

Estimating the Causal Effects of Education on Wage Inequality Using IV Methods and Sample Selection Models

Stacey H. Chen, SUNY at Albany*
Shakeeb Khan, Duke University

February 15, 2007

Abstract

We develop semiparametric and instrumental-variables approaches to self-selection problems in comparisons of wage inequality between educational groups. We propose new estimators to identify the causal effect of college education on scale parameters of wage distributions, using symmetry conditions on the joint distribution of outcome and selection errors, along with kernel weighting schemes. A simulation study indicates that the proposed estimators perform well in finite samples. We illustrate these methods with a well-known study by Card (1995), using college proximity as an instrument for schooling. Estimation results suggest that college education significantly increased the degree of wage inequality in 1976.

JEL Classification: C24, C14, C13.

Key Words: schooling choice, average treatment effect, quantile regression, instrumental variable method, scale ratio.

*Corresponding author: Stacey H. Chen. Assistant Professor in SUNY at Albany. Mailing address: 412 NBER, 1050 Massachusetts Ave., Cambridge, MA 02138. E-mail: schen@albany.edu. Thanks to Alberta Abadie, Josh Angrist, Bernd Fitzenberger, Ed Vytlačil, and three anonymous referees for valuable comments and suggestions. Thanks also to seminar participants at Gothe University, Queens University, Syracuse University, the 2005 Conference of Econometric Evaluation of Public Policy in Paris, and the 2005 Econometric Society World Congress in London.

1 Introduction

One of the most important topics in labor economics is the relationship between education and inequality. A substantial part of increasing inequality in the U.S. since the 1980s can be explained by the increased returns to education. At least as important are the growing gap in within-group inequality between educational levels. While wage inequality among college graduates have increased since the 1980s, within-group inequality among high school graduates has exhibited an inverted-U shape, rising sharply during the 1980s but falling since the 1990s. This phenomenon has been explored by Autor, Katz and Kearney (2006) and Lemieux (2006) using the Current Population Survey, and by Angrist, Chernozhukov and Fernandez-Val (2006) using the U.S. Census. Since the gap in within-group inequality between college and high school is widening and the workforce is more and more educated, the current research attention has turned to determining whether the recent growth in inequality can be explained by the increasing importance of education in terms of skilled-biased technical changes (see, e.g., Acemoglu 2002).

Most of the studies in this area take educational choices as exogenous. However, if education is endogenous – correlated with an individual’s ability or taste, then the observed differences in wage inequality across education levels can be a result of self selection, not necessarily caused by the choice itself. Identification of the effect of education on inequality is challenging because it involves inferences about *potential wages* for different schooling levels among each person in the sample. Potential wages of high school graduates, for example, would not be observed if they had decided to attend college. Heckman and Honore (1990) and Acemoglu (2002) have suggested that self-selected education truncates wage distributions, leading the inequality of observed wages to understate the degree of potential wage inequality at each level of schooling. Thus the growing difference in within-group inequality of observed wages between college and high school does not necessarily capture the increasing importance of education in rising inequality of potential wages.

Existing empirical evidence for this theory is built on Roy’s selection models, assuming that outcome and selection errors are normally distributed. Examples include Heckman and Sedlacek (1985), Gould (2002), and Chen (2006). The normality assumption has been considered necessary for identification of Roy’s selection model since Heckman and Honore

(1990). This paper also uses Roy's model but replaces the normality condition with alternative assumptions to address the problems of self selection. We develop and implement new estimators to identify the effect of college education on wage inequality, measured by the *ratio of scale parameters* of wage distributions between an economy with all college-educated workers and an economy with no college-educated workers. The estimators can also be applied to inequality comparisons by other dichotomous causal variables, such as union-nonunion, skilled-unskilled, and urban-rural indicators.

A large theoretical and empirical literature that addresses endogeneity problems focuses primarily on mean effects (e.g., the economic return to education or the schooling coefficient in wage regressions). Semiparametric estimation methods for mean effects have been proposed by Heckman (1990), Ahn and Powell (1993), and Donald (1995). Card (1999) provides an extensive review of related empirical results. More recently, progress has been made on semiparametric estimation of distributional effects. Examples include Abadie's (2002, 2003) instrumental variables (IV) methods and Abadie, Angrist and Imbens' (2002) quantile treatment effects (QTE) estimators. Both methods are built on the *local average treatment effect* (LATE) framework, developed by Imbens and Angrist (1994) and Angrist, Imbens and Rubin (1996). Also, Chernozhukov and Hansen's (2001, 2004, 2005) IV strategy for the estimation of causal effects on quantiles allows for multinomial treatments using different distributional assumptions from the LATE framework's.

The existing IV results under the LATE framework may not be very satisfying for empirical studies on inequality comparisons between educational groups, however, since identification comes from an instrument that induces exogenous selection into education for a specific subpopulation. The resulting LATE estimates, therefore, provide no representative interpretations especially for those whose educational choices are unaffected by the instrumental variable. In contrast, general interest of the inequality literature is concerning wage distributions in a random sample of workers who may or may not be affected by a given instrument. To link LATE results with the average treatment effect (ATE) for the entire population or for a covariate-specific subpopulation, Heckman and Vytlacil (2000) and Angrist (2004) have suggested the use of symmetry conditions. Under the assumption of symmetry, the LATE estimators of *mean effects* can identify the population effects of treatment on mean. How-

ever, the population effect or the ATE of education on *inequality* or *scale parameters* remains unidentified by LATE methods, even with the symmetry conditions, as this paper will point out.

Our initial IV approach to identification of the ATE of education on inequality (that is, the scale ratio for college versus high school groups) is a modification of Abadie's (2002, 2003) LATE estimators. We find sufficient conditions under which the scale ratio can be identified by the variance ratio using Abadie's LATE estimators. The second and third approaches invoke symmetry conditions to develop a *pairwise-matching* method. The pairwise-matching method is designed to find *symmetric pairs* – who satisfy symmetry conditions – from the treated and the untreated groups. We show how nuisance terms among symmetric pairs can be eliminated based on their propensity scores. Previous work by Songnian Chen (1999) has used symmetry conditions and pairwise matching to identify the population effect on *means* but the population effect on *scales* is not his parameter of interest.

In addition to outlining asymptotic properties and presenting the results of a small Monte Carlo study, we apply these estimators to estimate the causal link between college education and wage inequality. We replicate empirical results by Card (1995) and Kling (2001), using the same data and the same instrumental variable – an indicator for proximity to a four-year college 10 years before the survey year of 1976. Distance instruments of this sort have also been used by others, including Kane and Rouse (1993, 1995), Rouse (1995), Currie and Moretti (2003), and Cameron and Taber (2004). Our pairwise-matching estimates suggest a significant gap in potential wage inequality in 1976 between the college and high school groups. The results also show that the proposed IV estimates are very imprecise due to the limitation of sample size. The other two estimators, in contrast, are much more precise because pairwise matching generates many symmetric pairs even with small samples.

The paper is organized as follows: Section 2 introduces the basic model and establishes the key identification conditions; Section 3 describes the proposed estimators and corresponding procedures; Section 4 characterizes the asymptotic distributions of the estimators; Section 5 explores the finite sample properties of the estimators through a small-scale simulation study; Section 6 applies the new methods to estimate the effect of college education on potential wage inequality; and Section 7 concludes.

2 Identification

2.1 The Identification Problem

Suppose that we are interested in the effect of *treatment*, say college education, represented by a binary variable d_i , on the degree of inequality of a continuously distributed outcome variable y_i , such as log wages. For convenience of exposition, we consider a case with a binary choice although the model and the identification strategies can be generalized to cases with multiple choices. As in Imbens and Angrist’s (1994) LATE framework, the causal relationship between the treatment and the outcome is characterized by *potential outcomes*, y_{1i} and y_{0i} , which an individual would obtain with or without being exposed to the treatment. In the application to the causal effect of college education on wage inequality, y_1 is the potential log wage (or briefly “potential wage”) of a college-educated worker while y_0 is the potential wage of a non-college-educated worker.

Suppose that potential outcomes are determined by a location-scale-shift model:

$$y_{1i} = \mu_1(x_i) + \sigma_1(x_i)\epsilon_{1i}, \quad (2.1)$$

$$y_{0i} = \mu_0(x_i) + \sigma_0(x_i)\epsilon_{0i}, \quad (2.2)$$

where x_i is a set of exogenous *covariates*; $\mu_d(x_i)$ and $\sigma_d(x_i)$ are the *conditional location* and the *conditional scale* of potential outcomes by treatment status $d_i = d$ for given covariates x_i . In the inequality literature, conditional scale parameters are also referred to as *within-group inequality* among individuals with the same covariate values. Our parameter of interest - the ATE on within-group inequality - is $\sigma_1(x)/\sigma_0(x)$ the *scale ratio* for a covariate-specific subpopulation with $x_i = x$. Unlike the LATE literature, which focuses extensively on identification of mean effects for individuals who are affected by the instrumental variable, our goal of identification is the average treatment effect on inequality for randomly assigned individuals with certain covariates x_i . In addition, we note that although nonparametric, the additive and multiplicative separability we impose does impose restrictions on the conditional distributions of the potential outcome variables. For example, it imposes that ratios conditional interquantile spreads be independent of the regressors.

Identification of the scale ratio is challenging because it requires knowledge of potential wage distributions for each treatment status but we cannot observe both of the potential

outcomes for the same individual. In particular, comparison of variance in observed wages does not necessarily provide a correct answer:

$$\frac{Var[y_i|x_i = x, d_i = 1]}{Var[y_j|x_j = x, d_j = 0]} = \frac{\sigma_1^2(x)Var[\epsilon_{1i}|d_i = 1]}{\sigma_0^2(x)Var[\epsilon_{0j}|d_j = 0]} \neq \frac{\sigma_1^2(x)}{\sigma_0^2(x)}$$

because $Var[\epsilon_{1i}|d_i = 1]$ may not equal $Var[\epsilon_{0j}|d_j = 0]$ if the distribution of ϵ_{di} depends on treatment status d_i .

