2012), and simultaneous equations (Lewbel, 2012). The primary contribution of this paper is threefold. First, it enables the construction of bounds on the true effect under any restriction on the sensitivity parameter, rather than just point estimates for a single value of this parameter. This is important in applied work because the bounds can potentially change dramatically with small changes in the sensitivity parameter. Second, it applies an explicit partial identification framework to sensitivity analysis in the linear regression setting. This framework clarifies several issues of identification and inference. For example, it enables the construction of Imbens-Manski (2004) confidence intervals for partially identified parameters. Finally, the paper shows two example applications that demonstrate the methodology's potential to extract useful new information from data.

The first application is to a field experiment (Krueger, 1999) in which there are small deviations from pure random assignment. The second application is to an observational study (Bleakley, 2010b) in which the claim of exogeneity is more controversial. I find that the experimental study is substantially more robust than the observational study. In particular, Krueger's estimate of the effect of smaller classes on kindergarten test scores has narrow bounds and remains statistically significant even when the correlation between class size and unobservables is several times as large as the correlation observed between class size and the observed control variables. In contrast, the bounds on Bleakley's estimate of the effect of malaria on labour productivity are wide and include zero if the correlation between malaria and unobservables is as much as 30% of the correlation between malaria and the observed control variables. These findings are entirely based on patterns in the data, but correspond to what we would expect given knowledge of each study's research design. The results thus provide some evidence of this method's potential to generate an informative sensitivity analysis.

1.1 Related literature

Empirical researchers in economics have long augmented their main results with some form of informal sensitivity analysis. Leamer (1978) was an early and forceful proponent of formalizing and expanding the use of sensitivity analysis in parametric models, and developed Bayesian-influenced methods for systematic sensitivity analysis of measurement error (Klepper and Leamer, 1984), model selection (Leamer, 1978), and other common empirical problems. Manski (1994, 2003) adopts a mostly nonparametric approach, and recasts sensitivity analysis as estimation under assumptions that yield only partial identification of the parameter or estimand of interest. Manski's research has also inspired an extensive theoretical literature on inference under partial identification.

The particular type of sensitivity parameter used in this paper is similar in spirit to those seen in a number of recent papers, in that it restricts the unmeasurable deviation from conditional exogeneity in terms that are relative to some related measurable quantity. Rosenbaum (2002) develops a treatment-effects framework in which there is an unobserved binary variable affecting both outcomes and selection into treatment. Rosenbaum's sensitivity parameter is defined as the maximum odds ratio of (unobserved) treatment probabilities among pairs of cases that have been matched on observed characteristics. Imbens (2003) considers a parametric treatment effects model and uses as a sensitivity parameter the proportion of otherwise unexplained variation in the outcome that could be explained by the unobserved term in the treatment selection equation. Altonji, Elder, and Taber (2005b) consider a bivariate probit model where the cross-equation correlation of unobservable terms is proportional to the correlation in observable terms. Krauth (2007) considers a multiple-equation probit peer effects model in which the within-group correlation in unobservables is proportional to the within-group correlation in observables. Lewbel (2012) exploits a cross-equation covariance restriction to bound the parameters of a heteroskedastic simultaneous equations model without using instruments. Kreider and Hill (2009) and Kreider (2010) provide sensitivity analysis for treatment effects when treatment is subject to classification error. Conley, Hansen and Rossi (2012), Nevo and Rosen (2012) and Kraay (2012) develop tools for sensitivity analysis of instrumental variables regression in which the conventional IV exclusion restriction is “almost” true.

The most closely related papers are Altonji et al. (2005a) and Oster (2014) Altonji et al. (AET) propose an estimator of the bias in OLS or 2SLS estimation when the relationship between the variable of interest (for OLS) or its instrument (for IV) and unobservables is proportional to the relationship between that variable and the other observables. There are several important differences between the estimator in their paper and the one presented here. First, the relationship in AET is parameterized in terms of a ratio of linear projection coefficients, where the relationship here is parameterized as a ratio of correlations. The choice of parameterization is a matter of convenience here as there is a simple proportional relationship between these two ratios. Second, AET perform two calculations: a point estimate or bias correction under the assumption that the ratio of linear projection coefficients is exactly one, and a point estimate of the ratio under the null of no effect. In contrast, the approach described here allows both estimation of bounds and construction of Imbens-Manski (2004) confidence intervals under any assumption about the possible range of values for the sensitivity parameter. Oster (2014) proposes an extension of AET’s model in which unexplained variation in the outcome can be divided up into variation that could be explained by unobserved variables, and variation that is purely idiosyncratic (e.g., classical measurement error) and can be treated as truly exogenous. The model is then parameterized in terms of two sensitivity parameters: the proportional selection parameter as in AET ($\tilde{\delta}$), and the R-squared from the regression of the outcome on all non-idiosyncratic variables (R_{max}). Oster then uses results from several random-assignment studies to suggest an empirically plausible rule for setting $\tilde{\delta}$ and R_{max}. This approach has the advantage of providing a more empirically-grounded basis for setting appropriate values of the sensitivity parameters, at a cost of added complexity since there are two sensitivity parameters rather than one. In contrast to AET, Oster explicitly treats the resulting estimates as bounds on partially identified parameters, but does not take the additional step of developing inference procedures as done in this paper.

2 Overview

Consider an empirical researcher who is estimating a model of the form:

$$y = x\beta_x + \mathbf{c}\beta_{\mathbf{c}} + \varepsilon \tag{1}$$

where y is an outcome variable of interest, x is an explanatory variable of interest, and $\mathbf{c}$ is a vector of control variables. The researcher wishes to estimate β_x, which is interpreted as the effect of x on y. The control variables are treated as exogenous:[1]

$$\operatorname{corr}(\mathbf{c}, \varepsilon) = 0 \tag{2}$$

but the variable of interest x is potentially correlated with ε. In the absence of a suitable instrument for x, the researcher's only conventional option is to assume that the explanatory variable of interest is also exogenous (i.e., $\operatorname{corr}(x, \varepsilon) = 0$) and estimate the model by OLS, or to accept that the effect of interest is not identified.

The middle-ground alternative developed in this paper is to define a *relative correlation* parameter λ that satisfies:

$$\operatorname{corr}(x, \varepsilon) = \lambda \operatorname{corr}(x, \mathbf{c}\beta_{\mathbf{c}}) \tag{3}$$

and thus describes the correlation between the variable of interest and unobservables relative to the correlation between the variable of interest and a particular index of the observed control variables. The choice of index is entirely a matter of convenience, as another index would just imply a different value of λ. The particular index used in this paper weights elements

[1] As long as no causal interpretation is imposed on $\beta_{\mathbf{c}}$, it can be defined so that $\mathbf{c}$ is exogenous by construction. However, this means that when referring to variation in ε as variation in "unobservables" we really mean the portion of variation in unobservables that is orthogonal to variation in observables. When $\beta_{\mathbf{c}}$ is given a causal interpretation, exogeneity of $\mathbf{c}$ is a nontrivial assumption.

of $\mathbf{c}$ based on their statistical relationship with the outcome variable (the corresponding elements of $\beta_{\mathbf{c}}$), and has the useful property of invariance to arbitrary linear transformations of $\mathbf{c}$. Alternative indices that use factor analysis or similar dimension reduction techniques can be easily accommodated in the framework developed here.

On its own, equation (3) imposes almost no restrictions on $\text{corr}(x, \varepsilon)$ and can be interpreted merely as a definition for λ. Without further restrictions, the model parameters are not identified. While the conventional exogeneity assumption ($\lambda = 0$) is sufficient for point identification and consistent estimation by OLS, this paper considers a weaker *relative correlation restriction* (RCR) of the form:

$$\lambda^L \leq \lambda \leq \lambda^H \tag{4}$$

for some particular λ^L and λ^H. Given a sufficiently strong restriction, it is possible to estimate bounds on β_x as well as conduct hypothesis tests and construct confidence intervals. Alternatively, λ could be used as a sensitivity parameter for the OLS estimates. That is, a researcher could estimate β_x by OLS, and then estimate the smallest value of λ that implies the OLS estimate is not robust (either in the sense that the bounds contain zero or in the sense that the confidence interval contains zero). Both of these approaches are demonstrated in the empirical examples in Section 5.

The usefulness of this model rests on the idea that a researcher can impose plausible *a priori* bounds on λ. This in turn requires that the magnitude and sign of the correlation between x and $\mathbf{c}\beta_{\mathbf{c}}$ provides at least some information about the magnitude and sign of the correlation between x and ε. For example, the restriction $-1 \leq \lambda \leq 1$ would imply that the correlation between the explanatory variable of interest is no more correlated with unobservables than it is with the observable control variables, while the restriction $0 \leq \lambda \leq 1$ would imply the additional assumption that the two correlations have the same sign.

It is common practice in applied work to use patterns in observed explanatory variables as evidence in favor of ultimately untestable assumptions about unobserved variables. Papers using an experimental design, including the Krueger (1999) study used as an example application in Section 5.1, usually include a table showing that observed pre-treatment variables are roughly balanced between treatment and control groups, and interpret this balance as evidence for the balance in unobserved covariates that is required for identification. Observational studies using control variables, including the Bleakley (2010b) study used as an example application in Section 5.2, often report a simple regression, a "preferred specification" that includes the researcher's preferred control variables, and then some "robustness check" specifications that include additional control variables. The researcher then shows that the estimated effect changes substantially from the simple regression to the preferred specification, but does not change much between the preferred specification and the robustness checks. This is then used to argue that the identification problem has been solved, i.e., the researcher has found the exact set of control variables such that the remaining omitted variables are uncorrelated with the explanatory variable of interest. In other words, it is common in both experimental and observational studies to informally use low correlation between the explanatory variable of interest and some control variables as evidence in support of the identifying assumption of zero correlation between the explanatory variable of interest and the regression error term. This inference is usually implicit, and takes an "all or nothing" form: if the observed correlation is low enough, then it is assumed that the unobserved correlation can be taken as exactly zero. By making this inference explicit, this implicit decision rule can be replaced with a more plausible one: a low observable correlation suggests a low (but not necessarily zero) unobservable correlation, while a higher observable correlation suggests a higher unobservable correlation.

While the approach presented here emphasizes calculating bounds under multiple plausible assumptions on the sensitivity parameters, other work has suggested attention to particular assumptions. Altonji et al. (2005a, 2005b) argue in favor of a particular assumption which they call "equal selection on observables and unobservables" as providing an upper bound on the bias from OLS/2SLS. The argument takes the form of an explicit formal model with outcome and selection equations in which a large set of observed variables are selected randomly from a much larger set of possible variables. As a result of this random selection, the observed variables and unobserved variables (once normalized by their coefficients in the outcome equation) are likely to have similar coefficients in the selection equation, with equality in the limit as one adds variables. The bias estimated under the assumption of equal selection can be considered an upper bound if observed variables are selected specifically because of their potential to reduce bias in the OLS/2SLS estimates. The equal selection condition in Altonji et al.'s model is analogous but not identical to the restriction of equal correlation ($\lambda = 1$) in the model presented here. Oster (2014) takes Altonji et al.'s argument further by noting that some fraction of variation in outcomes is truly idiosyncratic (for example, classical measurement error), and so the bound can further be narrowed by imposing an empirically plausible upper bound on the R^2 from the regression of the outcome on both the observed and non-idiosyncratic unobserved variables.

