Bounding a linear causal effect using relative correlation restrictions*

Brian Krauth
Department of Economics
Simon Fraser University

September 2007

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.

*Preliminary research, comments are appreciated. An earlier version of this paper was presented at the 2006 Joint Statistical Meetings under the title "Forming better guesses about neighborhood effects on health: Estimating community effects using conditional modeling of unobservables." 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 policy 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 numerous applications where the standard solutions are inapplicable, but the policy question needs the best answer available. Conventionally, researchers in this situation make the weakest set of assumptions needed to achieve point identification of the causal effect, and hope that those assumptions are close enough to the truth. An alternative approach to this problem, argued most forcefully by Manski (1994), rejects overemphasis on point identification at the cost of the credibility of the identifying assumptions. Instead point identification is viewed as a special case of set or interval identification, and there is an explicit trade-off between the strength of the researcher's assumptions and the size of the identified set. Stronger assumptions narrow the range of the identified set, eventually to a single point. This set identification approach to identification questions in econometrics is closely related to the long-standing literature in statistics on sensitivity analysis. In classical sensitivity analysis, the strength of assumptions is parameterized by a single sensitivity parameter whose value is zero in the standard case. Interval restrictions on the sensitivity parameter lead to interval identification of the parameter of interest, with a direct and positive relationship in size between the two intervals.

The approach developed in this paper is in this spirit. A very simple linear model is parameterized with a single policy variable and a set of observed control variables. The outcome is linear in the policy variable, but is also affected by other factors that may be correlated with the policy variable. The sensitivity parameter describes the correlation between the policy variable and these unobservable other factors relative to the correlation between the policy variable and the control variables. This type of sensitivity parameter has been used recently in some specific applications using nonlinear econometric models (Altonji, Elder and Taber 2005, Imbens 2003, Krauth 2006), but not in the simple linear context. After developing the model and estimation method, I derive some of its statistical properties, and report the results from two applications.

The first application is to data from a natural or designed experiment in which there are small deviations from true random assignment. The example dataset used is from Project STAR, a well-known experimental study of the effect of smaller class size on student outcomes. Researchers anlyzing data from imperfect experiments will often show extensive tables of summary statistics for the particiant's background variables broken down into treatment and control groups. If the two groups are sufficiently similar in these background variables, the researcher might then argue that they are likely to be similar in outcome-relevant unobserved variables, and thus we can treat the experiment

as following true random assignment. Researchers will also sometimes add these variables as control variables in the regression estimating the treatment effect.

The second application is to provide a more conservative approach to non-experimental data in which a researcher cannot credibly claim exogeneity of the policy variable, and in which standard techniques for solving the endogeneity problem are inapplicable. The policy question in this application is the effect of income inequality on individual health. Public health researchers have investigated this question extensively, and have generally found a negative relationship between state-level inequality and health, even after controlling for individual income. These findings have been criticized by economists and other researchers in public health on various methodological grounds, but economists have yet to make much ground on producing more convincing answers to the question. As will be discussed, most of the standard techniques used in applied microeconometrics to solve endogeneity problems are inappropriate in this particular case.

The paper is organized as follows. Section 2 describes the model and derives the estimator and its statistical properties. Section 3 describes the application to the Project STAR study of the effect of class size on academic outcomes. Section 4 describes the application using CPS data to study the effect of income inequality on health. Section 5 concludes and notes avenues for further research.

2 Methodology

2.1 Model

The model is as follows. Let y be a scalar random variable representing the outcome of interest, and let z be a scalar random variable representing the policy or treatment of interest. The structural model is linear in z.

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

where the parameter θ measures the true effect of changes in z on y, and the scalar random variable u represents the effect of all variables other than z. In general, $E(u|z) \neq 0$.

Now, let X be a k-vector of control variables (including an intercept) such that $E(X'X)$ is positive definite. Define:

$$\begin{aligned} \beta &\equiv E(X'X)^{-1}E(X'u) \\ v &\equiv u - X\beta \end{aligned} \tag{2}$$

That is, $X\beta$ is the best linear predictor of u given X. By construction,

$$y = \theta z + X\beta + v$$

and

$$E(X'v) = 0$$

Note that β is not assumed to have a structural or causal interpretation, and is not of direct interest. In addition, we have not imposed any assumption of linearity in the relationship between y and X.

