Bounding a linear causal effect using relative correlation restrictions*

Brian Krauth
Department of Economics
Simon Fraser University

October 2009

Abstract

This paper describes and implements a simple and relatively conservative 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 use the sample correlation between the variable of interest and observed control variables to suggest a range of reasonable values for the correlation between the variable of interest and relevant unobserved variables. It is then possible to construct a range of parameter estimates consistent with that range of correlation values. In addition to establishing the estimation method and its properties, the paper demonstrates two applications. The first uses data from the Project STAR class size experiment, and demonstrates application to experiments with imperfect randomization. In this application, I find that the correlation between treatment and unobserved outcome-relevant factors would need to be 300% to 1000% as large as the correlation between treatment and observed outcome-relevant factors in order to eliminate or reverse the sign of the effect from that estimated by OLS. The second application uses CPS to study the relationship between state-level income inequality and self-reported health, and demonstrates application of the method to observational data when there is uncorrectible endogeneity of the variable of interest. In this application, I find that the estimated effect reverses sign if the correlation between inequality and health-relevant unobservables is at least 23% as large as the correlation between inequality and health-relevant observables.

*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`, web `http://www.sfu.ca/~bkrauth/`

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 would be valuable 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 being estimated by OLS regression of the outcome of interest on the explanatory variable of interest and a set of control variables. Deviations from the exogeneity assumption necessary for identification of the causal effect are modeled in terms of a single parameter that measures the correlation between the explanatory variable of interest and the unobservables, relative to the correlation between that variable and the control variables. The traditional exogeneity assumption imposes a value of zero for this sensitivity parameter. However, the model can be estimated under alternative values, with a range of plausible values for the sensitivity parameter being associated with a range of values for the parameter of interest. While traditional point estimates are not generally possible, these ranges can be used to construct hypothesis tests and confidence intervals that have the usual interpretation.

Two example applications show the potential informativeness 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. The experiment in question considers 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 (i.e., non-experimental) study in which the claim of conditional exogeneity is controversial, but standard techniques for solving this problem are equally unappealing. The observational study in question (based on Subramanian and Kawachi 2003) aims to measure the effect of income inequality on individual health.

Applying the methods developed in this paper, I find that the Project STAR results are quite robust. For example, the positive effect of smaller classes on kindergarten test scores found by Krueger 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. The inequality-and-health results are much less robust. Specifically, 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.

The paper is organized as follows. Section 2 describes the model and derives the estimator and its statistical properties. Section 3 analyzes Project STAR, and Section 4 analyzes the inequality and health data.

Section 5 concludes and notes avenues for further research.

1.1 Related literature

The methodology described in this paper is closely related to the literature on sensitivity analysis in statistics. The seminal contribution is Cornfield et al. (1959), while Rosenbaum (2002) provides a relatively recent survey by a major contributor. Work in this area has often focused on the estimation of binary treatment effects under some particular hypothesized omitted variable (e.g., smoking as an omitted variable in the measurement of the effect of occupational toxin exposure on lung cancer).

Another closely related area is the econometric literature on estimation and inference for parameters that are set identified rather than point identified. Manski (1994) argues that the assumptions needed to point identify parameters of interest from standard data are often not credible, but that the data can provide some restrictions on parameter values under weaker and more justifiable assumptions. The theoretical literature on set estimation has advanced substantially in recent years. Manski (2003) provides an overview and develops a framework for understanding set identification. Imbens and Manski (2004), Stoye (2009), and several others have developed tools for inference and the construction of confidence intervals.

The particular type of sensitivity parameter used in this paper is similar in spirit to those seen in a number of recent papers. Altonji, Elder, and Taber (2005) perform a sensitivity analysis of standard measures of the Catholic school effect by incorporating a selection model in which the degree of selection on unobservables is proportional to the degree of selection on observables. Krauth (2007) estimates peer effects in youth smoking by modeling the within-group correlation in unobservables as a proportion of the within-group correlation in observables. 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. Finally, Conley, Hansen and Rossi (2008) develop a systematic means of sensitivity analysis in instrumental variables regression that allow the conventional IV exclusion restriction to be "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

The model is similar to a standard linear regression model, with a scalar outcome of interest y, a scalar explanatory variable of interest z, and a vector of additional control variables X.

The structural/causal model of interest is linear in z:

$$y = \theta_0 z + u \tag{1}$$

where the parameter of interest θ_0 represents the actual effect of z on y, and the unobserved random variable u represents the effect of all other factors. The variables u and z are potentially correlated, implying that θ_0 cannot be estimated by the OLS regression of y on Z.

It is common practice to add a set of control variables to the regression, in the hope that doing so will produce a coefficient on z that is closer to the true value of θ_0. Let X be this k-vector of control variables (including an intercept) and let $X\beta_0$ be the best linear predictor of u given X, i.e.:

$$\beta_0 \in \arg\min_b E\left((u - Xb)^2\right) \tag{2}$$

We can then rewrite the structural model as:

$$y = \theta_0 z + X\beta_0 + v \tag{3}$$

where by construction:

$$E(X'v) = 0 \tag{4}$$

Note that β_0 has no particular structural or causal interpretation.

In order for the OLS regression of y on (z, X) to provide a consistent estimate of θ_0, it is necessary to add the assumption that $\text{corr}(z, v) = 0$. This assumption appears so often in applied research that it is easy to forget what it implies. Although it is clear that the variable of interest is probably correlated with many other outcome-relevant factors (i.e., $\text{corr}(z, u) \neq 0$) - or else including the control variables would be unnecessary - the researcher asserts that *all* such factors are in the set of control variables. This is a strong and often difficult-to-justify assumption, but there are many applications where the alternatives are even less appealing. There are also many applications, for example experiments with imperfect randomization, where this assumption may be close enough to the truth.

The goal of this paper is to provide a simple and intuitive means of relaxing this standard assumption. Instead of imposing $\text{corr}(z, v) = 0$, we define a sensitivity parameter λ_0 such that

$$\lambda_0 = \frac{\text{corr}(z, v)}{\text{corr}(z, X\beta_0)}$$

and suppose that $\lambda_0 \in \Lambda$ for some nonempty closed interval Λ. The expression above is not well-defined if $var(v)$ or $cov(z, X\beta_0)$ is exactly zero, so the exact condition is:

$$\exists \lambda_0 \in \Lambda : cov(z, v)\sqrt{var(X\beta_0)} = \lambda_0 cov(z, X\beta_0)\sqrt{var(v)} \tag{5}$$

That is, λ_0 is the correlation between z and unobservables (v) relative to the correlation between z and observable control variables ($X\beta_0$).

Finally, we impose a few conditions on the data:

$$var(z) > 0 \tag{6}$$

$$L(y|z, X) \neq y \tag{7}$$

$$var(L(y|X)) > 0 \tag{8}$$

Condition (6) holds if there is any variation in z in the data, while condition (7) holds as long as y is not an exact linear function of (z, X). Condition (8) can be tested by an ordinary coefficient significance test.