Section 3 below describes the model more precisely and develops estimation and inference methods. In the interest of space, tractability and clarity the model is kept simple: the variable of interest is assumed to have a constant linear effect on the outcome variable, and is a scalar. These simplifications can be relaxed. Section 4.2 extends the analysis to cover heterogeneous effects. The other simplifications can also be relaxed, but doing so is beyond the scope of this paper. Nonlinear effects have been addressed by previous authors (Altonji et al., 2005b, Krauth, 2007), and require much more detailed parametric restrictions on the relationships among model variables. The case where x is a k-vector can in principle be handled by making λ a k-vector as well, and is left for future research.

3 Detailed methodology

3.1 Model

Let $\mathbf{d} \equiv [y\,x\,\mathbf{c}]$, where y is a scalar outcome, x is a scalar explanatory variable of interest, and $\mathbf{c}$ is a k-length row vector of additional control variables including an intercept. The causal model is:

$$\text{ASSUMPTION 1:} \qquad y = y(x) = x\beta_x + u$$

where the random function $y(\cdot)$ is a potential outcome function giving the outcome associated with each possible value of x, the parameter of interest β_x represents the effect of x on y, and the unobserved random variable u represents the effect of all other factors. These other factors are not affected by x but may be correlated with it. Section 4.2 considers an extension in which the effect of x on y is heterogeneous across individuals.

The control variables in $\mathbf{c}$ are of interest primarily as an aid to identification of β_x, and so are not included explicitly in the structural model. They may or may not have a direct effect on y, and that effect may or may not be linear. Let $u^p = \mathbf{c}\beta_\mathbf{c}$ be the best linear predictor of u given $\mathbf{c}$, i.e.:

$$\begin{aligned} \beta_\mathbf{c} &\equiv E(\mathbf{c}'\mathbf{c})^{-1}E(\mathbf{c}'u) \qquad (5)\\ &= E(\mathbf{c}'\mathbf{c})^{-1}E(\mathbf{c}'y) - \beta_x E(\mathbf{c}'\mathbf{c})^{-1}E(\mathbf{c}'x) \\ \underbrace{\mathbf{c}\beta_\mathbf{c}}_{u^p} &= \underbrace{\mathbf{c}E(\mathbf{c}'\mathbf{c})^{-1}E(\mathbf{c}'y)}_{y^p} - \beta_x \underbrace{\mathbf{c}E(\mathbf{c}'\mathbf{c})^{-1}E(\mathbf{c}'x)}_{x^p} \end{aligned}$$

(where y^p and x^p are the best linear predictors of y and x, respectively, given $\mathbf{c}$) and let ε be the corresponding residual:

$$\varepsilon \equiv u - \mathbf{c}\beta_\mathbf{c} \qquad (6)$$

Note that these are just definitions and that $\beta_{\mathbf{c}}$ has no particular causal interpretation. Putting (5) and (6) together, we get:

$$y = x\beta_x + \mathbf{c}\beta_{\mathbf{c}} + \varepsilon \qquad \text{where } E(\mathbf{c}'\varepsilon) = 0 \tag{7}$$

which looks like the usual OLS regression equation, but is missing the necessary assumption that $E(x\varepsilon) = 0$, or equivalently that $\text{corr}(x, \varepsilon) = 0$.

That assumption is replaced by a weaker relative correlation restriction, which is defined as a nonempty and closed interval Λ that is known by the econometrician to satisfy:

$$\text{ASSUMPTION 2:} \qquad cov(x,\varepsilon)\sqrt{var(\mathbf{c}\beta_{\mathbf{c}})} = \lambda cov(x,\mathbf{c}\beta_{\mathbf{c}})\sqrt{var(\varepsilon)} \quad \text{for some } \lambda \in \Lambda$$

Assumption 2 is written in a somewhat nonintuitive fashion to allow maximum generality. In the usual case where $cov(x,\mathbf{c}\beta_{\mathbf{c}})$ and $var(\varepsilon)$ are both nonzero and Λ is the finite interval $[\lambda^L, \lambda^H]$, Assumption 2 simplifies to:

$$\lambda^L \leq \underbrace{\frac{\text{corr}(x,\varepsilon)}{\text{corr}(x,\mathbf{c}\beta_{\mathbf{c}})}}_{\lambda} \leq \lambda^H \tag{8}$$

That is, we are assuming that the correlation of the variable of interest (x) with unobservables (ε) relative to its correlation with observables ($\mathbf{c}\beta_{\mathbf{c}}$) can be restricted to lie within some known range (Λ).

The data takes the form of a sample of size n on $\mathbf{d}$ that can be used to construct a consistent and asymptotically normal estimator $\hat{m}_n$ of its first two moments. That is, let:

$$m_0 \equiv \text{vech}(E(\mathbf{d}'\mathbf{d}))$$

where $\text{vech}(\cdot)$ is the half-vectorization function (i.e., given a symmetric matrix it returns a column vector of its unique elements). We assume that our estimator of m_0 satisfies:

$$\text{ASSUMPTION 3:} \qquad \sqrt{n}\,(\hat{m}_n - m_0) \xrightarrow{D} N(0,\Sigma)$$

where: Since m_0 is just a vector of expected values, Assumption 3 is satisfied in a random sample by the corresponding sample average $\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n}\mathbf{d}_i'\mathbf{d}_i$

Finally, a few convenient and easily-verified conditions are imposed on m_0. First, all variables exhibit nontrivial variation:

$$\text{ASSUMPTION 4:} \qquad E(\mathbf{d}'\mathbf{d}) \text{ is finite and positive definite}$$

Positive-definiteness of $E(\mathbf{d}'\mathbf{d})$ is easily verified in data, and guarantees for example that $\beta_{\mathbf{c}}$ is well-defined. Next, at least one of the control variables is useful in forecasting y:

$$\text{ASSUMPTION 5:} \qquad var(y^p) > 0$$

Assumption 5 can be tested by an ordinary coefficient significance test.

The final assumption, made primarily for convenience, is that at least one of the control variables is useful in forecasting x:

$$\text{ASSUMPTION 6:} \qquad var(x^p) > 0$$

Assumption 6 is also easily testable, and simplifies the description of the estimation method and its properties in the remainder of this section. Section 4.1 relaxes this assumption and shows that the key properties of the estimation method are unaffected.

3.2 Identification

In general, it is not possible in this setting to identify the true value of β_x, but it is possible to identify a nontrivial set B_x that must contain β_x. This set is known as the *identified set* for the true effect, and includes ordinary point identification ($B_x = \{\beta_x\}$), partial identification ($B_x \subsetneq \mathbb{R}$), and nonidentification ($B_x = \mathbb{R}$) as special cases. This section characterizes the identified set.

First, note that the linear structure of the model implies that identification can be discussed entirely in terms of the relative correlation restriction Λ and the vector of second moments m_0. Estimation will then be based on a plug-in estimator that substitutes $\hat{m}_n$ for the unknown m_0. Let an *allowable second moment vector* be defined as an arbitrary vector m the same length as m_0 such that:

$$E_m(\mathbf{d}'\mathbf{d}) \text{ is finite and positive definite} \tag{9}$$

$$var_m(y^p) > 0 \tag{10}$$

$$var_m(x^p) > 0 \tag{11}$$

where the subscript m indicates that the expected values in question are calculated as if the unknown vector of second moments m_0 were equal to m (i.e., $E_m(\mathbf{d}'\mathbf{d}) = \text{vech}^{-1}(m)$). This notation will be useful in describing estimators for the parameters of interest that are based on $\hat{m}_n$, which in a sufficiently large sample will be close to m_0 but not identical. The model's assumptions described in Section 3.1 imply that m_0 satisfies (9)-(11) and is thus an allowable second moment vector. Since $E_m(\mathbf{d}'\mathbf{d})$ is a continuous function of m, these conditions are also satisfied by any m sufficiently close to m_0. This will in turn imply that since $\hat{m}_n \xrightarrow{p} m_0$, the probability that $\hat{m}_n$ satisfies these conditions will be going to one as n goes to infinity.

Next, note that both $\beta_{\mathbf{c}}$ and λ would be identified if β_x were known. Ignoring for the moment the possibility of singular matrices or division by zero, let:

$$\beta_{\mathbf{c}}(b_x;m) \equiv E_m(\mathbf{c}'\mathbf{c})^{-1}E_m(\mathbf{c}'y) - b_x E_m(\mathbf{c}'\mathbf{c})^{-1}E_m(\mathbf{c}'x) \tag{12}$$

and:

$$\lambda(b_x;m) \equiv \frac{\mathrm{corr}_m(x, y - b_x x - \mathbf{c}\beta_{\mathbf{c}}(b_x;m))}{\mathrm{corr}_m(x, \mathbf{c}\beta_{\mathbf{c}}(b_x;m))} \tag{13}$$

Equations (12) and (13) can be used to express the unknown parameters $\beta_{\mathbf{c}}$ and λ as known functions of the unknown structural parameter and vector of second moments, i.e.: $\beta_{\mathbf{c}} = \beta_{\mathbf{c}}(\beta_x;m_0)$ and $\lambda = \lambda(\beta_x;m_0)$.

Finally, let $B_x(\Lambda;m)$ be defined as the set of all b_x satisfying:

$$\begin{aligned} & cov_m(x, y - b_x x - \mathbf{c}\beta_{\mathbf{c}}(b_x;m))\sqrt{var_m(\mathbf{c}\beta_{\mathbf{c}}(b_x;m))} \\ & \qquad = \lambda cov_m(x, \mathbf{c}\beta_{\mathbf{c}}(b_x;m))\sqrt{var_m(y - b_x x - \mathbf{c}\beta_{\mathbf{c}}(b_x;m))} \end{aligned} \tag{14}$$

for some $\lambda \in \Lambda$. By construction, $B_x(\Lambda;m_0)$ is the identified set for the true effect. It will usually be more convenient to work with its upper and lower bounds:

$$\beta_x^L(\Lambda;m) = \inf B_x(\Lambda;m) \tag{15}$$

$$\beta_x^H(\Lambda;m) = \sup B_x(\Lambda;m) \tag{16}$$

The identified set is not always convex, so these bounds are not necessarily sharp.

Figure 1 shows a typical example of what the $\lambda(b_x;m)$ function looks like, while Proposition 1 below describes its most important features more formally.