In order for the OLS regression of y on (X, z) to consistently estimate the true effect θ, we would need $E(zv) = 0$ or equivalently, $\text{corr}(z, v) = 0$. Instead we suppose that:

$$\text{corr}(z, v) = \lambda \text{corr}(z, X\beta) \tag{3}$$

for some $\lambda \in R$. In order to for the correlations in (3) to exist, we need:

$$\begin{aligned} var(z) &> 0 \\ var(v) &> 0 \\ var(X\beta) &> 0 \end{aligned}$$

The first condition simply says that the variable of intrest has variation in the population, and can be verified directly from the data. The second condition also can be verified in the data, as its violation implies that y can be written as an exact linear combination of z and X. The third condition is somewhat more difficult to verify, but it holds as long as at β has at least one nonzero element other than the intercept.

The sensitivity parameter λ represents the correlation between our policy variable and unobservable factors, relative to the correlation between the policy variable and observable factors. The only restriction on the data generating process imposed directly in equation (3) is finiteness of λ, i.e. if $\text{corr}(z, X\beta) = 0$ then $\text{corr}(z, v) = 0$.

Without further restrictions on λ, θ is not identified. Alternatively, if we assume $\lambda = 0$, then OLS will consistently estimate θ. However, this assumption is often difficult to justify in applied work using observational data. This paper considers weaker assumptions of the form $\lambda \in \Lambda$, where Λ is some interval. Under this category of assumptions, we can place consistent bounds on θ. The informativeness of those bounds will depend on the choice of Λ.

2.2 Identification

First we define what can be identified from the joint distribution of (y, z, X). Identification is based on the $k + 1$ equations:

$$\begin{aligned} E(X'(y - \theta z - X\beta)) &= 0 \\ \text{corr}(z, (y - \theta z - X\beta)) - \lambda \text{corr}(z, X\beta) &= 0 \end{aligned} \tag{4}$$

and on the restriction that

$$\lambda \in \Lambda \equiv [\lambda_L, \lambda_H] \tag{5}$$

for some interval Λ. The first k equations in (4) are linear in all parameters, while the last equation is linear in λ but nonlinear in θ and β.

Define the function $\lambda : R \to R$ such that:

$$\lambda(\theta) \equiv \frac{\text{corr}(z, y - \theta z - P(y - \theta z|X))}{\text{corr}(z, P(y - \theta z|X))}$$

where $P(.|X)$ is used to denote the best linear predictor of a random variable given X. Proposition 1 below outlines some key properties of $\lambda(.)$.

Proposition 1 (Properties of $\lambda(.)$) *Let:*

$$\lambda^* \equiv \sqrt{\frac{1}{R^2_{zx}} - 1}$$

where R^2_{zx} is the R^2 from the best linear predictor of z given X. and let:

$$\theta^* \equiv \frac{cov(z, P(y|X))}{cov(z, P(z|X))}$$

where $P(.|X)$ is the best linear predictor of its argument given X. Then:

1. *$\lambda(.)$ has the property that:*

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

2. *$\lambda(\theta)$ exists and is differentiable for all $\theta \neq \theta^*$.*
3. *$\lambda(.)$ has the property that:*

$$\lim_{\theta \to \theta^*} |\lambda(\theta)| = \infty$$

for almost all joint distributions of (X, y, z).

Proof: See appendix.

Next, define the set function:

$$\Theta(\Lambda) \equiv \begin{cases} \{\theta : \lambda(\theta) \in \Lambda\} \cup \theta^* & \text{if } \operatorname{corr}(z, (y - \theta^* z - X\beta)) = 0 \\ \{\theta : \lambda(\theta) \in \Lambda\} & \text{if not} \end{cases}$$

The set $\Theta(\Lambda)$ is the set of values for θ that are consistent with the joint distribution of (y, x, z) and the restriction that $\lambda \in \Lambda$. In many cases we will be interested primarily in the range of $\Theta(\Lambda)$, so let:

$$\begin{aligned} \theta_L(\Lambda) &\equiv \inf \Theta(\Lambda) \\ \theta_H(\Lambda) &\equiv \sup \Theta(\Lambda) \end{aligned}$$

For a given Λ, $\theta_L(\Lambda)$ and $\theta_H(\Lambda)$ are numbers rather than sets.