2.2 Discussion

The λ parameter in this model can be interpreted as a sensitivity parameter for the standard OLS analysis. That is, it has four features:

1. If the sensitivity parameter's value were known, the parameter of interest could be estimated from the available data.
2. There is no way of estimating the sensitivity parameter from the available data.
3. Some conventional technique for estimating the parameter of interest (in this case, OLS regression with a set of control variables) can be interpreted as imposing a particular value for the sensitivity parameter.
4. The sensitivity parameter is in units such that we can think of plausible or implausible deviations from the value assumed in the conventional technique.

Given these features, we can use the model in this paper to provide a simple means of quantifiying the sensitivity of OLS regression findings to the assumption of conditional exogeneity. Features 1-3 above are demonstrated in Section 2.3. Feature 4 is a more subjective claim, and requires some additional discussion before proceeding.

In general, the sensitivity parameter λ represents the correlation between the variable of interest (z) and unobserved variables (v), relative to the correlation between the variable of interest and the control variables ($X\beta$). For example the statement $\lambda = 0.25$ would mean that z's correlation with the unobserved variable u is of the same sign as its correlation with the control variables $X\beta$, and 25% as large. Note that the $X\beta$ in equation (5) is an index constructed from the control variables rather than the control variables themselves, with more relevant (in the sense of having predictive power) variables receiving more weight. In addition, elements of β may be negative if a particular variable in X is negatively associated with the outcome. These two properties imply that estimation of θ is invariant to arbitrary linear transformations of the control variables. The v term in equation (5) is also a scalar representation of the total contribution[1] of multiple unobserved variables to the outcome being modeled.

In principle λ can be any[2] value, positive or negative. However, the usefulness of the methodology described here depends on the idea that some values for λ are more plausible than others.

- The usual assumption needed for consistency of OLS ($\text{corr}(z, v) = 0$) corresponds to the case $\lambda = 0$. That is, there is no omitted variables bias.

[1] Note that this is different from the most common methods of sensitivity analysis in the statistics literature. There, it is more common to identify and model a particular candidate omitted variable. However, econometricians more often confront a situation in which (due to simultaneous determination of z and y) there are a multitude of potential omitted variables.

[2] Because λ is finite, equation (5) does impose an important restriction on the data generating process: if if it turns out that z is uncorrelated with the observed variables ($\text{corr}(z, X\beta) = 0$) then the model implies that it is also uncorrelated with the unobserved variables ($\text{corr}(z, v) = 0$).

- If $\lambda > 0$, then $\text{corr}(z, v)$ is of the same sign as $\text{corr}(z, X\beta)$. That is, controlling for X removes some but not all bias.
- If $\lambda < 0$, then $\text{corr}(z, v)$ is of the oppposite sign as $\text{corr}(z, X\beta)$. That is, controlling for X "overcorrects" the bias from the bivariate regression, so that the true value of θ will lie somewhere between the results obtained in the regressions with and without controls.
- If $\lambda = 1$, then $\text{corr}(z, v)$ is of both the same sign and magnitude as $\text{corr}(z, X\beta)$. That is, the omitted variables bias is of the same sign and magnitude (relative to $var(v)/var(X\beta)$) as the bias reduction achieved by controlling for X.

In other words, λ can be interpreted 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 $\lambda = 1$ is what one would expect on average if the control variables were chosen randomly from a large set of plausible explanatory variables.

An alternative approach comes from another direction, and has been applied elsewhere in the sensitivity analysis literature. Given the OLS results, we could determine the smallest value of λ that overturns those results in some meaningful way. For example, if the OLS estimate of θ is positive, we might determine the smallest value of λ at which the estimated θ is negative, or the smallest value of λ at which it is statistically insignificant.

2.3 Identification

Given the model defined in Section 2.1, what can we identify about θ from the joint distribution of (y, z, X)? Let:

$$M \equiv \left[\begin{array}{cccc} 1 & y & z & X \end{array} \right] \tag{9}$$

and let

$$m \equiv \text{vech}(E(M'M)) \tag{10}$$

where vech(.) is the half-vectorization function (i.e., given a symmetric matrix it returns a column vector of its unique elements). As shown below, the vector of second moments m is the only feature of the distribution that is useful in estimating the model parameters.

The model defined in Section 2.1 implies k linear moment conditions:

$$E_m(X'(y - \theta_0 z - X\beta_0)) = 0 \tag{11}$$

and one nonlinear moment condition

$$\text{corr}_m(z, (y - \theta_0 z - X\beta_0)) - \lambda_0 \text{corr}_m(z, X\beta_0) = 0 \tag{12}$$

where E_m and corr_m have the subscript m to indicate that they can be recovered directly from the vector m.

The k equations in (11) can be solved for β given θ:

$$\beta(\theta; m) = E_m(X'X)^{-1} E_m(X'y) - \theta E_m(X'X)^{-1} E_m(X'z) \tag{13}$$

Let:

$$\begin{aligned}\theta_L(\Lambda;m) &\equiv \inf\{\theta : (\exists\lambda\in\Lambda : \text{corr}_m(z, y-\theta z - X\beta(\theta;m)) = \lambda \text{corr}_m(z, X\beta(\theta;m))\}\\ \theta_H(\Lambda;m) &\equiv \sup\{\theta : (\exists\lambda\in\Lambda : \text{corr}_m(z, y-\theta z - X\beta(\theta;m)) = \lambda \text{corr}_m(z, X\beta(\theta;m))\}\end{aligned}$$

The interval $[\theta_L(\Lambda;m), \theta_H(\Lambda;m)]$ is called the identified set for θ. By construction, the identified set is the smallest set that always contains θ_0 if $\lambda_0 \in \Lambda$. However, the identified set is potentially unbounded (i.e., the data provides no information on θ) or empty (i.e., the data reject the model). Proposition 2 provide conditions under which the identified set is bounded and nonempty.

First, however, it is useful to define a function giving the unique value of λ that is consistent with a given value of θ. Ignoring for the moment the possibility of dividing by zero, we can plug (13) into (12) and solve for λ to get:

$$\lambda(\theta;m) \equiv \frac{\text{corr}_m(z, y-\theta z - X\beta(\theta;m))}{\text{corr}_m(z, X\beta(\theta;m))} \tag{14}$$

Proposition 1 below outlines some relevant characteristics of this function.

