cemmap
centre for microdata methods and practice

Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality

Pedro Carneiro
Sokbae Lee

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP01/09

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Estimating Distributions of Potential Outcomes Using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality

Pedro Carneiro and Sokbae Lee*
University College London, Institute for Fiscal Studies
and Centre for Microdata Methods and Practice

January 2009

Abstract

This paper extends the method of local instrumental variables developed by Heckman and Vytlacil (1999, 2001, 2005) to the estimation of not only means, but also distributions of potential outcomes. The newly developed method is illustrated by applying it to changes in college enrollment and wage inequality using data from the National Longitudinal Survey of Youth of 1979. Increases in college enrollment cause changes in the distribution of ability among college and high school graduates. This paper estimates a semiparametric selection model of schooling and wages to show that, for fixed skill prices, a 14% increase in college participation (analogous to the increase observed in the 1980s), reduces the college premium by 12% and increases the 90-10 percentile ratio among college graduates by 2%.

Keywords: Comparative advantage, composition effects, local instrumental variables, marginal treatment effect, semiparametric estimation, wage inequality.

JEL classification codes: C14; C31; J31.

*Corresponding address: Department of Economics, University College London, London, WC1E 6BT, UK; emails: p.carneiro@ucl.ac.uk (Carneiro) and l.simon@ucl.ac.uk (Lee). We thank an editor, an associate editor, and three anonymous referees for helpful comments. Especially, we thank one anonymous referee for pointing out a mistake in the previous version of this paper. We also thank participants in numerous seminars for useful comments, especially Rita Almeida, Joe Altonji, David Autor, Richard Blundell, Olympia Bover, Peter Gottschalk, Alan Krueger, Alan Manning, Costas Meghir, Steve Pischke and Tom Stoker. This research is supported in part by the Economic and Social Research Council (ESRC) Research Grant RES-000-22-2542. We thank the Leverhulme Trust and ESRC (RES-589-28-0001) through the funding of the Centre for Microdata Methods and Practice (Carneiro and Lee) and of the research programme *Evidence, Inference and Inquiry* (Lee). Carneiro also thanks the Poverty Unit of the World Bank Research Group and Georgetown for their hospitality. Earlier versions of the paper were circulated under the titles "Changes in College Enrollment and Wage Inequality: Distinguishing Price and Composition Effects" and "Ability, Sorting and Wage Inequality".

1 Introduction

The potential outcomes framework has been increasingly popular in applied research. In a series of papers, Heckman and Vytlacil (1999, 2001, 2005) developed the method of local instrumental variables in nonparametric selection models using potential outcomes. Heckman, Urzua, and Vytlacil (2006) further extended the method of local instrumental variables and Aakvik, Heckman, and Vytlacil (2005) used factor structures for the analysis of the latent variable framework of Heckman and Vytlacil (1999, 2001, 2005). Vytlacil and Yildiz (2007) considered marginal means of potential outcomes in weakly separable models. Moffitt (2008) proposed a nonparametric series estimation method of estimating marginal treatment effects in heterogeneous populations. This paper makes two new contributions to this literature: first, we show how to extend the method of local instrumental variables of Heckman and Vytlacil to identify the distributions of potential outcomes; second, we develop a semiparametric method for estimating the entire marginal distributions of potential outcomes.

Distributions of potential outcomes are useful for policy makers who care about distributional effects of policies. To our best knowledge, estimation of marginal distributions of potential outcomes has been considered by Imbens and Rubin (1997) and Abadie (2002, 2003).[1] However, these three papers develop estimators under the local average treatment effect (LATE) framework of Imbens and Angrist (1994). They are useful for evaluating the effects of polices in place, but not for forecasting those of new polices. One could estimate a structural econometric model that describes individual choices and corresponding outcomes to predict the distributional effects of new polices, but this would involve stringent parametric and functional-form assumptions on the econometric model. In this paper, we provide an alternative method that can be used to evaluate the distributional effect of a new policy without specifying a complete parametric model. Moreover, since quantile treatment effects are defined as the differences between quantiles of marginal distributions of potential outcomes, we also contribute to the literature on quantile treatment effects and on instrumental variables estimation of quantile regression models.[2]

We apply our method to investigate changes in college enrollment and wage inequality in the United States. College enrollment doubled from 30% to 60% between 1960 and 2000 in the United States. Such a large increase in college enrollment rates is bound to cause changes in the quality of college and high school workers. As a result, we cannot compare measures of the college premium and within group inequality across different periods. Trends in the college premium and wage inequality confound changes in prices and changes in composition, and it is important to separate the two.

