10ISSN 1183-1057

SIMON FRASER UNIVERSITY

Department of Economics

Working Papers

11-02

“Bounding a linear casual effect using relative correlation restrictions”

Brian Krauth

August, 2011

SFU Economics

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

Brian Krauth

Department of Economics

Simon Fraser University

August 9, 2011

Abstract

This paper describes and implements a simple approach to the most common problem in applied microeconometrics: estimating a linear causal effect when the explanatory variable of interest might be correlated with relevant unobserved variables. The main idea is to place 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. These relative correlation restrictions allow a researcher to construct informative bounds on parameter estimates, and to assess the sensitivity of conventional estimates to plausible deviations from the identifying assumptions. The estimation method and its properties are described, and two empirical applications are demonstrated.

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/~bkrauth/code`.

1 Introduction

This paper describes a simple approach to the most common problem in applied microeconometrics: estimating a linear causal effect when the variable of interest might be correlated with relevant unobserved variables. The microeconometrician's standard methods - natural experiments, instrumental variables, fixed effects, and simply adding control variables - are all designed to solve this problem. However, there are many cases where the assumptions needed to identify the effect of interest are plausible but not necessarily exactly true. In this case it is useful to have a means of determining how sensitive one's results are to small or moderate deviations from the identifying assumptions.

This paper provides a simple means of performing such a sensitivity analysis when the causal effect is estimated by OLS regression of the outcome of interest on the explanatory variable of interest and a set of control variables. The validity of doing so depends on the strong assumption of conditional exogeneity, i.e., that the explanatory variable of interest is uncorrelated with the unobservable factors in the regression. This paper models deviations from conditional exogeneity in terms of a single parameter that measures the correlation between the explanatory variable of interest and the unobservable factors, relative to the correlation between the explanatory variable of interest and the control variables. In this framework, the conditional exogeneity assumption can be interpreted as a point restriction on the relative correlation parameter (i.e., that it is exactly zero) that yields consistent point estimates of the effect. When this point restriction is replaced by a weaker interval restriction, the effect is partially identified and one can construct consistent bounds on its true value. Hypothesis tests and confidence intervals can also be constructed and have the usual interpretation.

Two example applications show the potential usefulness of the methods developed here. The first application is to data from a natural or designed experiment in which there are small deviations from true random assignment. It is based on Krueger's (1999) analysis of data from Project STAR, a well-known study of the effect of smaller class size on student outcomes. The second application is to an observational study in which the claim of conditional exogeneity is controversial, but the usual tricks of applied microeconometrics are equally unappealing. It is based on Subramanian and Kawachi's (2003) study using CPS data to measure the effect of income inequality on individual health.

The methods developed in this paper find that the experimental Project STAR results are much more robust than the observational results on inequality and health. While this finding is not surprising, what matters here is that the difference is found entirely in the data, and can be quantified in a relatively straightforward manner. For example, the positive effect of smaller classes on kindergarten test scores remains even if the correlation between class size and unobservables is as much as ten times the correlation observed between class size and the observed control variables. In contrast, the positive relationship between state-level income inequality and an individual's probability of being in fair to poor health disappears if the correlation between inequality and unobservables is as much as 23% of the correlation between inequality and the observed control variables.

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. Manksi (1994; 2003) adopts a mostly nonparametric frequentist approach, and recasts sensitivity analysis as estimation and inference under assumptions that yield only partial identification of the parameter of interest. Manski's research has also led to 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 characterizes the unmeasurable deviation from conditional exogeneity in terms that are proportional to some related measurable quantity. Rosenbaum (2002), following a tradition of sensitivity analysis in the statistics literature dating back to Cornfield et al. (1959), develops a treatment-effects framework in which there is an unobserved binary variable affecting both outcomes and selection into treatment. The 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) uses as a sensitivity parameter the proportion of otherwise unexplained variation in the outcome that could be explained by the unobserved term in a treatment selection equation. Altonji, Elder, and Taber (2005) evaluate the sensitivity of the estimated Catholic school effect to endogenous school selection by incorporating a parametric selection model in which the degree of selection on unobservables is proportional to the degree of selection on observables. Krauth (2007) evaluates the sensitivity of estimated peer effects in youth smoking to nonrandom peer selection by modeling the within-group correlation in unobservables as a proportion of the within-group correlation in observables. Lewbel (2011) exploits a cross-equation covariance restriction to bound the parameters of a heteroskedastic simultaneous equations model without using instruments. Conley, Hansen and Rossi (2010), Kraay (2010), and Nevo and Rosen (2010) develop systematic methods of sensitivity analysis in instrumental variables regression in which the conventional IV exclusion restriction is “almost” true. Like these earlier studies, this paper models deviations from the standard approach in relative terms. Unlike those earlier papers, the analysis here is applicable to a simple OLS-based research design.

2 Methodology

2.1 Model

Let $\mathbf{D} \equiv [\mathbf{x}\ y\ z]$, where y is a scalar outcome, z is a scalar explanatory variable of interest, and $\mathbf{x}$ is a k-length row vector of additional control variables including an intercept. Our goal is to measure the effect of z on y, where the causal model assumes this effect is constant and linear. That is:

Assumption 1: $y = y(z) = \theta_0 z + u$

where the random function $y(.)$ is a potential outcome function giving the outcome associated with each possible value of z, the parameter of interest θ_0 represents the effect of z on y, and the unobserved random variable u represents the effect of all other factors. These other factors are not affected by z but may be correlated with it. Section 3.1 considers an extension in which the effect of z on y is heterogeneous across individuals.

The control variables do not enter into the causal model, and are only of interest to the extent they aid in the estimation of θ_0. Let $u^p = \mathbf{x}\beta_0$ be the best linear predictor of u given $\mathbf{x}$, i.e.:

$$\begin{aligned}\beta_0 &\equiv E(\mathbf{x}'\mathbf{x})^{-1}E(\mathbf{x}'u) \qquad (1)\\ &= E(\mathbf{x}'\mathbf{x})^{-1}E(\mathbf{x}'y) - \theta_0 E(\mathbf{x}'\mathbf{x})^{-1}E(\mathbf{x}'z)\\ \underbrace{\mathbf{x}\beta_0}_{u^p} &= \underbrace{\mathbf{x}E(\mathbf{x}'\mathbf{x})^{-1}E(\mathbf{x}'y)}_{y^p} - \theta_0 \underbrace{\mathbf{x}E(\mathbf{x}'\mathbf{x})^{-1}E(\mathbf{x}'z)}_{z^p}\end{aligned}$$

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

$$v \equiv u - \mathbf{x}\beta_0 \qquad (2)$$

Note that these are just definitions and that β_0 has no particular causal interpretation. Putting (1) and (2) together, we get:

$$y = \theta_0 z + \mathbf{x}\beta_0 + v \qquad \text{where } E(\mathbf{x}'v) = 0 \qquad (3)$$

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

Instead of the conventional practice of assuming this correlation is exactly zero, we impose a weaker *relative correlation restriction.* A relative correlation restriction is defined as a nonempty and closed interval Λ that is known by the econometrician to satisfy:

Assumption 2: $cov(z, v)\sqrt{var(\mathbf{x}\beta_0)} = \lambda_0 cov(z, \mathbf{x}\beta_0)\sqrt{var(v)}$
for some $\lambda_0 \in \Lambda$

As long as both $cov(z, \mathbf{x}\beta_0)$ and $var(v)$ are nonzero, Assumption 2 is equivalent to a simpler and more intuitive condition:

$$\lambda_0 = \frac{\mathrm{corr}(z, v)}{\mathrm{corr}(z, \mathbf{x}\beta_0)} \in \Lambda \qquad (4)$$

That is, we are assuming that the correlation of the variable of interest (z) with unobservables (v) relative to its correlation with observables ($\mathbf{x}\beta_0$) can be restricted to lie within some known range (Λ). That range can be very wide ($\Lambda = \mathbb{R}$) in which case Assumption 2 implies almost no restrictions on the model, or it can be vary narrow (e.g. conditional exogeneity, which can be written as $\Lambda = \{0\}$). Section 2.2 discusses the interpretation of relative correlation restrictions, and the considerations relevant to selecting an appropriate relative correlation restriction for empirical work. In this section, Λ is taken as given.

In order to discuss identification, estimation and inference, suppose we have a sample of size n on $\mathbf{D}$ that can be used to construct a consistent and asymtpotically normal estimator of its first two moments. That is, we have a random vector $\hat{m}_n$ such that:

Assumption 3: $\sqrt{n}\,(\hat{m}_n - m_0) \overset{D}{\rightarrow} N(0, \Sigma)$

where:

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

and vech(.) is the half-vectorization function (i.e., given a symmetric matrix it returns a column vector of its unique elements). Since m_0 is just a vector of first and second moments, Assumption 3 is satisfied by the corresponding sample average from a random sample.

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

Assumption 4: $E(\mathbf{D}'\mathbf{D})$ is finite and positive definite

Positive-definiteness of $E(\mathbf{D}'\mathbf{D})$ is easily verified in data, and guarantees for example that β_0 is well-defined.

Next, at least one of the control variables is useful in forecasting y:

Assumption 5: $var(y^p) > 0$

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 z:

Assumption 6: $var(z^p) > 0$

Assumption 6 allows for a simple description of the estimation method and its properties in the remainder of Section 2. It is easily testable by an ordinary coefficient significance test, and is likely to hold in most cases of interest. However, Assumption 6 is violated in an important special case: when z is assigned completely at random. Section 3.1 covers this special case, and shows that the results are similar to those presented in Section 2.