Proposition 1 (Properties of $\lambda(.)$) *Suppose that (3)-(8) hold, and that $var_m(L_m(z|X)) \neq 0$. Then $\lambda(.;m)$ has the following properties:*

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

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

2. *If there is an exact linear relationship between $L_m(y|X)$ and $L_m(z|X)$, i.e.,:*

$$\exists\tilde{\theta} : L_m(y|X) - \tilde{\theta}L(z|X)) \text{ is constant} \tag{15}$$

then $\theta^*(m) = \tilde{\theta}$, equation (5) is satisfied for all λ when $\theta = \tilde{\theta}$, and:

$$\begin{aligned}\lim_{\theta\uparrow\theta^*(m)} \lambda(\theta;m) &= \frac{cov(z,y) - \tilde{\theta}var(z)}{\sqrt{var(L_m(z|X))var(y-\tilde{\theta}z)}}\\ \lim_{\theta\uparrow\theta^*(m)} \lambda(\theta;m) &= -\frac{cov(z,y) - \tilde{\theta}var(z)}{\sqrt{var(L_m(z|X))var(y-\tilde{\theta}z)}}\end{aligned}$$

3. *If:*

$$\frac{cov_m(y,z)}{var_m(z)} = \frac{cov_m(y, L_m(z|X))}{var_m(L_m(z|X))} \tag{16}$$

then equation (5) is satisfied for all λ when $\theta = \theta^*$, and:

$$\lim_{\theta\to\theta^*(m)} \lambda(\theta;m) = \frac{\frac{var_m(z)}{var_m(L_m(z|X))} - 1}{\sqrt{\frac{var_m(y) - cov_m^2(z,y)/var_m(z)}{var_m(L_m(y|X)) - cov_m^2(L_m(z|X), L_m(y|X))/var_m(L_m(z|X))}} - 1}$$

4. *If neither (15) nor (16) hold, then there is no λ that satisfies equation (5) for $\theta = \theta^*$, and*

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

and

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

5. *Let:*

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

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

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

6. *For any* $\lambda \neq \lambda^*$ *there exists a* θ *such that* (λ, θ) *satisfy equation (5).*

Proposition 2 is the primary identification result in this paper. In a set-identification setting, the relevant identification question is not existence of the identified set but rather its size. In other words, if the identified set is $(-\infty, \infty)$, then data on (y, z, X) provides no information about the value of θ. If the identified set is bounded, then it is possible to learn something about θ from the data.

Proposition 2 (Size of the identified set) *Suppose that (3)-(8) hold. Then if* $var(L(z|X)) > 0$*:*

1. *The identified set* $[\theta_L(\Lambda; m), \theta_H(\Lambda; m)]$ *is empty only if* $\Lambda = \{\lambda^*(m)\}$.
2. *The identified set is bounded if and only if* $\lambda^*(m) \notin \Lambda$.

If $var(L(z|X)) = 0$*, the identified set is nonempty and bounded for any* Λ.

As Proposition 2 indicates, the value of $\lambda^*(m)$ can be thought of an upper bound on Λ above which identification breaks down.

2.4 Estimation

The model can be estimated by a conceptually straightforward plug-in method derived from the identification results in Section 2.3. Let $\hat{m}_n$ be a consistent estimator of m from a sample of size n:

$$\hat{m}_n \xrightarrow{p} m \tag{17}$$

Since m is just a vector of second moments, it can be consistently estimated from a random sample on (y, z, X). However the approach in this paper can be applied to any sample and estimation scheme that implies (17).

Replacing all of the quantities of interest in Section 2.3 with their sample analogs, we have:

$$\begin{aligned} \hat{\lambda}(\theta) &\equiv \lambda(\theta; \hat{m}) \\ \hat{\lambda}^* &\equiv \lambda^*(\hat{m}) \\ \hat{\theta}^* &\equiv \theta^*(\hat{m}) \\ \hat{\theta}_L(\Lambda) &\equiv \inf\{\theta : \lambda(\theta; \hat{m}) \in \Lambda\} \\ \hat{\theta}_H(\Lambda) &\equiv \sup\{\theta : \lambda(\theta; \hat{m}) \in \Lambda\} \end{aligned} \tag{18}$$

Consistency follows from the Slutsky theorem provided that the relevant quantity is continuous in m at the true parameter values.

Proposition 3 (Consistency) *Suppose that (3)-(8) and (17) hold. Then the estimators defined in (18) are consistent in the sense that:*

$$\begin{aligned} \hat{\theta}^* &\xrightarrow{p} \theta^*(m) && \text{if } \theta^*(m) \text{ exists} \\ \hat{\lambda}^* &\xrightarrow{p} \lambda^*(m) && \text{if } \lambda^*(m) \text{ exists} \\ \hat{\lambda}(\theta) &\xrightarrow{p} \lambda(\theta; m) && \forall \theta \neq \theta^*(m) \end{aligned}$$

If $\theta_L(\Lambda; m)$ is finite, then:

$$\hat{\theta}_L(\Lambda) \xrightarrow{p} \theta_L(\Lambda; m) \qquad \text{if } \tfrac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_L(\Lambda)} \neq 0$$

otherwise:

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

If $\theta_H(\Lambda; m)$ is finite, then:

$$\hat{\theta}_H(\Lambda) \xrightarrow{p} \theta_H(\Lambda; m) \qquad \text{if } \tfrac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_H(\Lambda)} \neq 0$$

otherwise:

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

2.5 Inference

While the estimation of set-identified parameters looks very different from traditional estimation methods, inference does not. That is, hypothesis tests and confidence intervals can be constructed that look very much like ordinary tests and confidence intervals.

We start by supposing that $\hat{m}$ is asymptotically normal:

$$\sqrt{n}\,(\hat{m} - m) \xrightarrow{D} N(0, \Sigma) \tag{19}$$

Again, this property would hold for the ordinary sample average from a random sample, as well as for many other estimation and sampling schemes.

Again, the quantities we are estimating are in most cases differentiable functions of m, so their asymptotic distribution can be obtained through straightforward application of the delta method. Proposition 4 below states this more explicitly for the identified set.

Proposition 4 (Asymptotic distribution of the identified set) *Suppose that (3)-(8) and (19) hold. In addition, suppose that $\lambda^* \notin \Lambda$, that $\nabla_m \lambda(\theta, m)|_{\theta=\theta_L(\Lambda)}$ and $\nabla_m \lambda(\theta, m)|_{\theta=\theta_H(\Lambda)}$ exist, $\frac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_L(\Lambda)} \neq 0$, and $\frac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_H(\Lambda)} \neq 0$. Then:*

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

where

$$A = - \begin{bmatrix} \frac{\nabla_m \lambda(\theta,m)}{\partial \lambda(\theta;m)/\partial \theta}|_{\theta=\theta_L, m=\hat{m}} \\ \frac{\nabla_m \lambda(\theta,m)}{\partial \lambda(\theta;m)/\partial \theta}|_{\theta=\theta_H, m=\hat{m}} \end{bmatrix}$$

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

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

$$\lim_{n\to\infty} \Pr([\theta_L, \theta_H] \in CI) = 1 - \alpha$$

and a confidence interval for the parameter of interest:

$$\lim_{n\to\infty} \inf_{\theta\in[\theta_L,\theta_H]} \Pr(\theta \in CI) = 1 - \alpha$$

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 one method of constructing such a confidence interval by reducing the critical values to account for the width of the identified set. Stoye (2009) proposes a modification to the Imbens-Manski confidence interval that has better properties as the width of the identified set converges to zero.

3 Application #1: Experiments with incomplete randomization

Next, we consider two applications. The applications have been chosen to illustrate the two primary uses of the methodology: analysis of experiments with incomplete randomization, and somewhat more conservative than usual analysis of observational data with potential endogeneity that cannot be corrected through use of instrumental variables or fixed effects. They have also been chosen with an eye towards applied questions that have been extensively researched with well-known data sources.