Proposition 2 describes some of the properties of the identified set $\Theta(\Lambda)$.

Proposition 2 (Properties of $\Theta(.)$) *The set $\Theta(\Lambda)$ is bounded if and only if $\lambda^* \notin \Lambda$.*

Proof: See appendix.

To understand the intuition for Proposition 2, note that R^2_{zx} is the R^2 from the OLS regression of z on X. When $R^2_{zx} = 1$, i.e., z is an exact linear function of X, we have $\lambda^* = 0$ and the model is unidentified even under exogeneity. This is of course the standard rank condition for OLS regression. As $R^2_{zx} \to 0$, the value of λ at which θ is unidentified increases.

Propositions 1 and 2 lead to the main identification result in Proposition 3.

Proposition 3 (Identification) *Let (λ_0, θ_0) be the true values of λ and θ. Then θ_0 is identified from the joint probability distribution of (X, y, z) in the sense that:*

$$\begin{aligned} \lambda(\theta_0) &= \lambda_0 \qquad \text{if } \theta_0 \neq \theta^* \\ \lambda_0 \in \Lambda &\Rightarrow \theta_0 \in \Theta(\Lambda) \end{aligned}$$

Proof: See appendix.

2.3 Estimation

Estimation proceeds in two steps. First, we calculate the sample analog to $\lambda(\theta)$ over a fine grid of values for θ:

$$\hat{\lambda}(\theta) \equiv \frac{\hat{\text{corr}}(z, y - \theta z - \hat{P}(y - \theta z|X))}{\hat{\text{corr}}(z, \hat{P}(y - \theta z|X))}$$

where $\hat{\text{corr}}$ is the sample correlation function, and $\hat{P}(y - \theta z|X)$ is the vector of fitted values from the OLS regression of $y - \theta z$ on X.

Then we calculate the sample analog to $\Theta(\Lambda)$. First, let:

$$\hat{\theta}^* \equiv \frac{\hat{cov}(z, \hat{P}(y|X))}{\hat{cov}(z, \hat{P}(z|X))}$$

and let:

$$\hat{\lambda}^* \equiv \sqrt{\frac{1}{\hat{R}^2_{zx}} - 1}$$

Then let:

$$\hat{\Theta}(\Lambda) \equiv \begin{cases} \{\theta : \hat{\lambda}(\theta) \in \Lambda\} \cup \hat{\theta}^* & \text{if } |\hat{\text{corr}}\left(z, \left(y - \hat{\theta}^* z - X\beta\right)\right)| < \epsilon_N \\ \{\theta : \hat{\lambda}(\theta) \in \Lambda\} & \text{if not} \end{cases}$$

where $\epsilon_N > 0$ is a small number, and let:

$$\begin{aligned} \hat{\theta}_L(\Lambda) &\equiv \inf \hat{\Theta}(\Lambda) \\ \hat{\theta}_H(\Lambda) &\equiv \sup \hat{\Theta}(\Lambda) \end{aligned}$$

Proposition 4 (Consistency of $\hat{\lambda}(.)$ and $\hat{\Theta}(.)$) *Let $\{x_i, y_i, z_i\}_{i=1}^N$ be a random sample of size N, and let $\hat{\lambda}(\theta)$ and $\hat{\Theta}(\Lambda)$ be defined over that sample Then:*

$$\begin{aligned} \operatorname*{plim}_{N\to\infty} \hat{\lambda}(\theta_0) &= \lambda(\theta_0) \qquad \text{if } \theta_0 \neq \theta^* \\ \operatorname*{plim}_{N\to\infty} \hat{\theta}_L(\Lambda) &= \theta_L(\Lambda) \qquad \text{if } \tfrac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_L(\Lambda)} \neq 0 \\ \operatorname*{plim}_{N\to\infty} \hat{\theta}_H(\Lambda) &= \theta_H(\Lambda) \qquad \text{if } \tfrac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_H(\Lambda)} \neq 0 \end{aligned}$$

Proof (incomplete): See appendix.