2.2 Identification by Instrumental Variables Using Compliers

Our first approach builds on LATE methods developed by Imbens and Angrist (1994), Abadie (2002, 2003) and Angrist (2004). The key identification strategy is to use an *instrumental variable* z_i to induce exogenous variation in treatment status. The dependence between the instrument and the treatment status is recognized by *potential treatment* indicator d_{zi} given $z_i = z$. For instance, z_i can be an indicator of college proximity as an instrument for education in wage regressions. Individuals with $d_{1i} > d_{0i}$ (or equivalently, $d_{0i} = 0$ and $d_{1i} = 1$) are called *compliers*, who would attend college if living nearby a college at the end of high school but would not attend otherwise. Although our identification methods proposed below can be generalized to cases with multi-valued instruments, we focus on a case with a binary instrument for convenience of exposition.

Vytlacil (2002) has shown that the required identification conditions in the LATE framework are equivalent to the *latent-index assignment* model, where treatment status d_i is determined by an index crossing a threshold:

$$d_i = I\{\eta_i \leq m(x_i, z_i)\}, \tag{2.3}$$

where $I\{\cdot\}$ is the indicator function. The normalized *latent index* η_i is independent of instrument z_i conditional on covariates x_i . The *selection index function* $m(x_i, z_i)$ is measurable but otherwise unspecified. The potential treatment indicator can be rewritten as

$$d_{1i} = I\{\eta_i \leq m(x_i, 1)\}, \quad d_{0i} = I\{\eta_i \leq m(x_i, 0)\}.$$

Thus compliers are those with $m(x, 0) < \eta \leq m(x, 1)$.

This equivalence result suggests that the identification conditions of the LATE approach can be written in terms of the latent index η_i and the selection index $m(x_i, z_i)$ as follows:

Assumption 1

LATE Assumptions:

(i) INDEPENDENCE: $(\epsilon_{1i}, \epsilon_{0i}, \eta_i)$ is independent of z_i conditional on x_i .

(ii) FIRST STAGE: $0 < \Pr\{z_i = 1|x_i\} < 1$ and $\Pr\{\eta_i > m(x_i, z_i)|x_i\} < \Pr\{\eta_i \leq m(x_i, z_i)|x_i\}$.

(iii) MONOTONICITY: $\Pr\{m(x_i, 0) \leq m(x_i, 1)|x_i\} = 1$.

Using these assumptions, Imbens and Angrist (1994) have provided an estimator of the average treatment effect on the means of potential outcomes for compliers; that is, $E[y_1 - y_0|d_1 > d_0]$. Abadie (2002, 2003) extended this to a case with arbitrary functions of potential outcomes, $h(y_1)$ and $h(y_0)$. His method suggests that the estimators of the conditional expectation functions of $h(y_1)$ and $h(y_0)$ can be written as follows:

$$\hat{E}[h(y_1)|d_1 > d_0, x] = \frac{\hat{E}[h(y)d(z - \hat{z}(x))|x]}{\hat{E}[d(z - \hat{z}(x))|x]}, \quad (2.4)$$

$$\hat{E}[h(y_0)|d_1 > d_0, x] = \frac{\hat{E}[h(y)(1 - d)(z - \hat{z}(x))|x]}{\hat{E}[(1 - d)(z - \hat{z}(x))|x]}, \quad (2.5)$$

where \hat{E} is a nonparametric estimator of a conditional expectation function, and $\hat{z}(x)$ denotes an estimator of $E[z_i|x_i = x]$, i.e.

$$\hat{z}(x) = \frac{\sum_{i=1}^n z_i K(x_i - x)}{\sum_{i=1}^n K(x_i - x)}$$

with $K(\cdot)$ be a kernel function. For the problem studied in this paper, working with first and second moments, i.e. $h(y) = y$ and $h(y) = y^2$, one can estimate the variance of potential outcomes for each treatment status conditional on the group of compliers.

When $h(y)$ equals y , Abadie's approach estimates the average y_1 and y_0 for compliers. Using the notation established in (2.1), (2.2) and (2.3), these parameters are

$$E[y_d|d_1 > d_0, x] = \mu_d(x) + \sigma_d(x)E[\epsilon_d|m(x, 0) < \eta \leq m(x, 1)], \quad (2.6)$$

for $d = 0, 1$. Using an analogous approach to estimate effects on first and second moments, we obtain estimates of covariate-specific variance in potential outcomes for compliers:

$$V[y_d|d_1 > d_0, x] = \sigma_d^2(x)V[\epsilon_d|m(x, 0) < \eta \leq m(x, 1)]. \quad (2.7)$$

While these parameters are of value for predicting effects on variance specifically for compliers, they are not necessarily have predictive value for randomly chosen individuals with characteristics x .

2.2.1 Proposed IV Estimator for the Scale Ratio – From LATE to ATE

Previous work by Angrist (2004) and Heckman and Vytlacil (2000) invokes a joint symmetry condition to identify ATE on mean, $\mu_1(x) - \mu_0(x)$, using LATE on mean. Specifically, letting $f_d(\epsilon_d, \eta)$ denote the probability density function of the normalized error terms in outcome and selection equations for a given treatment status $d = 0$, or 1, they assume:

Assumption 2

JOINT SYMMETRY: *Conditional on x and z , we have $f_d(\epsilon_d, \eta) = f_d(-\epsilon_d, -\eta)$ for $d = 0, 1$.*

Under joint symmetry, the nuisance term $E[\epsilon_d | m(x, 0) < \eta \leq m(x, 1)]$ in eq. (2.6) equals zero if $m(x, 1) = -m(x, 0)$ for given covariates x . Thus, the covariate-specific ATE on *mean* is identified by LATE. However, the ATE on *inequality cannot* be identified by LATE even with joint symmetry. In particular, consider the variance ratio for compliers:

$$\frac{V[y_1 | d_1 > d_0, x]}{V[y_0 | d_1 > d_0, x]} = \frac{\sigma_1^2(x) V[\epsilon_1 | m(x, 0) < \eta \leq m(x, 1)]}{\sigma_0^2(x) V[\epsilon_0 | m(x, 0) < \eta \leq m(x, 1)]}, \quad (2.8)$$

where the nuisance term $V[\epsilon_1 | m(x, 0) < \eta \leq m(x, 1)] / V[\epsilon_0 | m(x, 0) < \eta \leq m(x, 1)]$ does not equal one in general, even under the condition of joint symmetry. Unlike the ATE for the mean effect is identified under joint symmetry, identification of the ATE for the scale ratio suggests conditions other than joint symmetry. Here we invoke an equality assumption to eliminate the nuisance term such that the variance ratio for compliers identifies the covariate-specific scale ratio, $\sigma_1^2(x) / \sigma_0^2(x)$. Under equality, the condition of joint symmetry is *not* required for identification of the scale ratio, though as we show later on, the symmetry assumption will be useful in relaxing the monotonicity assumption. Throughout this paper, our identification for scale ratios will impose the following equality assumption.

Assumption 3

EQUALITY: *Conditional on x and z , the unit random terms ϵ_1 and ϵ_0 in the outcome equation have the same distribution.*

We first note that the above condition does *not* impose that the outcome error terms have the same distribution for the treated and untreated, since we are still allowing the scale factors to vary across treatment status. However, under our condition, the identification of the

scale ratio is achieved with a somewhat restrictive condition. Although it allows a complete flexibility in the conditional scales given weak distributional assumptions, it implicitly imposes restrictions on the conditional distributions of outcome errors. In particular, it requires the same sign of the coefficient of correlation between selection and outcome errors across treatment statuses, though we do allow the value of the coefficient to differ. However, this rules out a general case of Roy’s model, where the coefficient of correlation between outcome and selection errors can vary arbitrarily with treatment status. Important examples include Willis and Rosen (1979), Heckman and Sedlacek (1985) and Gould (2002), all of which require the assumption of normality.

To summarize our identification results in this section, we showed that although identification cannot be achieved under both assumptions of monotonicity and joint symmetry, it can be achieved by replacing the condition of joint symmetry with the Equality assumption, still maintaining the monotonicity condition. In the following section we establish that a symmetry assumption does facilitate identification in a certain respect. Specifically, we will show that it allows one to replace the monotonicity condition with the Equality assumption imposed here. See Table 1 for a summary of our identification results.

2.3 Identification of the Scale Ratio under Joint Symmetry

Here we show how identification of the scale ratio can be achieved by the condition of joint symmetry. This extra shape restriction will enable us to relax the monotonicity assumption imposed in the previous section if we still maintain the Equality assumption. Chen (1999) has noted that the condition of joint symmetry is useful to identify the covariate-specific ATE on the *mean*, not only restricted to compliers. This subsection shows that under joint symmetry, the covariate-specific ATE on *inequality* can also be identified without imposing the monotonicity condition of the previous section.

The key idea behind our identification strategy is to select *symmetric pairs* with $(d_i, d_j) = (1, 0)$ and $m(w_i) = -m(w_j)$, letting $w \equiv (x, z)$. Under joint symmetry, these pairs are equivalent to what Angrist (2004) called the *Symmetric Subsample*, which have the sum of probabilities of selection equal to one. That is:

$$m(w_i) = -m(w_j) \Leftrightarrow p(w_i) + p(w_j) = 1, \tag{2.9}$$

where $p(w) \equiv \Pr\{d = 1|w\}$ denotes the *probability of selection* (or the *propensity score*) for given values of covariates and instruments. We call this step the symmetric first stage, and assume such pairs exist in the data.

Assumption 4

SYMMETRIC FIRST STAGE: *Letting $w \equiv (x, z)$ we assume the distribution of w is “rich enough” in the sense that $p(w)$ has positive density on the interval $(0, 1)$. This ensures positive density for pairs of values (w_1, w_0) that satisfy $p(w_1) + p(w_0) = 1$.*

Like conventional methods of propensity score matching, symmetric first stage is satisfied conditional on a given bandwidth if propensity scores and selection indexes are continuously distributed. In implementation, we select a bandwidth parameter using Silverman’s (1986) rule-of-thumb method.