3.1 Background: Project STAR and the effect of smaller classes on student achievement

The effect of class size on student achievement has been extensively studied in the economics of education literature. Class size reductions are a commonly proposed and implemented policy aimed at improving student outcomes, and are one of the most costly. Despite this, a number of researchers (Hanushek 1986, for example) have found that class size does not have an important effect on student outcomes. However, many of these studies are based on observational data and are thus plagued by endogeneity issues. Project STAR (Student/Teacher Achievement Ratio) is a well-known experimental study implemented in Tennessee in the late 1980's, aiming to measure the effect of class size on academic outcomes.

The design of Project STAR is as follows. A total of 79 schools were selected by the researchers for participation, based on willingness to participate and various criteria for the suitability of the school for the study. Within each school, students entering kindergarten in 1985 were randomly assigned to one of three experimental groups: the small class (S) group,

the regular class (R) group, and 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 experiment continued through grade 3, with students in group S kept in small classes through grade 3, etc. Students were given achievement tests in each year of the experiment, and have been subject to several follow-up data collections through their high school years.

The Project STAR research team has published numerous papers in education journals over the years describing their findings that small classes are associated with better outcomes along several dimensions. These findings received more attention among economists beginning with the work of Krueger (1999). The primary contribution of that article over previous work in the education literature is the extensive investigation of the consequences of difficult-to-avoid deviations from the experimental design. In particular:

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 quite common in the analysis of data from field experiments, and follows two steps. First, he investigates whether the deviations from randomization produce statistically significant differences in observed background variables between the experimental groups. Krueger finds that there are not large differences, though they are occasionally statistically significant. Second, instead of simply comparing means across treatment and control groups (with adjustment for school-level fixed effects), he also estimates regression models that include these observed background variables as controls. Krueger finds that including these variables does not substantially change the estimated treatment effect.

This two-step procedure is common enough in the econometric analysis of experiments that it bears some exploration. First, note that one of the two steps is redundant. If there are no differences in the distribution of observed characteristics between the treatment and control groups, controlling for those characteristics necessarily has no effect on the estimated treatment effect (save for any new bias introduced by misspecification of functional form). If there are differences in the distribution of observed characteristics, then this is easily addressed by simply controlling for them in a standard regression framework. Deviations from random assignment create problems identifying the treatment effect when they lead to differences in the distribution of relevant unobserved characteristics, and not when they lead to differences in observed characteristics. Second, note

that the supposed null hypothesis - a perfectly implemented experimental assignment - is surely false in this case, as the project team has records of specific deviations from the experimental protocol. Failure to reject this null is in some sense simply a matter of insufficient sample size.

So why is this procedure followed? One possible explanaton is that the researchers are implicitly following a model in which the distribution of observed characteristics between treatment and control groups provides information on the distribution of relevant unobserved characteristics between the groups. Specifically, if the difference in observed characteristics is shown to be small, then the researcher is safe assuming the difference in unobserved characteristics is also small enough to be ignored. In the first part of the procedure, in which individual characteristics are compared one-by-one across the groups, each characteristic is essentially given equal importance. In the second part of the procedure, in which the characteristics are used as control variables in a linear regression, characteristics are given importance based on their association with the outcome.

This implicit and informal argument is at least somewhat plausible. However, as applied in practice it has some weaknesses that the current paper addresses. First, the argument is made explicit rather than implicit, and so can be discussed in context. Second, the conventional procedure is too binary: if one can show that the assignment looks mostly random on the basis of observables, one can credibly assume randomness of unobservables report the point estimate from OLS as a consistent estimate of the true effect. However, there are experiments in which observables are somewhat associated with the treatment, and researchers are faced in this situation with the option of either assuming randomness of unobservables or giving up on measuring the treatment effect. The alternative suggested in the current paper is to parameterize the amount of selection on unobservables relative to the measured selection on observables.

3.2 Data and methodology

The analysis in this paper is based on the longitudinal records from kindergarten through high school of the 11,601 students that participated in the experiment (Finn, Boyd-Zaharias, Fish and Gerber 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. Here I followed 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.

For his benchmark regression results, Krueger estimates a regression with school-level fixed effects:

$$y = \theta z + X\beta + S + v \tag{20}$$

where y is the test score outcome, z is an indicator of the class-size treatment, X is a vector of covariates, and S is an unobserved school-level

	Grade			
Variable	K	1	2	3
Class size	20.3	21.0	21.1	21.3
	(4.0)	(4.0)	(4.1)	(4.4)
Percentile score avg. SAT	51.4	51.8	51.3	51.3
	(26.6)	(26.9)	(26.5)	(27.0)
Free lunch	0.48	0.52	0.51	0.51
White	0.67	0.67	0.65	0.66
Girl	0.49	0.48	0.48	0.48
Age on September 1st	5.43	6.57	7.66	8.70
	(0.35)	(0.49)	(0.56)	(0.59)
Exited sample	0.29	0.26	0.21	
% of teachers with MA+ degree	0.35	0.35	0.37	0.44
% of teachers who are White	0.84	0.83	0.80	0.79
% of teachers who are male	0.00	0.00	0.01	0.03
# schools	79	76	75	75
# students	6325	6829	6840	6802
# small classes	127	124	133	140
# regular classes	99	115	100	90
# reg./aide classes	99	100	107	107

Table 1: Summary statistics, Project STAR data.

fixed effect. The school-level fixed effect is necessary in this case because students were randomly assigned within schools, but assignment probabilities differed across schools. The school effects can be incorporated into our framework by applying the standard within transformation and making a small modification to our assumption. First, subtract school-level averages from both sides of the equation:

$$y - \bar{y}_s = \theta(z - \bar{z}_s) + (X - \bar{X}_s)\beta + (v - \bar{v}_s) \tag{21}$$

Then assume that:

$$\text{corr}\left(\tilde{z}, \tilde{v}\right) = \lambda \text{corr}\left(\tilde{z}, \tilde{X}\beta\right) \tag{22}$$

where $\tilde{z} \equiv (z - \bar{z}_s)$, $\tilde{v} \equiv (v - \bar{v}_s)$, and $\tilde{X} \equiv (X - \bar{X}_s)$. We can then apply the methods described in Section 2.

3.3 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 but is otherwise identical to specification (1). This is done because the approach described in this paper is designed to evaluate a single policy variable. As

the results show, the regular-aide treatment is nearly irrelevant to student outcomes, and so can be omitted as an explanatory variable. This result corresponds to the findings of both Krueger and the original Project STAR research team. The results in Table 2 suggest that the small-class treatment increases test scores by about 5-7 percentile points. Note that the gap between the small and regular class groups does not generally increase over years.

