Forming better guesses about neighborhood effects on health: Estimating community effects using conditional modeling of unobservables

Brian Krauth
Simon Fraser University

Presented at Joint Statistical Meetings, Seattle WA
August 9, 2006

Abstract

A recent stream of econometric research (Altonji, Elder, and Taber 2005, Krauth 2006) uses the statistical relationship among the observed explanatory variables in a regression to generate a plausible guess about the relationship between observed and unobserved variables. The usual assumption of orthogonality between observed and unobserved variables is a special case - i.e., a bad guess. This paper extends this approach to the standard linear omitted variables problem, and demonstrates that this approach may be useful for the measurement of community effects on health.

1 Introduction

The empirical analysis of "neighborhood effects" and other social influences on individual behavioral outcomes has a long history in the social sciences. The standard exercise in this literature is to include the characteristics of a person's community[1] as an explanatory variable in an otherwise-standard individual-level regression model. The coefficient on the community-level variable is then taken to be an estimate of the effect of community characteristics on the individual outcome. In recent years, methodological research in economics (Manski 1993) and epidemiology (Oakes 2004) has taken a skeptical view of this interpretation. As these researchers note, nonrandom or "endogenous" community selection will tend to generate nonzero correlation between relevant unobserved characteristics of individuals and the observed characteristics of their communities. Simple regression analysis, which assumes that community characteristics are uncorrelated with or even independent of unobservables, may dramatically overstate the importance of communities.

[1] For the purposes of this paper I will refer to an individual's social group as his or her "community;" the discussion applies generally to a social group of any size or nature.

Applied researchers, particularly in economics, have responded to this critique with a variety of alternative research designs. Most use a design that is based on identifying some plausibly exogenous source of variation in community composition within a given observational data set, then exploiting this exogenous variation through instrumental variables or other techniques to measure the overall effect of community composition. However, there remain many applications of interest where such exogenous sources of variation are difficult if not impossible to find. In these applications, applied researchers face an unappealing choice of either maintaining the indefensible assumption of exogenous community selection (perhaps with careful selection of words to avoid making an explicit causal interpretation of results) or abandoning empirical research on community effects.

One example which will be considered in this paper is the vast literature on the effect of state-level income inequality on health outcomes. Measured income inequality is not easily controlled by policymakers, it changes quite slowly over time, it is measured with a great deal of error, and it affects individual health (if at all) through a complex mechanism that may include lags of many decades. Each of these features militates against the possibility of exploiting exogenous variation in inequality to identify the effect of inequality on health.

This paper explores an alternative approach to the estimation of community effects that may help in such cases. It is based on the idea of accepting that the explanatory variables in a simple regression model are likely to be correlated with outcome-relevant unobservables, and using the statistical relationship of observable explanatory variables to one another as a guide to their relationship with any relevant unobserved variables. This idea has been previously developed by Altonji et al. (2005) in modeling the selection of students into Catholic schooling, and by Krauth (2006a, 2006b) in modeling peer effects in youth smoking. Implementation requires that the econometrician posits an explicit *conditional model of unobservables* (CMU) that places sufficient restrictions on the distribution of unobservables given observables that the structural parameters are identified if the CMU is correctly specified.

The paper is organized into 3 parts. Section 2 reviews the relevant literature, with particular focus on (1) describing the specific solutions used by previous applied researchers to address the primary identification problems in studies of community effects; (2) discussing the limitations on applicability of these solutions in the estimation of community effects on health. Section 3 describes the CMU methodology. Finally, Section 4 provides the application to measuring the health effects of income inequality.

2 Related literature

2.1 Candidate solutions to the endogenous group selection problem

Following the methodological critiques of Manski (1993) and Brock and Durlauf (2000), the development of possible solutions to the problem of endogenous group selection has been an active area of research in economics. Research designs in this literature generally look for an exogenous source of variation in group selection, and leverage this exogenous variation to estimate the overall effect of groups. They can be divided into three basic categories: exogenous community assignment, exogenous cohort assignment, and exogenous individual-level treatment assignment.