Using symmetric pairs (i, j) , Assumptions (2)-(4) imply that the first and second moments of $(\epsilon_{1i}, \epsilon_{0j})$ must satisfy the following equalities:

$$E[\epsilon_{1i}|\eta_i \leq m(w_i), w_i] = E[\epsilon_{0j}|\eta_j > m(w_j), w_j] \quad (2.10)$$

$$E[\epsilon_{1i}^2|\eta_i \leq m(w_i), w_i] = E[\epsilon_{0j}^2|\eta_j > m(w_j), w_j]. \quad (2.11)$$

This leads the conditional variance of $(\epsilon_{1i}, \epsilon_{0j})$ for symmetric pairs to be equal between the treated and the untreated. Thus the scale ratio is identified by the ratio of the conditional variances in observed outcomes of symmetric pairs because the nuisance term is canceled out:

$$\frac{Var[y_i|m(w_i), d_i = 1, w_i = w]}{Var[y_j|m(w_j), d_j = 0, w_j = w]} = \frac{\sigma_1^2(x)Var[\epsilon_{1i}|\eta_i \leq m(w_i), w_i = w]}{\sigma_0^2(x)Var[\epsilon_{0j}|\eta_j > m(w_j), w_j = w]} = \frac{\sigma_1^2(x)}{\sigma_0^2(x)}. \quad (2.12)$$

While this approach allows a full flexibility in the conditional scales, it requires the existence of the first and second moments, and will not be suitable for data with heavy tails in their distributions; for example, wage distributions. Consequently, a quantile-based approach may be more desirable.

Maintaining Assumptions (2) and (3), we note that conditional quantiles of y_d are related to quantiles of ϵ_d by eq. (2.1) and (2.2). Letting $q_\tau(y|\cdot) \equiv F_{y|\cdot}^{-1}(\tau)$, we have $q_\tau(y|d, w) = \mu_1(x) + \sigma_1(x)q_\tau(\epsilon_1|d, w)$ at quantile τ for each covariate and instrumental value $w = (x, z)$. Thus for any $\tau \in (0, 1)$, *interquantile spread* IQ_τ from τ to $1 - \tau$ must satisfy the following

equalities:

$$IQ_\tau(y_i|d_i = 1, w_i = w) = \sigma_1(x)IQ_\tau(\epsilon_{1i}|\eta_i \leq m(w_i), w_i = w), \quad (2.13)$$

$$IQ_\tau(y_j|d_j = 0, w_j = w) = \sigma_0(x)IQ_\tau(\epsilon_{0j}|\eta_j > m(w_j), w_j = w), \quad (2.14)$$

for the treated and untreated groups, respectively.

The above equalities suggest a way to identify $\sigma_1(x)/\sigma_0(x)$ under Assumptions (2)-(4). To see that, Assumptions (2) and (3) collectively imply that $q_\tau(\epsilon_1|\eta \leq m, w) = q_{1-\tau}(-\epsilon_1|-\eta \leq m, w) = -q_{1-\tau}(\epsilon_0|\eta > -m, w)$, suggesting that the following equalities will hold for any quantile $\tau \in (0, 1)$ given a real number m :

$$q_\tau(\epsilon_{1i}|\eta_i \leq m, w_i) = -q_{1-\tau}(\epsilon_{0j}|\eta_j > -m, w_j), \quad (2.15)$$

$$q_{1-\tau}(\epsilon_{1i}|\eta_i \leq m, w_i) = -q_\tau(\epsilon_{0j}|\eta_j > -m, w_j), \quad (2.16)$$

which immediately imply an equality of conditional interquantile spreads of normalized outcome errors between symmetric quantiles τ and $1 - \tau$:

$$IQ_\tau(\epsilon_{1i}|\eta_i \leq m, w_i) = IQ_\tau(\epsilon_{0j}|\eta_j > -m, w_j). \quad (2.17)$$

This illustrates our strategy to identify the scale ratio using eq. (2.13) and (2.14). For a given pair of observations, one selects and the other does not, if a variation in the instrument (or covariates) within the pair can switch the treatment status and change the selection index from m to $-m$, then the ratio of interquantile spreads of observed outcomes identifies the scale ratio. Pairs of this sort are exactly those satisfying symmetric first stage in Assumption (4).

We note that the pairwise-matching methods can be applied to cases with a binary or continuous instrument. Intuitively, a continuously distributed instrumental variable is more likely to provide “rich enough” support that will facilitate finding sufficiently many symmetric pairs for identification.

3 Estimation

This section translates the identification strategies detailed in the previous section into tractable estimation procedures. All proposed procedures below aim to identify the scale ratio for a given set of covariate values x . The population average of scale ratios, which may

be considered a useful summary parameter, can be obtained by averaging across the values of x .

The variance-based IV estimator introduced in Section 2.2.1 can be easily implemented as follows. First, we estimate the variance of outcomes for compliers $Var[y_d | d_1 > d_0, x]$ by treatment status d , using eq. (2.4) and (2.5). This can be done for each regressor value x using a basic kernel method. Then, calculate the variance ratio for compliers to identify the scale ratio using eq. (2.7).

The other two estimators, which require the conditions of joint symmetry and symmetric first stage, use pairwise matching methods to search for and assign weights to symmetric pairs in the data set, as eq. (2.9) suggests. This requires first-stage nonparametric estimates of probability of treatment, $\hat{p}(w_i)$ and $\hat{p}(w_j)$, letting $w = (x, z)$. Since such estimates rarely, if ever, sum to exactly one, we give a higher weight to pairs who have the sum of propensity scores closer to one. To do so, we propose a *kernel weighting* scheme:

$$\hat{\omega}_{ij} = K_{h_{1n}}(1 - \hat{p}(w_i) - \hat{p}(w_j)), \quad (3.18)$$

where $K_{h_{1n}}(\cdot)$ is a kernel function for a given bandwidth h_{1n} .

With these weights we aim to construct variance-based and quantile-based estimators of scale ratios, using the weighted average of the ratios of conditional variances or interquantile functions for each regressor and instrumental variable value w . Because w is continuously distributed, we use conventional kernel functions and smoothing parameters. To construct the variance-based estimator, we first derive the ratio of observed variances between the treated and the untreated at each value of w , using the first and second moment estimators in eq. (2.10) and (2.11). For construction of the quantile-based estimator, we calculate the interquantile spread between quantiles τ and $1 - \tau$ for each value of w with the estimates of local quantile functions for the treated and the untreated. One can derive pointwise nonparametric estimates of quantile functions using the local polynomial procedure proposed in Chaudhuri (1991a, 1991b). See those papers for a detailed description of implementing that procedure.

Finally, the scale ratio estimators are formulated as the weighted average of the ratios of interquantile functions (or conditional variances) for the treated versus the untreated groups, weighted by first-stage estimates $\hat{\omega}_{ij}$ defined in (3.18). Precisely, the proposed estimator of

the scale ratio is written as follows:

$$\hat{r}(x) = \frac{\sum_{i,j} d_i(1-d_j)\hat{\omega}_{ij}I\hat{Q}_\tau(y_i|d_i=1, x_i=x, z_i)/I\hat{Q}_\tau(y_j|d_j=0, x_j=x, z_j)}{\sum_{i,j} d_i(1-d_j)\hat{\omega}_{ij}} \quad (3.19)$$

These estimation procedures have a disadvantage of requiring multiple semiparametric steps which involves selection of smoothing parameters, making implementation in finite samples difficult. This is a consequence of generality of the model considered because location functions $m(\cdot)$, $\mu_1(\cdot)$ and $\mu_0(\cdot)$ are left unspecified. Imposing parametric conditions on one of these functions will reduce the number of smoothing parameters needed, allow a simpler implementation in finite samples. For example, if we imposed $m(w) = w'\delta$, the existing estimation procedures can be used to estimate δ and propensity scores. Examples of such include Han (1987) and Powell, Stock and Stoker (1989). Furthermore, if the scale parameters only depend on treatment status (not regressors), the interquantile functions and the conditional variances will only depend on (one-dimensional) propensity score functions. This further reduces the dimensionality of the problem.

4 Large Sample Behavior

In the previous sections we established identification results for a conditional scale ratio, and discussed estimation procedures based on these identification strategies. In this section, we explore the asymptotic properties of an *averaged scale ratio* which is obtained by averaging across the regressor values of the proposed estimators in the previous section. Here we will only focus on the IV and symmetric quantile estimators, simply noting that the properties of the other proposed estimators follow from similar arguments and thus are omitted here. As this section will show the average estimators will converge at the parametric rate, a result that is loosely analogous to the asymptotic properties of (weighted) average derivative estimators (see, e.g. Powell *et al.* 1989).

Both theorems are proved in the appendix, which also contains the regularity conditions the proofs are based on. For the (averaged) variance IV estimator based on the monotonicity and equality assumptions, our theorem is as follows.