2.2 Interpreting relative correlation restrictions

The relative correlation restriction Λ is the primary identifying assumption in this model. As a result, the usefulness of the model in applied work depends on whether one can construct plausible relative correlation restrictions. This section discusses that issue.

The model presented in Section 2.1 could be parameterized in terms of the absolute correlation, i.e., the value of $\operatorname{corr}(z, v)$. Instead, it is parameterized in terms of relative correlation, i.e., the ratio of $\operatorname{corr}(z, v)$ to $\operatorname{corr}(z, \mathbf{x}\beta_0)$. This is done to reflect the common practice of using patterns in observed explanatory variables as evidence in favor of ultimately untestable assumptions about unobserved variables.

The clearest example of this is the standard practice in experimental studies of demonstrating covariate balance. Most economics papers using an experimental design present a table showing that pre-treatment variables are roughly balanced between treatment and control groups. This evidence of covariate balance is often cited in support of the identification scheme. Yet balance in *observed* covariates has no direct consequences for identification: any imbalance in observed covariates can be addressed in principle by regression, matching, and/or weighting. In contrast, balance in *unobserved* pretreatment covariates is a necessary and untestable condition for identification. In other words, the researcher is using the joint distribution of observed covariates to make inferences about the joint distribution of unobserved covariates.

In observational studies using control variables, a related common procedure is to 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 effect estimate 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 the 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.

To interpret the sign and scale of λ_0 it is useful to consider the omitted variables bias formula for the simple linear regression of y on z with no control variables:

$$\begin{aligned}\frac{cov(y,z)}{var(z)} &= \frac{cov(\theta_0 z + \mathbf{x}\beta + v, z)}{var(z)} \\ &= \theta_0 + \frac{cov(z,\mathbf{x}\beta)}{var(z)} + \frac{cov(z,v)}{var(z)} \\ &= \theta_0 + \underbrace{\mathrm{corr}(z,\mathbf{x}\beta)\sqrt{\frac{var(\mathbf{x}\beta)}{var(z)}}}_{\text{bias from omitting } \mathbf{x}} + \underbrace{\lambda_0 \mathrm{corr}(z,\mathbf{x}\beta)\sqrt{\frac{var(v)}{var(z)}}}_{\text{bias from omitting } v}\end{aligned}$$

This implies that:

- If $\lambda_0 = 0$, then $\mathrm{corr}(z, v)$ is also zero. That is, there is no omitted variables bias in the OLS regression once we control for $\mathbf{x}$.
- If $\lambda_0 > 0$, then $\mathrm{corr}(z, v)$ has the same sign as $\mathrm{corr}(z, \mathbf{x}\beta_0)$. That is, controlling for $\mathbf{x}$ reduces but does not eliminate bias.
- If $\lambda_0 < 0$, then $\mathrm{corr}(z, v)$ has the oppposite sign of $\mathrm{corr}(z, \mathbf{x}\beta_0)$. That is, controlling for $\mathbf{x}$ may reduce or increase bias.
- If $\lambda_0 = 1$, then $\mathrm{corr}(z, v)$ is of both the same sign and magnitude as $\mathrm{corr}(z, \mathbf{x}\beta_0)$.

We can thus interpret λ_0 as an index of how well-selected the control variables are for reducing the bias in OLS estimation. In a slightly different setting, Altonji, Elder and Taber (2005) make the argument that equal correlation ($\lambda_0 = 1$ here) is what one would expect on average if the control variables were chosen randomly from a large set of plausible explanatory variables. This argument has clear limitations – few researchers would select control variables at random – but it at least suggests that $\lambda_0 = 1$ is something of a benchmark value. Presumably, researchers would attempt to select precisely those available control variables that are most likely to reduce bias, and so a value of λ_0 between zero and one can be interpreted as somewhere between a perfectly chosen set of control variables and a randomly chosen set of control variables.

It may also be useful to answer the reverse question: what value of λ_0 would overturn the OLS results? For example, if the OLS estimate is positive, we may want to know how big λ_0 would have to be in order to imply that the true effect is zero or negative. Alternatively, one might be interested in how big a relative correlation would be needed to reduce the implied effect by some percentage or amount relative to the OLS estimate, as in Imbens (2003).

2.3 Identification

In general, it is not possible in this setting to identify the true value of θ_0 but it is possible to identify a nontrivial set Θ_0 that must contain θ_0. This set is known as the *identified set* for the true effect, and includes ordinary point identification ($\Theta_0 = \{\theta_0\}$), partial identification (Θ_0 is a proper subset of $\mathbb{R}$), and nonidentification ($\Theta_0 = \mathbb{R}$) as special cases. This section characterizes the identified set for θ_0 and how it can be constructed.

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:

$$E_m(\mathbf{D}'\mathbf{D}) \text{ is finite and positive definite} \tag{5}$$

$$var_m(y^p) > 0 \tag{6}$$

$$var_m(z^p) > 0 \tag{7}$$

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 sufficiently large sample will be close to m_0 but not identical. The model's assumptions described in Section 2.1 imply that m_0 satisfies (5)-(7) 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 β_0 and λ_0 would be identified if θ_0 were known. Ignoring for the moment the possibility of singular matrices or division by zero, let:

$$\beta(\theta; m) \equiv E_m(\mathbf{x}'\mathbf{x})^{-1}E_m(\mathbf{x}'y) - \theta E_m(\mathbf{x}'\mathbf{x})^{-1}E_m(\mathbf{x}'z) \tag{8}$$

and:

$$\lambda(\theta; m) \equiv \frac{\text{corr}_m(z, y - \theta z - \mathbf{x}\beta(\theta; m))}{\text{corr}_m(z, \mathbf{x}\beta(\theta; m))} \tag{9}$$

Equations (8) and (9) can be used to express the unknown parameters β_0 and λ_0 as known functions of the unknown structural parameter and vector of second moments, i.e.:

$$\begin{aligned}
\beta(\theta_0; m_0) &= E(\mathbf{x}'\mathbf{x})^{-1}E(\mathbf{x}'y) - \theta_0 E(\mathbf{x}'\mathbf{x})^{-1}E(\mathbf{x}'z) \\
&= \beta_0 \\
\lambda(\theta_0; m_0) &= \frac{\text{corr}(z, y - \theta_0 z - \mathbf{x}\beta_0)}{\text{corr}(z, \mathbf{x}\beta_0)} \\
&= \frac{\text{corr}(z, v)}{\text{corr}(z, \mathbf{x}\beta_0)} \\
&= \lambda_0
\end{aligned}$$

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

Finally, let $\Theta_0(\Lambda; m)$ be defined as the set of all θ satisfying:

$$\begin{aligned}
&cov_m(z, y - \theta z - \mathbf{x}\beta(\theta; m))\sqrt{var_m(\mathbf{x}\beta(\theta; m))} \\
&\qquad = \lambda cov_m(z, \mathbf{x}\beta(\theta; m))\sqrt{var_m(y - \theta z - \mathbf{x}\beta(\theta; m))}
\end{aligned} \tag{10}$$

for some $\lambda \in \Lambda$. By construction, $\Theta_0(\Lambda; m_0)$ is the set of all θ_0 satisfying Assumption 2, i.e., the identified set for the true effect. Figure 1 shows how $\Theta_0(\Lambda; m_0)$ can be found from $\lambda(\theta; m_0)$.

Proposition 1 (Properties of $\lambda(.)$) *Let m satisfy (5)-(7). Then the function $\lambda(.; m)$ has the following properties:*

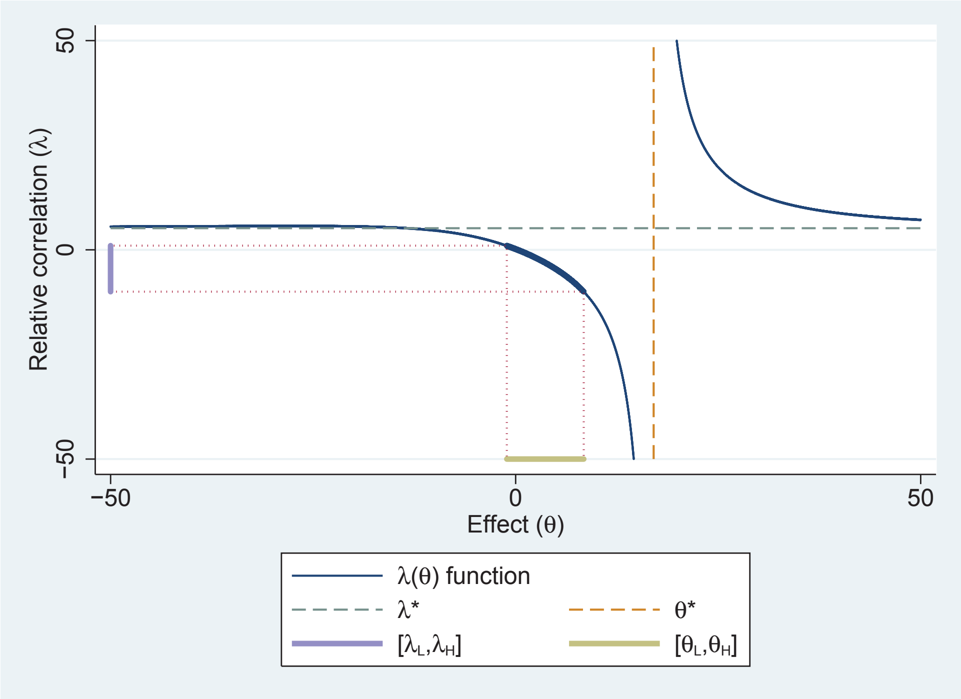

Figure 1: A typical $\lambda(\theta; m)$ function giving the relative corrrelation (λ) as a function of the assumed value θ for the effect of interest. The function exists and is differentiable in θ everywhere but at θ^* (the value of θ at which $\mathrm{corr}(z, \mathbf{x}\beta(\theta)) = 0$). Its limit as θ approaches positive or negative infinity is λ^*. Near θ^*, the function goes towards positive or negative infinity. Both θ^* and λ^* are easily identified from the data. The identified set $\Theta_0 = [\theta_L, \theta_H]$ given the relative correlation restriction $\Lambda = [\lambda_L, \lambda_H]$ can be found by inverting $\lambda(\theta; m)$.