[1] A recent working paper by Chen and Khan (2007) develops estimators of the scale ratio between potential outcomes under some symmetry conditions on the joint distribution of outcome and selection errors.

[2] Some recent papers in this literature include: Abadie, Angrist, and Imbens (2002), Chesher (2003), Imbens and Newey (2003), Chernozhukov and Hansen (2005, 2006), Chernozhukov, Imbens, and Newey (2007), Ma and Koenker (2006), Lee (2006), and Horowitz and Lee (2007) among others.

The goal of our empirical exercise is to uncover the empirical magnitude of this problem, generally called composition effect. In order to do so, we estimate a semiparametric model of heterogeneous agents self-selecting into college, and uncover the magnitude of selection observed in the data.[3] We use the resulting estimates to characterize the distributions of wages for individuals enrolling either in college or in high school at a given point in time, and how they change in response to changes in college enrollment. We find that, for fixed skill prices, an increase of 14% in the proportion of college-goers (of similar magnitude to the one observed in the 1980s) leads to: i) a reduction of 12% in the college premium; ii) a 2% increase in the ratio of the 90th to the 10th percentile (P90-P10) of the college wage distribution; iii) and no change in the P90-P10 ratio of the high school wage distribution.

The remainder of this paper is organized as follows. In section 2 we present a simple econometric model which underlies our empirical work. In section 3, we describe the basic ideas behind a semiparametric estimation procedure based on Section 2, and in section 4 we describe in detail the way we apply it. Secion 5 presents asymptotic distributions for our estimators. The data we use are described in Section 6. In section 7 we apply our model to the study of wage inequality using white males in the NLSY. Using our estimates we document the patterns of sorting of individuals to different levels of schooling and the empirical importance of selection bias and composition effects. Section 8 gives some concluding remarks. In the Appendix we provide a detailed description of the data, further details of our estimation procedure, and proofs of theorems given in Section 5.

2 The Econometric Model and Identification of Potential Outcome Distributions

The econometric model we consider is that of Heckman and Vytlacil (1999, 2001, 2005).[4] Let Y_1 and Y_0 be potential individual outcomes in two states, 1 and 0. In this paper, Y_1 is the log college wage and Y_0 is the log high school wage, as in Willis and Rosen (1979).

We assume

$$Y_1 = \mu_1(X, U_1) \quad \text{and} \quad Y_0 = \mu_0(X, U_0), \tag{2.1}$$

where X is a vector of observed random variables influencing potential outcomes, μ_1 and μ_0 are unknown functions, and U_1 and U_0 are unobserved random variables.

[3] One measure of the importance of selection is, say, the OLS-IV gap in estimates of the returns to schooling. As discussed in Card (2001), the usual finding is that instrumental variables (IV) estimates of the return to one year of schooling are above ordinary least squares (OLS) estimates of the same parameter by 2 to 3 percentage points (corresponding to 25 to 50% of the size of the OLS coefficient). Much of the literature on inequality studies the evolution of the college premium, estimated as the difference in log wages of individuals with 12 and 16 years of education. If we extrapolated the reported OLS-IV gap to four years of schooling we would get something on the order of 8-12% percentage points. This corresponds to roughly 20 to 25% of the college premium in 1980, and almost 40 to 50% of its increase in between 1980 and 1990 (Katz and Murphy, 1992).

[4] For example, see Section 2 of Heckman and Vytlacil (2005).

We assume that individuals choose to be in state 1 or 0 (prior to the realizations of the outcomes of interest) according to the following equation:

$$S = 1 \text{ if } \mu_S(Z) - U_S > 0, \tag{2.2}$$

where Z is a vector of observed random variables influencing the decision equation, μ_S is an unknown function of Z, and U_S is an unobserved random variable. In this paper, equation (2.2) can be interpreted as the reduced form of an economic model of college attendance.[5] The advantage of specifying it this way is the relatively little structure it imposes on the model. In particular, Vytlacil (2002) shows that the independence and monotonicity assumptions needed to interpret instrumental variables estimates in a model of heterogeneous returns (e.g., Imbens and Angrist, 1994) imply that the data can be rationalized with the model of equations (2.1) and (2.2) (as long as one does not impose parametric functional forms and distributional assumptions on the model). This result guarantees that our model is consistent with the IV estimates of the returns to college that can be produced in our data.