Theorem 1 *Our variance-based IV estimator, $\hat{r}_{IV} \equiv \frac{1}{n} \sum_{i=1}^n \frac{\hat{\sigma}_1^2(x_i)}{\hat{\sigma}_0^2(x_i)}$ of*

$r_{IV} = E_X[\sigma_1^2(x_i)/\sigma_0^2(x_i)]$ has the following linear representation:

$$\hat{r}_{IV} - r_{IV} = \frac{1}{n} \sum_{i=1}^n \psi_{ai} + \psi_{bi}$$

where $\psi_{ai} = \sigma_1^2(x_i)/\sigma_0^2(x_i) - E_X[\sigma_1^2(x_i)/\sigma_0^2(x_i)]$ and

$$\begin{aligned} \psi_{bi} &= \sigma_{0i}^{-2}(\psi_{12i} - 2\mu_{1i}\psi_{11i}) \\ &- \frac{\sigma_{1i}^2}{\sigma_{0i}^4}(\psi_{01i} - 2\mu_{0i}\psi_{02i}) + o_p(n^{-1/2}), \end{aligned} \quad (4.20)$$

where

$$\sigma_{1i}^2 = E[y_{1i}^2|x_i, d_{1i} > d_{0i}] - (E[y_{1i}|x_i, d_{1i} > d_{0i}])^2$$

and σ_{0i}^2 is defined analogously, for $d = 0, 1$ μ_{di} above is $E[y_{di}|d_{1i} > d_{0i}, x_i]$, and for $d = 0, 1$ and $k = 1, 2$, ψ_{dki} is of the form

$$(E[I[d_i = d]|z_i = 1, x_i] - E[I[d_i = d]|z_i = 0, x_i = x])^{-1} \times \quad (4.21)$$

$$\begin{aligned} &\left(y_i^k I[d_i = d] z_i - E[y_i^k I[d_i = d]|z_i = 1, x_i] + y_i^k I[d_i = d](1 - z_i) - E[y_i^k I[d_i = d]|z_i = 0, x_i] \right) - \\ &(E[y_i^k I[d_i = d]|z_i = 1, x_i](E[I[d_i = d]|z_i = 1, x_i] - E[I[d_i = d]|z_i = 0, x_i = x])^{-2} \times \quad (4.22) \\ &\left(I[d_i = d] z_i - E[I[d_i = d]|z_i = 1, x_i] + I[d_i = d](1 - z_i) - E[I[d_i = d]|z_i = 0, x_i] \right) \end{aligned}$$

The root- n consistency and asymptotic normality of the estimator follows from this linear representation.

Turning attention to the quantile estimator, while the regularity conditions for the quantile-based estimator are standard when compared to existing work (Ahn and Powell 1993, Chen and Khan 2003), they are still quite detailed, particularly as multiple semiparametric steps are involved.

To ease notational burdens, letting $w = (x, z)$, we impose the parametric restriction $m(w) = w'\delta \equiv v$, and assume the scale parameters depend on treatment but not covariates- i.e. $\sigma_1(x_i) = \sigma_1$, $\sigma_0(x_i) = \sigma_0$. It is important to note that a nonparametric $m(\cdot)$ will still allow our estimator of r to be root- n consistent and asymptotically normal, analogous to Ahn and Powell's (1993) results.

In addition, define $q_\tau^{(d)}(w) = q_\tau(y_i|d_i = d, (x_i, z_i) = w)$,

$\Delta q_\tau^{(d)}(w) = IQ_\tau(y_i|d_i = d, (x_i, z_i) = w)$ and

$\Delta \bar{q}_\tau^{(d)}(w) = E[\Delta q_\tau^{(d)}(w)|v]$ for $d = 0, 1$. The asymptotic property of the quantile-based estimator of the scale ratio, defined here as $r = \sigma_1/\sigma_0$.

Theorem 2 *Under regularity conditions H1,S0,RD2,S2,KH1 and I in the Appendix, if the selection equation estimator has the following linear representation, composed by ψ_i^+ :*

$$\hat{\delta} - \delta = \frac{1}{n} \sum_{i=1}^n \psi_i^+ + o_p(n^{-1/2}),$$

then we have

$$\sqrt{n}(\hat{r} - r) \Rightarrow N(0, \Sigma_0^{-2} E[(\psi_i^- + \mathcal{M}\psi_i^+)^2]),$$

where $\Sigma_0 = E[p(v_i)^2 f(p(v_i))]$, $p(v_i)$ is the propensity score, $f(p(v_i))$ its density. Furthermore, $\psi_i^- = (1 - p(v_i))^2 f_V(v_i) f_W(w_i) \left[d_i \Delta q_\tau^{(0)}(w_i)^{-1} \phi_{1i} - (1 - d_i) \frac{\Delta q_\tau^{(1)}(w_i)}{\Delta q_\tau^{(0)}(w_i)^2} \phi_{0i} \right]$,

where we denote ϕ_{di} as

$f_{U_{2d}|W}(0|w_i)^{-1} (I[y_i \leq q_{\tau_2}^{(d)}(w_i)] - \tau_2) - f_{U_{1d}|W}(0|w_i)^{-1} (I[y_i \leq q_{\tau_1}^{(d)}(w_i)] - \tau_1)$ for $d = 0, 1$, and $\Delta q_\tau^{(d)}$ denotes the conditional interquantile spread for treatment status d . Finally, define \mathcal{M} as

$$E[(1 - p(v_i)) \cdot [\mathcal{G}_2(v_i, v_i) p(-v_i) f_V(-v_i) + \mathcal{G}(v_i, v_i) p'(-v_i) f_V(-v_i) + \mathcal{G}(v_i, v_i) p(-v_i) f_V'(-v_i)]], \quad (4.23)$$

where

$$\mathcal{G}(v_i, v_j) \equiv \int \int \left[\Delta q_\tau^{(1)}(w_i) - \Delta q_\tau^{(0)}(w_j) \right] (w_i + w_j)' dF(w_i|v_i) dF(w_j|v_j)$$

and $\mathcal{G}_2(\cdot, \cdot)$ is the partial derivative of $\mathcal{G}(\cdot)$ with respect to its second argument.

5 Monte Carlo Study

5.1 The Design and Specifications

The previous section has explored the conditions under which the proposed estimators are well-behaved in a large sample. In this section we assess their small sample performance through Monte Carlo simulations. In our design, outcome errors $(\epsilon_{1i}, \epsilon_{0i})$ are drawn from the same distribution to satisfy the Equality condition (Assumption 3). To assure dependence between outcome and selection errors (ϵ_{di}, η_i) , we assume their correlation coefficient to be 0.5. To satisfy the symmetry condition (Assumption 2), we generate a sample (ϵ_{di}, η_i) of size n from each of the following distributions: Bivariate Normal, Student-t with 10 degrees of freedom, Student-t with 3 degrees of freedom, and the Cauchy distribution. We note that the last three distributions have increasingly heavy tails. We then increase the sample size from $n = 100$ to 800 and iterate the estimation process 801 times.

Each sample is divided into 2 subsamples, the college and non-college groups ($d_i = 1, 0$), by a selection equation:

$$d_i = I\{x_i + z_i \geq \eta_i\}, \quad (5.24)$$

where a regressor x_i is uniformly distributed between -1 and 1, and an instrument z_i is a binary variable with a 50-50 chance to equal 1 or 0. This design will be referred to as *Model (A)*. By construction, the monotonicity condition (in Assumption 1) is satisfied in this design. In *Model (B)*, we experiment a case where the monotonicity condition is violated:

$$d_i = I\{x_i + z_i + 4x_i z_i \geq \eta_i\}. \quad (5.25)$$

The violation of the monotonicity condition is caused by the fact that

$$\Pr\{d_{1i} \geq d_{0i} | x_i\} = \Pr\{x_i < \eta_i \leq 5x_i + 1\} = [F(6) - F(1)] - [F(-4) - F(-1)]/2 < 1,$$

where F is the cumulative distribution function of η_i .

Both models satisfy the symmetric first stage (Assumption 4), which requires the density of propensity score to be positive, for a given bandwidth. In *Model (A)*, for example, propensity scores for $z_i = 1$ and $z_j = 0$ are $\Pr\{\eta_i \leq x_i + 1\}$ and $\Pr\{\eta_j \leq x_j\}$, respectively, both of which are strictly positive.

Throughout this Monte Carlo study, the outcome equation is assumed to be

$$y_i = d_i[2 + x_i + \sigma_1(x_i)\epsilon_{1i_i}] + (1 - d_i)[x_i + \sigma_0(x_i)\epsilon_{0i_i}].$$

We consider two cases: First, we assume that the scale parameters are constant within educational groups, by letting $(\sigma_0, \sigma_1) = (1, e^{0.2})$. Next, we allow the scale parameters to vary with x , by assuming $(\sigma_0(x), \sigma_1(x)) = (1.08, 1.1e^{0.2})$ for positive x and $(0.92, 0.9e^{0.2})$ for non-positive x . For both cases, the average ratio of scales approximately equals 1.22. We use uniform kernels and select the bandwidths by Silverman's Rule of Thumb. We also experiment with other kernel functions and find that the results using uniform kernels are reliable.

5.2 Simulation Results

We use four assessment measures: Mean Bias, Median Bias, Root Mean-Squared Errors (RMSE), and Mean Absolute Deviations (MAD). While consistency of the quantile-based

estimator requires root- n convergence of MAD, both of the variance-based estimators (that is, the IV estimator and the variance-based pairwise-matching method) require root- n convergence of RMSE. The following analysis therefore focuses on the convergence of MAD and RMSE, and refers the convergence of an estimator as the root- n convergence of the corresponding assessment measure. Using the above simulation designs, which satisfy the identification conditions described in Section 2, the results suggest that the proposed estimators converge in small samples. In particular, the quantile-based estimator converges even when underlying distributions have heavy tails. We further compare the proposed IV estimator with the pairwise-matching methods using a design where the monotonicity condition is violated. The results show that violation of monotonicity leads the IV estimator to be inconsistent but does not affect the convergence of the pairwise-matching methods. Our findings are summarized in Tables 2A and 2B and discussed below.

To establish a benchmark, we first estimate the models by MLE, which is correctly specified for the Bivariate Normal specification and is misspecified for the other specifications. The results show that if the model is correctly specified, the assessment measures of MLE are much smaller than the proposed estimators'. For example, Mean Bias and Median Bias of MLE are about 1/10 of the proposed estimators'. In contrast, if the model is misspecified (for instance, in Bivariate Student- $t(3)$ or Cauchy distributions), as expected, MLE may generate larger biases than the proposed methods.

Table 2A suggests that the proposed estimators converge at the rate of root- n for both constant and heterogeneous scale functions. As Panel-A shows, the estimators converge rapidly at least at the rate of root- n when scale parameters do not depend on covariates x . In Panel-B when scale parameters vary with x , the estimators still converge approximately at the rate of root- n although the convergence is slower than before. Notably, the above results require finite variances of both outcome and selection errors. If the values of variances are indefinite, such as Bivariate Cauchy, the variance-based estimators fail to converge but the quantile-based estimator still converges at the rate of root- n . This highlights an advantage of the quantile-based estimator over the variance-based methods for distributions with heavy tails.