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

$$\theta^*(m) \equiv \frac{cov_m(z^p, y^p)}{var_m(z^p)}$$

2. *Let:*

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

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

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

3. *For any $\lambda \neq \lambda^*(m)$ there exists at least one θ satisfying (10).*

4. *Let:*

$$\tilde{\Theta}_0(\Lambda; m) = \{\theta : \lambda(\theta; m) \in \Lambda\}$$

Then:

$$\tilde{\Theta}_0(\Lambda; m) \subset \Theta_0(\Lambda; m) \subset \tilde{\Theta}_0(\Lambda; m) \cup \{\theta^*(m)\}$$

Proof: See Appendix A.1.

The identified set is not necessarily convex, so it will usually be more convenient to work with its upper and lower bounds:

$$\theta_L(\Lambda; m) = \inf \Theta_0(\Lambda; m) \tag{11}$$

$$\theta_H(\Lambda; m) = \sup \Theta_0(\Lambda; m) \tag{12}$$

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 $\Theta_0(\Lambda; m_0)$ is nonempty and bounded if $\lambda^*(m_0) \notin \Lambda$.*

Proof: See Appendix A.2.

2.4 Estimation

The identified features of the model can be estimated by substituting $\hat{m}_n$ for m_0 in the quantities defined in Section 2.3. Let:

$$\begin{aligned}
\hat{\lambda}(\theta) &\equiv \lambda(\theta; \hat{m}_n) \\
\hat{\lambda}^* &\equiv \lambda^*(\hat{m}_n) \\
\hat{\theta}^* &\equiv \theta^*(\hat{m}_n) \\
\hat{\theta}_L(\Lambda) &\equiv \inf\{\theta : \lambda(\theta; \hat{m}_n) \in \Lambda\} \\
\hat{\theta}_H(\Lambda) &\equiv \sup\{\theta : \lambda(\theta; \hat{m}_n) \in \Lambda\}
\end{aligned} \tag{13}$$

An important complication in characterizing the asymptotic properties of these estimators is the possibility of nonidentification. That is, if the identified set is unbounded, a good estimator for the identified set should also be unbounded with high probability for a sufficiently large sample size. Proposition 3 below shows this to be the case.

Proposition 3 (Consistency) *The estimators defined in (13) are consistent. That is:*

$$\begin{aligned}
\hat{\theta}^* &\xrightarrow{p} \theta^*(m_0) \\
\hat{\lambda}^* &\xrightarrow{p} \lambda^*(m_0) \\
\hat{\lambda}(\theta) &\xrightarrow{p} \lambda(\theta; m_0) \qquad \textit{for all } \theta \neq \theta^*(m_0)
\end{aligned}$$

If $\Theta_0(\Lambda; m_0)$ *is bounded then:*

$$\begin{aligned}
\hat{\theta}_L(\Lambda) &\xrightarrow{p} \theta_L(\Lambda; m_0) \qquad \textit{if } \frac{d\lambda(\theta; m_0)}{d\theta}|_{\theta=\theta_L(\Lambda; m_0)} \neq 0 \\
\hat{\theta}_H(\Lambda) &\xrightarrow{p} \theta_H(\Lambda; m_0) \qquad \textit{if } \frac{d\lambda(\theta; m_0)}{d\theta}|_{\theta=\theta_H(\Lambda; m_0)} \neq 0
\end{aligned}$$

and if $\Theta_0(\Lambda; m_0) = \mathbb{R}$ *then for any* B*:*

$$\lim_{n\to\infty} \Pr((\hat{\theta}_H(\Lambda) > B) = \lim_{n\to\infty} \Pr((\hat{\theta}_L(\Lambda) < B) = 1$$

Proof: See Appendix A.3.

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

2.5 Inference

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

The quantities to be estimated are in most cases differentiable functions of m_0, so they will be asymptotically normal with a covariance matrix that can be obtained through straightforward application of the delta method. Proposition 4 below states this more explicitly for the endpoints of the identified set.

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

$$A \equiv -\begin{bmatrix} \left.\dfrac{\nabla_m \lambda(\theta; m)}{\partial\lambda(\theta; m)/\partial\theta}\right|_{\theta=\theta_L(\Lambda; m_0), m=m_0} \\ \left.\dfrac{\nabla_m \lambda(\theta; m)}{\partial\lambda(\theta; m)/\partial\theta}\right|_{\theta=\theta_H(\Lambda; m_0), m=m_0} \end{bmatrix}$$

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

$$\sqrt{n}\begin{bmatrix} \hat{\theta}_L(\Lambda) - \theta_L(\Lambda; m_0) \\ \hat{\theta}_H(\Lambda) - \theta_H(\Lambda; m_0) \end{bmatrix} \overset{D}{\to} N\left(0, A\Sigma A'\right)$$

Proof: See Appendix A.4.

Note that existence of A requires two conditions to be satisfied: that neither θ_L nor θ_H is identical to $\theta^*(m_0)$ (guaranteeing existence of the derivatives), and that $\partial\lambda(\theta; m)/\partial\theta$ is nonzero when evaluated at θ_L or θ_H. By analogy, the asymptotic distribution is likely to provide a poor approximation to the finite sample distribution when either θ_L or θ_H is close to θ^*, or when λ^* is close to Λ.

The asymptotic distribution described in Proposition 4 can be used to construct Wald-type hypothesis tests and confidence intervals for θ_0. In constructing confidence intervals under partial identification, Imbens and Manski (2004) note the necessity of distinguishing between a confidence interval for the identified set:

$$\lim_{n\to\infty} \Pr(\Theta_0(\Lambda) \subset CI^{set}) = 1 - \alpha$$

and a confidence interval for the true parameter value:

$$\lim_{n\to\infty} \inf_{\theta\in\Theta_0(\Lambda)} \Pr(\theta \in CI^{par}) = 1 - \alpha$$

A confidence interval for the identified set can be constructed using the lower and upper bounds, respectively, of the ordinary confidence intervals for $\hat{\theta}_L(\Lambda)$ and $\hat{\theta}_H(\Lambda)$.

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.

3 Extensions

3.1 Perfect experiments

Assumption 6 of the model says that the explanatory variable of interest is at least slightly correlated with the control variables. This assumption is made strictly for convenience in presenting the results in Sections 2.3 - 2.5, as it guarantees existence of the intermediate quantities $\lambda(\theta, m_0)$, $\lambda^*(m_0)$, and $\theta^*(m_0)$, and avoids the need to discuss various exceptions and special cases. While Assumption 6 is likely to hold in most cases of interest, it will not hold in the special case of pure random assignment. That is, if the treatment z really is independent of the pretreatment control variables $\mathbf{x}$ it will also be uncorrelated with those variables, and Assumption 6 will not hold. This section considers the implications of replacing Assumption 6 with its opposite:

Assumption $6'$: $var(z^p) = 0$

Assumption $6'$ implies that $\lambda(\theta, m_0)$, $\lambda^*(m_0)$, and $\theta^*(m_0)$ are undefined. However, Propositions 5 and 6 below show that the identified set is still well-defined, and can be estimated consistently by the estimators described in Section 2.4.

Proposition 5 (Size of the identified set) *The identified set $\Theta_0(\Lambda; m_0)$ is nonempty and bounded. In particular, $\Theta_0(\Lambda; m_0) = \left\{\frac{cov(z,y)}{var(z)}\right\} = \{\theta_0\}$.*

Proof: See Appendix A.5.

Proposition 6 (Consistency) *Suppose that Λ is bounded and includes zero. Then: $\hat{\theta}_L(\Lambda) \overset{p}{\to} \theta_0$ and $\hat{\theta}_H(\Lambda) \overset{p}{\to} \theta_0$.*

Proof: See Appendix A.6.

3.2 Heterogeneity in response

The model presented in Section 2 assumes a constant marginal effect of z on y. This section describes how the estimation method would apply to the case of heterogeneous response.

Replace Assumption 1 with:

Assumption $1'$: $y = y(z) = tz + u \qquad \text{where } E(t) = \theta_0$

where t is the individual-specific marginal effect of z on y and θ_0 is a parameter representing the average marginal effect in the population. If z 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 θ_0 the average treatment effect. For notational simplicity, normalize y, z, and $\mathbf{x}$ to mean zero so that $\mathbf{x}$ does not need to have an intercept.