Proposition 1 (Properties of $\lambda(\cdot)$) *Let m satisfy (9)-(11). Then the function $\lambda(.;m)$ has the following properties:*

1. *$\lambda(b_x;m)$ exists and is differentiable for all $b_x \neq \beta_x^\infty(m)$, where:*

$$\beta_x^\infty(m) \equiv \frac{cov_m(x^p, y^p)}{var_m(x^p)}$$

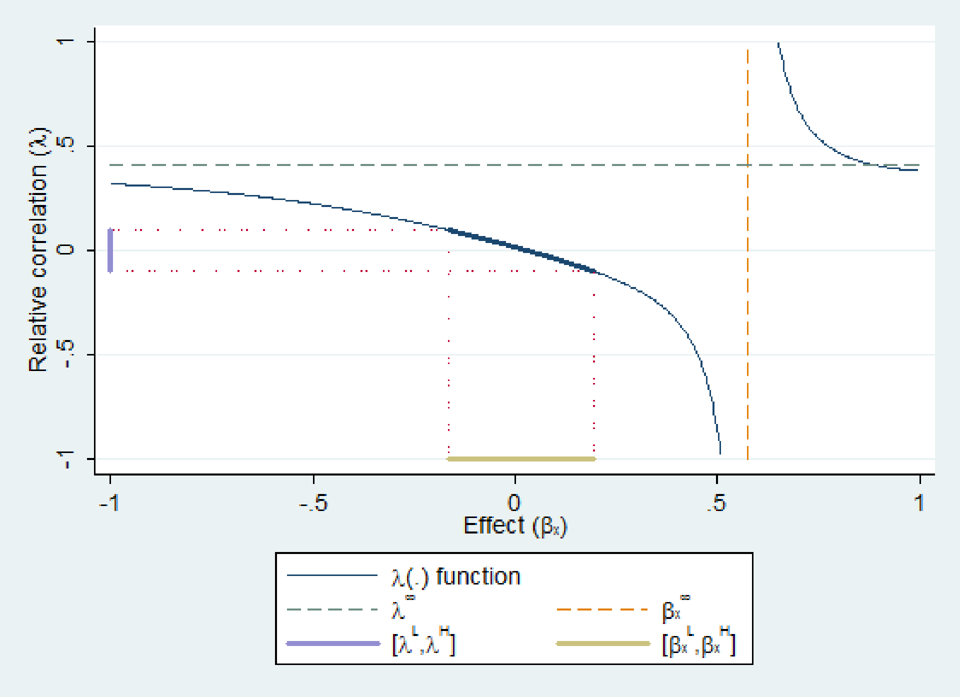

Figure 1: A typical $\lambda(b_x;m)$ function giving the relative correlation (λ) under the assumption that $\beta_x = b_x$. In general, the function exists and is differentiable in b_x everywhere but at β_x^∞ (the value of b_x at which $\text{corr}(x, \mathbf{c}\beta_{\mathbf{c}}(b_x)) = 0$). Its limit as b_x approaches positive or negative infinity is λ^∞. Near β_x^∞, the function goes towards positive or negative infinity. Both β_x^∞ and λ^∞ are easily identified from the data. Given the relative correlation restriction $\Lambda = [\lambda^L, \lambda^H]$, the bounds $[\beta_x^L, \beta_x^H]$ can be found by inverting $\lambda(b_x;m)$.

2. *Let:*

$$\lambda^\infty(m) \equiv \sqrt{\frac{var_m(x)}{var_m(x^p)} - 1}$$

Then $\lambda^\infty(m) \geq 0$ *and:*

$$\lim_{b_x \to \infty} \lambda(b_x; m) = \lim_{b_x \to -\infty} \lambda(b_x; m) = \lambda^\infty(m)$$

3. *For any* $\lambda \neq \lambda^\infty(m)$ *there exists at least one* b_x *satisfying (14).*

4. *Let:*

$$\tilde{B}(\Lambda; m) = \{b_x : \lambda(b_x; m) \in \Lambda\}$$

Then:

$$\tilde{B}(\Lambda; m) \subset B_x(\Lambda; m) \subset \tilde{B}(\Lambda; m) \cup \{\beta_x^\infty(m)\}$$

Proof: See Appendix A.1.

Although $\lambda(b_x; m)$ is continuous and differentiable for all $b_x \neq \beta_x^\infty(m)$ it is not always monotonic and can contain local maxima or minima. This non-monotonicity implies that the correct bounds as defined in equations (15) and (16) are not necessarily equal to $\inf(\{b : \lambda(b_x; m) \in \{\lambda^L, \lambda^H\}\})$ and $\sup(\{b : \lambda(b_x; m) \in \{\lambda^L, \lambda^H\}\})$. That is, the bounds cannot necessarily be constructed by just finding the point estimates associated with λ^L and λ^H. The estimation method described in Section 3.3 accounts for this issue and estimates the bounds as defined in equations (15) and (16).

Proposition 2 is the primary identification result of the paper, and describes conditions under which the identified set is both nonempty and bounded. Under these conditions, data can be used to estimate nontrivial bounds on the true effect.

Proposition 2 (Size of the identified set) *The identified set* $B_x(\Lambda; m_0)$ *is nonempty and bounded if* $\lambda^\infty(m_0) \notin \Lambda$.

Proof: See Appendix A.2.

3.3 Estimation

The identified features of the model can be estimated by substituting $\hat{m}_n$ for m_0 in the quantities defined in Section 3.2. Let:

$$\begin{aligned}
\hat{\lambda}(b_x) &\equiv \lambda(b_x;\hat{m}_n) \\
\hat{\lambda}^{\infty} &\equiv \lambda^{\infty}(\hat{m}_n) \\
\hat{\beta}_x^{\infty} &\equiv \beta_x^{\infty}(\hat{m}_n) \\
\hat{\beta}^L(\Lambda) &\equiv \inf\{b_x : \lambda(b_x;\hat{m}_n) \in \Lambda\} \\
\hat{\beta}^H(\Lambda) &\equiv \sup\{b_x : \lambda(b_x;\hat{m}_n) \in \Lambda\}
\end{aligned} \tag{17}$$

An important complication in characterizing the asymptotic properties of these estimators is the possibility of nonidentification. That is, the estimated bounds should go to infinity when the identified set is unbounded. Proposition 3 below shows this to be the case.

Proposition 3 (Consistency) *The estimators defined in (17) are consistent. That is:*

$$\begin{aligned}
&\hat{\beta}_x^{\infty} \xrightarrow{p} \beta_x^{\infty}(m_0) \\
&\hat{\lambda}^{\infty} \xrightarrow{p} \lambda^{\infty}(m_0) \\
&\hat{\lambda}(b_x) \xrightarrow{p} \lambda(b_x;m_0) \qquad \textit{for all } b_x \neq \beta_x^{\infty}(m_0)
\end{aligned}$$

If $B_x(\Lambda;m_0)$ is bounded then:

$$\begin{aligned}
\hat{\beta}^L(\Lambda) &\xrightarrow{p} \beta_x^L(\Lambda;m_0) \qquad \textit{if } \frac{d\lambda(b_x;m_0)}{db_x}\Big|_{b_x=\beta_x^L(\Lambda;m_0)} \neq 0 \\
\hat{\beta}^H(\Lambda) &\xrightarrow{p} \beta_x^H(\Lambda;m_0) \qquad \textit{if } \frac{d\lambda(b_x;m_0)}{db_x}\Big|_{b_x=\beta_x^H(\Lambda;m_0)} \neq 0
\end{aligned}$$

and if $B_x(\Lambda;m_0) = \mathbb{R}$ then for any B:

$$\lim_{n\to\infty} \Pr((\hat{\beta}^H(\Lambda) > B) = \lim_{n\to\infty} \Pr((\hat{\beta}^L(\Lambda) < B) = 1$$

Proof: See Appendix A.3.

Note that consistency of $\hat{\beta}^L(\Lambda)$ (for example) requires two conditions to be satisfied: that $\beta_x^L(\Lambda;m_0) \neq \beta_x^{\infty}(m_0)$ (guaranteeing existence of the derivative $\partial\lambda(b_x;m)/\partial b_x$ when evaluated at $\beta_x^L(\Lambda;m_0)$), and that $\partial\lambda(b_x;m)/\partial b_x$ is nonzero when evaluated at $\beta_x^L(\Lambda;m_0)$. By analogy, $\hat{\beta}^L(\Lambda)$ is likely to be a noisy estimator when either $\beta_x^L(\Lambda;m_0)$ is close to $\beta_x^{\infty}(m_0)$, or when $\lambda^{\infty}(m_0)$ is close to Λ (since result 2 of Proposition 1 implies that $\partial\lambda(b_x;m)/\partial b_x \to 0$ as $|b_x| \to \infty$).

3.4 Inference

Hypothesis tests and confidence intervals can be constructed for partially identified parameters, with no change in interpretation.

A first step in developing inference is to obtain an asymptotic distribution for the estimators defined in equation (17) . These estimators are in most cases differentiable functions of $\hat{m}_n$, so they will be asymptotically normal with a covariance matrix that can be obtained through straightforward application of the delta method. Proposition 4 below describes the asymptotic distribution for the bounds. This asymptotic distribution can be used to construct Wald-type hypothesis tests and confidence intervals for β_x.

Proposition 4 (Asymptotic distribution for estimated bounds) *Let:*

$$A \equiv - \begin{bmatrix} \left. \dfrac{\nabla_m \lambda(b_x;m)}{\partial \lambda(b_x;m)/\partial b_x} \right|_{b_x=\beta_x^L(\Lambda;m_0), m=m_0} \\ \left. \dfrac{\nabla_m \lambda(b_x;m)}{\partial \lambda(b_x;m)/\partial b_x} \right|_{b_x=\beta_x^H(\Lambda;m_0), m=m_0} \end{bmatrix}$$

where the row vector $\nabla_m \lambda(b_x,m)$ is the gradient of $\lambda(b_x,m)$ with respect to m. If A exists, then:

$$\sqrt{n} \begin{bmatrix} \hat{\beta}^L(\Lambda) - \beta_x^L(\Lambda;m_0) \\ \hat{\beta}^H(\Lambda) - \beta_x^H(\Lambda;m_0) \end{bmatrix} \xrightarrow{D} N\left(0, A\Sigma A'\right)$$

Proof: See Appendix A.4.

Existence of the matrix A requires two conditions to be satisfied: that neither β_x^L nor β_x^H is identical to $\beta_x^\infty(m_0)$ (guaranteeing existence of the derivatives), and that $\partial \lambda(b_x;m)/\partial b_x$ is nonzero when evaluated at β_x^L or β_x^H. By analogy, the asymptotic distribution is likely to provide a poor approximation to the finite sample distribution when either β_x^L or β_x^H is close to β_x^∞, or when λ^∞ is close to Λ.

In constructing confidence intervals under partial identification, Imbens and Manski (2004) note the necessity of distinguishing between a confidence interval for the identified set:

$$\lim_{n\to\infty} \Pr(B_x(\Lambda) \subset CI^{set}) = 1 - \alpha$$

and a confidence interval for the true parameter value:

$$\lim_{n\to\infty} \inf_{b_x \in B_x(\Lambda)} \Pr(b_x \in CI^{par}) = 1 - \alpha$$

A confidence interval for the identified set can be constructed using the lower and upper bounds, respectively, of the ordinary confidence intervals for $\hat{\beta}^L(\Lambda)$ and $\hat{\beta}^H(\Lambda)$. A confidence interval for the true parameter value is generally narrower than one for the identified set. Imbens and Manski describe a method of constructing such a confidence interval by reducing the critical values to account for the width of the identified set. Stoye (2009) notes that validity of the Imbens-Manski procedure requires a strong assumption of superefficient estimation for the width of the identified set. However, he also shows that superefficiency will hold if the estimators of the bounds are jointly asymptotically normal and ordered by construction (Stoye, 2009, Lemma 3). These criteria are satisfied in the setting of this paper, and so Stoye's more elaborate procedure is not required.