We note that the monotonicity condition is required only for the IV estimator but not

for the pairwise matching methods. To compare these two approaches, we estimate *Model (B)* in Table 2B, where the monotonicity condition is violated and the other conditions (e.g., joint symmetry) are still satisfied. As expected, the results indicate that the IV estimator is no longer converging but pairwise-matching estimators still converge at the rate of root- n .

It is worth emphasizing that the condition of joint symmetry is neither stronger nor weaker than the monotonicity condition. These two conditions of identification use different subpopulations to identify the ATE on inequality: While the pairwise-matching methods use *symmetric pairs*, the IV estimator uses *compliers*. Since both methods use a subpopulation for identification, precision of estimation depends on the size of the subpopulation. Pairwise-matching estimates based on symmetric pairs can be much more efficient than the IV estimates using compliers because the number of symmetric pairs can be considerably more than the number of compliers, as we will see below.

6 Causal Effects of Education on Inequality

We apply the proposed methods to estimate the causal effect of schooling on wage inequality. Economists have been interested in understanding the link between college education and wage inequality at least since Juhn, Murphy and Pierce (1993). Yet few empirical studies in the literature address the problems of endogenous schooling. Heckman and Honore (1990) suggest that *under the assumption of normality*, self-selected education truncates wage distributions, causing the degree of inequality within an educational sector to be underestimated. This has been echoed by Heckman and Sedlacek (1985, 1990) and Gould (2002) concerning the choices of industrial sectors and occupations, based on the normality assumption. More recently, Chen’s (2006) parametric estimates suggest that potential wage inequality is more understated among college graduates than among high school graduates. Consequently, the observed difference in inequality between high school and college graduates understates the causal effect of college education on potential wage inequality.

Unlike the previous studies using the assumption of normality to identify the relationship between sector choices and wage inequality, we estimate this causal link without assuming normality. The assumption that log wages are normally distributed, in fact, has been rejected by Heckman and Sedlacek’s (1985,1990) and Lee’s (1982) studies. By applying our proposed

semiparametric estimators to a study on wage and education, we demonstrate below the estimation procedure and summarize the empirical results.

6.1 Determination of Wage and Wage Inequality

Consider a location-scale shift model of log wages, which has been used by Blau and Kahn (1994) and Lemieux (1998):

$$y = \alpha d + \mu(x) + \sigma_d(x)\epsilon_d, \quad (6.26)$$

The outcome variable y is log hourly wage (or briefly “wage”) and the treatment variable d is the decision to attend college. The coefficient on college attendance α is a measure of college wage premium. The covariate-specific location and scale parameters, $\mu(x)$ and $\sigma_d(x)$, are unknown functions of x . *Potential wage inequality* in a choice group with covariate x is measured by scale parameter $\sigma_d(x)$. We use a latent-index selection model to characterize schooling choice $d = I\{\eta \leq m(x, z)\}$, where $m(\cdot)$ remains unspecified and z is an instrument for education. η and ϵ_d are normalized random variables, independent of z conditional on x . The objective of our empirical study is to identify the *average ratio of scales* by averaging the scale ratios over covariate values.

6.2 Data and Replicated IV Estimates of Returns to Schooling

Our empirical application starts with a replication of well-known studies on the return to education by Card (1995) and Kling (2001). We apply exactly the same data used in their studies to our estimation procedures. The data in their studies - downloadable from the *Journal of Business and Economic Statistics* web site - are drawn from the *National Longitudinal Survey of Young Men* (NLSYM) in 1976. The NLSYM provides detailed individual controls (e.g., demographics, parental education, location of residence) and a plausible instrument for education; that is, proximity to a four-year college in the county of residence in 1966. Our sample contains all 3010 observations of the 1976 survey used in their analysis, including 1489 workers with high school diploma or less (with less than or equal to 12 years of schooling) and 1521 with some college education (with more than 12 years of schooling). For the purpose of illustration, we focus on the self-selected decision to attend college and omits other potential choices, such as dropping out of high school. We will examine the importance of omitting this in future research using more general models; e.g., ordered discrete-choice selection models.

Descriptive statistics in Table 3 show that workers with some college education have average wages 20 percent higher than the less educated. In contrast, the degree of within-group inequality is nearly constant across educational levels. However, the small differences in observed wage inequality between educational levels can be a result of self selection. For instance, if investing in college education induces a high conditional variance in potential wages, risk-averse individuals would opt for lower levels of schooling than they would otherwise would.

Following Card and Kling, we use the college proximity dummy to instrument for education. As noted in their studies, men who were raised in local labor markets with a nearby four-year college have significantly higher levels of education. Row (b) of Table 3 indicates that 73 percent of workers with some college education have a four-year college in county, about 10 percentage points higher than the fraction of workers in the high school group. This differential persists even after controlling for regional and family background variables, as shown in their studies.

We replicate Card's (1995) and Kling's (2001) results exactly. As Part (a) of Table 4 shows, the estimated return to one additional year of schooling increases to 13 percent from 7 percent, after the problem of endogenous schooling is addressed. Due to the limitation of sample size, our application uses fewer regional background controls than Card's and Kling's studies. Using the subset of the covariates, we find similar estimates to their results. The subset of covariates used in our application includes work experience, a Black indicator, residence in Southern states in 1976, residence in a metropolitan area in 1976 and in 1966, and college attendance of both parents. Following Kling's suggestion, we exclude the quadratic term for work experience, take work experience as an endogenous variable in our model, and use age as an excluded instrument.

Using this subset of covariates, our results in Columns (4) and (5) in Part (b) show that 2SLS estimates of the return to *college attendance* are more than triple of OLS estimates, consistent with the results in Columns (1)-(3) using the full set of covariates. This highlights the self selection issues caused by college attendance, which may also lead the estimated effect of college education on wage inequality to be biased.

6.3 Estimation Results on Inequality

OLS and Heckman two-stage estimates in Part (A) of Table 5 show that the variance in wages of the college group is 8 percent to 9 percent higher than that of the high school group. Adjustments for selection by Heckman two-stage increase the variance in wage by more than one third for each educational group. Because selection correction increases the estimates of scales proportionally across schooling levels, the results show that the estimates of the scale ratios do not change after adjusting for selection.

Heckman two-stage estimates are established based on the assumption of normality. The mean and variance of observed wages conditional on education and covariates are given by

$$\begin{aligned} E[y_i|d_i = d; x_i, z_i] &= \alpha d + \mu(x) + \beta_d \lambda_{di}, \\ \text{Var}[y_i|d_i = d; x_i, z_i] &= \sigma_d^2(x) - \beta_d^2 \delta_{di}. \end{aligned}$$

Letting $m_i = m(x_i, z_i)$, define $\delta_{di} \equiv 1 - \text{Var}[\eta_i|d_i = d] = \lambda_{di}^2 - m_i \lambda_{di}$ and denote λ_{di} as the inverse Mills ratios for educational level d . Because $\beta_d^2 = \rho_d^2 \delta_{di}^2$ must be non-negative, the variance of observed wages must be no higher than the variance of potential wages; i.e. $\text{Var}[y_i|d_i = d; x_i, z_i] \leq \sigma_d^2(x)$, suggesting that OLS estimates of the wage variance must be no larger than Heckman two-stage estimates.

Unlike OLS and Heckman two-stage identify the variance in wages for each educational level thereby the average ratio of scales is identified, the proposed semiparametric methods only identify the average ratio of scales but not the value of the scale for a given educational level. This is because the nature of multiplicative heteroscedasticity allows us to eliminate nuisance terms by taking ratios. However, the nuisance term of scale parameters within educational groups cannot be identified without additional distributional assumptions if the condition of normality is not satisfied.

The first proposed method is an IV procedure, assuming that the LATE conditions are valid. Implementation of the method begins with calculation of conditional variance in wages for each schooling level, using Abadie's approach in eq. (2.4) and (2.5). The estimate of the average ratio of scales is then obtained by averaging the ratios of conditional variances over all possible covariate values.

The proposed IV estimates in Part (B) of Table 5 are extremely imprecise because there

are few compliers in the sample. Using Abadie’s (2003, Lemma 1) method, we calculate the fraction of compliers $Pr\{d_1 > d_0|x\} = E[d|z = 1, x] - E[d|z = 0, x]$ conditional on covariate values x and then take average over x . Estimates show that compliers only account for 10 percent of the observations, about 300 compliers in the sample.

In contrast, the pairwise-matching methods built upon symmetry conditions generate much more precise estimates than the IV methods. This efficiency gain is a consequence of the pairwise matching scheme, which draws symmetric pairs from all possible $1,521 \times 1,489 = 2,264,769$ matches from the college and high school groups. The estimation procedure starts with selecting symmetric pairs so that the condition of symmetric first stage is satisfied. We assign pairs with positive kernel weights when the sum of probabilities of selection is sufficiently close to one. As Part (C) shows, symmetric pairs account for about 12 percent of all matches from the college and high school groups. That is, about 0.2 million pairs are available for estimation.

Pairwise-matching estimates in Part (C) show that the average ratio of scales ranges from 1.16 to 1.31 at the 5 percent significance level, suggesting that college education causes a 16 percent to 31 percent increase in the degree of wage inequality. This is roughly 10 to 20 percentage points higher than the benchmark results using OLS and Heckman two-stage. This difference, however, is not statistically significant at the 5 percent level.

We note that the results based on quantile-based pairwise-matching are specific to given quantiles. For instance, the estimated average scale ratio using the 75-25 interquantile range is about 1.31, higher than the estimate of 1.2 based on the 90-10 interquantile range. This suggests that this method *over-identifies* the average ratio of scales and its efficiency can be improved by averaging the estimates over interquantile ranges.