2.4 Inference

For most parameter values, the lower and upper bounds of $\hat{\Theta}(\Lambda)$ can be written as continuous but nonlinear functions of a set of sample moments. As a result, these bounds are $\sqrt{N}$-consistent and asymptotically normal, and an asymptotic covariance matrix can be derived through straightforward application of the delta method. Proposition 5 below states this more explicitly.

Proposition 5 (Asymptotic distribution of $\hat{\Theta}(\Lambda)$) *Suppose that $\theta_0 \neq \theta^*$, $\lambda^* \notin \Lambda$, $\frac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_L(\Lambda)} \neq 0$, and $\frac{d\lambda(\theta)}{d\theta}|_{\theta=\theta_H(\Lambda)} \neq 0$. Then:*

$$\sqrt{N}\begin{bmatrix} \hat{\theta}_L(\Lambda) - \theta_L(\Lambda) \\ \hat{\theta}_H(\Lambda) - \theta_H(\Lambda) \end{bmatrix} \xrightarrow{D} N\left(0, \begin{bmatrix} \sigma_L^2 & \sigma_{LH} \\ \sigma_{LH} & \sigma_H^2 \end{bmatrix}\right)$$

where the asymptotic covariance matrix can be derived using the delta method.

Proof (incomplete): See appendix.

Proposition 5 can be used to perform simple asymptotic hypothesis tests on θ_0. In constructing confidence intervals, Imbens and Manski (2004) note the necessity of distinguishing between a confidence interval for the identified set $\Theta(\Lambda)$ and for the true parameter value θ_0. A confidence interval for the identified set can be constructed using the lower and upper bounds, respectively, of the ordinary confidence intervals of $\theta_L(\Lambda)$ and $\theta_H(\Lambda)$. The confidence interval for the true parameter value is generally narrower than that for the identified set, and its construction is described in Imbens and Manski.

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.

	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.

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

$$y = \theta z + X\beta + S + v \tag{6}$$

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 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:

$$y - \bar{y}_s = \theta(z - \bar{z}_s) + (X - \bar{X}_s)\beta + (v - \bar{v}_s) \tag{7}$$

Then assume that:

$$\text{corr}\left(\tilde{z}, \tilde{v}\right) = \lambda \text{corr}\left(\tilde{z}, \tilde{X}\beta\right) \tag{8}$$

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[1] 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. There

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.

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.

The results reported in Table 3 can also be presented graphically, and this mode of presentation provides a bit more insight into where the results are coming from. Figure 1 shows the results graphically for kindergarten and grade 1, while Figure 2 shows results for grades 2 and 3. In each graph, the solid line is the estimated $\hat{\lambda}(\theta)$ function. The dashed vertical line is the estimated $\hat{\lambda}^*$. The shaded region shows the range of $\hat{\Theta}([0, \lambda])$ for each positive value of λ. The dot shows the OLS point estimate of the effect.

[1]The standard errors reported in Table 2 are not yet corrected for classroom-level clustering, as are the standard errors in Krueger's Table 5. Applying the correction would raise most standard errors by a factor of about 1.5 to 2.0.

	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
	(0.74)	(0.64)	(0.71)	(0.63)	(0.75)	(0.65)	(0.81)	(0.68)
Regular/aide class	0.26		1.77		1.54		-0.51	
	(0.71)		(0.69)		(0.72)		(0.78)	
White/Asian	8.39	8.39	6.94	6.98	6.45	6.48	6.05	6.05
	(1.24)	(1.24)	(1.09)	(1.09)	(1.16)	(1.16)	(1.26)	(1.26)
Girl	4.38	4.38	3.83	3.82	3.42	3.41	4.19	4.20
	(0.59)	(0.59)	(0.56)	(0.56)	(0.59)	(0.59)	(0.62)	(0.62)
Free lunch	-13.08	-13.08	-13.55	-13.55	-13.62	-13.64	-12.95	-12.94
	(0.71)	(0.71)	(0.68)	(0.68)	(0.71)	(0.71)	(0.74)	(0.74)
White teacher	-1.13	-1.09	-4.02	-4.23	0.43	0.61	0.28	0.27
	(1.18)	(1.17)	(1.02)	(1.01)	(0.93)	(0.92)	(1.06)	(1.06)
Teacher experience	0.26	0.27	0.06	0.07	0.10	0.11	0.05	0.05
	(0.06)	(0.06)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)	(0.04)
Master's degree	-0.59	-0.60	0.44	0.55	-1.06	-0.92	0.93	0.89
	(0.77)	(0.77)	(0.70)	(0.70)	(0.72)	(0.72)	(0.76)	(0.76)
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.