Hypothesis tests are subject to similar considerations. A one-sided hypothesis test on β_x can be implemented by a conventional one-sided test on either β_x^L or β_x^H. A two-sided hypothesis test is potentially more complex because the simple null $H_0 : \beta_x = b_x$ can be rejected if and only if the joint null $H_0 : (\beta_x^L \leq b_x, \beta_x^H \geq b_x)$ can be rejected. While there exist many options for hypothesis testing of multiple inequalities, a simple solution is to just invert the Imbens-Manski confidence interval. For example, we reject the null $H_0 : \beta_x = b_x$ at 5% significance if b_x is outside of the 95% confidence interval.

4 Extensions

4.1 Pure random assignment

The main results in Section 3 are derived under the assumption (Assumption 6) that the explanatory variable of interest is at least slightly correlated with the control variables. This assumption is made strictly for convenience in presenting results, as it guarantees existence of the intermediate quantities $\lambda(b_x, m_0)$, $\lambda^\infty(m_0)$, and $\beta_x^\infty(m_0)$, and avoids the need to discuss various exceptions and special cases. This section replaces Assumption 6 ith its opposite:

$$\text{ASSUMPTION } 6': \qquad var(x^p) = 0$$

Assumption $6'$ is an important special case because it will hold in pure random assignment.

Assumption $6'$ implies that $\lambda(b_x, m_0)$, $\lambda^\infty(m_0)$, and $\beta_x^\infty(m_0)$ are undefined and so the results in Section 3 do not apply. However, the identified set is still well-defined, and can be estimated consistently by the estimators described in Section 3.3. The reason for this is that $var(x^p) = 0$ implies $cov(x, \mathbf{c}\beta_\mathbf{c})$ is also zero, and since λ is finite by Assumption 2 this implies that $cov(x, \varepsilon)$ is zero. Given that, Proposition 5 below shows that β_x is point-identified and can be estimated by OLS. Proposition 6 below shows that the RCR bounds will also consistently estimate β_x.

Proposition 5 (Size of the identified set) *The identified set $B_x(\Lambda; m_0)$ is nonempty and bounded for any Λ. In particular, $B_x(\Lambda; m_0) = \left\{ \frac{cov(x,y)}{var(x)} \right\} = \{\beta_x\}$.*

Proof: See Appendix A.5.

Proposition 6 (Consistency) *Suppose that Λ is bounded and includes zero. Then: $\hat{\beta}^L(\Lambda) \xrightarrow{p} \beta_x$ and $\hat{\beta}^H(\Lambda) \xrightarrow{p} \beta_x$.*

Proof: See Appendix A.6.

4.2 Heterogeneity in response

The model presented in Section 3 assumes a constant marginal effect of x on y. This section describes how the estimation method would apply to the case of heterogeneous response.

To do this, replace Assumption 1 with:

$$\text{ASSUMPTION } 1': \qquad y = y(x) = tx + u \qquad \text{where } E(t) = \beta_x$$

where t is the individual-specific marginal effect of x on y, and β_x is a parameter representing the average marginal effect in the population. If x is a binary treatment indicator, then Assumption $1'$ fits the standard treatment effects framework, with u the untreated outcome, $t + u$ the treated outcome, t the individual-specific treatment effect, and β_x the average treatment effect. For notational simplicity, normalize y, x, and $\mathbf{c}$ to mean zero so that $\mathbf{c}$ does not need to have an intercept.

The average marginal effect β_x is point-identified if the potential outcomes are mean-independent of x conditional on $\mathbf{c}$, i.e., if $E(t|x, \mathbf{c}) = E(t|\mathbf{c})$ and $E(u|x, \mathbf{c}) = E(u|\mathbf{c})$. If these conditional expectation functions happen to be linear in $\mathbf{c}$, then β_x is consistently estimated by the OLS regression of y on $(x, \mathbf{c}, x\mathbf{c})$. Note that conditional mean independence is the important assumption here, as one can always choose $\mathbf{c}$ to make the conditional expectations linear (e.g. by making $\mathbf{c}$ binary).

Next I derive a version of equation (18) that replaces the mean-independence and linear CEF assumptions with relative correlation restrictions. Without loss of generality, let $x\mathbf{c}\beta_{x\mathbf{c}} + \mathbf{c}\beta_{\mathbf{c}}$ be the best linear predictor of $y - x\beta_x$ given $(x\mathbf{c}, \mathbf{c})$. Let $\varepsilon \equiv y - x\beta_x - x\mathbf{c}\beta_{x\mathbf{c}} - \mathbf{c}\beta_{\mathbf{c}}$ be the corresponding residual. Then

$$y = x\beta_x + x\mathbf{c}\beta_{x\mathbf{c}} + \mathbf{c}\beta_{\mathbf{c}} + \varepsilon \qquad \text{where } E(x\mathbf{c}'\varepsilon) = E(\mathbf{c}'\varepsilon) = 0. \tag{18}$$

Consistent OLS estimation of equation (18) requires the additional assumption that $E(x\varepsilon) = 0$ or equivalently $\text{corr}(x, \varepsilon) = 0$. This condition would hold under the conditional mean-independence and linear CEF assumptions, but the goal here is to relax those assumptions. As in Section 3.1, this is done by replacing the absolute correlation restriction $\text{corr}(x, \varepsilon) = 0$ with a relative correlation restriction Λ such that:

$$\lambda = \frac{\text{corr}(x, \varepsilon)}{\text{corr}(x, x\mathbf{c}\beta_{x\mathbf{c}} + \mathbf{c}\beta_{\mathbf{c}})} \in \Lambda \tag{19}$$

In this version of the model, the relative correlation parameter λ can be interpreted as the correlation between the treatment and unobserved heterogeneity (in both untreated outcome and treatment response) relative to the correlation between the treatment and observed heterogeneity (in both untreated outcome and treated response).

Equations (18) and (19) are identical to equations (7) and (8) in Section 3.1, with $(x\mathbf{c}, \mathbf{c})$ as the control variables instead of just $\mathbf{c}$. Therefore this model can be fit into the framework of Section 3.1, and the results from Section 3 apply directly.

5 Applications

The two applications described in this section have been chosen to illustrate the two primary settings in which OLS regression is used to estimate causal effects: random-assignment experiments, and observational studies using control variables.

5.1 Application #1: Project STAR

Project STAR (Student/Teacher Achievement Ratio) is an influential class size experiment implemented in Tennessee in the late 1980's. Class size reductions are a common and expensive initiative for improving schools, but their effect on academic achievement is controversial (Hanushek, 1986). As is often the case with field experiments involving human subjects, Project STAR's implementation deviated slightly from the original random-assignment design. The application here shows that relative correlation restrictions are useful for analyzing such deviations.

5.1.1 Background

The analysis here is based on Krueger (1999). A total of 79 schools were nonrandomly selected for participation in Project STAR. Within each school, students entering kindergarten in 1985 were randomly assigned to the small class (S) group, the regular class (R) group, or the regular class with full-time teacher aide (RA) group. Each school had at least one class of each type. Students in group S were organized into classes with 13 to 17 students, while students in the R and RA groups were organized into classes with 22-25 students. Teachers were also randomly assigned. The experimental treatment continued through grade 3, and students were given achievement tests each year. The implementation deviated from the experimental design in several ways. For example, some students were moved between the small and regular class groups as a result of behavioral issues and/or parental request. Other students left their original schools, and Krueger notes that there is some evidence that students in the small class treatment were less likely to change schools.

Krueger's approach to the problem of imperfect randomization is similar to that described in Section 2. That is, he shows that observed pretreatment variables are similar (within-school) in the treated and control groups, and uses this observation to argue that it is reasonable to treat the data as arising from true random assignment:

> "None of the three background variables displays a statistically significant association with class-type assignment at the 10 percent level, which suggests that random assignment produced relatively similar groups in each class size, on average. As an overall test of random assignment, I regressed a dummy variable indicating assignment to a small class on the three background measures in rows 1-3 and school dummies. For each wave, the student characteristics had no more than a chance association with class-type assignment." (Krueger, 1999, page 504)

While Krueger presents these results as a "test" of the null hypothesis of random assignment, the deviations from the experimental design described above already imply that this null is false. Krueger says as much earlier in the paper: "As in any experiment, there were deviations from the ideal experimental design in the actual implementation of Project STAR." (Krueger, 1999, p. 499). Failure to reject this null only means that the sample size is not big enough to reveal what is already known to be true.

An alternative interpretation of this procedure is that it aims to show that deviations from random assignment produce small (if nonzero) differences in observable pre-treatment characteristics between treated and control groups, and therefore can plausibly be assumed to produce small differences in unobservable pre-treatment characteristics. This interpretation can be made more explicitly and quantitatively by using relative correlation restrictions.

5.1.2 Data and methodology

The data are from Finn, Boyd-Zaharias, Fish, and Gerber (2007). The outcome variable is the student's score on the Stanford achievement tests, translated into percentile units based on the distribution of test scores in the control group. The treatment variable is an indicator for whether the student was assigned to a small class (13-17 students) or a regular class with or without a teacher's aide (22-25 students). Control variables include school fixed effects as well as participation in free/subsidized lunch program, race, gender, age, teacher race, teacher experience, and whether the teacher has a graduate degree. Krueger's regressions include school-level fixed effects to account for the fact that class size was randomly assigned within schools, but assignment probabilities differed across schools. These fixed effects can be incorporated into the framework of this paper by applying the standard within transformation and defining y, x and $\mathbf{c}$ in terms of deviations from the corresponding school-level averages.

5.1.3 OLS and RCR results

The first two rows in Table 1 show OLS regression estimates for the effect of smaller classes, using a minor modification of the specification from Table 5 in Krueger (1999). The modification is that the regular/aide class indicator has been dropped so that there is only one treatment effect of interest. Both Krueger and the original Project STAR research team found that the aide treatment was nearly irrelevant to student outcomes. The next few rows in Table 1 report bounds on the treatment effect under a series of relative correlation restrictions. Asymptotic 95% confidence intervals are reported in parentheses for both the OLS point estimates and RCR bounds. All confidence intervals are robust to clustering by teacher, and the RCR confidence intervals are calculated based on the method described in Imbens and Manski (2004). Statistical significance for the RCR bounds is assessed by inverting the Imbens-Manski confidence intervals.

In addition to reporting bounds and confidence intervals for the class size effect, Table 1 reports three auxiliary statistics. The first is the estimated value of λ^∞, the relative correlation at which identification breaks down. The second is the estimated value of $\lambda(0)$, the relative correlation implied by a class size effect of exactly zero. The third is the lowest value of λ for which the RCR bounds include zero:

$$\inf\left\{\lambda : \hat{\beta}^L((-\infty,\lambda]) \leq 0 \leq \hat{\beta}^H((-\infty,\lambda])\right\}$$

Because the identified set is not necessarily convex, this value is not necessarily equal to $\hat{\lambda}(0)$. However, because the identified set is monotonic in Λ (in the sense that $\Lambda' \subset \Lambda \Rightarrow B_x(\Lambda') \subset B_x(\Lambda)$) the value can be found by simple iteration starting from $\min(\hat{\lambda}(0), \hat{\lambda}^\infty)$.