Our proposed estimators require bandwidths and kernel functions to derive propensity scores, to select symmetric pairs, and to define “closeness” among covariate values. Throughout the above empirical application, we select bandwidths according to Silverman’s Rule of Thumb conditional on covariate values. In particular, we set the bandwidth as $1.06sn^{1/5}$, where n is the sample size and s is the standard deviation of the given covariate. We use uniform kernel functions, which generate very similar results as those based on normal kernel functions.

7 Concluding Remarks

This paper proposes IV-type and semiparametric estimators to test whether the contrast in wage inequality between the college educated and high school graduates is statistically significant. Using weak distributional and functional form assumptions, we provide three ways to assess the causal effect of schooling on wage inequality. The first is built on the previous LATE results, using conditional variance functions. The second and the third ones are based on pairwise-matching methods, using either conditional variance functions or conditional quantile regressions. While the first approach requires monotonicity but does not require joint symmetry, the last two methods need joint symmetry but do not need monotonicity. Because joint symmetry is not weaker or stronger than monotonicity, the proposed pairwise-matching methods provide alternative strategies other than IV methods for identification.

A key ingredient of our pairwise-matching methods is the kernel weighting scheme that selects pairs satisfying symmetry conditions. In our application, for example, college-high school pairs are assigned positive weights if the sum of their probabilities of college-going is close enough to one.

The simulation studies suggest that the proposed estimators perform well in finite samples with the number of observations as small as 100. All of the three estimators converge at the rate of root-n if the bivariate distribution has a finite second moment. For bivariate distributions with the first and the second moments undefined, e.g. Bivariate Cauchy, the variance-based estimators are not consistent but the quantile-based method still converges rapidly at the rate of root-n.

Although the IV estimator is simpler and faster in computation than the pairwise-matching methods, it requires large data sets to derive enough many compliers in order to provide precise estimates. Furthermore, in the case where the instrument is continuous, the IV estimator does not apply but the pairwise-matching methods still do.

We apply the proposed IV and semiparametric procedures to the NLSYM66 sample in 1976, which was used by Card (1995) and Kling (2001) to instrument schooling with a college-proximity indicator. After replicating their well-known results about the causal effect of schooling on average wages, we use the same sample to estimate the causal effect of college education on inequality. On one hand, the results based on pairwise matching suggest that

the decision to attend college significantly increased the degree of wage inequality among workers in 1976. This result is similar to what OLS Heckman two stage suggest. On the other hand, the proposed IV estimates are very imprecise and fail to reject the hypothesis that college earnings and high school earnings can be equally dispersed. This imprecise result is primarily due to the limitation of sample size.

In light of our empirical results, we believe that the endogeneity issues caused by schooling choice is important in estimating the educational contrast in inequality, especially in large samples such as the U.S. Census. As previous work by Autor, Katz and Kearney (2006) and Lemieux (2006) has noted, it is important to determine whether the recent growth of wage inequality can be explained by the growing contrast in inequality between college and high school. The methods proposed in this paper can potentially help us reassess the role of college education in rising inequality during the recent decades. We leave this empirical task for future research.

Although the proposed estimators prove consistent, possible improvements in efficiency could be achieved in several ways. For instance, while this paper demonstrates how symmetry conditions identify the ATE on *scale parameters* by pairwise-matching schemes, Songnian Chen (1999) has provided similar methods for estimating the ATE on *location parameters* under similar conditions. If the treatment effects on both location and scale parameters can be estimated simultaneously, then the efficiency of the proposed estimators can be achieved using a combined estimation procedure. In addition, we note that the proposed quantile-based estimator over-identifies the average scale ratio for multiple pairs of symmetric quantiles (see eq. (3.19)). This suggests that the degree of efficiency can be improved further by averaging the proposed estimates over various pairs of symmetric quantiles. We leave the potential enhancement in efficiency to future studies.

Acknowledgements

We would like to thank Alberto Abadie, Josh Angrist, Jonah Gelbach, and Ed Vytlačil for valuable suggestions. Thanks to seminar participants at Queens University, Academia Sinica, Gothe University, University of Toronto, the Evaluation Conference at the Centre for European Economic Research, and the Conference of Econometric Evaluation of Public

Policies at CREST-INSEE. All errors are ours.

Appendix A: Regularity Conditions and Proofs

The distribution theory of our proposed IV estimator is based on the following regularity conditions:

Assumption RS (Random sampling) The vector $(y_i, z_i, d_i, x_i)'$ is i.i.d.

Assumption RD (Regressor distribution) The regressor vector x_i has support which is a compact subset of \mathbf{R}^k . x_i may have discrete and continuous components, and we let k_c denote the number of continuous components. We assume the conditional density function of the continuous components given the discrete components is continuously differentiable of order p , where $p > 5k_c/2$.

Assumption K (Kernel function and bandwidth) The kernel function is of order p and the bandwidth, satisfies $\sqrt{nh_n^p} \rightarrow 0$ and $nh_n^{d_x} \rightarrow \infty$.

Assumption MF (Moment functions) Define the functions $m_{kl}(\cdot)$ $m = 0, 1$ $l = 1, 2$ as $E[y_i^l | x_i, d_i = m, z_i = 1]$. Then these functions are p times continuously differentiable.

With these regularity conditions we will establish the following lemma, from which the proof of Theorem 1 will follow immediately.

Lemma 7.1 Let $\hat{r}_{IV} \equiv \frac{1}{n} \sum_{i=1}^n \hat{\sigma}_1^2(x_i) / \hat{\sigma}_0^2(x_i)$ denote our estimator, and let $r_{nIV} \equiv \frac{1}{n} \sum_{i=1}^n \sigma_1^2(x_i) / \sigma_0^2(x_i)$

Under the above Assumptions, our variance IV estimator has the following linear representation:

$$\hat{r}_{IV} - r_{nIV} = \frac{1}{n} \sum_{i=1}^n \psi_{bi} + o_p(n^{-1/2}) \quad (\text{A-1})$$

where recall the terms ψ_{bi} is defined in the text in the statement of Theorem 1.

Proof:

Note we can linearize the ratio as in e.g. page 2204 of Newey and McFadden.

We will establish a linear representation for

$$\hat{r}_{IV} - r_{nIV} = \frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} ((\hat{\sigma}_{1i}^2 - \sigma_{1i}^2)) - \quad (\text{A-2})$$

$$\frac{1}{n} \sum_{i=1}^n \sigma_{1i}^2 \sigma_{0i}^{-4} ((\hat{\sigma}_{0i}^2 - \sigma_{0i}^2)) \quad (\text{A-3})$$

where recall

$$\sigma_{1i}^2 = E[y_{1i}^2 | x_i, d_{1i} > d_{0i}] - (E[y_{1i} | x_i, d_{1i} > d_{0i}])^2$$

and σ_{0i}^2 is defined analogously.

Turning attention to (A-2), here we will establish a linear representation for

$$\frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (\hat{\sigma}_{1i}^2 - \sigma_{1i}^2) \quad (\text{A-4})$$

Here we will only formally establish a linear representation for the second moment of the outcome variable component of the variance, noting the square of the first moment term can be handled very similarly. Recall the term $E[y_{1i}^2 | x_i, d_{1i} > d_{0i}]$ and its kernel estimator each involve the ratio of differences. Our first step will be to linearize the ratio. To ease notation, let the true values of $E[y_{1i}^2 | x_i, d_{1i} > d_{0i}]$ be denoted as $\frac{a_1 - a_2}{b_1 - b_2}$, where a_1, a_2, b_1, b_2 denote $E[y_i^2 d_i | z_i = 1, x_i], E[y_i^2 d_i | z_i = 0, x_i], E[d_i | z_i = 1, x_i], E[d_i | z_i = 0, x_i]$ respectively. Let the kernel estimator be denoted as $\frac{\hat{a}_1 - \hat{a}_2}{\hat{b}_1 - \hat{b}_2}$. By the fourth root consistency of the kernel estimators, which follows from Assumptions RS, RD, MF (see Chen and Khan 2003) the linearization of ratio implies we can derive linear representations for

$$\frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (b_1 - b_2)^{-1} (\hat{a}_1 - a_1 - \hat{a}_2 - a_2) - \frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (a_1 - a_2) (b_1 - b_2)^{-2} (\hat{b}_1 - b_1 - \hat{b}_2 - b_2) \quad (\text{A-5})$$

since the remainder term is $o_p(n^{-1/2})$.

Here, we will focus on the term

$$\frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (b_1 - b_2)^{-1} (\hat{a}_1 - a_1) \quad (\text{A-6})$$

as similar terms may be used for the other components. Here, we notice that \hat{a}_1 is simply a kernel estimator for the regression function $E[y_i d_i | z_i = 1, x_i = x]$. Consequently, we can apply results from Newey and McFadden (1994) or Chen and Khan 2003) to represent

$$\sigma_{0i}^{-2} (b_1 - b_2)^{-1} (\hat{a}_1 - a_1) \quad (\text{A-7})$$

as (now expressing the definitions of a_1, a_2, b_1, b_2)

$$\frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (E[d_i|z_i = 1, x_i] - E[d_i|z_i = 0, x_i])^{-1} (y_i^2 d_i z_i - \quad (\text{A-8})$$

$$E[y_i^2 d_i|z_i = 1, x_i] + y_i^2 d_i(1 - z_i) - E[y_i^2 d_i|z_i = 0, x_i]) + o_p(n^{-1/2})$$

which takes care of the first term in the linearization in (A-5). The second term is of the form:

$$\frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (E[y_i^2 d_i|z_i = 1, x_i] (E[d_i|z_i = 1, x_i] - E[d_i|z_i = 0, x_i])^{-2} (d_i z_i - \quad (\text{A-9})$$

$$E[d_i|z_i = 1, x_i] + d_i(1 - z_i) - E[d_i|z_i = 0, x_i]) + o_p(n^{-1/2})$$

and subtracting (A-9) from (A-8) establishes the form of the linearization for the second moment of the outcome variable. We will denote the term in the resulting summation (excluding σ_{0i}^{-2}) by ψ_{12i} . Turning attention to the square of the first moment, the linear representation for the first moment would be the same as above simply replacing y_i^2 with y_i . Let ψ_{11i} denote this term. Now let μ_{1i} denote $E[y_{1i}|d_{1i} > d_{0i}, x_i]$ then the linear representation for the square of the first moment can be denoted by

$$\frac{1}{n} \sum_{i=1}^n 2\sigma_{0i}^{-2} \mu_{1i} \psi_{11i} + o_p(n^{-1/2}) \quad (\text{A-10})$$

which is a straight forward application of the delta method.