	Grade							
Explanatory	K		1		2		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
	(1.19)	(1.03)	(1.16)	(1.05)	(1.22)	(1.05)	(1.21)	(1.04)
Regular/aide class	0.26		1.77		1.54		-0.51	
	(1.06)		(0.97)		(1.06)		(1.09)	
White/Asian	8.39	8.39	6.94	6.98	6.45	6.48	6.05	6.05
	(1.35)	(1.35)	(1.18)	(1.19)	(1.18)	(1.18)	(1.43)	(1.43)
Girl	4.38	4.38	3.83	3.82	3.42	3.41	4.19	4.20
	(0.62)	(0.63)	(0.56)	(0.56)	(0.60)	(0.60)	(0.65)	(0.65)
Free lunch	-13.08	-13.08	-13.55	-13.55	-13.62	-13.64	-12.95	-12.94
	(0.77)	(0.77)	(0.87)	(0.87)	(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
	(2.15)	(2.17)	(1.94)	(1.95)	(1.74)	(1.74)	(1.79)	(1.78)
Teacher experience	0.26	0.27	0.06	0.07	0.10	0.11	0.05	0.05
	(0.10)	(0.10)	(0.06)	(0.06)	(0.06)	(0.06)	(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)	(1.07)	(1.05)	(1.04)	(1.17)	(1.17)
School fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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.

Next, we apply the methodology described in Section 2. The outcome variable y is the average percentile SAT score, the policy variable z is the small class treatment, and the set of control variables X are those teacher and student background variables included in specification (2) of Table 2. Table 3 reports the resulting interval estimate of the treatment effect θ for various choices of Λ. The point estimates of $\theta_L(\Lambda)$ and $\theta_H(\Lambda)$ are reported in square brackets, while the 95% asymptotic confidence intervals are reported in round brackets. Confidence intervals are calculated based on the method described in Stoye (2009), and are based on cluster-robust standard errors.

For the kindergarten data, Table 3 indicates that the estimated effect of small classes remains similar in magnitude even if the correlation between the treatment and unobservables is ten times as large as the correlation between the treatment and observables. For the other grades,

the results remain strong but somewhat less so. For the grade 1 data, the range of 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 Application #2: Observational data with "uncorrectable" unobserved heterogeneity

The second type of application of this approach is to situations where the relevant data comes from an observational rather than experimental setting, and there are no apparent solutions to the endogeneity problem. In this case, a researcher faces the choice between providing no estimates of the effect of interest at all, or providing OLS estimates and hoping for the best.

4.1 Background: Income inequality and health

An extensive literature in public health considers the question of whether a higher level of income inequality has a substantial negative impact on individual health outcomes in industrialized countries. The best known proponent of the "inequality hypothesis" is the British epidemiologist Richard G. Wilkinson (1996), who has identified several mechanisms by which inequality may have a negative effect on health. The first and most obvious mechanism is through health expenditures: if health is a normal good and health expenditures have a positive but declining marginal product in health outcomes, then a mean-preserving spread in income within a society will tend to reduce the average health outcome. A second category of mechanisms is more psychological and behavioral in nature, and will lead to a negative relationship between inequality and health even controlling for a person's own level of income. The low social status associated with low *relative* income may lead to increased stress, which has been shown in experimental animal studies to have 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 substantially more important than health expenditures in explaining cross-sectional variation in health outcomes.

Deaton (2003) provides a thorough review from the economist's perspective on the empirical literature evaluating the inequality hypothesis. That literature dates back to Rodgers (1979), who finds that more unequal countries have higher age-adjusted mortality rates after controlling for the country's average income. Numerous researchers subsequently studied the health-inequality relationship using cross-country or cross-state data, and with findings that also tended to support the inequality hypothesis. However, these aggregate studies have been heavily criticized on methodological grounds of data quality/comparability, likelihood of omitted variables

	$[\hat{\theta}_L(\Lambda), \hat{\theta}_H(\Lambda)]$ by grade			
Λ	K	1	2	3
$\{0.00\}$	5.20	6.72	4.97	5.30
	$(3.16, 7.22)$	$(4.69, 8.86)$	$(2.89, 7.03)$	$(3.28, 7.43)$
$[0.00, 0.25]$	$[5.18, 5.20]$	$[6.17, 6.72]$	$[4.61, 4.97]$	$[5.01, 5.30]$
	$(3.13, 7.22)$	$(4.31, 8.58)$	$(2.73, 6.86)$	$(2.90, 7.20)$
$[0.00, 0.50]$	$[5.16, 5.20]$	$[5.60, 6.72]$	$[4.25, 4.97]$	$[4.72, 5.30]$
	$(3.02, 7.21)$	$(3.74, 8.49)$	$(2.44, 6.79)$	$(2.48, 7.18)$
$[0.00, 0.75]$	$[5.15, 5.20]$	$[5.05, 6.72]$	$[3.90, 4.97]$	$[4.40, 5.30]$
	$(2.85, 7.27)$	$(3.05, 8.46)$	$(2.07, 6.74)$	$(1.90, 7.15)$
$[0.00, 1.00]$	$[5.13, 5.20]$	$[4.49, 6.72]$	$[3.54, 4.97]$	$[4.07, 5.30]$
	$(2.62, 7.30)$	$(2.28, 8.45)$	$(1.60, 6.72)$	$(1.22, 7.13)$
$[0.00, 2.00]$	$[5.06, 5.20]$	$[2.21, 6.72]$	$[2.08, 4.97]$	$[2.51, 5.30]$
	$(1.23, 7.35)$	$(-1.19, 8.45)$	$(-0.73, 6.70)$	$(-2.50, 7.08)$
$[0.00, 3.00]$	$[4.99, 5.20]$	$[-0.15, 6.72]$	$[0.56, 4.97]$	$[0.43, 5.30]$
	$(-0.31, 7.50)$	$(-5.10, 8.45)$	$(-3.46, 6.70)$	$(-7.77, 7.07)$
$[0.00, 4.00]$	$[4.91, 5.20]$	$[-2.63, 6.72]$	$[-1.00, 4.97]$	$[-2.47, 5.30]$
	$(-2.09, 7.52)$	$(-9.42, 8.45)$	$(-6.45, 6.70)$	$(-15.70, 7.07)$
$[0.00, 5.00]$	$[4.83, 5.20]$	$[-5.21, 6.72]$	$[-2.62, 4.97]$	$[-6.87, 5.30]$
	$(-3.94, 7.60)$	$(-14.20, 8.45)$	$(-9.69, 6.70)$	$(-29.97, 7.07)$
$[0.00, 7.50]$	$[4.61, 5.20]$	$[-12.44, 6.72]$	$[-7.01, 4.97]$	$(-\infty, \infty)$
	$(9.09, 7.66)$	$(-29.22, 8.45)$	$(-19.14, 6.70)$	$(-\infty, \infty)$
$[0.00, 10.00]$	$[4.36, 5.20]$	$[-21.60, 6.72]$	$[-12.05, 4.97]$	$(-\infty, \infty)$
	$(-15.01, 7.70)$	$(-53.14, 8.45)$	$(-31.56, 6.70)$	$(-\infty, \infty)$
$[0.00, 15.00]$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
	(∞, ∞)	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
$[0.00, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
	(∞, ∞)	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
$(-\infty, 0.00]$	$[5.20, 8.17]$	$[6.72, 134.57]$	$[4.97, 96.33]$	$[5.30, 15.12]$
	$(3.16, 25.12)$	$(5.00, 217.24)$	$(3.06, 406.27)$	$(2.72, 239.90)$
$\hat{\lambda}^*$	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: Interval estimates of the treatment effect of small class sizes on average percentile SAT score, given interval restrictions on relative correlation of treatment with unobservables. Intervals in square brackets are the set estimates themselves, while the intervals in the round brackets are 95% asyptotic confidence intervals calculated by the method of Stoye (2009).