The average marginal effect θ_0 is point-identified if the potential outcomes are mean-independent of z conditional on $\mathbf{x}$, i.e., if $E(t|z, \mathbf{x}) = E(t|\mathbf{x})$ and $E(u|z, \mathbf{x}) = E(u|\mathbf{x})$. If these conditional expectation functions happen to be linear in $\mathbf{x}$, then θ_0 is consistently estimated by the OLS regression of y on $(z, \mathbf{x}, z\mathbf{x})$. Note that conditional mean independence is the important assumption here, as one can always choose $\mathbf{x}$ to make the conditional expectations linear (e.g. by making $\mathbf{x}$ binary).

Next we derive a version of equation (14) that replaces the mean-independence and linear CEF assumptions with relative correlation restrictions. Without loss of generality, let $z\mathbf{x}\Gamma_0+\mathbf{x}\beta_0$ be the best linear predictor of $y - \theta_0 z$ given $(z\mathbf{x}, \mathbf{x})$. Let $v \equiv y - \theta_0 z - z\mathbf{x}\Gamma_0 - \mathbf{x}\beta_0$ be the corresponding residual. Then

$$y = \theta_0 z + z\mathbf{x}\Gamma_0 + \mathbf{x}\beta_0 + v \qquad \text{where } E(z\mathbf{x}'v) = E(\mathbf{x}'v) = 0. \tag{14}$$

Consistent OLS estimation of equation (14) requires the additional assumption that $E(zv) = 0$ or equivalently $\text{corr}(z, v) = 0$. 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 2.1, this is done by replacing the absolute correlation restriction $\text{corr}(z, v) = 0$ with a relative correlation restriction Λ such that:

$$\lambda_0 = \frac{\text{corr}(z, v)}{\text{corr}(z, z\mathbf{x}\Gamma_0 + \mathbf{x}\beta_0)} \in \Lambda \tag{15}$$

In this version of the model, the relative correlation parameter λ_0 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 (14) and (15) are identical to equations (3) and (4) in Section 2.1, with $(z\mathbf{x}, \mathbf{x})$ as the control variables instead of just $\mathbf{x}$. Therefore this model can be fit into the framework of Section 2.1, and the results from Section 2 apply directly.

4 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.

4.1 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.

4.1.1 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 most important deviations from the experimental design were:

1. Between grades, some students were moved between the small and regular class groups as a result of behavioral issues and/or possibly pressure by parents.
2. New students entered Project STAR schools during the experiment, and were randomly assigned to one of the experimental groups.
3. Some students moved out of their original schools. Krueger notes that there is some evidence that students in the small class treatment are less likely to change schools.

Krueger's approach to the problem of imperfect randomization is similar to that described in Section 2.2. 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 this provides evidence for 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 is in some sense simply a matter of insufficient sample size.

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.

4.1.2 Data

The data are from Finn et al. (2007). Table 1 reports summary statistics and is a partial reconstruction of the table in the appendix of Krueger (1999). Most table entries are self-explanatory, with the exception of the test score variables. The test score variables are constructed according to the procedure described by Krueger: raw scores on each of the individual subject tests in a given year are converted into percentiles based on the distribution of scores among students in the control group. Each student's percentile scores are then averaged across subjects. The resulting score thus has a potential range of zero to 100, has a mean and median close to 50, and can be roughly though not exactly interpreted in percentile units. Krueger leaves out some details in describing how the test score variable was constructed (e.g., how ties were broken in calculating percentiles), so the average test scores reported in Table 1 differ by up to 0.3 of a percentage point from those reported by Krueger.

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, z and $\mathbf{x}$ in terms of deviations from the corresponding school-level averages.

	Grade			
Variable	K	1	2	3
Average class size	20.3	21.0	21.1	21.3
(std. dev.)	(4.0)	(4.0)	(4.1)	(4.4)
Average percentile score, SAT	51.4	51.8	51.3	51.3
(std. dev.)	(26.7)	(26.9)	(26.5)	(27.0)
% Free lunch	48	52	51	51
% White/Asian	67	67	65	67
% Female	49	48	48	48
Average Age on September 1st	5.43	6.58	7.66	8.70
(std. dev.)	(0.35)	(0.49)	(0.56)	(0.59)
% Exited sample	29	26	21	
% of teachers with MA+ degree	35	35	37	44
% of teachers who are White	84	83	80	79
% of teachers who are male	0	0	1	3
Teacher experience, years	9.26	11.63	13.14	13.93
(std. dev.)	(5.81)	(8.94)	(8.65)	(8.61)
# schools	79	76	75	75
# students	6325	6829	6840	6802
# small classes	127	124	133	140
# regular classes	99	115	100	89
# reg./aide classes	99	100	107	107

Table 1: Summary statistics, Project STAR data.

4.1.3 OLS results

Table 2 shows OLS regression results, and is a partial reconstruction of Table 5 in Krueger (1999). For each grade, two specifications are reported. Specification (1) corresponds to specification (4) in Krueger's Table 5, while specification (2) omits the regular/aide class indicator. This is done because the approach described in this paper is designed to evaluate the effect of a single explanatory variable. Both Krueger and the original Project STAR research team found that the regular-aide treatment was nearly irrelevant to student outcomes. The results in Table 2 suggest that the small-class treatment increases test scores by five to seven percentile points.

Explanatory	Kindergarten		Grade 1		Grade 2		Grade 3	
Variable	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Small class	5.33	5.20	7.55	6.72	5.76	4.97	5.01	5.30
	(1.20)	(1.04)	(1.17)	(1.05)	(1.22)	(1.05)	(1.22)	(1.05)
Regular/aide class	0.26		1.77		1.54		-0.51	
	(1.07)		(0.97)		(1.06)		(1.10)	
White/Asian	8.39	8.39	6.94	6.98	6.45	6.48	6.05	6.05
	(1.36)	(1.36)	(1.19)	(1.19)	(1.19)	(1.19)	(1.44)	(1.44)
Girl	4.38	4.38	3.83	3.82	3.42	3.41	4.19	4.20
	(0.63)	(0.63)	(0.56)	(0.56)	(0.60)	(0.60)	(0.66)	(0.66)
Free lunch	-13.08	-13.08	-13.55	-13.55	-13.62	-13.64	-12.95	-12.94
	(0.77)	(0.77)	(0.88)	(0.88)	(0.72)	(0.72)	(0.81)	(0.81)
White teacher	-1.13	-1.09	-4.02	-4.23	0.43	0.61	0.28	0.27
	(2.17)	(2.18)	(1.95)	(1.96)	(1.75)	(1.75)	(1.80)	(1.80)
Teacher experience	0.26	0.27	0.06	0.07	0.10	0.11	0.05	0.05
	(0.11)	(0.10)	(0.06)	(0.06)	(0.06)	(0.07)	(0.06)	(0.06)
Master's degree	-0.59	-0.60	0.44	0.55	-1.06	-0.92	0.93	0.89
	(1.05)	(1.05)	(1.07)	(1.08)	(1.06)	(1.04)	(1.18)	(1.18)
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of observations	5,839	5,839	6,452	6,452	5,953	5,953	6,100	6,100

Table 2: OLS estimates of effect of class sizes on average percentile rank on Stanford Achievement Test. Standard errors (robust to clustering by teacher) are in parentheses.

4.1.4 RCR results

Table 3 reports the results from estimating the treatment effect under a series of relative correlation restrictions. As in the OLS results, the outcome variable y is the average percentile SAT score, the explanatory variable of interest z is the small class treatment, and the set of control variables $\mathbf{x}$ are those teacher and student background variables included in specification (2) of Table 2. To account for school fixed effects, each variable is expressed in terms of deviation from the corresponding school-level average. The point estimates of $\theta_L(\Lambda)$ and $\theta_H(\Lambda)$ are reported in square brackets, while the 95% asymptotic confidence intervals for θ_0 are reported in round brackets. Confidence intervals are calculated based on the method described in Imbens and Manski (2004), and are robust to clustering by teacher. For the relative correlation restriction $\Lambda = (-\infty, 0.0]$, the function $\hat{\lambda}(\theta)$ does not exist at $\hat{\theta}_H(\Lambda) = \hat{\theta}^*$ and so the confidence intervals reported are one-tailed confidence intervals for $\theta_L(\Lambda)$ only.

The results in Table 3 suggest that Krueger's original findings are quite robust. The estimated effect of small classes in kindergarten remains similar in magnitude even if the correlation between the treatment and unobservables is as much as ten times as large as the correlation between the treatment and observables. The results are slightly less robust for the later grades. The range of grade 1 treatment effects consistent with the data is strictly positive as long as the correlation between treatment and unobservables is somewhat less than three times as large as the correlation between the treatment and unobservables. For the grade 2 and 3 data, the range of estimated treatment effects is positive for a relative correlation of slightly more than three, but not for a relative correlation of 3.5 or above.

4.2 Application #2: Inequality and health

The second type of application of the RCR approach is to studies using observational data in which causal effects are estimated by OLS regression using a carefully selected set of control variables. Despite the increased use of methods like natural experiments, instrumental variables, regression discontinuity, and difference-in-differences, the OLS-with-controls regression remains a staple of applied work. The reason for this is simple: there are many interesting questions for which there exists no credible source of exogenous variation in the explanatory variable of interest. One example of such a question is the relationship between inequality and health.