Collecting all our results we can conclude that we have the following linear representation for the variance of the treated group.

$$\frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (\hat{\sigma}_{1i}^2 - \sigma_{1i}^2) = \frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (\psi_{12i} - 2\mu_{1i} \psi_{11i}) + o_p(n^{-1/2}) \quad (\text{A-11})$$

Note we can derive an analogous linear representation for $\hat{\sigma}_{0i}^2 - \sigma_{0i}^2$ with analogous terms in the summation, which we will denote here by ψ_{01i}, ψ_{02i} .

Next linearize the ratio to conclude

$$\frac{1}{n} \sum_{i=1}^n \frac{\hat{\sigma}_{1i}^2}{\hat{\sigma}_{0i}^2} - \frac{\sigma_{1i}^2}{\sigma_{0i}^2} = \frac{1}{n} \sum_{i=1}^n \sigma_{0i}^{-2} (\psi_{12i} - 2\mu_{1i} \psi_{11i}) - \frac{\sigma_{1i}^2}{\sigma_{0i}^4} (\psi_{01i} - 2\mu_{0i} \psi_{02i}) + o_p(n^{-1/2}) \quad (\text{A-12})$$

where μ_{0i} above is $E[y_{0i}|d_{1i} > d_{0i}, x_i]$.

This completes the proof of the theorem.

REGULARITY CONDITIONS OF THEOREM 2: Let h_{0n} and h_{1n} denote the bandwidths for the selection equation estimation and the pairwise matching kernel weighing scheme in the first stage, and let h_{2n} denote the bandwidth for the linear polynomial quantile regressions in the second stage. The regularity conditions for Theorem 2 is summarized below.

Assumption I (Identification) $\Sigma_0 > 0$.

We next impose conditions on the kernel function used to match propensity score values and its bandwidth sequence:

Assumption KH1 The kernel function $K_{1n}(\cdot)$ is assumed to have the following properties:

- i) $K_{1n}(\cdot)$ is twice continuously differentiable with a bounded second derivative and has a compact support; (ii) symmetric about zero; and (iii) a fourth-order kernel with $\int u^l K_{1n}(u) du = 0$ for $l = 1, 2, 3$ and $\int u^4 K_{1n}(u) du \neq 0$. The bandwidth sequence h_{1n} is of the form: $h_{1n} = c_1 n^{-\gamma_1}$, where c_1 is a constant and $\gamma_1 \in (\frac{1}{8}, \frac{1}{6})$.

The following assumption characterizes the smoothness of the density of and the conditional expectation functions of the selection index:

Assumption S0 The function $f_V(\cdot)$ has an order of differentiability of four, with the fourth-order derivative bounded.

We next impose three conditions associated with the estimation of interquartile spreads. This involves smoothness assumptions on the conditional quantile functions and on the distributions of $w_i = (x_i, z_i)$ and the residuals associated with the quantile functions. For notational convenience, we describe the conditions in terms of w , whose support is denoted by \mathcal{W} .

Assumption RD2 (Distribution of regressors and instruments) The vector w can be decomposed as $w = (w^{(c)'}, w^{(ds)'})'$ where the k_c -dimensional vector $w^{(c)}$ is continuously distributed, and the k_{ds} -dimensional vector $w^{(ds)}$ is discretely distributed. Letting $f_{W^{(c)}|W^{(ds)}}(\cdot|w^{(ds)})$ denote the conditional density function of $w_i^{(c)}$, we assume it is bounded away from zero and is Lipschitz continuous on \mathcal{W} . Letting $f_{W^{(ds)}}(\cdot)$ denote the mass function of $w^{(ds)}$, we assume that there is a finite number of mass points on \mathcal{W} . Finally, we let $f_W(\cdot)$ denote $f_{W^{(c)}|W^{(ds)}}(\cdot|\cdot)f_{W^{(ds)}}(\cdot)$.

Assumption S2 (Smoothness of conditional quantile functions)

S2.1 The polynomial used for the second-stage quantile function estimators is of order m .

S2.2 For all values of $w^{(ds)}$, the quantile functions $q_{\tau_1}^{(d)}(\cdot)$ and $q_{\tau_2}^{(d)}(\cdot)$ $d = 0, 1$ are bounded and m times continuously differentiable with bounded m^{th} derivatives with respect to $w^{(c)}$ on \mathcal{W} .

Assumption H1 (Second-stage bandwidth sequence for interquartile spread estimation).

The bandwidth sequence used to estimate the conditional interquantile spread is of the form: $h_{2n} = c_2 n^{-\gamma_2}$, where c_2 is a constant, and $\gamma_2 \in \left(\frac{\gamma_1 + \frac{1}{2}}{m}, \frac{1 - 4\gamma_1}{3k_c}\right)$, where γ_1, m and k_c are given in Assumptions KH1, S2 and RD2 respectively.

PROOF OF THEOREM 2: The arguments used to derive the limiting distribution theory are very similar to those used in Chen and Khan (2003), hereafter referred to as CK. We thus only provide a sketch of the main arguments, referring readers interested in technical details to CK.

We note we can write $\hat{r} = \hat{\Sigma}_1 / \hat{\Sigma}_0$, where

$$\begin{aligned}\hat{\Sigma}_1 &= \frac{1}{n(n-1)} \sum_{i \neq j} d_j (1 - d_i) \hat{\omega}_{ij} \Delta \hat{q}_{\tau}^{(1)}(w_i) / \Delta \hat{q}_{\tau}^{(0)}(w_i), \\ \hat{\Sigma}_0 &= \frac{1}{n(n-1)} \sum_{i \neq j} d_j (1 - d_i) \hat{\omega}_{ij}.\end{aligned}$$

Recall that here $r = \frac{\sigma_1}{\sigma_0}$. We will establish a linear representation for $\hat{r} - r$.

Our proof strategy is to establish the probability limit of the denominator and establish a linear representation for the numerator. The probability limit of the denominator follows from similar arguments used in proving Lemmas A.5 and A.6 in CK:

$$\hat{\Sigma}_0 \xrightarrow{p} \Sigma_0 \equiv E[p(v_i)^2 f(p(v_i))],$$

Turning attention to $\hat{\Sigma}_1 - r \hat{\Sigma}_0$, consider an expansion of $\hat{\omega}_{ij}$ around $\omega_{ij} = h_{2n}^{-1} K_2 \left(\left(w_i' \delta_0 + w_j' \delta_0 \right) / h_{2n} \right)$. After using this expansion, $\hat{\Sigma}_1 - r \hat{\Sigma}_0$ equals:

$$\frac{1}{n(n-1)} \sum_{i \neq j} d_j (1 - d_i) \omega_{ij} \left[\Delta \hat{q}_{\tau}^{(1)}(w_i) / \hat{q}_{\tau}^{(0)}(w_i) - r \right]$$

We note that if we replace $\Delta \hat{q}_{\tau}^{(1)}(w_i), \Delta \hat{q}_{\tau}^{(0)}(w_i)$ with $\Delta q_{\tau}^{(1)}(w_i), \Delta q_{\tau}^{(0)}(w_i)$ in the above expression, the term is $o_p(n^{-1/2})$ by arguments similar to those used in the steps used to prove

Lemma A.4 in CK. We can thus work with:

$$\frac{1}{n(n-1)} \sum_{i \neq j} d_j(1-d_i) \omega_{ij} \Delta q_\tau^{(0)}(w_i)^{-1} \left[\left(\Delta \hat{q}_\tau^{(1)}(w_i) - \Delta q_\tau^{(1)}(w_i) \right) \right] + \frac{\Delta q_\tau^{(1)}(w_i)}{\Delta q_\tau^{(0)}(w_i)^2} \left[\left(-\Delta \hat{q}_\tau^{(0)}(w_i) + \Delta q_\tau^{(0)}(w_i) \right) \right]$$

We establish a linear representation for the term involving $(\Delta \hat{q}_\tau^{(1)}(w_i) - \Delta q_\tau^{(1)}(w_i))$. It follows from the arguments used in Lemma A.4 in CK that

$$\frac{1}{n(n-1)} \sum_{i \neq j} d_j(1-d_i) \omega_{ij} \Delta q_\tau^{(0)}(w_i)^{-1} \left[\Delta \hat{q}_\tau^{(1)}(w_i) - \Delta q_\tau^{(1)}(w_i) \right]$$

can be expressed as

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n (1-p(v_i))^2 f_V(v_i) d_i f_W(w_i) \\ & \cdot \Delta q_\tau^{(0)}(w_i)^{-1} \left(f_{U_{21}|W}(0|w_i)^{-1} (I[y_i \leq q_{\tau_2}^{(1)}(w_i)] - \tau_2) - f_{U_{11}|W}(0|w_i)^{-1} (I[y_i \leq q_{\tau_1}^{(1)}(w_i)] - \tau_1) \right) + o_p(n^{-1/2}). \end{aligned}$$

An analogous linear representation can be derived for the term involving $(\Delta \hat{q}_\tau^{(0)}(w_i) - \Delta q_\tau^{(0)}(w_i))$, where we would replace d_i with $(1-d_i)$ in the above expression, and superscripts (1) with superscripts (0). Collecting both these terms we get this can be written as

$$\frac{1}{n} \sum_{i=1}^n \psi_i^- + o_p(n^{-1/2}).$$

We next consider the linear term of $\hat{\omega}_{ij}$ around ω_{ij} . This is of the form

$$\frac{1}{n(n-1)} \sum_{i \neq j} d_j(1-d_i) \omega'_{ij}(w_i + w_j)' (\hat{\delta} - \delta_0) \left[\frac{\Delta \hat{q}_\tau^{(1)}(w_i)}{\Delta \hat{q}_\tau^{(0)}(w_i)} - r \right]$$

where $\omega'_{ij} = h_{2n}^{-2} K_2' \left((w_i' \delta_0 + w_j' \delta_0) / h_{2n} \right)$.