Exogenous community assignment designs are those in which at least some aspect of the community assignment is controlled by a central authority that intentionally or accidentally randomizes. Examples include the assignment of first-year college roommates (Sacerdote 2001), the assignment of public housing applicants to buildings or neighborhoods (Katz, Kling and Liebman 2001, Oreopoulos 2003) and the assignment of soldiers to Army bases (Antecol and Cobb-Clark 2006). These studies are able to establish fairly cleanly whether communities affect individual outcomes; however they have two key limitations. First, the requirement of a central authority limits applicability to social interactions within groups that are formally delineated and among sub-populations whose autonomy is limited relative to the society as a whole. Second, if the number of communities is small and those communities differ along many dimensions, identification of a community effect does not imply one can identify which features of a community matter.

Designs based on exogenous cohort assignment often appear in the estimation of educational peer effects (Hoxby 2000, Hanushek, Kain, Markman and Rivkin 2003). Studies in this category usually assume that each school draws from an endogenously determined population of students, while each cohort within the school represents a random draw from that population. For example, a 4th grade class of 20 students drawn from a population with 50% males will sometimes have 9 male students, sometimes 10, etc. As a result, the small year-to-year variation in composition is random conditional on the identity of the school. Cohort-based studies have proved quite fruitful in the study of educational peer effects. They have two key limitations. First, they are limited to community types that are small enough that random selection will induce sufficient heterogeneity: the law of large numbers would eliminate any heterogeneity from random selection at the level of state, city, or even census tract. Second, because community effects are identified from very small cohort-to-cohort variations in composition, extrapolation to the larger variations that are of interest is far more sensitive to both nonlinearity and measurement error than is usually the case.

Designs based on exogenous individual-level treatment assignment take the endogenously selected social group as given, and exploit the

presence of some individual-level treatment that has been assigned randomly, often as part of some experimental design. For example, Duflo and Saez (2003) perform an experiment in which a randomly selected subset of faculty members were given a financial incentive to attend a retirement planning seminar; faculty members with a large proportion of departmental colleagues who received the incentive were more likely both to attend the seminar and to engage in active retirement planning, even after controlling for whether they received the incentive. Boozer and Cacciola (2001) use data from the Project STAR class size experiment; students in later grades with a high proportion of classmates who had received the low-class-size treatment made higher exam scores, even after controlling for the student's own class-size history. This category of studies generally uses an instrumental variables framework to estimate what are called endogenous[2] community effects. This is done by estimating a regression of the individual-level outcome on the individual-level treatment and the average outcome in the community, using the proportion of community members receiving the treatment as an instrument for the average outcome in the community. This design is a promising way to identify endogenous community effects if such effects are of primary interest. However, there are three main limitations. First, it requires the presence of the experimental treatment. Second, and this is a particular problem for applications related to health behavior, the treatment must be effective - completely ineffective interventions will fail to identify the effect at all, while interventions that are only marginally effective will run into the now well-established "weak instruments" problem. Third, this approach is limited to the measurement of endogenous effects rather than contextual effects.

2.2 Community effects on health

Epidemiologists have long been interested in the effects of communities on health-related behavior and health outcomes. One branch of the empirical literature considers neighborhood-level influences, with the average socioeconomic status of the community as the primary neighborhood-level variable. Another branch considers higher levels of aggregation – of aggregation such as counties, cities states or nations – and is most concerned with the effect of income inequality. Diez-Roux (2001) provides a recent survey of the literature on neighborhood effects, while Subramanian and Kawachi (2004) provide a recent survey on the literature on inequality and health. While the mechanisms of interest may vary across these two branches, the methodological issues are much the same.

Endogenous group selection is often acknowledged as an issue in this literature, but is given far less prominence than it receives in the related economics literature. Generally speaking, the possibility of endogenous

[2]In the terminology of Manski (1993), endogenous social effects are present when the behavior/outcome of an individual is affected by the distribution of that behavior/outcome in the social group, while contextual effects are present when the the behavior/outcome of an individual is affected by the distribution of background characteristics within the social group.

selection is acknowledged in some but not all papers, and is addressed primarily by adding more variables to the regression in hopes of reducing omitted variables bias. However, while adding variables might help reduce the bias in effect estimates, it does not generally solve the problem. To use the language of economists, note that *any* individual characteristic that affects individual health will also affect the marginal (utility) benefit associated with a community-induced change in health. As a result, if communities influence health and are the result of individual choices, every variable that affects the outcome has at least some effect on the choice of community.