For each individual, the observed outcome Y is

$$Y = SY_1 + (1 - S)Y_0.$$

The set of variables in X can be a subset of Z. For identification, assume that there is at least one variable in Z that is not in X (exclusion restriction). As in Heckman and Vytlacil (2001,2005), we can rewrite (2.2) as:

$$S = 1 \text{ if } P > V,$$

where $V = F_{U_S|X,Z}[U_S|X, Z]$, $P = F_{U_S|X,Z}[\mu_S(Z)|X, Z]$, and $F_{U_S|X,Z}(u_s|x, z)$ is the CDF of U_S conditional on $X = Z$ and $Z = z$.[6] Note that for any arbitrary distribution of U_S conditional on X and Z, by definition, $V \sim \text{Unif}[0, 1]$ conditional on X and Z.

We make the following assumptions as in Heckman and Vytlacil (2005).

Assumption 1. *Assume that (1) $\mu_S(Z)$ is a nondegenerate random variable conditional on X; (2) (U_1, U_S) and (U_0, U_S) are independent of Z conditional on X; (3) The distribution of U_S conditional on (X, Z) and that of $\mu_S(Z)$ conditional on X are absolutely continuous with respect to Lebesgue measure; and (4) For a measurable function G, $E|G(Y_1)| < \infty$, and $E|G(Y_0)| < \infty$.*

The following theorem provides identification of our objects of interest in the nonparametric model given by (2.1) and (2.2).

[5] Carneiro, Heckman and Vytlacil (2007) use this model to study heterogeneity in the returns to college and present an economic model that can justify the specification in (2.2).

[6] Throughout the paper, for any random vector X, $f_X(x)$ denotes the PDF of X and $F_X(x)$ denotes the CDF of X. In addition, for any random variables X and Y, $f_{Y,X}(y, x)$, $f_{Y|X}(y|x)$, and $F_{Y|X}(y|x)$ denote the joint PDF of Y and X and the conditional PDF and CDF of Y on $X = x$, respectively. We suppress subscripts in the notation whenever this can be done without causing confusion.

Theorem 1. *Consider the nonparametric selection model given by (2.1) and (2.2). Let* $V = F_{U_S|X,Z}[U_S|X,Z]$ *and* $P = F_{U_S|X,Z}[\mu_S(Z)|X,Z]$. *Let Assumption 1 hold. Then*

$$\begin{aligned} E\left[G(Y_1)|X=x,V=p\right] &= E\left[G(Y)|X=x,P=p,S=1\right] \\ &\quad + p\frac{\partial E\left[G(Y)|X=x,P=p,S=1\right]}{\partial p} \\ E\left[G(Y_0)|X=x,V=p\right] &= E\left[G(Y)|X=x,P=p,S=0\right] \\ &\quad - (1-p)\frac{\partial E\left[G(Y)|X=x,P=p,S=0\right]}{\partial p} \end{aligned}$$

provided that $E\left[G(Y)|X=x,P=p,S=1\right]$ *and* $E\left[G(Y)|X=x,P=p,S=0\right]$ *are continuously differentiable with respect to* p *for almost every* x.

Proof. Assumptions (1) and (3) ensure that P is a nondegenerate, continuously distributed random variable conditional on X. Assumption (4) is needed to ensure that expectations considered below are finite. Notice that

$$\begin{aligned} E\left[G(Y)|X=x,P=p,S=1\right] &= E\left[G(Y)|X=x,P=p,V<p\right] \\ &= \int_0^p E\left[G(Y_1)|X=x,V=v\right] f_{V|X}(v|x)dv/p \\ &= \int_0^p E\left[G(Y_1)|X=x,V=v\right] \int f_{V|X,Z}(v|x,z)f_{Z|X}(z|x)dz\,dv/p \\ &= \int_0^p E\left[G(Y_1)|X=x,V=v\right] dv/p, \end{aligned}$$

where the second equality follows from assumption (2), the fourth equality comes from the fact that V is uniformly distributed on $[0,1]$ conditional on X and Z. The first conclusion follows by multiplying both sides of the equation above by p and differentiating both sides with respect to p. The proof of the second conclusion is similar. ∎

This theorem extends the identification results of Heckman and Vytlacil (1999, 2001, 2005).[7] The conditional means of Y_1 and Y_0 given $X=x$ and $V=v$ are identified by taking $G(Y)=Y$ and therefore the marginal treatment effect (MTE), defined as $E\left(Y_1-Y_0|X=x,V=v\right)$, is identified. Furthermore, the conditional distributions of Y_1 and Y_0 given $X=x$ and $V=v$ are identified by choosing $G(Y)=1(Y\leq y)$, where $1(\cdot)$ is the standard indicator function, and therefore the conditional densities and quantiles are also identified.