bias and other problems (Deaton 2003), so more recent work in this literature uses linked individual-aggregate data with controls for individual income and background characteristics. Many of these studies exhibit a great deal of methodological sophistication and complexity, including the deployment of elaborate multilevel models. At the same time, almost none have done much to address the issue of endogenous community selection. For example, none of the 21 studies cited in the recent review article by Subramanian and Kawachi (2004) have a research design aimed at addressing endogenous community selection. Oakes (2004) argues that this failure implies their resulting estimates "will always be wrong" (p. 1941). Oakes argues that much of the lack of attention to community selection and other identification issues is misplaced priority on the use of elaborate multilevel models. An alternative explanation is provided by Diez Roux (2001):

> "To the extent that neighborhoods influence the life chances of individuals, neighborhood social and economic characteristics may be related to health through their effects on achieved income, education, and occupation, making these individual-level characteristics mediators (at least in part) rather than confounders. In addition, because socioeconomic position is one of the dimensions along which residential segregation occurs, living in disadvantaged neighborhoods may be one of the mechanisms leading to adverse health outcomes in persons of low socioeconomic status. For these reasons, although teasing apart the independent effects of both dimensions may be useful as part of the analytic process, it is also artificial." (Diez-Roux 2001, p. 1786)

In this view, the true effect that econometricians have gone to such great length to estimate is not the quantity of interest anyway. Because community composition is not under the direct control of policymakers, the neighborhood effect itself does not correspond to any policy response of interest.

An alternative explanation is that this question is particularly ill-suited for the typical methods by which microeconometricians deal with endogeneity. The inequality-health relationship has several relevant features:

1. Health outcomes, particularly the most important ones (mortality and life expectancy) are affected by events decades in the past. The hypothesized mechanisms by which inequality affects health (e.g., stress, depression, increased smoking, drinking, and drug use) include mechanisms that tend to operate over decades rather than months.

2. Aggregate measures of income inequality based on household surveys are notoriously noisy measures of the underlying quantity of interest (inequality in some form of permanent income, possibly adjusted for credit constraints). The underlying quantity of interest changes slowly over time, so most year-to-year variations in measured inequality are noise.

3. The current level of income inequality is the outcome of a complex

interaction of policies and historical accidents. There is no government policy that can have a substantial impact on inequality without also affecting other variables relevant to health outcomes.

The first two features make the use of panel data with cross-sectional fixed effects particularly unappealing.

There are also more specific ways in which the existing methods are unsuitable for the measurement of community effects on health. First, as Mellor and Milyo (2003) emphasize, particularly important health outcomes - mortality and life expectancy in particular - are affected by events decades in the past. As a result the connection between current community and current health may say little about the overall influence of community over the life cycle. Because most methods for overcoming endogenous community choice are based on small short-term changes in the social environment, these approaches might be limited to more rapidly-responding intermediate outcomes such as health behavior (smoking/drinking/etc.) and injuries. Another issue, particularly in the literature on inequality and health, is that community variables are measured with a great deal of noise. The fixed-effect model used for the cohort-based research design will be particularly problematic here - fixed effects models can dramatically amplify the bias associated with measurement error in explanatory variables.

4.2 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 literature on inequality and health (Blakely, Kennedy, Glass and Kawachi 2000, Blakely, Lochner and Kawachi 2002, Mellor and Milyo 2002, Mellor and Milyo 2003, Subramanian and Kawachi 2003, Subramanian and Kawachi 2004). Individual-level explanatory variables include age, sex, race (black/white/other), education in years, log equivalized total income (total household income divided by the square root of household size), employment status (employed/not employed) and health insurance status (insured/not insured). The community-level variable is the state-level Gini coefficient for 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 unweighted summary statistics.

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:	
Gini coefficient for household income	0.43
	(0.02)
# of individuals	187,760
# of states (including DC)	51

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

4.3 Results under assumption of exogeneity

Table 5 shows the basic regression results for the special case of exogeneity. These estimates can be considered a benchmark for the subsequent analysis that considers alternatives to exogeneity. The first set of estimates are for a linear model, and are estimated using OLS with cluster-robust estimates of standard errors. 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. The individual-level coefficients are estimated with great precision due to the large sample size, and are almost all statistically significant.

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.

Comparison between the linear and logistic model estimates is somewhat complicated by the fact that linear models produce constant marginal effects and variable odds ratios while logistic models produce variable marginal effects and constant odds ratios. To make a reasonable comparison we consider a representative case of an individual with characteristics

	Linear		Logistic	
Variable	(1)	(2)	(1)	(2)
State-level income inequality	0.903	0.299	8.564	4.608
	(0.159)	(0.122)	(1.226)	(1.173)
Log income		-0.031		-0.254
		(0.002)		(0.009)
Age (yrs)		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 (yrs)		-0.013		-0.093
		(0.002)		(0.003)
Not employed		0.129		1.089
		(0.003)		(0.017)
No health insurance		0.066		0.529
		(0.005)		(0.018)

Table 5: Regression results for model with assumption of exogeneity ($\lambda = 0$). Linear model estimated using OLS, with cluster-robust standard errors. Logistic model estimated as random-intercept multilevel model with maximum likelihood.

that imply a probability of self-rated fair/poor health of 15% (the average in the data). For this representative individual, the linear model implies a marginal effect of 0.299 while the logistic model implies a marginal effect of 0.588. The odds ratio for this representative individual associated with an increase in the state-level Gini coefficient of 0.05 is 1.26 for the logistic model and 1.12 for the linear model. As these results suggest, using a linear model results in a somewhat weaker but still statistically significant association between the state-level Gini coefficient and the probability of self-rated fair/poor health.

4.4 Results

The estimates reported in Table 5 are based on models in which exogeneity is assumed. As discussed in Section 2, this is a strong and somewhat indefensible assumption, so we evaluate the effect of deviations from exogeneity

The model to be estimated is the linear model (2) from Table 5. Table 6 reports the range of estimated coefficients on inequality $[\hat{\theta}_L(\Lambda), \hat{\theta}_H(\Lambda)]$ as a function of restriction on the relative correlation $\lambda \in \Lambda$.