As in Krueger's original paper, the OLS results in Table 1 suggest that the small-class treatment increases test scores by five to seven percentile points. The RCR results suggest that these findings are quite robust, especially for kindergarten test scores. If the correlation between the small-class treatment and unobservables is no larger than the correlation between the treatment and observables ($0 \leq \lambda \leq 1$), the RCR bounds on class size effects in all grades are quite narrow and the lower bounds of the confidence intervals are all far from zero. If the correlation with unobservables is allowed to be three times as large as the correlation with observables, the lower bounds for kindergarten, grade 2, and grade 3 effects are all strictly positive but statistically insignificant. The bounds on the class size effect remain strictly positive for any relative correlation below about twelve for Kindergarten, slightly less than three for grade 1, and slightly more than three for grades 2 and 3. The finding that the Kindergarten effects are more robust than the effects for other grades is consistent with the consensus in the literature (Schanzenbach, 2006, p. 206) that the later grades deviate more from the random assignment design as a result of attrition and transfers.

5.2 Application #2: Economic effects of malaria

Bleakley (2010b) uses malaria eradication campaigns in the United States, Brazil, Colombia, and Mexico as natural experiments to measure the effect of childhood exposure to malaria on labor productivity. The economic effects of malaria are both of direct interest for those tropical countries in which malaria remains an issue, and of interest in relation to the more general literature on the potential role of disease as a barrier to economic development.

Grade:	Kindergarten	Grade 1	Grade 2	Grade 3
OLS point estimate ($\lambda = 0$)	5.20^{***}	6.72^{***}	4.97^{***}	5.30^{***}
(95% CI)	$(3.17, 7.24)$	$(4.66, 8.78)$	$(2.90, 7.04)$	$(3.24, 7.36)$
Bounds, $0 \leq \lambda \leq 0.1$	$[5.19, 5.20]^{***}$	$[6.50, 6.72]^{***}$	$[4.83, 4.97]^{***}$	$[5.19, 5.30]^{***}$
(95% CI)	$(3.17, 7.23)$	$(4.53, 8.69)$	$(2.85, 6.98)$	$(3.14, 7.31)$
Bounds, $0 \leq \lambda \leq 0.5$	$[5.17, 5.20]^{***}$	$[5.62, 6.72]^{***}$	$[4.26, 4.97]^{***}$	$[4.72, 5.30]^{***}$
(95% CI)	$(3.00, 7.22)$	$(3.72, 8.49)$	$(2.44, 6.80)$	$(2.52, 7.17)$
Bounds, $0 \leq \lambda \leq 1$	$[5.14, 5.20]^{***}$	$[4.50, 6.72]^{***}$	$[3.55, 4.97]^{***}$	$[4.08, 5.30]^{***}$
(95% CI)	$(2.49, 7.21)$	$(2.27, 8.46)$	$(1.58, 6.73)$	$(1.30, 7.10)$
Bounds, $0 \leq \lambda \leq 3$	$[4.99, 5.20]$	$[-0.15, 6.72]$	$[0.57, 4.97]$	$[0.44, 5.30]$
(95% CI)	$(-1.00, 7.20)$	$(-5.11, 8.45)$	$(-3.49, 6.71)$	$(-7.27, 7.06)$
$\hat{\lambda}^{\infty}$	12.31	13.85	14.88	5.79
$\hat{\lambda}(0)$	28.94	2.94	3.37	3.18
Minimum λ for which bounds include zero	12.31	2.94	3.37	3.18

Table 1: OLS estimates and RCR bounds for the effect of small classes on average percentile rank on Stanford Achievement Test. Intervals in square brackets are the bounds themselves, while the intervals in the round brackets are 95% cluster-robust asymptotic confidence intervals. *** = statistically significant at 1%, ** = significant at 5%, * = significant at 10%.

5.2.1 Data and model

Bleakley's research design exploits two major malaria eradication campaigns, one in the southern United States in the 1920's and a worldwide campaign in the 1950's. The data are observed at the state level (for the United States, Brazil, and Mexico) or the municipio-level (for Colombia) for cohorts who reached adulthood before the campaign started (the pre-eradication cohort) and for cohorts born shortly after the campaign (the post-eradication cohort). The econometric model is:

$$\underbrace{Y_{j,post} - Y_{j,pre}}_{y} = \underbrace{\beta M_{j,pre}}_{x\beta_x} + \underbrace{X_{j,pre}\Gamma + \alpha}_{\mathbf{c}\beta_{\mathbf{c}}} + \underbrace{\varepsilon_{j,post}}_{\varepsilon} \tag{20}$$

where $Y_{j,pre}$ and $Y_{j,post}$ are proxies for pre-eradication and post-eradication labor productivity, $M_{j,pre}$ is a measure of pre-eradication malaria incidence or prevalence, and $X_{j,pre}$ is a set of pre-eradication state characteristics used as control variables. The parameter of interest is β, which is interpreted as the (negative) effect of malaria ($M_{j,pre}$) on productivity in the pre-eradication cohort ($Y_{j,pre}$). This model can be interpreted as a standard two-period fixed effects research design that has been modified to allow state-specific trends as long as those trends are exogenous conditional on the predetermined characteristics in $X_{j,pre}$. This model can be incorporated into the RCR framework with the relative correlation parameter $\lambda = \frac{\text{corr}(M_{j,pre},\varepsilon_{j,post})}{\text{corr}(M_{j,pre},X_{j,pre}\Gamma)}$. The OLS regression can then be interpreted as imposing the assumption $\lambda = 0$.

Data and code to generate the OLS results are obtained from Bleakley (2010a). Results are reported only for the U.S. and Colombia, the two countries with the largest sample sizes and strongest OLS results in the original paper.

5.2.2 Results

Table 2 shows results for U.S. states, using malaria mortality as a measure of malaria prevalence and Hong's index of malaria ecology as a measure of malaria incidence. The labor productivity variables are the occupational income score and the Duncan socioeconomic index, both of which use the state occupational distribution to approximate log average earnings. Table 2 reports results for Bleakley's "Basic" and "Full controls" specifications. The Basic specification includes the log of state average unskilled wages in 1899 and a dummy variable for southern states, while the Full controls specification also includes various indicators of health, education, demographic, and labour market conditions. For each specification, Table 2 reports the OLS point estimate of the effect of malaria on productivity. Each point estimate corresponds to a value reported in Table 1 in Bleakley (2010b). To facilitate comparison with the RCR results, Table 2 reports 95% asymptotic confidence intervals for the OLS estimates rather than standard errors as in the original paper. RCR bounds and confidence intervals are reported for the same ranges of the relative correlation parameter as in Table 1.

The OLS results for U.S. states suggest a strong and statistically significant effect of malaria. The RCR results suggest that this finding is robust to a mild correlation between pre-eradication malaria incidence or prevalence and unobserved factors relative to the correlation between malaria and the control variables. For example, consider the Basic specification estimate of the effect of malaria mortality on the occupational index score. The results in Table 2 imply that if the correlation between initial malaria mortality and unobserved factors is between zero and ten percent of the correlation between initial malaria mortality and the control variables ($0 \leq \lambda \leq 0.1$), then the bounds on the effect are $[0.08, 0.11]$ in comparison to the point estimate of 0.11. These bounds are statistically significant at 5%, and the 95% confidence interval of $(0.01, 0.19)$ is only slightly wider than the OLS confidence interval of $(0.03, 0.19)$. While the OLS results are robust to a mild relative correlation, they are not robust to a moderate or large relative correlation. If the correlation between malaria and unobserved factors is equal to or greater than 32% of the correlation between malaria and the control variables, then the RCR bounds include no effect at all. The estimated value for λ^{∞} implies that no RCR bounds can be placed on the effect if the correlation between malaria and unobserved factors is as much as 93% as large as the correlation between malaria and the control variables. Similar results are found using the Duncan socioeconomic index as the dependent variable, and using Hong's malaria ecology index as the malaria variable: the findings of a substantial negative effect of malaria are robust to a mild relative correlation but not to a moderate or high relative correlation. The results using the Full Controls specification are less clear: because of limited degrees of freedom, confidence intervals are substantially wider and the OLS results are not generally robust to even a mild relative correlation.

Table 3 reports results for Colombia. The OLS results suggest a large and significant effect of malaria on income, while the OLS results for literacy and schooling are weaker. For the Basic specification the RCR analysis implies that the effect of malaria ecology on income is robust to mild relative correlation for both measures, and is robust to moderate relative correlation for the Poveda measure of malaria ecology. These findings are further supported by the Full Controls specification: the OLS point estimates remain statistically significant and similar in magnitude to the Basic specification, and the RCR results for the Poveda measure suggest that the effect is robust to mild relative correlation.

Overall, Bleakley's results are robust only to mild relative correlation. Estimated bounds on the effect of malaria generally include no effect when the correlation between malaria and unobservables is more than 20-30% of the correlation between malaria and observables. Bleakley's findings thus depend very critically on the claim that the control variables in the "full specification" include almost all important variables that affect productivity and are correlated with malaria incidence or prevalence.

6 Conclusion

The methodology developed in this paper provides a simple means of providing bounds on causal parameters under relative correlation restrictions. In the Project STAR application using data from a random-assignment experiment, the bounds on the class size effect are narrow and the lower bound is strictly positive even if class size is several times more strongly correlated with unobserved factors than with the observed control variables. In the application using observational data to measure the effect of malaria on productivity, the bounds on the effect are much wider, and the lower bound is negative as long as the upper bound on the correlation between malaria and unobserved factors is at least 30% of the correlation between malaria and the observed control variables.

While it is not surprising that data from random assignment is more reliable than observational data, note that this finding of greater robustness comes entirely from the data itself and not from any information on the research design. While the method described in this paper is no substitute for careful evaluation of research design, it provides a systematic and straightforward means for that evaluation to be informed by the data itself.

The methodology can be advanced in future research along two main fronts. First, the model is intentionally kept simple here but might be usefully extended to accomodate common features like fixed effects or simple forms of nonlinearity. Second, the estimators used here are based on ratios and/or inverses, and the literature on weak identification emphasizes that standard delta-method asymptotics can provide a poor approximation in a finite sample when a relevant denominator is nearly zero. Alternative inference procedures that are robust to this possibility may be useful in this setting.