	$\hat{\Theta}(\Lambda)$ by grade			
Λ	K	1	2	3
$\{0.00\}$	5.20	6.72	4.97	5.30
$[0.00, 0.25]$	$[5.18, 5.20]$	$[6.17, 6.72]$	$[4.61, 4.97]$	$[5.01, 5.30]$
$[0.00, 0.50]$	$[5.16, 5.20]$	$[5.60, 6.72]$	$[4.25, 4.97]$	$[4.72, 5.30]$
$[0.00, 0.75]$	$[5.15, 5.20]$	$[5.05, 6.72]$	$[3.90, 4.97]$	$[4.40, 5.30]$
$[0.00, 1.00]$	$[5.13, 5.20]$	$[4.49, 6.72]$	$[3.54, 4.97]$	$[4.07, 5.30]$
$[0.00, 2.00]$	$[5.06, 5.20]$	$[2.21, 6.72]$	$[2.08, 4.97]$	$[2.51, 5.30]$
$[0.00, 3.00]$	$[4.99, 5.20]$	$[-0.15, 6.72]$	$[0.56, 4.97]$	$[0.43, 5.30]$
$[0.00, 4.00]$	$[4.91, 5.20]$	$[-2.63, 6.72]$	$[-1.00, 4.97]$	$[-2.47, 5.30]$
$[0.00, 5.00]$	$[4.83, 5.20]$	$[-5.21, 6.72]$	$[-2.62, 4.97]$	$[-6.87, 5.30]$
$[0.00, 7.50]$	$[4.61, 5.20]$	$[-12.44, 6.72]$	$[-7.01, 4.97]$	$(-\infty, \infty)$
$[0.00, 10.00]$	$[4.36, 5.20]$	$[-21.60, 6.72]$	$[-12.05, 4.97]$	$(-\infty, \infty)$
$[0.00, 15.00]$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
$[0.00, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$	$(-\infty, \infty)$
$(-\infty, 0.00]$	$[5.20, 8.17]$	$[6.72, 134.57]$	$[4.97, 96.33]$	$[5.30, 15.12]$
$\hat{\lambda}^*$	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.