Oakes (2004) argues that the failure of researchers in social epidemiology to more seriously address community selection implies their resulting estimates "will always be wrong" (p. 1941). In Oakes' view, much of the lack of attention to community selection and other identification issues is because researchers have placed excessive priority on the deployment of elaborate multilevel models. An alternative explanation is that the existing methods for addressing community selection outlined in Section 2.1 above are simply unsuitable for the particular questions of interest to social epidemiologists. One general argument 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.

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.

To an extent, the news here is not good: the prospects for point identification of the true causal effect of (for example) income inequality on health are poor. Fortunately, as Diez Roux's comment above indicates, it is not clear that this is necessary. Estimates of the effect of inequality on health are not of policy interest because there exists a policymaker somewhere with the ability to directly set inequality to some optimal level. Instead, the ultimate question is whether there is one more reason – beyond the obvious ones – to worry about inequality and to aim for its reduction. For this purpose a good guess, or range of guesses, about the relative magnitude of any effect is sufficient. Section 3 below does not describe a complex research design aimed at point identification of the causal effect, but rather outlines one method for generating improved guesses from a simple research design.

3 Methodology

This section outlines a simple linear model with omitted variables, defines a conditional model of unobservables, and establishes the basic statistical properties of model estimates using CMUs. Note that the model presented in this paper is linear, and has no explicit multilevel structure. This simplicity helps to keep the fundamental features of the approach more transparent. Extension of the approach to particular nonlinear models, or to multilevel models, is conceptually straightforward but involves some application-driven modeling choices See Altonji et al. (2005) or Krauth (2006a, 2006b) for examples.

3.1 Structural model and conditional model of unobservables

Let the structural/causal model of interest be given by:

$$\begin{aligned} Y &= y(X,U) \\ &= X\beta + U\delta \\ &= X\beta + u \qquad (\text{where } u \equiv U\delta) \end{aligned} \tag{1}$$

where the scalar outcome variable Y is a function $y(X,U)$ of the K_X-length vector X of characteristics that are observed in the data and K_U-length vector U of unobserved characteristics. Note that $y(.)$ is to be interpreted here as a function giving the individual's counterfactual outcome for given values of X and U, while Y is the realized outcome for the individual under his or her realized values for X and U. For this application $y(.)$ is assumed to be linear, so the parameter vectors β and δ represent true marginal effects, i.e., a one-unit change in X produces a β-unit change in Y. To simplify notation and without loss

of generality, all variables are normalized to have a mean of zero and a variance of one. As usual, it is also assumed that:

$$V \equiv E(X'X) \text{ is nonsingular} \tag{2}$$

The data takes the form of a sample $\{y_i, x_i\}_{i=1}^n$ from the population (Y, X), with some sampling design that can be used to construct consistent estimates of the population moments $E(X'X)$ and $E(X'y)$. A random sample is the most straightforward example, but stratified and cluster samples will usually meet this condition as well.

Identification of β generally requires the imposition of strong restrictions on u. The standard identifying assumption employed by researchers is exogeneity: $E(u|X) = 0$ or $E(X'u) = 0$. Exogeneity can be thought of as a special case of a more general category of *conditional models of unobservables* (CMU's). Let a CMU be defined as a known function $\rho(.)$ such that:

$$\begin{aligned} E(X'u) &= \sigma_u \rho(V, \beta, \theta) \qquad \text{or equivalently,} \\ corr(X, u) &= \rho(V, \beta, \theta) \end{aligned} \tag{3}$$

where $\sigma_u^2 = var(u)$ and θ is some parameter fixed by the econometrician (i.e., not to be estimated). The usual exogeneity assumption corresponds to the special case $\rho(V, \beta, \theta) = 0$.

The idea behind use of a CMU is that we do not know anything directly about the correlation between the observable and unobservable variables. However, we might be willing to suppose that the correlation between observable and unobservable explanatory variables is in some way similar to the correlation among observable explanatory variables. As a result, the $\rho(.)$ function takes three arguments: the covariance matrix of observables X, the vector of structural parameters β, and some fixed parameter θ. Section 3.3 describes an example CMU.