A Proofs of propositions

A.1 Proposition 1

Proof: To establish result 1, note that:

$$\begin{aligned}
\lambda(b_x;m) &= \frac{\mathrm{corr}_m\left(x, y-b_x x-(y-b_x x)^p\right)}{\mathrm{corr}_m\left(x,(y-b_x x)^p\right)} \\
&= \frac{\frac{cov_m(x,y-b_x x-(y-b_x x)^p)}{\sqrt{var_m(x)var_m(y-b_x x-(y-b_x x)^p)}}}{\frac{cov_m(x,(y-b_x x)^p)}{\sqrt{var_m(x)var_m((y-b_x x)^p)}}} \\
&= \frac{\left(\frac{cov_m(x,y)-b_x var_m(x)}{cov_m(x,y^p)-b_x cov_m(x,x^p)}-1\right)}{\sqrt{\frac{var_m(y-b_x x)}{var_m((y-b_x x)^p)}-1}}
\end{aligned}$$

We can apply several properties of the best linear predictor, specifically that $cov(x,y^p) = cov(x^p,y^p)$, $cov(x,x^p) = var(x^p)$ and $var(y-y^p) = var(y) - var(y^p)$, to further derive:

$$\begin{aligned}
\lambda(b_x;m) &= \frac{\left(\frac{cov_m(x,y)-b_x var_m(x)}{cov_m(x^p,y^p)-b_x var_m(x^p)}-1\right)}{\sqrt{\frac{var_m(y)-2b_x cov_m(x,y)+b_x^2 var_m(x)}{var_m(y^p)-2b_x cov_m(x^p,y^p)+b_x^2 var_m(x^p)}-1}} \\
&= \frac{\left(\frac{p_1}{p_2}-1\right)}{\sqrt{\frac{p_3}{p_4}-1}} \qquad (21)
\end{aligned}$$

where p_1, p_2, p_3, and p_4 are all polynomials (and thus differentiable) in b_x. They are also differentiable in m. Application of the quotient and product rules implies that $\lambda(b_x;m)$ is differentiable provided that (a) $p_2 \neq 0$, (b) $p_4 \neq 0$, and (c) $\frac{p_3}{p_4} > 1$. Condition (a) fails if and only if:

$$p_2 = cov_m(x^p,y^p) - b_x var_m(x^p) = 0$$

Since $var_m(x^p) > 0$ by equation (11), we can solve to get

$$b_x = \frac{cov_m(x^p,y^p)}{var_m(x^p)} = \beta_x^\infty(m)$$

Condition (b) fails if and only if:

$$p_4 = var_m(y^p - b_x x^p) = 0$$

which implies that $y^p - b_x x^p$ is constant. Since the covariance of any random variable with a constant is zero, this in turn implies that $cov(x^p, y^p - b_x x^p) = cov(x^p, y^p) - b_x var(x^p) = 0$. Again we can solve for b_x to get:

$$b_x = \frac{cov_m(x^p, y^p)}{var_m(x^p)} = \beta_x^\infty(m)$$

Condition (c) fails if and only if $p_3 \leq p_4$, or equivalently:

$$var_m(y - b_x x) \leq var_m(y^p - b_x x^p)$$

Note that $y^p - b_x x^p$ is the best linear predictor of $y - b_x x$, so:

$$var_m\,(y - b_x x) = var_m\,(y^p - b_x x^p) + var_m\,(y - b_x x - (y^p - b_x x^p))$$

This implies that $var_m\,(y - b_x x - (y^p - b_x x^p)) = 0$, which also implies that:

$$y - b_x x - (y^p - b_x x^p) = 0$$

Rearranging, we get:

$$y = y^p - b_x x^p + b_x x$$

which implies that y is an exact linear function of $(x, \mathbf{c})$ and equation (9) is violated. Therefore, condition (c) must hold. Since conditions (a), (b), and (c) hold for all $b_x \neq \beta_x^\infty(m)$, $\lambda(b_x; m)$ is differentiable at all $b_x \neq \beta_x^\infty(m)$.

To establish result 2, note that $var_m(x)$ is strictly positive by (9) and $var_m(x^p)$ is strictly positive by (11). Therefore:

$$\lim_{b_x \to \infty} (cov_m(x, y) - b_x var_m(x)) = -\infty$$

$$\lim_{b_x \to \infty} (cov_m(x^p, y^p) - b_x var_m(x^p)) = -\infty$$

So by L'Hospital's rule:

$$\lim_{b_x \to \infty} \frac{cov_m(x, y) - b_x var_m(x)}{cov_m(x^p, y^p) - b_x var_m(x^p)} = \frac{var_m(x)}{var_m(x^p)}$$

By the same reasoning:

$$\lim_{b_x \to \infty} (var_m(x) - 2b_x cov_m(x, y) + b_x^2 var_m(x)) = \infty$$

$$\lim_{b_x \to \infty} (var_m(y^p) - 2b_x cov_m(x^p, y^p) + b_x^2 var_m(x^p)) = \infty$$

$$\lim_{b_x \to \infty} (-2cov_m(x, y) + 2b_x var_m(x)) = \infty$$

$$\lim_{b_x \to \infty} (-2cov_m(x^p, y^p) + 2b_x var_m(x^p)) = \infty$$

So by two applications of L'Hospital's rule:

$$\lim_{b_x \to \infty} \frac{var_m(y) - 2b_x cov_m(x,y) + b_x^2 var_m(x)}{var_m(y^p) - 2b_x cov_m(x^p,y^p) + b_x^2 var_m(x^p)} = \frac{var_m(x)}{var_m(x^p)}$$

Result 2 can then be derived by substitution, and the argument repeated for $\lim_{b_x \to -\infty}$.

To prove result 3 I first show how the behavior of $\lambda(b_x;m)$ near $\beta_x^\infty(m)$ depends on some special cases:

Case A: Suppose that m implies an exact linear relationship between y^p and x^p, i.e.

$$E_m\left((y^p - a_m - b_m x^p)^2\right) = 0 \tag{22}$$

for some a_m and b_m. Then equation (14) is satisfied for all λ when $b_x = \beta_x^\infty(m) = b_m$.

Proof: To show that $\beta_x^\infty(m) = b_m$:

$$\begin{aligned}
\beta_x^\infty(m) &= \frac{cov_m(x^p,y^p)}{var_m(x^p)} \\
&= \frac{cov_m(x^p,a_m + b_m x^p) + cov_m(x^p, y^p - a_m - b_m x^p)}{var_m(x^p)} \\
&= \frac{b_m\, var_m(x^p) + 0}{var_m(x^p)} \\
&= b_m
\end{aligned}$$

To show that equation (14) is satisfied at $\beta_x^\infty(m)$ for all λ, note that $\mathbf{c}\beta_{\mathbf{c}}(b_x;m) = y^p - b_x x^p$. This implies that:

$$\begin{aligned}
var_m(\mathbf{c}\beta_{\mathbf{c}}(\beta_x^\infty(m);m)) &= var_m\left(y^p - \beta_x^\infty(m) x^p\right) \\
&= var_m\left(y^p - b_m x^p\right) \\
&= var_m\left(y^p - b_m x^p\right) - 2cov_m\left(y^p - b_m x^p, a_m\right) + var_m(a_m) \\
&= var_m\left(y^p - a_m - b_m x^p\right) \\
&= 0
\end{aligned}$$

and by the same argument $cov_m(x, \mathbf{c}\beta_{\mathbf{c}}(\beta_x^\infty(m);m)) = 0$. Equation (14) thus reduces to $0 = \lambda 0$, a condition that is satisfied by any λ.

Case B: Suppose that m implies:

$$\frac{cov_m(y,x)}{var_m(x)} = \frac{cov_m(y^p,x^p)}{var_m(x^p)} \tag{23}$$

Then equation (14) is satisfied for all λ when $b_x = \beta_x^\infty(m)$.

Proof: First, note that in this case:

$$\begin{aligned}
& cov_m(x, y - \beta_x^\infty(m)x - \mathbf{c}\beta_\mathbf{c}(\beta_x^\infty(m); m)) \\
&\qquad = cov_m(x, y - \beta_x^\infty(m)x - y^p + \beta_x^\infty(m)x^p) \\
&\qquad = cov_m(x,y) - \beta_x^\infty(m) var_m(x) - cov_m(y^p, x^p) + \beta_x^\infty(m) var_m(x^p) \\
&\qquad = cov_m(x,y) - \frac{cov_m(x,y)}{var_m(x)} var_m(x) - cov_m(y^p, x^p) + \frac{cov(x^p, y^p)}{var(x^p)} var(x^p) \\
&\qquad = 0
\end{aligned}$$

and:

$$\begin{aligned}
cov_m(x, \mathbf{c}\beta_\mathbf{c}(\beta_x^\infty(m); m)) &= cov_m(x, y^p - \beta_x^\infty(m)x^p) \\
&= cov_m(x^p, y^p) - \beta_x^\infty(m) var_m(x^p) \\
&= cov_m(x^p, y^p) - \frac{cov_m(x^p, y^p)}{var_m(x^p)} var_m(x^p) \\
&= 0
\end{aligned}$$

Equation (14) thus reduces to $0 = \lambda 0$, which is satisfied for all λ.

Case C: Suppose that neither (22) nor (23) hold. Then for any $\lambda \in (-\infty, \lambda^\infty(m)) \cup (\lambda^\infty(m), \infty)$ there is a b_x such that $\lambda(b_x; m) = \lambda$, i.e., that solves equation (14).

Proof: First, note that since $cov_m(x^p, y^p) - \beta_x^\infty(m) var_m(x^p) = 0$, the existence of a solution to equation (14) when $b_x = \beta_x^\infty(m)$ requires that either $var_m(y^p - \beta_x^\infty(m)x^p) = 0$, implying (22) holds, or $cov_m(x,y) - \beta_x^\infty(m) var_m(x) = 0$, implying (23) holds. Since neither holds, there is no λ that satisfies equation (14) for $b_x = \beta_x^\infty(m)$.

Next I characterize the behavior of $\lambda(b_x; m)$ near $\beta_x^\infty(m)$. Since $var_m(x^p) > 0$, p_2 is positive for $b_x < \beta_x^\infty(m)$, negative for $b_x > \beta_x^\infty(m)$, and zero when $b_x = \beta_x^\infty(m)$. Also note that $cov_m(x,y) - \beta_x^\infty(m) var_m(x) = cov_m(x,y) - \frac{cov_m(x^p, y^p)}{var_m(x^p)} var_m(x)$, so p_1 is strictly positive for all $b_x \approx \beta_x^\infty(m)$ if $\frac{cov_m(x,y)}{var_m(x)} > \frac{cov_m(x^p, y^p)}{var_m(x^p)}$, and strictly negative for all $b_x \approx \beta_x^\infty(m)$ if $\frac{cov_m(x,y)}{var_m(x)} < \frac{cov_m(x^p, y^p)}{var_m(x^p)}$. This implies that:

$$\lim_{b_x \uparrow \beta_x^\infty(m)} \lambda(b_x; m) = \begin{cases} \infty & \text{if } \frac{cov_m(y,x)}{var_m(x)} > \frac{cov_m(y^p, x^p)}{var_m(x^p)} \\ -\infty & \text{if } \frac{cov_m(y,x)}{var_m(x)} < \frac{cov_m(y^p, x^p)}{var_m(x^p)} \end{cases}$$

and

$$\lim_{b_x \downarrow \beta_x^\infty(m)} \lambda(b_x; m) = \begin{cases} -\infty & \text{if } \frac{cov_m(y,x)}{var_m(x)} > \frac{cov_m(y^p, x^p)}{var_m(x^p)} \\ \infty & \text{if } \frac{cov_m(y,x)}{var_m(x)} < \frac{cov_m(y^p, x^p)}{var_m(x^p)} \end{cases}$$