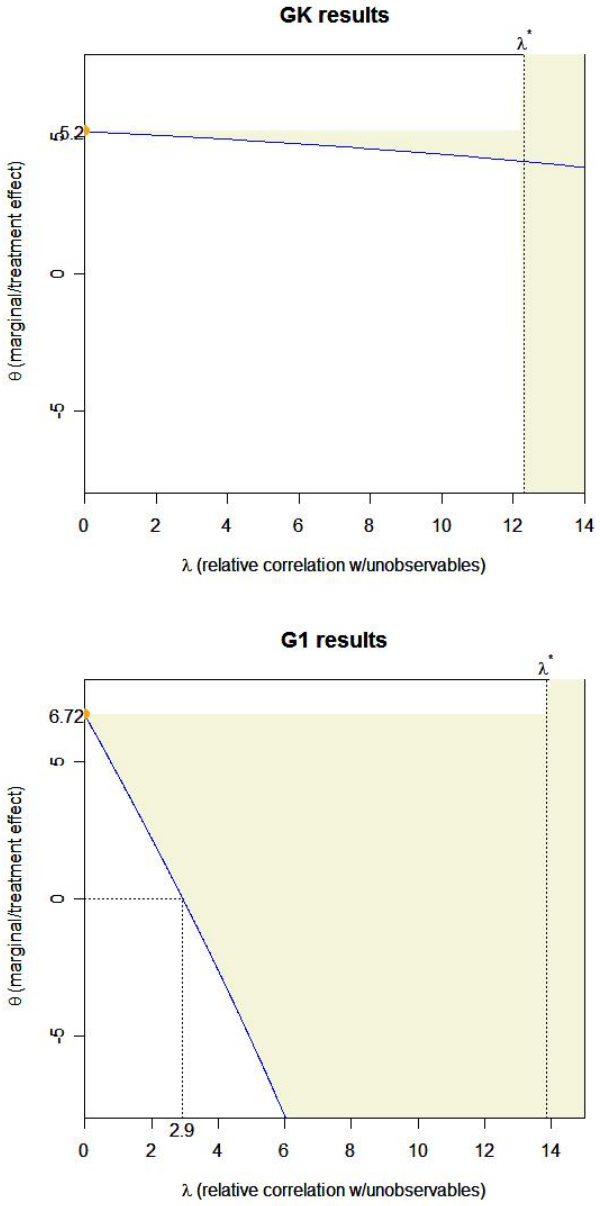

Figure 1: Estimated treatment effect of small class size (θ) on average percentile SAT exam scores for various restrictions on relative selection on unobservables (λ).

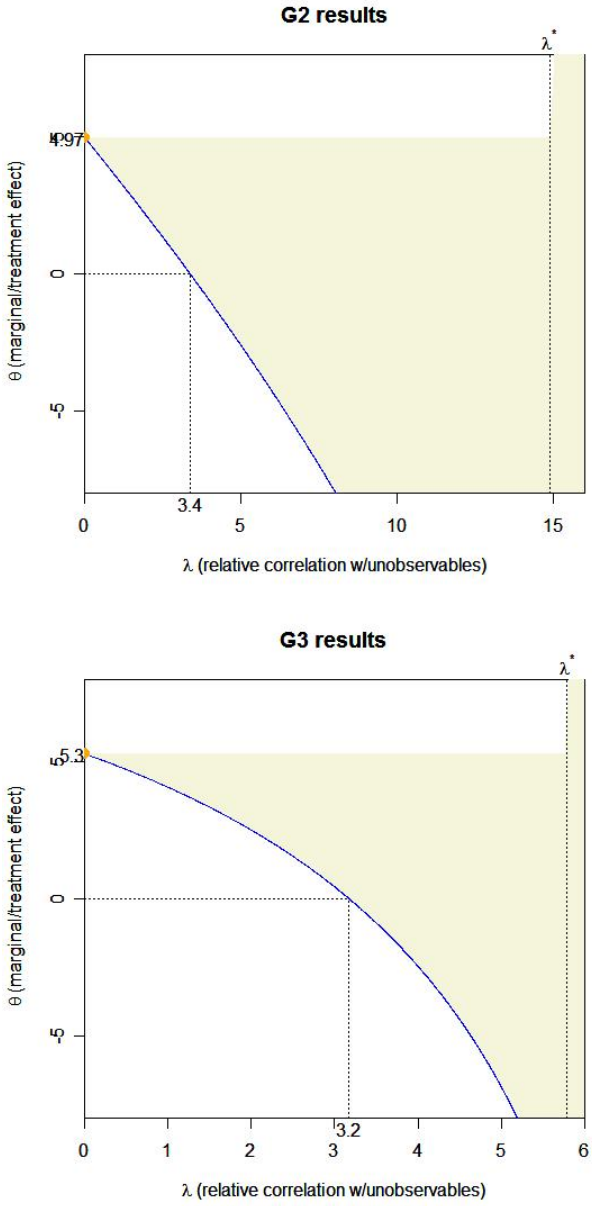

Figure 2: Estimated treatment effect of small class size (θ) on average percentile SAT exam scores for various restrictions on relative selection on unobservables (λ).

To summarize these results, accounting for deviations from random assignment in Project STAR would overturn the primary results only if these deviations affected the distribution of outcome-relevant unobservables across the treatment and control groups much more (by a factor of more than 10 for the kindergarten data, and by a factor of about 3 for the other grades) than it affected the distribution of outcome-relevant observables.

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 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 in-

terest 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.

	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.

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 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 $\Theta(\Lambda)$ as a function of restriction on the relative correlation $\lambda \in \Lambda$.