3.2 Estimation and identification

Estimation follows a simple two-step procedure. The first step is to estimate a linear regression of Y on X using ordinary least squares (or weighted least squares under stratified sampling). Let the reduced form parameters be given by:

$$\begin{aligned} V &\equiv E(X'X) \\ b &\equiv V^{-1}E(X'y) \\ \sigma_\epsilon^2 &\equiv E((y - Xb)^2) \end{aligned} \tag{4}$$

and let $(\hat{V}, \hat{b}, \hat{\sigma}_\epsilon^2)$ be consistent estimators:

$$\hat{V} \equiv \frac{1}{n}\sum_{i=1}^n w_i x_i' x_i \quad \text{where} \quad \text{plim } \hat{V} = V \tag{5}$$

$$\hat{b} \equiv \hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n w_i x_i' y_i \quad \text{where} \quad \text{plim } \hat{b} = b$$

$$\hat{\sigma}_\epsilon^2 \equiv \frac{1}{n - K_X}\sum_{i=1}^n w_i \left(y_i - x_i\hat{b}\right)^2 \quad \text{where} \quad \text{plim } \hat{\sigma}_\epsilon^2 = \sigma_\epsilon^2$$

where the weights w_i are equal to one for a random sample, and are the appropriate sampling weights for stratified samples.

As a caution, one might be tempted in applications with clustering or substantial heteroskedasticity to replace the OLS estimator of b in (5) with a lower variance generalized least squares (GLS/FGLS) estimator. However, it must be noted that equation (1) does not in general imply linearity of $E(y|X)$. One implication of this is that OLS will consistently estimate b under random sampling, but generalized least squares (GLS/FGLS) will not.

Once the reduced form parameters have been estimated, the second step is to solve for estimates of the structural parameters based on these reduced form parameters. Proposition 1 below describes the procedure and its statistical properties.

Proposition 1 *Let (1)-(5) hold. Then:*

1. *The reduced form parameters are related to the structural parameters by:*

$$\begin{bmatrix} b \\ \sigma_\epsilon^2 \end{bmatrix} = \begin{bmatrix} \beta \\ \sigma_u^2 \end{bmatrix} + \begin{bmatrix} \sigma_u V^{-1} \rho(V, \beta, \theta) \\ -\sigma_u^2 \rho(V, \beta, \theta)' V^{-1} \rho(V, \beta, \theta) \end{bmatrix} \quad (6)$$

2. *Suppose equation (6) has a unique and continuous solution in a neighborhood of the true parameter values. Let $(\hat{\beta}, \hat{\sigma}_u^2)$ solve:*

$$\begin{bmatrix} \hat{b} \\ \hat{\sigma}_\epsilon^2 \end{bmatrix} - \begin{bmatrix} \hat{\beta} \\ \hat{\sigma}_u^2 \end{bmatrix} - \begin{bmatrix} \sigma_u \hat{V}^{-1} \rho(\hat{V}, \hat{\beta}, \theta) \\ -\sigma_u^2 \rho(\hat{V}, \hat{\beta}, \theta)' \hat{V}^{-1} \rho(\hat{V}, \hat{\beta}, \theta) \end{bmatrix} = 0 \quad (7)$$

   Then $\hat{\beta}$ is a consistent estimator of β.

Proof: See Appendix.

Asymptotic standard errors and test statistics for $\hat{\beta}$ can be constructed using application of the delta method to the reduced form standard errors and other statistics. The estimated covariance matrix for $\hat{\beta}$ can be made robust to arbitrary heteroskedasticity, within-cluster correlation, etc. by using the appropriate covariance matrix estimator for the reduced form.

3.3 Example: Proportional correlation

The *proportional correlation* CMU is based on the idea that the correlation among unobservables and between observables and unobservables should roughly approximate that among observables. We start with a proportionality parameter θ. The simple proportional correlation model is:

$$corr(X_k, u) = \theta corr(X_k, X\beta - X_k\beta_k) \qquad \forall k \in 1, \ldots, K_X \quad (8)$$

The proportionality parameter θ can be interpreted as the correlation between X_k and unobservable factors (weighted by their contribution to the outcome) relative to the correlation between X_k and observable factors (weighted by their contribution to the outcome). The standard exogeneity assumption can be considered a special case ($\theta = 0$) of proportional correlation. Deriving the exact form of the $\rho(.)$ function

implied by equation (8) is straightforward if tedious algebra. Note that in order for the right side of (8) to be well-defined, we must have data on at least two explanatory variables with nonzero effects on the outcome.