> I have thus established that $\lim_{b_x \to -\infty} \lambda(b_x;m) = \lambda^\infty(m)$, that $\lim_{b_x \uparrow \beta_x^\infty} \lambda(b_x;m)$ is either $-\infty$ or ∞, and that $\lambda(b_x;m)$ is continuous on $(-\infty, \beta_x^\infty(m))$. Suppose for the moment that $\lim_{b_x \uparrow \beta_x^\infty(m)} \lambda(b_x;m) = -\infty$. By the intermediate value theorem, for any $\lambda \in (-\infty, \lambda^\infty(m))$, there exists some $b_x \in (-\infty, \beta_x^\infty(m))$ such that $\lambda(b_x;m) = \lambda$. This is a sufficient condition for b_x to solve equation (14). Since $\lim_{b_x \uparrow \beta_x^\infty(m)} = -\infty$, then $\lim_{b_x \downarrow \beta_x^\infty(m)} \lambda(b_x;m) = \infty$. Again, since $\lambda(b_x;m)$ is continuous on $(\beta_x^\infty(m), \infty)$, the intermediate value theorem implies that for any $\lambda \in (\lambda^\infty(m), \infty)$ there exists some $b_x \in (\beta_x^\infty(m), \infty)$ such that $\lambda(b_x;m) = \lambda$. Therefore, for any $\lambda \in (-\infty, \lambda^\infty(m)) \cup (\lambda^\infty(m), \infty)$ there is a b_x such that $\lambda(b_x;m) = \lambda$, i.e., that solves equation (14). The same argument can be duplicated for the case $\lim_{b_x \uparrow \beta_x^\infty(m)} \lambda(b_x) = \infty$. Note that there may or may not be a b_x such that $\lambda(b_x;m) = \lambda^\infty(m)$.

To prove result 4, pick any b_x and consider two cases. First, suppose that $b_x = \beta_x^\infty(m)$. Then $b_x \notin \tilde{B}(\Lambda;m)$ since $\lambda(b_x;m)$ does not exist. Next, suppose that $b_x \neq \beta_x^\infty(m)$. Then $\lambda(b_x;m)$ exists (by result 1 of this proposition) and provides the unique λ that solves equation (14) for that λ. Therefore,

$$b_x \in \tilde{B}(\Lambda;m) \text{ if and only if } b_x \in B_x(\Lambda;m) \text{ and } b_x \neq \beta_x^\infty(m)$$

which is another way of stating the result. □

A.2 Proposition 2

Proof: Since Λ is nonempty, $\lambda^\infty(m_0) \notin \Lambda$ implies that Λ must contain some $\lambda \neq \lambda^\infty(m_0)$. Result 3 of Proposition 1 says that there exists some b_x such that (λ, b_x) satisfy equation (14). Therefore, the identified set is nonempty.

Since Λ is closed, $\lambda^\infty(m_0) \notin \Lambda$ implies that there is some $\varepsilon > 0$ such that $(\lambda^\infty(m_0) - \varepsilon, \lambda^\infty(m_0) + \varepsilon)$ is disjoint from Λ. Result 2 of Proposition 1 says that $\lim_{b_x \to \infty} \lambda(b_x;m_0) = \lim_{b_x \to -\infty} \lambda(b_x;m_0) = \lambda^\infty(m_0)$. This means that given such an ε, there is some finite B_ε such that $B_\varepsilon > \beta_x^\infty(m_0)$ and:

$$\begin{aligned}
|b_x| > B_\varepsilon &\Rightarrow \lambda(b_x;m_0) \in (\lambda^\infty(m_0) - \varepsilon, \lambda^\infty(m_0) + \varepsilon) \qquad \text{(by result 2 of Proposition 1)} \\
&\Rightarrow \lambda(b_x;m_0) \notin \Lambda \qquad \text{(since } (\lambda^\infty(m_0) - \varepsilon, \lambda^\infty(m_0) + \varepsilon) \text{ is disjoint from } \Lambda\text{)} \\
&\Rightarrow b_x \notin \tilde{B}(\Lambda, m_0) \qquad \text{(by definition of } \tilde{B}\text{)} \\
&\Rightarrow b_x \notin \tilde{B}(\Lambda, m_0) \cup \{\beta_x^\infty(m_0)\} \qquad \text{(since } B_\varepsilon > \beta_x^\infty(m_0)\text{)} \\
&\Rightarrow b_x \notin B_x(\Lambda, m_0) \qquad \text{(by result 4 of Proposition 1)}
\end{aligned}$$

Therefore, the identified set is bounded. □

A.3 Proposition 3

Proof: Both $\beta_x^\infty(m)$ and $\lambda^\infty(m)$ are continuous in m by the quotient rule, given that $var_m(x^p) > 0$. Result 1 of Proposition 1 says that $\lambda(b_x;m)$ is continuous in m for all $b_x \neq \beta_x^\infty(m)$). So the first set of results follows from the straightforward application of Slutsky's theorem.

For the second result, note that the implicit function theorem implies that $\beta_x^L(\Lambda;m)$ is continuously differentiable in m if $\frac{d\lambda(b_x;m)}{db_x}|_{b_x=\beta_x^L(\Lambda;m)} \neq 0$. In that case, consistency of $\hat{b_{xL}}(\Lambda)$ follows from Slutsky's theorem. The same argument applies to $\hat{b_{xH}}(\Lambda)$.

For the third result, note that if $B_x(\Lambda;m_0) = \mathbb{R}$, then result 2 of Proposition 1 implies $\lambda^\infty(m_0)$ is in the interior of Λ. Therefore, there exists an $\varepsilon > 0$ and $B_1 < B$ such that $[\lambda(B_1;m_0) - \varepsilon, \lambda(B_1;m_0) + \varepsilon] \subset \Lambda$. Since $\hat{\lambda}(B_1) \xrightarrow{p} \lambda(B_1;m_0)$:

$$\lim_{n\to\infty} \Pr(\hat{b_{xL}} < B) \geq \lim_{n\to\infty} \Pr(\hat{\lambda}(B_1) \in \Lambda) = 1$$

The same argument applies to $\hat{b_{xH}}(\Lambda)$, with a change of sign. □

A.4 Proposition 4

Proof: Both $\beta_x^L(\Lambda;m)$ and $\beta_x^H(\Lambda;m)$ are differentiable in m under these conditions, so the result follows from direct application of the delta method, where:

$$A = \begin{bmatrix} \nabla_m \beta_x^L(\Lambda;m)|_{m=m_0} \\ \nabla_m \beta_x^H(\Lambda;m)|_{m=m_0} \end{bmatrix} \tag{24}$$

The expression for A given in the proposition comes from applying the implicit function theorem:

$$\begin{aligned} \nabla_m \beta_x^L(\Lambda;m) &= -\left.\frac{\nabla_m \lambda(b_x;m)}{\partial\lambda(b_x;m)/\partial b_x}\right|_{b_x=\beta_x^L(\Lambda;m)} \\ \nabla_m \beta_x^H(\Lambda;m) &= -\left.\frac{\nabla_m \lambda(b_x;m)}{\partial\lambda(b_x;m)/\partial b_x}\right|_{b_x=\beta_x^H(\Lambda;m)} \end{aligned} \tag{25}$$

and substituting. While mathematically unnecessary, this substitution is important computationally. Derivatives of $\lambda(b_x;m)$ – a closed form function with closed form derivatives – can be calculated much more accurately than derivatives of $\beta_x^L(\Lambda;m)$ – an implicit function that must be approximated by iterative methods. □

A.5 Proposition 5

Proof: If $var(x^p) = 0$, then $cov(x, y^p - b_x x^p) = 0$ for all b_x. This implies that (14) holds if and only if $cov(x, y - b_x x) = 0$, i.e., if $b_x = cov(x,y)/var(x)$. □

A.6 Proposition 6

Proof: First, rewrite:

$$\begin{aligned}\lambda(b_x;m) &= \frac{\text{corr}_m(x, y - b_x x - y^p + b_x x^p)}{\text{corr}_m(x, y^p - b_x x^p)} \\ &= \frac{cov_m(x, y - b_x x - y^p + b_x x^p)}{\text{corr}_m(x, y^p - b_x x^p)\sqrt{var_m(x) var_m(y - b_x x - y^p + b_x x^p)}} \\ &= \frac{q_1(b_x;m)}{q_2(b_x;m)}\end{aligned}$$

The numerator of $\lambda(b_x;\hat{m}_n)$ is:

$$q_1(b_x;\hat{m}_n) \xrightarrow{p} cov(x,y) - b_x var(x)$$

while the denominator is

$$q_2(b_x;\hat{m}_n) \xrightarrow{p} 0$$

In a given finite sample, $q_2(b_x;\hat{m}_n)$ will be nonzero with probability one if x or any of $\mathbf{c}$ is continuously distributed, and probability approaching one as $n \to \infty$ (WPA1) otherwise. So $\lambda(b_x;\hat{m}_n)$ will exist even though $\lambda(b_x;m_0)$ does not. Let $\beta_x^{OLS}(m)$ be the value of b_x that implies $q_1(b_x;m) = 0$, or equivalently:

$$\beta_x^{OLS}(m) = \frac{cov_m(x - x^p, y - y^p)}{var_m(z - x^p)}$$

Note that $\beta_x^{OLS}(\hat{m}_n)$ is just the coefficient on x from the OLS regression of y on x and $\mathbf{c}$, and that:

$$\beta_x^{OLS}(\hat{m}_n) \xrightarrow{p} \beta_x^{OLS}(m_0) = \frac{cov(x - x^p, y - y^p)}{var(x - x^p)} = \frac{cov(x,y)}{var(x)} = \beta_x \qquad (26)$$

Since $q_1(\beta_x^{OLS}(\hat{m}_n)) = 0$ by construction and $q_2(\beta_x^{OLS}(\hat{m}_n)) \neq 0$ WPA1:

$$\lambda(\beta_x^{OLS}(\hat{m}_n);\hat{m}_n) = 0 \in \Lambda \qquad \text{WPA1}$$

Therefore:

$$\hat{\beta}^L(\Lambda) \leq \beta_x^{OLS}(\hat{m}_n) \leq \hat{\beta}^H(\Lambda) \qquad \text{WPA1} \qquad (27)$$

Pick any $\varepsilon > 0$. The event $(|\beta_x^{OLS}(\hat{m}_n) - \beta_x| < \varepsilon)$ clearly implies $(\beta_x^{OLS}(\hat{m}_n) > \beta_x - \varepsilon)$, which itself implies $(\hat{\beta}^H(\Lambda) > \beta_x - \varepsilon)$ by equation (27). Therefore:

$$\Pr(|\beta_x^{OLS}(\hat{m}_n) - \beta_x| < \varepsilon) \leq \Pr(\hat{\beta}^H(\Lambda) > \beta_x - \varepsilon) \leq 1$$