Figure 3 displays the results from Table 6 graphically. The top graph in the figure shows the results for a wider range of λ values, while the bottom graph shows more detail for a narrower range of λ. In both graphs the line describes $\hat{\lambda}(\theta)$, the shaded area indicates the correspondence $\Theta([0, \lambda])$, and the dot indicates the OLS coefficient of 0.30 (as reported in Table 5.

As the figure and table show, 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 con-

Λ	$\hat{\Theta}(\Lambda)$
$\{0.00\}$	0.30
$[0.00, 0.10]$	$[0.16, 0.30]$
$[0.00, 0.20]$	$[0.03, 0.30]$
$[0.00, 0.30]$	$[-0.10, 0.30]$
$[0.00, 0.40]$	$[-0.24, 0.30]$
$[0.00, 0.50]$	$[-0.38, 0.30]$
$[0.00, 0.75]$	$[-0.73, 0.30]$
$[0.00, 1.00]$	$[-1.09, 0.30]$
$[0.00, 2.00]$	$[-2.71, 0.30]$
$[0.00, 3.00]$	$[-4.70, 0.30]$
$[0.00, 4.00]$	$[-7.37, 0.30]$
$[0.00, 5.00]$	$[-11.83, 0.30]$
$[0.00, 6.00]$	$(-\infty, \infty)$
$[0.00, \infty)$	$(-\infty, \infty)$
$(-\infty, 0.00]$	$[0.30, 17.04]$
$\hat{\lambda}^*$	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).

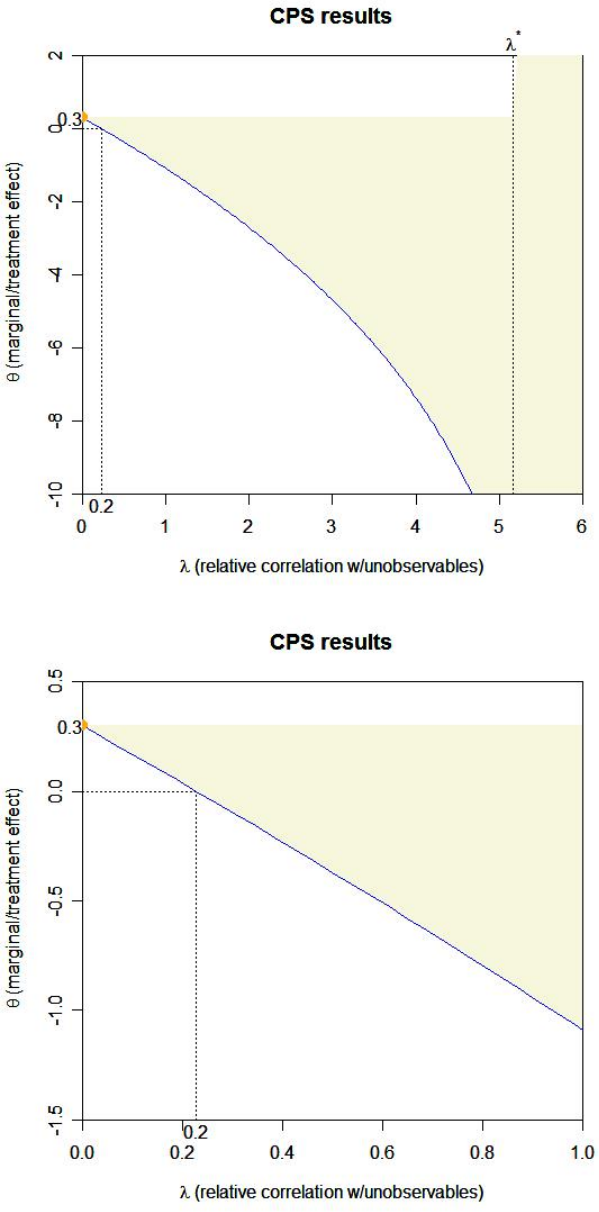

Figure 3: Estimated marginal effect of state-level income inequality on self-rated poor health for a proportional correlation model with varying values of the relative correlation λ.

trol 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 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.

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., "Simulation-based estimation of peer effects," *Journal of Econometrics*, 2006, *133*, 243–271.

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.

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.