Notice that we can only identify $E\left[G(Y_1)|X=x,V=p\right]$ over the support of P for individuals in $S=1$ conditional on $X=x$, and $E\left([G(Y_0)|X=x,V=p\right]$ over the support of P for individuals in $S=0$

[7] The identification results of Heckman and Vytlacil (1999, 2001, 2005) are mainly concerned with average treatment effects. Vytlacil and Yildiz (2007) develop identification results for the marginal means of potential outcomes in weakly separable models. We identify not only average treatment effects but also whole marginal distributions of potential outcomes.

conditional on $X = x$. As a consequence, we can only identify the MTE over the common support of P for individuals in $S = 1$ and $S = 0$ conditional on $X = x$.

The identification result in Theorem 1 is very general since it does not impose any restrictions on the functional forms of μ_1 and μ_0 in (2.1). However, such a flexible framework has some disadvantages that limit its practical usefulness. One important disadvantage is that the precision of a nonparametric estimator based on Theorem 1 decreases rapidly as the number of continuously distributed components of X increases (curse of dimensionality). Another disadvantage is that it is difficult to have full support of P for some observed values of X, thereby implying that treatment parameters such as the MTE or counterfactuals distributions such as $F_{Y_1}(y_1)$ and $F_{Y_0}(y_0)$ are not identified. To circumvent these disadvantages, we specify and estimate a separable version of (2.1) under a more stringent assumption on unobservables, but one that is relatively standard in empirical work: we assume that (U_1, U_S) is independent of Z as well as independent of X; likewise for (U_0, U_S). The assumption of separability implies the following modification in our model:

$$\begin{aligned} Y_1 &= \mu_1(X) + U_1 \\ Y_0 &= \mu_0(X) + U_0, \end{aligned} \tag{2.3}$$

as opposed to $Y_1 = \mu_1(X, U_1)$ and $Y_0 = \mu_0(X, U_0)$. In addition, we impose flexible but parametric forms for (μ_1, μ_0) and a semiparametric form for μ_S on the model so that estimating the model reduces to a feasible semiparametric estimation problem. The main reason why we adopt this particular semiparametric specification is that it is relatively more difficult to model parametric relationships among unobservables, (U_1, U_S) and (U_0, U_S) than those among observables. Exact specifications are given in section 7.1.[8]

3 Semiparametric Estimation

This section describes semiparametric estimators of the expectations, quantiles, and marginal distributions of Y_1 and Y_0 conditional on $X = x$ and $V = v$. We consider a semiparametric selection model given by (2.2) and (2.3). From now on, let $V = F_{U_S}[U_S]$ and $P = F_{U_S}[\mu_S(Z)]$. We now assume that:

Assumption 2. *Assume that (1) $\mu_S(Z)$ is a nondegenerate random variable conditional on X; (2) (U_1, U_S) and (U_0, U_S) are independent of (Z, X); (3) The distributions of U_S and $\mu_S(Z)$ are absolutely continuous with respect to Lebesgue measure; (4) $E|Y_1| < \infty$ and $E|Y_0| < \infty$; (5) $0 < Pr(S = 1|Z) < 1$;*

[8] One important advantage of estimating a flexible model over a complete parametric model is that the sources of variation in the data that identify the model are very clear, as are the types of simulations that can be performed. For example, suppose that, empirically, the variables Z are never high enough or low enough to induce full participation in college, or no participation in college. In this case we cannot estimate the full distribution of unobserved heterogeneity, and we cannot simulate economies where college participation rates are 0 or 1 without imposing more structure in the model (e.g., Heckman and Vytlacil, 2005, Ichimura and Taber, 2000). The use of semiparametric methods forces discipline both in the reporting of results, and in the construction of simulations within the range of the data.