4.2.1 Background

The relationship between economic inequality and health is the subject of an extensive literature in public health, surveyed by Deaton (2003), Subramanian and Kawachi (2004), and Wilkinson and Pickett (2006). The typical finding in this literature is that a higher level of income inequality has a substantial negative impact on individual health outcomes in industrialized countries,

Relative	Bounds on class size effect by grade			
correlation	$[\hat{\theta}_L(\Lambda), \hat{\theta}_H(\Lambda)]$			
restriction (Λ)	K	1	2	3
$\{0.00\}$	5.20	6.72	4.97	5.30
	$(3.17, 7.24)$	$(4.66, 8.78)$	$(2.90, 7.04)$	$(3.24, 7.36)$
$[0.00, 1.00]$	$[5.14, 5.20]$	$[4.50, 6.72]$	$[3.55, 4.97]$	$[4.08, 5.30]$
	$(2.49, 7.21)$	$(2.27, 8.46)$	$(1.58, 6.73)$	$(1.30, 7.10)$
$[0.00, 3.00]$	$[4.99, 5.20]$	$[-0.15, 6.72]$	$[0.57, 4.97]$	$[0.44, 5.30]$
	$(-1.00, 7.20)$	$(-5.11, 8.45)$	$(-3.49, 6.71)$	$(-7.27, 7.06)$
$[0.00, 5.00]$	$[4.84, 5.20]$	$[-5.19, 6.72]$	$[-2.61, 4.97]$	$[-6.87, 5.30]$
	$(-5.31, 7.20)$	$(-14.22, 8.45)$	$(-9.74, 6.71)$	$(-29.38, 7.07)$
$[0.00, 10.00]$	$[4.37, 5.20]$	$[-21.58, 6.72]$	$[-12.04, 4.97]$	$(-\infty, \infty)$
	$(-18.48, 7.20)$	$(-53.46, 8.46)$	$(-31.78, 6.72)$	$(-\infty, \infty)$
$[0.00, 15.00]$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
$[0.00, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
$(-\infty, 0.00]$	$[5.20, 8.17]$	$[6.72, 134.57]$	$[4.97, 96.33]$	$[5.30, 15.12]$
	$(3.37, \infty)$	$(4.83, \infty)$	$(2.90, \infty)$	$(3.27, \infty)$
Other parameter estimates:				
$\hat{\lambda}^*$	12.31	13.85	14.88	5.79
$\hat{\theta}^*$	8.17	134.57	96.33	15.12
$\hat{\lambda}(0)$	28.94	2.94	3.37	3.18

Table 3: Bounds on the effect of class sizes on average percentile rank on Stanford Achievement Test, given relative correlation restrictions. Intervals in square brackets are the bounds themselves, while the intervals in the round brackets are 95% cluster-robust asymptotic confidence intervals.

even after accounting for the individual's own income. Wilkinson and Pickett (2009) argue that the key mechanism for this effect is that the low social status associated with low relative income leads to increased stress, which has both a direct negative impact on health and an indirect effect through depression and unhealthy behaviors. In wealthy societies with extensive public healthcare systems, health behavior may be more important than health expenditures in explaining cross-sectional variation in health outcomes. Wilkinson and Pickett (2009), among others, use these findings to argue in favor of large-scale income redistribution in industrialized countries.

Most recent empirical work in this literature regresses individual health on regional or state-level income inequality, controlling for the respondent's own income and background characteristics. Although many of these studies exhibit a great deal of methodological sophistication and complexity, including the deployment of elaborate multilevel models, almost none have done much to address the issue of endogenous community selection. For example, none of the 21 studies cited in the review article by Subramanian and Kawachi (2004) have a research design aimed at addressing endogenous community selection. Researchers in this literature are aware of the issue, but this question is particularly ill-suited for the typical methods microeconometricians use to deal with endogeneity. The inequality-health relationship has several relevant features:

1. Most commonly hypothesized mechanisms by which inequality affects health (e.g., stress, depression, increased smoking, drinking, and drug use) operate with long and variable lags.
2. Inequality changes slowly over time, and is measured with a great deal of noise.
3. Government policies that affect the income distribution are also likely to affect relative prices, allocations, and other variables relevant to health outcomes.[1]

The first two features make the use of panel data with cross-sectional fixed effects particularly unappealing, while the third feature implies that suitable instrumental variables or natural experiments will be difficult to find.

[1] One can also argue that the "effect" of inequality on health is not clearly defined because inequality is not a policy but rather an outcome of policy. For the purposes of this example, the causal model in which inequality has a well-defined (if zero) effect on health will be taken as given.

4.2.2 Data

The primary data source is the pooled 1996 and 1998 Current Population Survey (CPS) March supplement (US Department of Labour, 1998). The sample consists of all CPS respondents at least 18 years of age, and the outcome variable is a binary indicator of self-reported poor health. Specifically, respondents were asked "Would you say your health in general is ..." and are coded as $y = 1$ if they reported "Fair" or "Poor" and $y = 0$ if they reported "Good," "Very Good," or "Excellent." This particular data source and outcome variable have been used extensively in the inequality and health literature (Blakely et al., 2000, 2002; Mellor and Milyo, 2002, 2003; Subramanian and Kawachi, 2003, 2004), and so have been selected for ease of comparison. Individual-level explanatory variables include age, sex, race, education, log equivalized household income (total household income divided by the square root of household size), employment status, and health insurance status. The community-level variable is the state-level Gini coefficient for equivalized household income, as calculated by the Census Bureau from the 1990 Census (US Census Bureau, 2000).

The pooled CPS sample includes 188,785 over-18 respondents, of which 1,015 reported zero or negative household income. In order to use log household income as an explanatory variable, these cases are dropped yielding 187,760 respondents in the sample. Table 4 reports summary statistics. All estimates using the CPS data are unweighted. Weighted results are similar.

4.2.3 OLS and related results

Table 5 shows regression results based on the standard assumption that inequality is conditionally exogenous. The first set of estimates are for a linear model, and are estimated using OLS. Standard errors are robust to clustering by state. The second set of estimates are for a logistic model with a state-level random effect, and are estimated by maximizing the restricted penalized quasi-likelihood.

In general, Table 5 shows a statistically significant association between measured state-level inequality and the probability of self-rated fair/poor health. To put the coefficient magnitudes in perspective, the linear regression with specification (2) implies that a one-standard-deviation increase in inequality is associated with an increase in the probability of fair/poor health of 0.6 percentage points. This is roughly the same predictive effect as a 20% increase in one's own income.

Variable	Unweighted mean (std. dev.)
Individual-level characteristics:	
Self-reported fair or poor health	0.15
Log equivalized household income	10.03
	(0.88)
Age, years	44.9
	(17.49)
Female	0.53
Black	0.09
Asian/other	0.05
Education, years	12.73
	(2.71)
Not employed	0.36
No health insurance	0.21
State-level characteristics:	
Income inequality (Gini coefficient)	0.43
	(0.02)
# of individuals	187,760
# of states (including DC)	51

Table 4: Summary statistics, linked CPS-Census data.

The logistic model estimates in Table 5 can be compared to those seen in previous research using this data source. The logistic coefficient estimate of 4.608 corresponds to an odds ratio of 1.26 associated with an increase in the state-level Gini coefficient of 0.05. This is similar in magnitude to the odds ratios of 1.31 to 1.39 reported by Subramanian and Kawachi (2003) also using CPS data. The corresponding odds ratio for the linear model would vary across individuals. For a representative individual whose characteristics imply a probability of self-rated fair/poor health of 15% (the average in the data), the odds ratio would be 1.12.

4.2.4 RCR results

Table 6 reports the results from estimating the effect of inequality on health under a series of relative correlation restrictions. As the table shows, increases in λ from the benchmark case of exogeneity are generally associated with decreases in the estimated marginal effect of inequality. A relative correlation of 23% or greater (i.e., $\lambda > 0.23$) implies that the range of point estimates for θ consistent with the data includes zero. That is, in order to interpret this data as demonstrating a positive causal relationship between inequality and poor health, one

Explanatory	Linear		Logistic	
Variable	(1)	(2)	(1)	(2)
Income inequality (Gini coef.)	0.903	0.299	8.564	4.608
	(0.160)	(0.124)	(1.226)	(1.173)
Log equivalized household income		-0.031		-0.254
		(0.001)		(0.009)
Age, years		0.005		0.036
		(<0.001)		(<0.001)
Female		-0.007		-0.082
		(0.001)		(0.015)
Black		0.050		0.437
		(0.007)		(0.024)
Asian/other		0.010		0.174
		(0.006)		(0.038)
Education, years		-0.013		-0.093
		(0.001)		(0.003)
Not employed		0.129		1.089
		(0.003)		(0.017)
No health insurance		0.066		0.529
		(0.005)		(0.018)
# of observations	187,760	187,760	187,760	187,760

Table 5: Estimated effect of inequality on self-reported fair or poor health under assumption of exogeneity. Linear model estimated using OLS, with standard errors robust to clustering by state. Logistic model estimated as random-intercept multilevel model with maximum likelihood.

would need to claim that the correlation between inequality and unobserved factors affecting health is no greater than 23% as large as the correlation between inequality and the observed factors that affect health.