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 (incomplete)

A.1 Proposition 1

To prove property (1), note that:

$$\begin{aligned}
\lambda(\theta) &= \frac{\text{corr}\,(z, y-\theta z - P(y-\theta z|X))}{\text{corr}\,(z, P(y-\theta z|X))} \\
&= \frac{\frac{cov(z,y-\theta z-P(y-\theta z|X))}{\sqrt{var(z)var(y-\theta z-P(y-\theta z|X))}}}{\frac{cov(z,P(y-\theta z|X))}{\sqrt{var(z)var(P(y-\theta z|X))}}} \\
&= \frac{cov(z,y)-\theta var(z)-cov(z,P(y|X))+\theta cov(z,P(z|X))}{cov(z,y)-\theta cov(z,P(z|X))} \\
&\times \sqrt{\frac{var(P(y-\theta z|X))}{var(y-\theta z-P(y-\theta z|X))}}
\end{aligned}$$

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

$$\begin{aligned}
\lambda(\theta) &= \frac{cov(z,y)-\theta var(z)-cov(z,P(y|X))+\theta var(P(z|X))}{cov(z,y)-\theta var(P(z|X))} \\
&\times \sqrt{\frac{var(P(y-\theta z|X))}{var(y-\theta z)-var(P(y-\theta z|X))}} \\
&= \frac{cov(z,y)-\theta var(z)-cov(z,P(y|X))+\theta var(P(z|X))}{cov(z,y)-\theta var(P(z|X))} \\
&\times \sqrt{\frac{var(P(y|X))-2\theta cov(P(y|X),P(z|X))+\theta^2 var(P(z|X))}{var(y)-2\theta cov(y,z)+\theta^2 var(z)-var(P(y|X))+2\theta cov(P(y|X),P(z|X))-\theta^2 var(P(z|X))}}
\end{aligned}$$

Taking limits and applying l'Hôpital's rule we get:

$$\begin{aligned}
\lim_{\theta\to\infty}\lambda(\theta) &= \frac{var(z)-var(P(z|X))}{var(P(z|X))} \times \sqrt{\frac{var(P(z|X))}{var(z)-var(P(z|X))}} \\
&= \sqrt{\frac{var(z)-var(P(z|X))}{var(P(z|X))}} \\
&= \sqrt{\frac{1}{R^2_{zx}}-1}
\end{aligned}$$

The same applies for $\lim_{\theta\to-\infty}\lambda(\theta)$

To demonstrate property (2), simply note that the numerator and denominator of $\lambda(\theta)$ are both differentiable in θ, so application of the quotient rule implies differentiability of $\lambda(\theta)$ unless its denominator of is zero. Its denominator is zero if:

$$\begin{aligned}
0 &= cov(z, P(y-\theta z|X)) \\
&= cov(z, P(y|X)) - \theta cov(z, P(z|X))
\end{aligned}$$

Solving for θ, we get property (2). For property (3), we need to show that the numerator of $\lambda(\theta)$ is nonzero when $\theta = \theta^*$:

$$\begin{aligned} cov(z, y - \theta^* z - P(y - \theta^* z|X)) &= cov(z, y) - \theta^* var(z) - cov(z, P(y - \theta^* z|X)) \\ &= cov(z, y) - \frac{cov(z, P(y|X))}{cov(z, P(z|X))} var(z) \end{aligned}$$

So $\lim_{\theta \to \theta^*} |\lambda(\theta)| = \infty$ unless $\frac{cov(z,y)}{var(z)} = \frac{cov(z,P(y|X))}{var(P(z|X))}$.

A.2 Proposition 2

This proposition follows directly from Proposition 1.

A.3 Proposition 3

The first result follows directly from substitution of equation (3) into the definition of $\lambda(.)$. The second result is true by the construction of $\Theta(.)$.

A.4 Proposition 4

The first result follows from the straightforward application of the law of large numbers (implying that the sample averages are consistent estimates of the corresponding population moments) and Slutsky's theorem (since $\lambda(\theta)$ is a continuous function of the population moments for all $\theta \neq \theta^*$).

For the second result, note that $\theta_L(\Lambda)$ is continuous in population moments 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.

A.5 Proposition 5

Note that $\theta_L(\Lambda)$ is differentiable in population moments under these conditions, so the result follows from application of the delta method.