Λ	$[\hat{\theta}_L(\Lambda), \hat{\theta}_H(\Lambda)]$	95% CI
$\{0.00\}$	0.30	$(0.06, 0.54)$
$[0.00, 0.10]$	$[0.16, 0.30]$	$(-0.07, 0.54)$
$[0.00, 0.20]$	$[0.03, 0.30]$	$(-0.20, 0.54)$
$[0.00, 0.30]$	$[-0.10, 0.30]$	$(-0.33, 0.54)$
$[0.00, 0.40]$	$[-0.24, 0.30]$	$(-0.47, 0.54)$
$[0.00, 0.50]$	$[-0.38, 0.30]$	$(-0.62, 0.54)$
$[0.00, 0.75]$	$[-0.73, 0.30]$	$(-1.01, 0.54)$
$[0.00, 1.00]$	$[-1.09, 0.30]$	$(-1.43, 0.54)$
$[0.00, 2.00]$	$[-2.71, 0.30]$	$(-3.45, 0.54)$
$[0.00, 3.00]$	$[-4.70, 0.30]$	$(-6.25, 0.54)$
$[0.00, 4.00]$	$[-7.37, 0.30]$	$(-10.82, 0.54)$
$[0.00, 5.00]$	$[-11.83, 0.30]$	$(-22.40, 0.54)$
$[0.00, 6.00]$	$(-\infty, \infty)$	$(-\infty, \infty)$
$[0.00, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
$(-\infty, 0.00]$	$[0.30, 17.04]$	$(0.06, 18, 600)$
$\hat{\lambda}^*$	5.17	
$\hat{\theta}^*$	17.04	
$\hat{\lambda}(0)$	0.23	

Table 6: Estimated effect of income inequality on health. Each row reports a range of estimates for the true effect (θ) consistent with a different range of possible values for the relative correlation of inequality with health-related unobservables (λ).

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, we 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.

5 Conclusion

The methodology developed in this paper provides a simple means of providing bounds on causal parameters under interval restrictions on the degree of endogeneity. 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.

Several areas remain for future work. The methodology can be advanced along two main fronts. First, the inference in the current paper is based on asymptotics that are known to provide a poor approximation in finite sample for estimators of the type under consideration. Second, the model developed in Section 2 is based on a simple linear model with random sampling, and many applications involve complications such as fixed effects, clustered samples, etc. Extending the model to handle such cases will provide greater applicability.

Additional applications should also be explored. The Project STAR data and the CPS inequality and health data to some extent represent opposite extremes. Project STAR is a relatively (though not perfectly) clean experiment and the inequality and health data are particularly plagued with endogeneity problems. It would be interesting to see how the results will be different for an experimental study with more extensive deviations from the experimental protocol than are seen in Project STAR, and for an observational study in which researchers more seriously argue for exogeneity of treatment than do researchers in the inequality and health literature.

References

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber, "Selection on observed and unobserved variables: Assessing the ef-

fectiveness 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 Hansen, and Peter E. Rossi, "Plausibly Exogenous," Working paper, Chicago Booth GSB 2008.

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

Diez-Roux, Ana V., "Investigating Neighborhood and Area Effects on Health," *American Journal of Public Health*, 2001, *91*, 1783–1789.

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

Hanushek, Eric, "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.

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.

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.

Oakes, J. Michael, "The (mis)estimation of neighborhood effects: Causal inference for a practicable social epidemiology," *Social Science & Medicine*, 2004, *58*, 1929–1952.

Rodgers, G.B., "Income and inequality as determinants of mortality: An international cross-section analysis," *Population Studies*, 1979, *33* (3), 343–351. Reprinted with comments in *International Journal of Epidemiology* 31:533-538, 2001.

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, *Unhealthy Societies: The Afflictions of Inequality*, London: Routledge, 1996.

A Proofs of propositions

The following proofs are still in progress, and may be incomplete.

A.1 Proposition 1

First, note that:

$$\begin{aligned}
\lambda(\theta) &= \frac{\text{corr}\left(z, y-\theta z - L(y-\theta z|X)\right)}{\text{corr}\left(z, L(y-\theta z|X)\right)} \\
&= \frac{\frac{cov(z,y-\theta z-L(y-\theta z|X))}{\sqrt{var(z)var(y-\theta z-L(y-\theta z|X))}}}{\frac{cov(z,L(y-\theta z|X))}{\sqrt{var(z)var(L(y-\theta z|X))}}} \\
&= \left(\frac{cov(z,y)-\theta var(z)}{cov(z,L(y|X))-\theta cov(z,L(z|X))} - 1\right) \\
&/ \sqrt{\frac{var(y-\theta z)}{var(L(y-\theta z|X))} - 1}
\end{aligned}$$

We can apply several properties of the linear projection, specifically that $cov(z, L(y|X)) = cov(L(z|X), L(y|X))$, $cov(z, L(z|X)) = var(L(z|X))$ and $var(y - L(y|X)) = var(y) - var(L(y|X))$, to further derive:

$$\begin{aligned}
\lambda(\theta) &= \left(\frac{cov(z,y) - \theta var(z))}{cov(L(z|X), L(y|X)) - \theta var(L(z|X))} - 1 \right) \\
&/ \sqrt{\frac{var(y) - 2\theta cov(z,y) + \theta^2 var(z)}{var(L(y|X)) - 2\theta cov(L(z|X), L(y|X)) + \theta^2 var(L(z|X))} - 1} \\
&= \left(\frac{p_1}{p_2} - 1 \right) / \sqrt{\frac{p_3}{p_4} - 1}
\end{aligned}$$

where p_1, p_2, p_3, and p_4 are all polynomials (and thus differentiable) in θ. Application of the quotient and product rules implies that $\lambda(\theta)$ is also differentiable provided that $p_2 \neq 0$, $p_4 \neq 0$, and $\frac{p_3}{p_4} > 1$. The first of these conditions fails if:

$$p_4 = cov(z, L(y|X)) - \theta cov(z, L(z|X)) = 0$$

i.e., if $\theta = \frac{cov(L(z|X), L(y|X))}{var(L(z|X))} = \theta^*$. The second condition fails if:

$$p_4 = var(L(y|X) - \theta L(z|X)) = 0$$

which implies that $\hat{y} - \theta \hat{z}$ is constant. This in turn implis that $cov(L(y|X), L(z|X)) = \theta var(L(z|X))$, i.e. $\theta = \theta^*(m)$ as defined above. Next, consider the condition $p_3 > p_4$. Note that p_3 is the variance of $(y - \theta z)$ while p_4 equals the variance of its best linear predictor given X. This means that $p_3 > p_4$ unless $y - \theta z = L(y|X) - \theta L(z|X)$ for some θ. Rearranging, this implies that $y = L(y|X) + \theta(z - L(z|X)))$ for some θ. But if there exists such a θ, then assumption (7 is violated. Therefore, $\lambda(\theta)$ is differentiable at every value of θ other than $\theta^*(m)$.

To establish result (2), suppose that $L(y|X) = \tilde{\theta} L(x|X)$. Then for any $\theta \neq \tilde{\theta}$:

$$\begin{aligned}
\lambda(\theta) &= \left(\frac{cov(z,y) - \theta var(z)}{(\tilde{\theta} - \theta) var(L(z|X))} - 1 \right) \\
&/ \sqrt{\frac{var(y) - 2\theta cov(z,y) + \theta^2 var(z)}{(\tilde{\theta} - \theta)^2 var(L(z|X))} - 1} \\
&= \frac{cov(z,y) - \theta var(z) - (\tilde{\theta} - \theta) var(L(z|X))}{\sqrt{var(L(z|X)) * \left(var(y) - 2\theta cov(z,y) + \theta^2 var(z) - (\tilde{\theta} - \theta) var(L(z|X)) \right)}} \\
&* \; sign(\tilde{\theta} - \theta)
\end{aligned}$$

The limits in result (2) then follow by substitution. To show that equation (5) is satisfied at $\tilde{\theta}$ for all λ, note that $var(L(y|X) - \tilde{\theta} L(z|X)) = cov(z, L(y|X) - \tilde{\theta} L(z|X)) = 0$. Equation (5) thus reduces to $0 = \lambda_0 * 0$, a condition that is satisfied by any $\lambda_0 \in (-\infty, \infty)$.