Relative correlation	Bounds on effect of income inequality	
restriction (Λ)	$[\hat{\theta}_L(\Lambda), \hat{\theta}_H(\Lambda)]$	95% CI
$\{0.00\}$	0.30	$(0.06, 0.54)$
$[0.00, 0.10]$	$[0.17, 0.30]$	$(-0.03, 0.51)$
$[0.00, 0.20]$	$[0.04, 0.30]$	$(-0.16, 0.50)$
$[0.00, 0.30]$	$[-0.10, 0.30]$	$(-0.29, 0.50)$
$[0.00, 0.50]$	$[-0.37, 0.30]$	$(-0.58, 0.50)$
$[0.00, 1.00]$	$[-1.09, 0.30]$	$(-1.37, 0.50)$
$[0.00, 3.00]$	$[-4.70, 0.30]$	$(-6.00, 0.50)$
$[0.00, 5.00]$	$[-11.83, 0.30]$	$(-20.69, 0.50)$
$[0.00, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
$(-\infty, 0.00]$	$[0.30, 17.04]$	$(0.06, \infty)$
Other parameter estimates:		
$\hat{\lambda}^*$	5.17	
$\hat{\theta}^*$	17.04	
$\hat{\lambda}(0)$	0.23	

Table 6: Bounds on the effect of income inequality on health. Bounds for the true effect are reported in square brackets, and 95% cluster-robust asymptotic confidence intervals are reported in parentheses.

5 Conclusion

The methodology developed in this paper provides a simple means of providing bounds on causal parameters under relative correlation restrictions. In the application using the experimental Project STAR data, 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 the observational CPS data, the bounds on the effect of income inequality on the prevalence of fair/poor health are much wider, and the lower bound is negative as long as the upper bound on the correlation between inequality and unobserved factors is at least 23% of the correlation between inequality and the observed control variables.

These two applications have been selected in part to represent two extremes. One would expect to find the Project STAR findings are more robust than the inequality-and-health findings, given the unavoidable differences in research design. The important thing to note here is that this finding of greater robustness comes entirely from the data. 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.

The methodology can be advanced in future research along two main fronts. First, the model is quite simple and might be usefully extended to accomodate common features like fixed effects or simple forms of nonlinearity. Second, the inference in the current paper is based on standard asymptotics. Because the underlying estimators are based on ratios/inverses, standard asymptotics can provide a poor approximation in finite sample when a relevant denominator is nearly zero. Alternative inference procedures may be more robust to this potential form of weak identification.

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools," *Journal of Political Economy*, 2005, *113*, 151–184.

Blakely, Tony A., Bruce P. Kennedy, Roberta Glass, and Ichiro Kawachi, "What is the lag time between income inequality and health status?," *Journal of Epidemiology and Community Health*, 2000, *54*, 318–319.

_ , Kimberly Lochner, and Ichiro Kawachi, "Metropolitan area income inequality and self-rated health: A multi-level study," *Social Science and Medicine*, 2002, *54*, 65–77.

Conley, Timothy G., Christian B. Hansen, and Peter E. Rossi, "Plausibly exogenous," *Review of Economics and Statistics*, 2010. Forthcoming, DOI: 10.1162/REST_a_00139.

Cornfield, J., W. Haenszel, E. Hammond, A. Lilienfield, M. Shimkin, and E. Wynder, "Smoking and lung cancer: Recent evidence and a discussion of some questions," *Journal of the National Cancer Institute*, 1959, *22*, 173–203.

Deaton, Angus S., "Health, inequality, and economic development," *Journal of Economic Literature*, 2003, *41*, 113–158.

Finn, Jeremy D., Jayne Boyd-Zaharias, Reva M. Fish, and Susan B. Gerber, "Project STAR and Beyond: Database User's Guide," Data set and documentation, HEROS, Inc. 2007. Retrieved from `http://www.heros-inc.org/data.htm`, 7/15/2007.

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

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

_ **and Charles F. Manski**, "Confidence intervals for partially identified parameters," *Econometrica*, 2004, *72*, 1845–1857.

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

Kraay, Aart, "Instrumental variables regressions with uncertain exclusion restrictions: A Bayesian approach," *Journal of Applied Econometrics*, 2010. Forthcoming, DOI: 10.1002/jae.1148.

Krauth, Brian V., "Peer effects and selection effects on youth smoking in California," *Journal of Business and Economic Statistics*, 2007, *25*, 288–298.

Krueger, Alan B., "Experimental estimates of education production functions," *Quarterly Journal of Economics*, 1999, *114*, 497–532.

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

Lewbel, Arthur, "Using Heteroskedasticity to Identify and Estimate Mismeasured and Endogenous Regressor Models," *Journal of Business and Economic Statistics*, 2011. Forthcoming.

Manski, Charles F., *Identification Problems in the Social Sciences*, Harvard University Press, 1994.

_ , *Partial Identification of Probability Distributions*, Springer-Verlag, 2003.

Mellor, Jennifer M. and Jeffrey Milyo, "Income inequality and health status in the United States: Evidence from the Current Population Survey," *Journal of Human Resources*, 2002, *37*, 510–539.

_ **and** _ , "Is exposure to income inequality a public health concern? Lagged effects of income inequality on individual and population health," *Health Services Research*, 2003, *38*, 137–151.

Nevo, Aviv and Adam M. Rosen, "Identification with imperfect instruments," *Review of Economics and Statistics*, 2010. Forthcoming, DOI: 10.1162/REST_a_00171.

Rosenbaum, Paul R., *Observational Studies, 2nd edition*, Springer, 2002.

Stoye, Jörg, "More on confidence intervals for partially identified parameters," *Econometrica*, 2009, *77*, 1299–1315.

Subramanian, S. V. and Ichiro Kawachi, "The association between state income inequality and worse health is not confounded by race," *International Journal of Epidemiology*, 2003, *32*, 1022–1028.

_ **and** _ , "Income inequality and health: What have we learned so far?," *Epidemiologic Reviews*, 2004, *26*, 78–91.

US Census Bureau, *Historical income tables for states: Table S4. Gini ratios by state: 1969, 1979, 1989.*, Washington, D.C.: Income Statistics Branch/Housing and Household Economic Statistics Division, 2000. Retrieved from `http://www.census.gov/hhes/income/histinc/state/statetoc.html`.

US Department of Labour, *Current Population Survey*, Washington, D.C.: Bureau of Labour Statistics, 1998. Retrieved from `http://www.bls.census.gov/cps/cpsmain.htm`.

Wilkinson, Richard G. and Kate E. Pickett, "Income inequality and health: A review and explanation of the evidence," *Social Science and Medicine*, 2006, *62*, 1768–1784.

_ **and** _ , *The spirit level : Why more equal societies almost always do better*, Allen Lane, 2009.

A Proofs of propositions

A.1 Proposition 1

Proof: To establish result 1, note that:

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

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

$$\begin{aligned}
\lambda(\theta; m) &= \frac{\left(\frac{cov_m(z,y)-\theta var_m(z)}{cov_m(z^p,y^p)-\theta var_m(z^p)} - 1\right)}{\sqrt{\frac{var_m(y)-2\theta cov_m(z,y)+\theta^2 var_m(z)}{var_m(y^p)-2\theta cov_m(z^p,y^p)+\theta^2 var_m(z^p)} - 1}} \\
&= \frac{\left(\frac{p_1}{p_2} - 1\right)}{\sqrt{\frac{p_3}{p_4} - 1}} \qquad (16)
\end{aligned}$$

where p_1, p_2, p_3, and p_4 are all polynomials (and thus differentiable) in θ. They are also differentiable in m. Application of the quotient and product rules implies that $\lambda(\theta; m)$ 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:

$$p_2 = cov_m(z^p, y^p) - \theta var_m(z^p) = 0$$

Since $var_m(z^p) > 0$ by equation (7), we can solve to get

$$\theta = \frac{cov_m(z^p, y^p)}{var_m(z^p)} = \theta^*(m)$$

Condition (b) fails if and only if:

$$p_4 = var_m(y^p - \theta z^p) = 0$$

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

$$\theta = \frac{cov_m(z^p, y^p)}{var_m(z^p)} = \theta^*(m)$$

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

$$var_m(y - \theta z) \leq var_m(y^p - \theta z^p)$$

Note that $y^p - \theta z^p$ is the best linear predictor of $y - \theta z$, so:

$$var_m\left(y - \theta z\right) = var_m\left(y^p - \theta z^p\right) + var_m\left(y - \theta z - (y^p - \theta z^p)\right)$$

This implies that $var_m\left(y - \theta z - (y^p - \theta z^p)\right) = 0$, which also implies that:

$$y - \theta z - (y^p - \theta z^p) = 0$$

Rearranging, we get:

$$y = y^p - \theta z^p + \theta z$$

which implies that y is an exact linear function of $(z, \mathbf{x})$ and equation (5) is violated. Therefore condition (c) must hold. Since conditions (a), (b), and (c) hold for all $\theta \neq \theta^*(m)$, $\lambda(\theta; m)$ is differentiable at all $\theta \neq \theta^*(m)$.