A generalization of the proportional correlation CMU is to have proportional correlation within and across particular categories of variables. For example, suppose that our data come from a multilevel structure; some variables are measured at the individual level while others are measured at a community level. In this case we might suppose:

$$\begin{aligned} u &= u^{group} + u^{ind} \qquad (9) \\ corr(X_k, u^{group}) &= \theta corr\left(X_k, \sum_{j \neq k} X_j \beta_j I\left[X_j \text{ is group-level}\right]\right) \\ corr(X_k, u^{indiv}) &= \theta corr\left(X_k, \sum_{j \neq k} X_j \beta_j I\left[X_j \text{ is individual-level}\right]\right) \\ &\text{etc.} \end{aligned}$$

In other words, the correlation between each observed level-A variable and the unobserved level-B variables (weighted by their effect on the outcome) is proportional to the correlation between that variable and the observed level-B variables (weighted by their effect on the outcome, and excluding the original variable if $A = B$).

Another example is where our data feature a binary treatment (T) along with some individual-level covariates (X) included to address non-experimental selection into treatment. In this case we might adjust the original model:

$$Y = X\beta + \gamma T + u \qquad (10)$$

and assume:

$$\begin{aligned} corr(T, u) &= \theta corr(T, X\beta) \qquad (11) \\ corr(X_k, u) &= \theta corr(X_k, X\beta - X_k\beta_k) \end{aligned}$$

In other words the correlation between the treatment and the (effect-weighted) unobservables is proportional to the correlation between the treatment and the effect-weighted observables.

4 Application: Inequality and self-reported health

An empirical literature relating income inequality to negative health outcomes dates back to Rodgers (1979). While much of the early research on the subject relied on aggregate data, more recent work has used linked individual-aggregate data to avoid some of the limitations of ecological inference. While many of these studies exhibit a great deal of methodological sophistication and complexity, almost none have done much to address the issue of endogenous community selection. For example, none of the 21 studies cited in Subramanian and Kawachi's

recent review article (2004) have a research design aimed at addressing endogenous selection. The application in this section aims to make progress in this direction.

4.1 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 the social determinants of 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 1 reports unweighted summary statistics.

4.2 Results under assumption of exogeneity

Table 2 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 2 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 2 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

Variable	Mean	Std. Dev
Individual-level characteristics:		
Self-reported poor health	0.15	0.35
Log income	10.03	0.88
Age, years	44.9	17.49
Female	0.53	0.50
Black	0.09	0.29
Asian/Other	0.05	0.21
Education, years	12.73	2.71
Not employed	0.36	0.48
No health insurance	0.21	0.41
State-level characteristics:		
Gini coefficient of household income	0.43	0.02
Number of individuals	187,760	
Number of states (incl DC)	51	

Table 1: Summary statistics for linked CPS-Census 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 2: Regression results for model with assumption of exogeneity ($\theta = 0$). Linear model estimated using OLS, with cluster-robust standard errors. Logistic model estimated as random-intercept multilevel model with maximum likelihood.

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 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.

4.3 Results under assumption of proportional correlation

The estimates reported in Table 2 are based on models in which exogeneity is assumed. As discussed in Section 3, this is a strong and somewhat indefensible assumption. This section uses the CMU approach to consider the effect of deviations from exogeneity on estimates of the effect of inequality on health.

The model to be estimated is the linear model (2) from Table 2. The conditional model of unobservables to be used will be the simple proportional correlation model defined in equation (8). The relative correlation parameter θ will be varied from $\theta = 0$ (exogeneity) to $\theta = 1$ (equal correlation among observables and unobservables).

Table 3 reports the estimated coefficient on inequality as a function of the relative correlation θ. Cluster-robust standard errors are also reported. Figure 1 displays the results from Table 3 graphically. The solid line in the figure depicts point estimates, while the dashed lines represent pointwise 95% asymptotic confidence bands.

As the figure and table show, increases in θ from the benchmark case of exogeneity are generally associated with decreases in the estimated marginal effect of inequality, though the relationship is not entirely monotonic. A relative correlation of 15% or greater (i.e., $\theta > 0.15$) implies that the marginal effect of inequality is no longer statistically significant at conventional levels. The decline in the point estimate is relatively moderate for relative correlations of 35% or less, but becomes much larger above that point. For relative correlations above 35%, the point estimates are usually negative, implying that greater inequality is associated with better health.