Note we can replace $\frac{\Delta \hat{q}_\tau^{(1)}(w_i)}{\Delta \hat{q}_\tau^{(0)}(w_i)}$ with $\frac{\Delta q_\tau^{(1)}(w_i)}{\Delta q_\tau^{(0)}(w_i)}$ in the above expression. The resulting remainder term is $o_p(n^{-1/2})$ by the root- n consistency of $\hat{\delta}$ and the uniform consistency of the quantile estimators. In what follows, we derive an expression for the probability limit of

$$\frac{1}{n(n-1)} \sum_{i \neq j} d_j(1-d_i) \omega'_{ij} \left[\frac{\Delta q_\tau^{(1)}(w_i)}{\Delta q_\tau^{(0)}(w_i)} - r \right] (w_i + w_j)'$$

Using standard U-statistic projection theorems and the change of variables, the above term converges in probability to \mathcal{M} , defined in (4.23). Thus, the linear term in the expansion has the linear representation:

$$\frac{1}{n} \sum_{i=1}^n \mathcal{M} \psi_{\delta_i} + o_p(n^{-1/2}).$$

Finally we note higher order terms in the expansion of $\hat{\omega}_{ij}$ around ω_{ij} are asymptotically negligible by the uniform rates of convergence of the quantile estimators and the root- n consistency of $\hat{\delta}$. This completes the linear representation $\hat{r} - r$.

References

- [1] Abadie, A. (2002), “Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models,” *Journal of the American Statistical Association*, 97 (457), 284–292.
- [2] — (2003), “Semiparametric Instrumental Variable Estimation of Treatment Response Models,” *Journal of Econometrics*, 113, 231–263.
- [3] Abadie, A., Angrist, J., and Imbens, G. W. (2002), “Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings,” *Econometrica*, 70 (1), 91–117.
- [4] Acemoglu, D., (2002), “Technical Change, Inequality, and the Labor Market,” *Journal of Economic Literature*, 40 (457), 7–72.
- [5] Ahn, H., and Powell J. L. (1993), “Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism,” *Journal of Econometrics*, 58 (1–2), 3–29.
- [6] Angrist, J. (2004), “Treatment Effect Heterogeneity in Theory and Practice,” *Economic Journal*, 114, C52–C83.
- [7] Angrist, J., Chernozhukov, V., and Fernandez-Val, I. (2006), “Quantile Regression under Misspecification with an Application to the U.S. Wage Structure,” *Econometrica*, 539–564.
- [8] Angrist, J., Imbens, G. W., and Rubin D. B. (1996), “Identification of Causal Effects Using Instrumental Variables,” *Journal of the American Statistical Association*, 91(434), 444–472.
- [9] Autor, D., Katz, L. F., and Kearney M. S. (2006), “The Polarization of the U.S. Labor Market.” *American Economic Review Papers and Proceedings*, 96(2), 189 - 194.
- [10] Blau, F. D. and Kahn, L. M. (1994), “Rising Wage Inequality and the U.S. Gender Gap (in Rising Wage Inequality in the United States: Causes and Consequences),” *American Economic Review*, 84(2), 23–28. *Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association*.
- [11] Cameron, S. V., and Taber, C. (2004), “Estimation of Educational Borrowing Constraints Using Returns to Schooling,” *Journal of Political Economy*, 112 (1), 132–182.
- [12] Card, D. (1995), “Using Geographic Variations in College Proximity to Estimate the Return to Schooling,” in Louis N. Christofides and E. Kenneth Grant and Robert Swidinsky, eds., *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp*, Toronto: University of Toronto Press, pp. 201–222.
- [13] — (1999), “The Causal Effect of Education on Earnings,” in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, Elsevier Science B.V.
- [14] Chaudhuri, P. (1991a), “Global Nonparametric Estimation of Conditional Quantiles and their Derivatives,” *Journal of Multivariate Analysis*, 39(2), 246–269.
- [15] — (1991b), “Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation,” *Annals of Statistics*, 19(2), 760–777.

- [16] Chen, S. (1999), “Distribution-Free Estimation of the Random Coefficient Dummy Endogenous Variable Model,” *Journal of Econometrics*, 91 (1), 171–199.
- [17] Chen, S., and Khan, S. (2003), “Semiparametric Estimation of a Heteroscedastic Sample Selection Model,” *Econometric Theory*, 19, 1040–1064.
- [18] Chen, S. H. (2006), “Estimating the Variance in Wages in the Presence of Selection and Unobservable Heterogeneity,” forthcoming, *the Review of Economics and Statistics*.
- [19] Chernozhukov, V., and Hansen C. (2001), “Inference on Instrumental Quantile Regression Process for Structural and Treatment Effect Models,” *Journal of Econometrics*, forthcoming.
- [20] — (2004), “The Impact of 401(K) Participation on Savings: An IV-QR Analysis.” *Review of Economics and Statistics*, 86(3), 735-751.
- [21] — (2005), “An IV Model of Quantile Treatment Effects,” *Econometrica*, 73(1), 245-261.
- [22] Currie, J., and Moretti, E. (2003), “Mother’s Education and the Intergenerational Transmission of Human Capital,” *Quarterly Journal of Economics*, 118(4), 1495 -1532.
- [23] Donald, S. G. (1995), “Two-step Estimation of Heteroscedastic Sample Selection Models,” *Journal of Econometrics*, 65(2), 347–380.
- [24] Gould, E. D. (2002), “Rising Wage Inequality, Comparative Advantage, and the Growing Importance of General Skills in the United States,” *Journal of Labor Economics*, 20(1), 105-147.
- [25] Han, A. K. (1987), “Non-parametric Analysis of a Generalized Regression Model: The Maximum Rank Correlation Estimator,” *Journal of Econometrics*, 35(2-3), 303–316.
- [26] Heckman, J. J. (1979), “Sample Selection Bias as a Specification Error,” *Econometrica*, 47(1), 153–162.
- [27] — (1990), “Varieties of Selection Bias,” *American Economic Review Papers and Proceedings*, 90(2), 313–318.
- [28] Heckman, J. J. and Honore, B. E. (1990), “The Empirical Content of the Roy Model,” *Econometrica*, 58(5), 1121-1149.
- [29] Heckman, J. J. and Sedlacek, G. L. (1985), “Heterogeneity, Aggregation and Market Wage Functions: An Empirical Model of Self Selection in the Labor Market.” *Journal of Political Economy*, 93, 1077-1125
- [30] — (1990), “Self-Selection and the Distribution of Hourly Wages.” *Journal of Labor Economics*, 8, S329-S363
- [31] Heckman, J. J., and Vytlacil, E. J. (2000), “Local Instrumental Variables,” NBER Working Paper, T0252.
- [32] Imbens, G. W., and Angrist, J. D. (1994), “Identification and Estimation of Local Average Treatment Effects (in Notes and Comments),” *Econometrica*, 62(2), 467–475.
- [33] Juhn, C., Murphy, K. M., and Pierce, B. (1993), “Wage Inequality and the Rise in Returns to Skill,” *Journal of Political Economy*, 101(3), 410–442.

- [34] Kane, T. J., and Rouse, C. E. (1993), “Labor-Market Returns to Two- and Four-Year College: Is a Credit a Credit and Do Degrees Matter?,” NBER Working Paper, W4268.
- [35] — (1995), “Labor-Market Returns to Two- and Four-Year College,” *American Economic Review*, 85(3), 600–614.
- [36] Katz, L. F., and Autor, D. (1999), “Changes in the Wage Structure and Earnings Inequality,” in O. Ashenfelter and D. Card (eds.), *Handbook of Labor Economics*, Volume 3A. Amsterdam: Elsevier Science.
- [37] Khan, S., and Powell, J. L. (2001), “Two Step Estimation of Semiparametric Censored Regression Models,” *Journal of Econometrics*, 103, 73-110.
- [38] Kling, J. R. (2001), “Interpreting Instrumental Variables Estimates of the Returns to Schooling,” *Journal of Business and Economic Statistics*, 19, 358-64.
- [39] Lee, L. F. (1982), “Some Approaches to the Correction of Selectivity Bias,” *Review of Economic Studies*, 49(3), 355-372
- [40] Lemieux, T. (1998), “Estimating the Effects of Unions on Wage Inequality in a Panel Data Model with Comparative Advantage and Nonrandom Selection,” *Journal of Labor Economics*, 16(2), 261–291.
- [41] — (2006), “Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill,” *American Economic Review*, 1-64.
- [42] Newey, W. K., and McFadden, D.. (1994), “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics*, chapter 4, 2111-2245.
- [43] Powell, J. L. (1989), “Semiparametric Estimation of Bivariate Latent Variable Models,” University of Wisconsin, unpublished manuscript.
- [44] Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), “Semiparametric Estimation of Index Coefficients,” *Econometrica*, 57, 1403–1430.
- [45] Rouse, C. E. (1995), “Democratization or Diversion? The Effect of Community Colleges on Educational Attainment,” *Journal of Business and Economic Statistics*, 13(2), 217–224.
- [46] Silverman, B. W. (1986), *Density Estimation*, Chapman and Hall, London.
- [47] Vytlačil E. (2002), “Independence, Monotonicity, and Latent Index Models: An Equivalence Result” *Econometrica*, 70(1), 331-341.
- [48] Willis, R.J. and Rosen, S. (1979), “Education and Self-Selection.” *Journal of Political Economy*, 87(5), S7-S36.