By (26), $\Pr(|\beta_x^{OLS}(\hat{m}_n) - \beta_x| < \varepsilon) \rightarrow 1$, so by the sandwich theorem:

$$\Pr(\hat{\beta}^H(\Lambda) > \beta_x - \varepsilon) \rightarrow 1 \tag{28}$$

Let λ^{max} satisfy $|\lambda| \leq \lambda^{max}$ for all $\lambda \in \Lambda$. Then $\lambda \in \Lambda$ implies $|\lambda| \leq \lambda^{max}$. Therefore:

$$\begin{aligned} 0 &\leq \Pr(\hat{\beta}^H(\Lambda) \geq \beta_x + \varepsilon) \\ &= \Pr(\lambda(b_x; \hat{m}_n) \in \Lambda \text{ for some } b_x > \beta_x + \varepsilon) \\ &\leq \Pr(|\lambda(b_x; \hat{m}_n)| \leq \lambda^{max} \text{ for some } b_x \geq \beta_x + \varepsilon) \end{aligned} \tag{29}$$

Now, for any $\delta \neq 0$

$$\begin{aligned} q_1(\beta_x + \delta; \hat{m}_n) &\xrightarrow{p} cov(x,y) - (\beta_x + \delta)var(x) = -\delta var(x) \neq 0 \\ q_2(\beta_x + \delta; \hat{m}_n) &\xrightarrow{p} 0 \end{aligned}$$

Therefore:

$$\Pr(|\lambda(b_x; \hat{m}_n)| \leq \lambda^{max} \text{for some } b_x \geq \beta_x + \varepsilon) \rightarrow 0 \tag{30}$$

By the sandwich theorem (29) and (30) imply $\Pr(\hat{\beta}^H(\Lambda) \geq \beta_x + \varepsilon) \rightarrow 0$, or equivalently that:

$$\Pr(\hat{\beta}^H(\Lambda) < \beta_x + \varepsilon) \rightarrow 1 \tag{31}$$

Taking (28) and (31) together produces:

$$\Pr(|\hat{\beta}^H(\Lambda) - \beta_x| < \varepsilon) \rightarrow 1 \tag{32}$$

which is the result stated in the proposition. The same argument applies to β_x^L. □

References

Altonji, J. G., T. E. Elder, and C. R. Taber (2005a): "An evaluation of instrumental variable strategies for estimating the effects of Catholic schooling," *Journal of Human Resources*, 40, 791–821.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005b): "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 113, 151–184.

Bleakley, H. (2010a): "Data and code for 'malaria eradication in the Americas: A retrospective analysis of childhood exposure'," `doi://10.1257/app.2.2.1`, accessed 3/7/2012.

Bleakley, H. (2010b): "Malaria eradication in the Americas: A retrospective analysis of childhood exposure," *American Economic Journal: Applied Economics*, 2, 1–45.

Conley, T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly exogenous," *Review of Economics and Statistics*, 94, 260–272.

Finn, J. D., J. Boyd-Zaharias, R. M. Fish, and S. B. Gerber (2007): "Project star and beyond: Database user's guide," Data set and documentation, HEROS, Inc., retrieved from `http://www.heros-inc.org/data.htm`, 7/15/2007.

Hanushek, E. A. (1986): "The economics of schooling: Production and efficency in public schools," *Journal of Economic Literature*, 24, 1141–1177.

Imbens, G. W. (2003): "Sensitivity to exogeneity assumptions in program evaluation," *American Economic Review*, 93, 126–132.

Imbens, G. W. and C. F. Manski (2004): "Confidence intervals for partially identified parameters," *Econometrica*, 72, 1845–1857.

Klepper, S. and E. E. Leamer (1984): "Consistent sets of estimates for regressions with errors in all variables," *Econometrica*, 52, 163–184.

Kraay, A. (2012): "Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach," *Journal of Applied Econometrics*, 27, 108–128.

Krauth, B. V. (2007): "Peer effects and selection effects on youth smoking in california," *Journal of Business and Economic Statistics*, 25, 288–298.

Kreider, B. (2010): "Regression coefficient identification decay in the presence of infrequent classification errors," *Review of Economics and Statistics*, 92, 1017–1023.

Kreider, B. and S. C. Hill (2009): “Partially identifying treatment effects with an application to covering the uninsured,” *Journal of Human Resources*, 44, 409–449.

Krueger, A. B. (1999): “Experimental estimates of education production functions,” *Quarterly Journal of Economics*, 114, 497–532.

Leamer, E. E. (1978): *Specification Searches: Ad Hoc Inference with Non Experimental Data*, John Wiley and Sons.

Lewbel, A. (2012): “Using heteroskedasticity to identify and estimate mismeasured and endogenous regressor models,” *Journal of Business and Economic Statistics*, 30, 67–80.

Manski, C. F. (1994): *Identification Problems in the Social Sciences*, Harvard University Press.

Manski, C. F. (2003): *Partial Identification of Probability Distributions*, Springer-Verlag.

Nevo, A. and A. M. Rosen (2012): “Identification with imperfect instruments,” *Review of Economics and Statistics*, 94, 659–671.

Oster, E. (2014): “Unobservable selection and coefficient stability: Theory and evidence,” Working paper, University of Chicago.

Rosenbaum, P. R. (2002): *Observational Studies, 2nd edition*, Springer.

Schanzenbach, D. W. (2006): “What have researchers learned from Project STAR?” *Brookings Papers on Education Policy*, 9, 205–228.

Stoye, J. (2009): “More on confidence intervals for partially identified parameters,” *Econometrica*, 77, 1299–1315.

Malaria measure:	Malaria mortality		Malaria ecology (Hong)	
	Occupational	Duncan	Occupational	Duncan
Dependent variable:	Income	Socioeconomic	Income	Socioeconomic
	Score	Index	Score	Index
Basic specification:				
OLS point estimate	0.11***	0.13**	0.24***	0.22***
(95% CI)	(.03,.19)	(.00,.27)	(.17,.30)	(.11,.32)
Bounds, $0 \leq \lambda \leq 0.1$	[.08,.11]**	[.09,.13]	[.24,.26]***	[.22,.28]***
(95% CI)	(.01,.19)	(-.06,.28)	(.18,.35)	(.12,.37)
Bounds, $0 \leq \lambda \leq 0.5$	[-.09,.11]	[-.17,.41]	[.02,.45]	[-.19,.63]
(95% CI)	(-.21,.18)	(-.36,.92)	(-.16,.60)	(-.66,.98)
Bounds, $0 \leq \lambda \leq 1$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$
Bounds, $0 \leq \lambda \leq 3$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$
$\hat{\lambda}^{\infty}$	0.93	0.93	0.76	0.76
$\hat{\lambda}(0)$	0.32	0.26	0.53	0.32
Minimum λ for which bounds include zero	0.32	0.26	0.53	0.32
Full controls:				
OLS point estimate	0.11**	0.17*	0.21***	0.26***
(95% CI)	(.01,.21)	(-.02,.36)	(.11,.32)	(.07,.46)
Bounds, $0 \leq \lambda \leq 0.1$	[.08,.11]*	[.17,.17]	[.17,.21]*	[.26,.39]***
(95% CI)	(-.02,.19)	(-.09,.34)	(-.02, .29)	(.11,.68)
Bounds, $0 \leq \lambda \leq 0.5$	[-.43,.11]	[-.90,1.23]	$(-\infty,\infty)$	$(-\infty,\infty)$
(95% CI)	(-2.02,.19)	(-4.29,4.67)	$(-\infty,\infty)$	$(-\infty,\infty)$
Bounds, $0 \leq \lambda \leq 1$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$
Bounds, $0 \leq \lambda \leq 3$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$
$\hat{\lambda}^{\infty}$	0.53	0.53	0.35	0.35
$\hat{\lambda}(0)$	0.27	0.32	0.22	0.20
Minimum λ for which bounds include zero	0.27	0.32	0.22	0.20

Table 2: OLS and RCR estimates of the effect of malaria mortality and ecology on indicators of labor productivity in U.S. states. 95% confidence intervals in parentheses, *** = statistically significant at 1%, ** = significant at 5%, * = significant at 10%. OLS estimates are reproduced from Bleakley (2010b), Table 1.

	Malaria ecology (Poveda)			Malaria ecology (Mellinge		
Dependent variable:		Years of	Income		Years of	Inc
	Literacy	schooling	index	Literacy	schooling	in
Basic specification:						
OLS point estimate	0.04***	0.17*	0.06***	0.07***	0.06	0.0
(95% CI)	(.01,.06)	(.00,.34)	(.04,.09)	(.04,.10)	(-.15,.28)	(.02
Bounds, $0 \leq \lambda \leq 0.1$	[.02,.04]	[.10,.17]	[.06,.07]***	[.06,.07]***	[-.02,.06]	[.04,
(95% CI)	(-.01,.06)	(-.09,.32)	(.04,.10)	(.03,.10)	(-.23,.26)	(.01
Bounds, $0 \leq \lambda \leq 0.5$	[-.08,.04]	[-.56,.17]	[.06,.16]***	[-.01,.07]	[-.48,.06]	[.01
(95% CI)	(-.12,.06)	(-.94,.32)	(.04,.21)	(-.05,.10)	(-.81,.25)	(-.0
Bounds, $0 \leq \lambda \leq 1$	[-.63,.68]	[-4.71,4.84]	[-.54,.69]	[-.16,.07]	[-1.54,.06]	[-.1
(95% CI)	(-1.35,1.41)	(-10.08,10.32)	(-1.24,1.37)	(-.25,.10)	(-2.29,.25)	(-.2
Bounds, $0 \leq \lambda \leq 3$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-$
$\hat{\lambda}^{\infty}$	1.06	1.06	1.06	1.49	1.49	1
$\hat{\lambda}(0)$	0.19	0.20	0.93	0.43	0.08	0
Minimum λ for which bounds include zero	0.19	0.20	0.73	0.43	0.08	0
Full controls:						
OLS point estimate	0.01	0.16*	0.06***	0.05***	0.08	0.0
(95% CI)	(-.02,.03)	(-.02,.35)	(.04,.09)	(.02,.07)	(-.15,.31)	(.00
Bounds, $0 \leq \lambda \leq 0.1$	[-.02,.01]	[.07,.16]	[.06,.08]***	[.03,.05]*	[-.03,.08]	[.02
(95% CI)	(-.05,.03)	(-.17,.33)	(.04,.11)	(.00,.07)	(-.28,.29)	(-.0
Bounds, $0 \leq \lambda \leq 0.5$	[-.22,.01]	[-1.64,1.65]	[-.16,.31]	[-.11,.05]	[-1.02,.08]	[-.1
(95% CI)	(-.29,.03)	(-2.21,2.63)	(-.25,.38)	(-.16,.07)	(-1.57,.28)	(-.2
Bounds, $0 \leq \lambda \leq 1$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-$
Bounds, $0 \leq \lambda \leq 3$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-\infty,\infty)$	$(-$
$\hat{\lambda}^{\infty}$	0.73	0.73	0.73	0.95	0.95	0
$\hat{\lambda}(0)$	0.02	0.14	0.42	0.20	0.08	0
Minimum λ for which bounds include zero	0.02	0.14	0.39	0.20	0.08	0

Table 3: OLS and RCR estimates of the effect of malaria ecology on indicators of labor productivity in Colombian municipios. 95% confidence intervals in parentheses, *** = statistically significant at 1%, ** = significant at 5%, * = significant at 10%. OLS estimates are reproduced from Bleakley (2010), Table 3.