To establish result (3), note that since $var(L(z|X)) > 0$, p_2 is positive for $\theta < \theta^*$, negative for $\theta > \theta^*$, and zero when $\theta = \theta^*$. When

$\frac{cov(z,y)}{var(z)} = \frac{cov(L(z|X),L(y|X))}{var(L(z|X))}$, we also have $p_1 = 0$ when $\theta = \theta^*$ so we apply L'Hospital's rule to get the limit:

$$\lim_{\theta \to \theta^*} \frac{cov(z,y) - \theta var(z)}{cov(L(z|X), L(y|X)) - \theta var(L(z|X))} = \frac{var(z)}{var(L(z|X))} \tag{23}$$

To establish result (4), again note that since $var(L(z|X)) > 0$, p_2 is positive for $\theta < \theta^*$, negative for $\theta > \theta^*$, and zero when $\theta = \theta^*$. Also note that $cov(z,y) - \theta^* var(z) = cov(z,y) - \frac{cov(L(z|X),L(y|X))}{var(L(z|X))} var(z)$, so p_1 is strictly positive for all $\theta \approx \theta^*$ if $\frac{cov(z,y)}{var(z)} > \frac{cov(L(z|X),L(y|X))var(z)}{var(L(z|X))}$, and strictly negative for all $\theta \approx \theta^*$ if $\frac{cov(z,y)}{var(z)} < \frac{cov(L(z|X),L(y|X))var(z)}{var(L(z|X))}$. The results on limits then follow directly. Next, note that since $cov(L(z|X), L(y|X)) - \theta^* var(L(z|X)) = 0$, equation (5) implies that either $var(L(y|X) - \theta^* L(z|X)) = 0$, implying (15) holds, or $cov(z,y) - \theta^* var(z) = 0$, implying (16) holds. Since neither holds, there is no λ that satisfies equation (5) for $\theta = \theta^*$

To establish result (5), note that:

$$\lim_{\theta \to \infty} (cov(z,y) - \theta var(z)) = -\infty \tag{24}$$

$$\lim_{\theta \to \infty} (cov(L(z|X), L(y|X)) - \theta var(L(z|X))) = -\infty \tag{25}$$

So by L'Hospital's rule:

$$\lim_{\theta \to \infty} \frac{cov(z,y) - \theta var(z)}{cov(L(z|X), L(y|X)) - \theta var(L(z|X))} = \frac{var(z)}{var(L(z|X))} \tag{26}$$

Also note that:

$$\begin{aligned}
\lim_{\theta \to \infty} (var(y) - 2\theta cov(z,y) + \theta^2 var(z)) &= \infty \\
\lim_{\theta \to \infty} (var(L(y|X)) - 2\theta cov(L(z|X), L(y|X)) + \theta^2 var(L(z|X))) &= \infty \\
\lim_{\theta \to \infty} (-2cov(z,y) + 2\theta var(z)) &= \infty \\
\lim_{\theta \to \infty} (-2cov(L(z|X), L(y|X)) + 2\theta var(L(z|X))) &= \infty
\end{aligned}$$

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

$$\lim_{\theta \to \infty} \frac{var(y) - 2\theta cov(z,y) + \theta^2 var(z)}{var(L(y|X)) - 2\theta cov(L(z|X), L(y|X)) + \theta^2 var(L(z|X))} = \frac{var(z)}{var(L(z|X))} \tag{27}$$

Result (5) can then be derived by substitution, and the argument repeated for $\lim_{\theta \to -\infty}$.

To establish result (6), note first that if (15) or (16) hold, then (λ, θ^*) always satsifies equation (5) for any λ. If neither of these conditions hold, then we have established that $\lim_{\theta \to \infty} \lambda(\theta; m) = \lambda^*$, that $\lim_{\theta \uparrow \theta^*} \lambda(\theta; m)$ is either $-\infty$ or ∞, and that $\lambda(\theta; m)$ is continuous on (∞, θ^*). Suppose for the moment that $\lim_{\theta \uparrow \theta^*} = -\infty$. By the intermediate value theorem, for any $\lambda \in (-\infty, \lambda^*)$, there exists be some $\theta \in (-\infty, \theta^*)$ such that $\lambda(\theta, m) = \lambda$. If $\lim_{\theta \uparrow \theta^*} = -\infty$, then property (1) implies that $\lim_{\theta \downarrow \theta^*} = \infty$. Again, since $\lambda(\theta; \infty)$ is continuous on (θ^*, ∞), the intermediate value theorem implies that for any $\lambda \in (\lambda^*, \infty)$ there exists some $\theta \in (\theta^*, \infty)$ such that $\lambda(\theta; m) = \lambda$). The same argument can be duplicated for the case $\lim_{\theta \uparrow \theta^*} = \infty$. Note that there may or may not be a θ such that $\lambda(\theta) = \lambda^*$.

A.2 Proposition 2

The first part of this proposition (nonemptiness) follows directly from property (6) of Proposition 1, and the second part follows from property (5) of Proposition 1.

If $var(L(z|X)) = 0$, then $cov(z, L(y|X) - \theta L(z|X)) = 0$ for all θ. This implies that (5) holds (for any λ) if and only if $cov(z, y - \theta z) = 0$, i.e., if $\theta = cov(z, y)/var(z)$. The identified set is thus identical to the regression coefficient from the OLS regression of y on z.

A.3 Proposition 3

The first set of results follows from the straightforward application of Slutsky's theorem (since the functions described are continuous in m for all $\theta \neq \theta^*$).

For the second result, note that the implicit function theorem implies that $\theta_L(\Lambda)$ is differentiable in m if $\frac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_L(\Lambda)} \neq 0$. 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 $\theta_L(\Lambda; m) = -\infty$, then Λ is nondegenerate and $\exists \epsilon > 0, B_1 < B : [\lambda(B_1; m) - \epsilon, \lambda(B_1; m) + \epsilon] \subset \Lambda$. Since $\hat{\lambda}(B_1) \xrightarrow{p} \lambda(B_1; m)$, we have:

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

A.4 Proposition 4

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

$$A = \begin{bmatrix} \nabla_m \theta_L(\Lambda; m) \\ \nabla_m \theta_H(\Lambda; m) \end{bmatrix} \tag{28}$$

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

$$\begin{aligned} \nabla_m \theta_L(\Lambda; m) &= -\left. \frac{\nabla_m \lambda(\theta; m)}{\partial \lambda(\theta; m)/\partial \theta} \right|_{\theta=\theta_L(\Lambda;m)} \\ \nabla_m \theta_H(\Lambda; m) &= -\left. \frac{\nabla_m \lambda(\theta; m)}{\partial \lambda(\theta; m)/\partial \theta} \right|_{\theta=\theta_H(\Lambda;m)} \end{aligned} \tag{29}$$

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.