To establish result 2, note that $var_m(z)$ is strictly positive by (5) and $var_m(z^p)$ is strictly positive by (7). Therefore:

$$\lim_{\theta \to \infty} \left(cov_m(z, y) - \theta var_m(z)\right) = -\infty$$
$$\lim_{\theta \to \infty} \left(cov_m(z^p, y^p) - \theta var_m(z^p)\right) = -\infty$$

So by L'Hospital's rule:

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

By the same reasoning:

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

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

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

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

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

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

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

To prove result 3 we first show how the behavior of $\lambda(\theta;m)$ near $\theta^*(m)$ depends on some special cases:

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

$$E_m\left((y^p - a_m - b_m z^p)^2\right) = 0 \tag{17}$$

for some a_m and b_m. Then equation (10) is satisfied for all λ when $\theta = \theta^*(m) = b_m$.

Proof: To show that $\theta^*(m) = b_m$:

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

To show that equation (10) is satisfied at $\theta^*(m)$ for all λ, note that $\mathbf{x}\beta(\theta;m) = y^p - \theta z^p$.

This implies that:

$$\begin{aligned} var_m(\mathbf{x}\beta(\theta^*(m);m)) &= var_m\left(y^p - \theta^*(m)z^p\right) \\ &= var_m\left(y^p - b_m z^p\right) \\ &= var_m\left(y^p - b_m z^p\right) - 2cov_m\left(y^p - b_m z^p, a_m\right) + var_m(a_m) \\ &= var_m\left(y^p - a_m - b_m z^p\right) \\ &= 0 \end{aligned}$$

and by the same argument $cov_m(z, \mathbf{x}\beta(\theta^*(m);m)) = 0$. Equation (10) thus reduces to $0 = \lambda 0$, a condition that is satisfied by any λ.

Case B: Suppose that m implies:

$$\frac{cov_m(y,z)}{var_m(z)} = \frac{cov_m(y^p,z^p)}{var_m(z^p)} \tag{18}$$

Then equation (10) is satisfied for all λ when $\theta = \theta^*(m)$.

Proof: First, note that in this case:

$$\begin{aligned} &cov_m(z, y - \theta^*(m)z - \mathbf{x}\beta(\theta^*(m);m)) \\ &\qquad = cov_m(z, y - \theta^*(m)z - y^p + \theta^*(m)z^p) \\ &\qquad = cov_m(z,y) - \theta^*(m)var_m(z) - cov_m(y^p,z^p) + \theta^*(m)var_m(z^p) \\ &\qquad = cov_m(z,y) - \frac{cov_m(z,y)}{var_m(z)}var_m(z) - cov_m(y^p,z^p) + \frac{cov(z^p,y^p)}{var(z^p)}var(z^p) \\ &\qquad = 0 \end{aligned}$$

and:

$$\begin{aligned} cov_m(z, \mathbf{x}\beta(\theta^*(m);m)) &= cov_m(z, y^p - \theta^*(m)z^p) \\ &= cov_m(z^p,y^p) - \theta^*(m)var_m(z^p) \\ &= cov_m(z^p,y^p) - \frac{cov_m(z^p,y^p)}{var_m(z^p)}var_m(z^p) \\ &= 0 \end{aligned}$$

Equation (10) thus reduces to $0 = \lambda 0$, which is satisfied for all λ.

Case C: Suppose that neither (17) nor (18) hold. Then for any $\lambda \in (-\infty, \lambda^*(m)) \cup (\lambda^*(m), \infty)$ we can find a θ such that $\lambda(\theta; m) = \lambda$, i.e., that solves equation (10).

Proof: First, note that since $cov_m(z^p, y^p) - \theta^*(m) var_m(z^p) = 0$, the existence of a solution to equation (10) when $\theta = \theta^*(m)$ requires that either $var_m(y^p - \theta^*(m) z^p) = 0$, implying (17) holds, or $cov_m(z, y) - \theta^*(m) var_m(z) = 0$, implying (18) holds. Since neither holds, there is no λ that satisfies equation (10) for $\theta = \theta^*(m)$.

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

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

and

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

We have thus established that $\lim_{\theta \to -\infty} \lambda(\theta; m) = \lambda^*(m)$, that $\lim_{\theta \uparrow \theta^*} \lambda(\theta; m)$ is either $-\infty$ or ∞, and that $\lambda(\theta; m)$ is continuous on $(-\infty, \theta^*(m))$. Suppose for the moment that $\lim_{\theta \uparrow \theta^*(m)} \lambda(\theta; m) = -\infty$. By the intermediate value theorem, for any $\lambda \in (-\infty, \lambda^*(m))$, there exists some $\theta \in (-\infty, \theta^*(m))$ such that $\lambda(\theta; m) = \lambda$. This is a sufficient condition for θ to solve equation (10). Since $\lim_{\theta \uparrow \theta^*(m)} = -\infty$, then $\lim_{\theta \downarrow \theta^*(m)} \lambda(\theta; m) = \infty$. Again, since $\lambda(\theta; m)$ is continuous on $(\theta^*(m), \infty)$, the intermediate value theorem implies that for any $\lambda \in (\lambda^*(m), \infty)$ there exists some $\theta \in (\theta^*(m), \infty)$ such that $\lambda(\theta; m) = \lambda$. Therefore, for any $\lambda \in (-\infty, \lambda^*(m)) \cup (\lambda^*(m), \infty)$ we can find a θ such that $\lambda(\theta; m) = \lambda$, i.e., that solves equation (10). The same argument can be duplicated for the case $\lim_{\theta \uparrow \theta^*(m)} \lambda(\theta) = \infty$. Note that there may or may not be a θ such that $\lambda(\theta; m) = \lambda^*(m)$.

To prove result 4, pick any θ and consider two cases. First, suppose that $\theta = \theta^*(m)$. Then $\theta \notin \tilde{\Theta}_0(\Lambda; m)$ since $\lambda(\theta; m)$ does not exist. Next, suppose that $\theta \neq \theta^*(m)$. Then $\lambda(\theta; m)$ exists (by result 1 of this proposition) and provides the unique λ that solves equation (10) for that λ.

Therefore,

$$\theta \in \tilde{\Theta}_0(\Lambda; m) \text{ if and only if } \theta \in \Theta_0(\Lambda; m) \text{ and } \theta \neq \theta^*(m)$$

which is another way of stating the result. □

A.2 Proposition 2

Proof: Since Λ is nonempty, $\lambda^*(m_0) \notin \Lambda$ implies that Λ must contain some $\lambda \neq \lambda^*(m_0)$. Result 3 of Proposition 1 says that there exists some θ such that (λ, θ) satisfy equation (10). Therefore the identified set is nonempty.

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

$$\begin{aligned}
|\theta| > B_\epsilon &\Rightarrow \lambda(\theta; m_0) \in (\lambda^*(m_0) - \epsilon, \lambda^*(m_0) + \epsilon) \qquad \text{(by result 2 of Proposition 1)} \\
&\Rightarrow \lambda(\theta; m_0) \notin \Lambda \qquad \text{(since } (\lambda^*(m_0) - \epsilon, \lambda^*(m_0) + \epsilon) \text{ is disjoint from } \Lambda) \\
&\Rightarrow \theta \notin \tilde{\Theta}_0(\Lambda, m_0) \qquad \text{(by definition of } \tilde{\Theta}_0) \\
&\Rightarrow \theta \notin \tilde{\Theta}_0(\Lambda, m_0) \cup \{\theta^*(m_0)\} \qquad \text{(since } B_\epsilon > \theta^*(m_0)) \\
&\Rightarrow \theta \notin \Theta_0(\Lambda, m_0) \qquad \text{(by result 4 of Proposition 1)}
\end{aligned}$$

Therefore, the identified set is bounded. □

A.3 Proposition 3

Proof: Both $\theta^*(m)$ and $\lambda^*(m)$ are continuous in m by the quotient rule, given that $var_m(z^p) > 0$. Result 1 of Proposition 1 says that $\lambda(\theta; m)$ is continuous in m for all $\theta \neq \theta^*(m)$). 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 $\theta_L(\Lambda; m)$ is continuously differentiable in m if $\frac{d\lambda(\theta;m)}{d\theta}|_{\theta=\theta_L(\Lambda;m)} \neq 0$. In that case, consistency of $\hat{\theta}_L(\Lambda)$ follows from Slutsky's theorem. The same argument applies to $\hat{\theta}_H(\Lambda)$.