There are two related but distinct approaches to interpreting the results in Figure 1. One is as a sensitivity analysis of the benchmark results in Table 2. By this approach, Figure 1 shows that the point estimate is not sensitive to small deviations from exogeneity, but is affected by larger deviations. Statistical significance of the parameter

Relative Correlation	Estimated marginal Effect
0.0 (exogeneity)	0.299
	(0.122)
0.1	0.247
	(0.123)
0.2	0.196
	(0.125)
0.3	0.179
	(0.131)
0.4	-0.194
	(0.143)
0.5	0.072
	(0.126)
0.6	-0.138
	(0.146)
0.7	-0.451
	(0.149)
0.8	-0.728
	(0.204)
0.9	-1.086
	(0.155)
1.0	-1.397
	(0.158)

Table 3: Regression results for model with proportional correlation. Cluster-robust standard errors in parentheses.

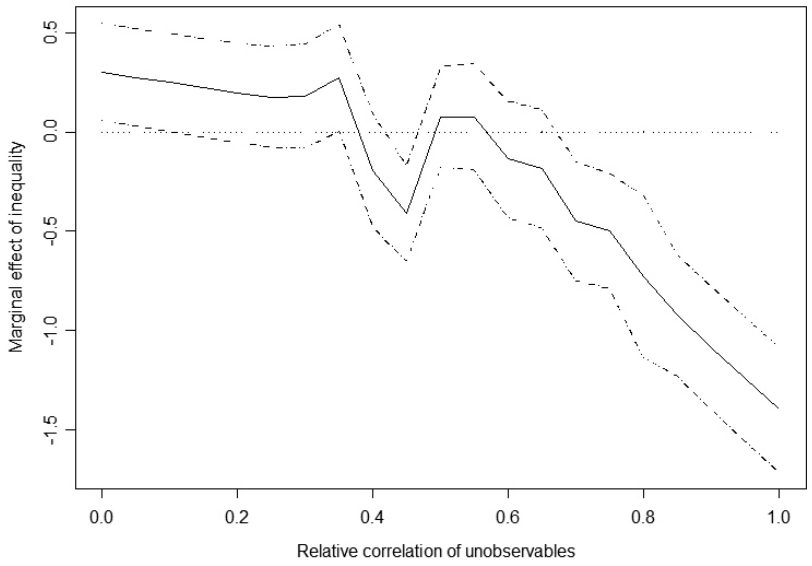

Figure 1: Estimated marginal effect of state-level income inequality on self-rated poor health for a proportional correlation model with varying values of the relative correlation θ.

estimate is sensitive to small deviations from exogeneity, implying that the data only provide strong evidence for a positive relationship between inequality and poor health if one believes the relative correlation to be 15% or lower.

A related but distinct approach is to see Figure 1 as providing a means for readers to construct bounds on the effect of inequality based on the bounds they are willing to place on the relative correlation. For example, if one is willing to assume that $\theta \in [0, 0.3]$ then (ignoring sampling error for the moment) the data imply $\beta_{ineq} \in [0.179, 0.299]$. If a more conservative reader is only willing to assume that $\theta \in [0, 1]$, then the data imply $\beta_{ineq} \in [-1.397, 0.299]$. One can also choose a point estimate based on one's best guess about θ. Note that in this interpretation, $\theta = 0$ has no privileged position except possibly as a plausible lower bound.

5 Conclusion

This paper has demonstrated a simple application of conditional modeling of unobservables to the problem of estimating the effect of state-level inequality on the individual-level probability of self-reported poor health. The results indicate that inequality has a negative effect on health if the correlation between observables and unobservables is less than 40% as large as the correlation among observables. Otherwise, inequality appears to have a positive effect on health, all else being equal.

The development of practical research designs based on conditional modeling of unobservables is still in an early phase. One area remaining for future research is the question of which CMU's are plausible. For example, Altonji et al. (2005) show that an equal-correlation model (i.e., the proportional correlation model described here with $\theta = 1$) holds in expected value if the explanatory variables are selected randomly from a large set of candidate variables. Other models of variable selection may yield different implications. Another area for future work is in the extension of the approach to models with more specific causal foundations and less restrictive functional form assumptions. This paper uses a linear model, while Krauth (2006a, 2006b) applies the method to an equilibrium discrete choice model, and Altonji et al. (2005) use a Heckman-type parametric selection model. Extensions to semiparametric methods such as propensity score matching represent a promising next step.