For the third result, note that if $\Theta_0(\Lambda; m_0) = \mathbb{R}$, then result 2 of Proposition 1 implies $\lambda^*(m_0)$ is in the interior of Λ. Therefore, there exists an $\epsilon > 0$ and $B_1 < B$ such that $[\lambda(B_1; m_0) -$

$\epsilon, \lambda(B_1; m_0) + \epsilon] \subset \Lambda$. Since $\hat{\lambda}(B_1) \xrightarrow{p} \lambda(B_1; m_0)$, we have:

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

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

A.4 Proposition 4

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

$$A = \begin{bmatrix} \nabla_m \theta_L(\Lambda; m)|_{m=m_0} \\ \nabla_m \theta_H(\Lambda; m)|_{m=m_0} \end{bmatrix} \tag{19}$$

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

$$\begin{aligned} \nabla_m \theta_L(\Lambda; m) &= -\left. \frac{\nabla_m \lambda(\theta; m)}{\partial \lambda(\theta; m)/\partial \theta} \right|_{\theta=\theta_L(\Lambda;m)} \\ \nabla_m \theta_H(\Lambda; m) &= -\left. \frac{\nabla_m \lambda(\theta; m)}{\partial \lambda(\theta; m)/\partial \theta} \right|_{\theta=\theta_H(\Lambda;m)} \end{aligned} \tag{20}$$

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

A.5 Proposition 5

Proof: If $var(z^p) = 0$, then $cov(z, y^p - \theta z^p) = 0$ for all θ. This implies that (10) holds if and only if $cov(z, y - \theta z) = 0$, i.e., if $\theta = cov(z, y)/var(z)$. □

A.6 Proposition 6

Proof: First, we rewrite:

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

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

$$q_1(\theta;\hat{m}_n) \overset{p}{\to} cov(z,y) - \theta var(z)$$

while the denominator is

$$q_2(\theta;\hat{m}_n) \overset{p}{\to} 0$$

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

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

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

$$\theta_{OLS}(\hat{m}_n) \overset{p}{\to} \theta_{OLS}(m_0) = \frac{cov(z - z^p, y - y^p)}{var(z - z^p)} = \frac{cov(z,y)}{var(z)} = \theta_0 \tag{21}$$

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

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

Therefore:

$$\hat{\theta}_L(\Lambda) \leq \theta_{OLS}(\hat{m}_n) \leq \hat{\theta}_H(\Lambda) \qquad \text{WPA1} \tag{22}$$

Pick any $\epsilon > 0$. The event $(|\theta_{OLS}(\hat{m}_n) - \theta_0| < \epsilon)$ clearly implies $(\theta_{OLS}(\hat{m}_n) > \theta_0 - \epsilon)$, which itself implies $(\hat{\theta}_H(\Lambda) > \theta_0 - \epsilon)$ by equation (22). Therefore:

$$\Pr(|\theta_{OLS}(\hat{m}_n) - \theta_0| < \epsilon) \leq \Pr(\hat{\theta}_H(\Lambda) > \theta_0 - \epsilon) \leq 1$$

By (21), $\Pr(|\theta_{OLS}(\hat{m}_n) - \theta_0| < \epsilon) \to 1$, so by the sandwich theorem:

$$\Pr(\hat{\theta}_H(\Lambda) > \theta_0 - \epsilon) \to 1 \tag{23}$$

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

$$\begin{aligned} 0 &\leq \Pr(\hat{\theta}_H(\Lambda) \geq \theta_0 + \epsilon) \qquad (24) \\ &= \Pr(\lambda(\theta; \hat{m}_n) \in \Lambda \text{ for some } \theta > \theta_0 + \epsilon) \\ &\leq \Pr(|\lambda(\theta; \hat{m}_n)| \leq \lambda^{max} \text{ for some } \theta \geq \theta_0 + \epsilon) \end{aligned}$$

Now, for any $\delta \neq 0$

$$\begin{aligned} &q_1(\theta_0 + \delta; \hat{m}_n) \xrightarrow{p} cov(z, y) - (\theta_0 + \delta)var(z) = -\delta var(z) \neq 0 \\ &q_2(\theta_0 + \delta; \hat{m}_n) \xrightarrow{p} 0 \end{aligned}$$

Therefore,

$$\Pr(|\lambda(\theta; \hat{m}_n)| \leq \lambda^{max} \text{for some } \theta \geq \theta_0 + \epsilon) \to 0 \tag{25}$$

By the sandwich theorem (24) and (25) imply $\Pr(\hat{\theta}_H(\Lambda) \geq \theta_0 + \epsilon) \to 0$, or equivalently that:

$$\Pr(\hat{\theta}_H(\Lambda) < \theta_0 + \epsilon) \to 1 \tag{26}$$

Taking (23) and (26) together we get:

$$\Pr(|\hat{\theta}_H(\Lambda) - \theta_0| < \epsilon) \to 1 \tag{27}$$

which is the result stated in the proposition. The same argument applies to θ_L. □

B Monte Carlo results

This section reports the results from some simple Monte Carlo experiments. In each experiment, a sample of size $n = 1,000$ is generated from the model:

$$y = \theta_0 z + \beta_1 x_1 + \beta_2 x_2 + v \qquad \text{where } E(x_1 v) = E(x_2 v) = 0 \tag{28}$$

where $\text{corr}(z, \beta_1 x_1 + \beta_2 x_2) = \rho_{z,x\beta}$ and $\text{corr}(z, v) = \lambda_0 \rho_{z,x\beta}$. For convenience, (z, x_1, x_2, v) are jointly normal with mean zero and unit variance, $\text{corr}(x_1, x_2)$ is set to zero, and β is set so that $var(\beta_1 x_1 + \beta_2 x_2) = 1$, i.e., $\beta_1 = \beta_2 = \sqrt{0.5}$. Rather than assuming z is equally correlated with x_1 and x_2, we set $\text{corr}(z, x_2) = 0$ and $\text{corr}(z, x_1) = \rho_{z,x\beta}\sqrt{2}$, which implies $\text{corr}(z, \beta_1 x_1 + \beta_2 x_2) = \rho_{z,x\beta}$. Given the simulated data, the RCR model is then estimated for the relative correlation restriction $\Lambda = [0, \lambda_H]$ The parameters $(\theta_0, \lambda_0, \rho_{z,x\beta}, \lambda_H)$ are varied across experiments.

Table 7 shows the main results. The first six columns show the true values set for the model parameters $(\theta_0, \lambda_0, \rho_{z,x\beta}, \lambda_H)$ and the related quantities (θ^*, λ^*). The next four columns show the average values of the estimators $(\hat{\theta}^*, \hat{\lambda}^*, \hat{\theta}_L, \hat{\theta}_H)$. The final column shows the actual coverage rate of the Imbens-Manski confidence interval for θ_0 with a nominal coverage of 95%. As the table shows, the estimator performs well in this setting. In all of the 14 cases in which the assumed relative correlation restriction actually holds (i.e., when $\lambda_0 \in [0, \lambda_H]$) the average bounds contain or come very close to containing the true parameter value of zero and the coverage probabilities are close to the nominal coverage of 0.95.

As one would expect, the estimator performs less well when the assumed relative correlation restriction is misspecified (i.e., when $\lambda_H = 0.1$ and λ_0 is either 0.5 or 1.0) and this misspecification is quantitatively important (i.e., when $\rho_{z,x\beta}$ is not very small). The estimated bounds are generally biased upwards, and the coverage probablilties are low. Estimates of θ^* and λ^* are substantially biased (and highly variable) when $\rho_{z,\beta} \approx 0$, but are much more well-behaved when $\rho_{z,\beta}$ is larger. Note that this does not substantially affect estimates for the parameter of interest.

True values						Average values				CI coverage
θ_0	λ_0	$\rho_{z,x\beta}$	λ_H	θ^*	λ^*	$\hat{\theta}^*$	$\hat{\lambda}^*$	$\hat{\theta}_L$	$\hat{\theta}_H$	$\Pr(\theta_0 \in CI)$
0.0	any	0.0	0.1	undefined	∞	-0.13	39.17	-0.0015	0.0010	0.949
0.0	any	0.0	1.0	undefined	∞	-0.13	39.17	-0.0127	0.0124	0.959
0.0	0.0	0.001	0.1	500	707.11	0.62	39.64	-0.0015	0.0009	0.949
0.0	0.0	0.001	1.0	500	707.11	0.62	39.64	-0.0132	0.0119	0.959
0.0	0.5	0.001	0.1	500	707.11	0.62	39.64	*-0.0011*	*0.0014*	*0.949*
0.0	0.5	0.001	1.0	500	707.11	0.62	39.64	-0.0127	0.0124	0.959
0.0	1.0	0.001	0.1	500	707.11	0.62	39.64	*-0.0006*	*0.0006*	*0.948*
0.0	1.0	0.001	1.0	500	707.11	0.62	39.64	-0.0122	0.0129	0.959
0.0	0.0	0.1	0.1	5	7.00	4.99	7.19	-0.0104	-0.0002	0.949
0.0	0.0	0.1	1.0	5	7.00	4.99	7.19	-0.1040	-0.0002	0.947
0.0	0.5	0.1	0.1	5	7.00	4.99	7.19	*0.0406*	*0.0507*	*0.709*
0.0	0.5	0.1	1.0	5	7.00	4.99	7.19	-0.0523	0.0508	0.997
0.0	1.0	0.1	0.1	5	7.00	4.99	7.19	*0.0918*	*0.1018*	*0.151*
0.0	1.0	0.1	1.0	5	7.00	4.99	7.19	-0.0004	0.1018	0.954
0.0	0.0	0.2	0.1	2.5	3.39	2.50	3.41	-0.0220	-0.0002	0.948
0.0	0.0	0.2	1.0	2.5	3.39	2.50	3.41	-0.2334	-0.0002	0.948
0.0	0.5	0.2	0.1	2.5	3.39	2.50	3.41	*0.0873*	*0.1085*	*0.190*
0.0	0.5	0.2	1.0	2.5	3.39	2.50	3.41	-0.1186	0.1085	1.000
0.0	1.0	0.2	0.1	2.5	3.39	2.50	3.41	*0.1968*	*0.2172*	*0.000*
0.0	1.0	0.2	1.0	2.5	3.39	2.50	3.41	-0.0009	0.2172	0.949

Table 7: Monte Carlo results, 10,000 replications per experiment. Italics indicate bounds or confidence intervals based on invalid relative correlation restrictions (i.e. $\lambda_0 \notin [0, \lambda_H]$).