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.

Antecol, Heather and Deborah Cobb-Clark, "Identity and racial harassment," Working Paper, Claremont-McKenna College 2006.

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.

Boozer, Michael A. and Stephen E. Cacciola, "Inside the black box of Project STAR: Estimation of peer effects using experimental data," Discussion Paper 832, Yale University Economic Growth Center 2001.

Brock, William A. and Steven N. Durlauf, "Interactions-based models," in James J. Heckman and Edward Leamer, eds., *Handbook of Econometrics, Volume 5*, North-Holland, 2000.

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

Duflo, Esther and Emmanuel Saez, "The role of information and social interactions in a retirement plan decision: Evidence from a randomized experiment," *Quarterly Journal of Economics*, 2003, *118*, 815–842.

Hanushek, Eric A., John F. Kain, Jacob M. Markman, and Steven G. Rivkin, "Does peer ability affect student achievement?," *Journal of Applied Econometrics*, 2003, *18* (5), 527–544.

Hoxby, Caroline M., "Peer effects in the classroom: Learning from gender and race variation," Working Paper 7867, NBER 2000.

Katz, Lawrence F., Jeffrey R. Kling, and Jeffrey B. Liebman, "Moving to Opportunity in Boston: Early results of a randomized mobility experiment," *Quarterly Journal of Economics*, 2001, *116* (2), 607–654.

Krauth, Brian V., "Peer effects and selection effects on youth smoking in California," *Journal of Business and Economic Statistics*, 2006. Forthcoming.

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

Manski, Charles F., "Identification of endogenous social effects: The reflection problem," *Review of Economic Studies*, 1993, *60* (3), 531–542.

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.

Oreopoulos, Phillip, "The long-run consequences of growing up in a

poor neighborhood," *Quarterly Journal of Economics*, 2003, *118* (4), 1533–1575.

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.

Sacerdote, Bruce, "Peer effects with random assignment: Results for Dartmouth roommates," *Quarterly Journal of Economics*, May 2001, *116* (2), 681–704.

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. Available at `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. Available at `http://www.bls.census.gov/cps/cpsmain.htm`.

A Proofs

A.1 Proposition 1

Let $L(.|.)$ be the linear projection operator:

$$L(y|X) = X\beta + L(u|X) \tag{12}$$

which implies that:

$$\begin{aligned} b &= \beta + E\left(X'X\right)^{-1} E(X'u) \\ &= \beta + \sigma_u V^{-1}\rho(V,\beta,\theta) \end{aligned} \tag{13}$$

Equation (13) gives a fairly standard omitted variables bias formula. Next, let $\epsilon \equiv y - L(y|X)$ be the reduced form residual and let $u \equiv U\delta$ be the structural residual.

$$\begin{aligned} \epsilon &= y - L(y|X) \\ &= (X\beta + u) - (X\beta + L(u|X)) \\ &= u - L(u|X) \\ &= u - \sigma_u X V^{-1}\rho(V,\beta,\theta) \end{aligned} \tag{14}$$

Now, because the projection residual is orthogonal to the projection, we have:

$$\begin{aligned} \sigma_u^2 &\equiv var(u) \\ &= var(u - L(u|X)) + var(L(u|X)) + 2cov(u - L(u|X), L(u|X)) \\ &= var(\epsilon) + var(L(u|X)) + 0 \\ &= \sigma_\epsilon^2 + E\left(\left(\sigma_u X V^{-1}\rho(V,\beta,\theta)\right)'\left(\sigma_u X V^{-1}\rho(V,\beta,\theta)\right)\right) \\ &= \sigma_\epsilon^2 + \sigma_u^2 E\left(\rho(V,\beta,\theta)'V^{-1}X'XV^{-1}\rho(V,\beta,\theta)\right) \\ &= \sigma_\epsilon^2 + \sigma_u^2\rho(V,\beta,\theta)'V^{-1}\rho(V,\beta,\theta) \end{aligned} \tag{15}$$

Rearranging equations (13) and (15) yields equation (6). Part 2 of the proposition follows from application of Slutsky's theorem.