and (6) $E[U_1|P=p,S=1]$*,* $E[U_0|P=p,S=0]$*,* $f_{U_1|P,S=1}(u_1|p)$ *and* $f_{U_0|P,S=0}(u_0|p)$ *are continuously differentiable with respect to* p*.*

We first consider estimation of $E[Y_1|X=x,V=v]$ and $E[Y_0|X=x,V=v]$. Under the assumption that (U_1,U_S) and (U_0,U_S) are independent of X,

$$E[Y_1|X=x,V=v]=\mu_1(x,\beta_1)+E[U_1|V=v],$$

and

$$E[Y_0|X=x,V=v]=\mu_0(x,\beta_0)+E[U_0|V=v],$$

where the functional forms of μ_1 and μ_0 are specified up to finite dimensional parameters β_1 and β_0. Thus, estimates of $E[Y_1|X=x,V=v]$ and $E[Y_0|X=x,V=v]$ can be obtained by estimating β_1, β_0, $E[U_1|V=v]$, and $E[U_0|V=v]$.

First, we estimate β_1 and β_0 using a semiparametric version of the sample selection estimator of Das, Newey, and Vella (2003). Notice that under the assumption that U_1 and V are independent of X and Z, we have

$$E[Y|X=x,P=p,S=1]=\mu_1(x,\beta_1)+\lambda_1(p), \tag{3.1}$$

where $\lambda_1(\cdot)$ is an unknown function of P. Equation (3.1) suggests that β_1 can be estimated by a partially linear regression of Y on X and P using only observations with $S=1$. Since P is unobserved, Das, Newey, and Vella (2003) suggest a two-step procedure. The first step consists in the construction of the estimated P and the second step consists in the estimation of β_1 using the estimated P. In this paper, the first step is carried out by a series regression of S on Z. In particular, we approximate $\mu_S(z)$ by some linear parts and some nonparametric parts. The second step is accomplished using a Robinson (1988)-type estimator with the estimated P.[9] Analogously, β_0 can be estimated by a partially linear regression of Y on X and estimated P using only observations with $S=0$. See Section 4.1 for a detailed description of our estimators of P, β_1 and β_0.

We now consider estimation of $E[U_j|V=v]$ for $j=0,1$. It follows from directly applying Theorem 1 with $G(u)=u$:

$$E[U_1|V=v] = E[U_1|P=v,S=1]+v\frac{\partial E[U_1|P=v,S=1]}{\partial p} \tag{3.2}$$

$$E[U_0|V=v] = E[U_0|P=v,S=0]-(1-v)\frac{\partial E[U_0|P=v,S=0]}{\partial p}. \tag{3.3}$$

Equations (3.2) and (3.3) are the basis for nonparametric estimators of $E[U_1|V=v]$ and $E[U_0|V=v]$ proposed in this paper.

[9] Series estimation is used in Das, Newey, and Vella (2003) for both the first and second steps. See also Heckman, Ichimura, Smith and Todd (1998).

Local polynomial estimation is used in the paper to estimate $E(U_1|P=v,S=1)$ (which corresponds to $\lambda_1(p)|_{p=v}$ in equation (3.1)), $E(U_0|P=v,S=0)$ and their partial derivatives with respect to P. This is because local polynomial estimation not only provides a unified framework for estimating both a function and its derivative but also has a variety of desirable properties in comparison to other available nonparametric methods.[10] See Section 4.2 for detailed description of nonparametric estimators of $E[U_1|V=v]$ and $E[U_0|V=v]$.

Finally, notice that $f_{Y_1|X,V}(y_1|x,v)$ and $f_{Y_0|X,V}(y_0|x,v)$ can be obtained by location shifts from $f_{U_1|V}(u_1|v)$ and $f_{U_0|V}(u_0|v)$, i.e.,

$$f_{Y_1|X,V}(y_1|x,v) = f_{U_1|V}(y_1 - \mu_1(x,\beta_1)|v) \quad \text{and}$$
$$f_{Y_0|X,V}(y_0|x,v) = f_{U_0|V}(y_0 - \mu_0(x,\beta_0)|v).$$

To obtain $f_{Y_1|X,V}(y_1|x,v)$ and $f_{Y_0|X,V}(y_0|x,v)$, once we know β_1 and β_0 we only need to estimate $f_{U_1|V}(u_1|v)$ and $f_{U_0|V}(u_0|v)$. As in (3.2) and (3.3), we can obtain identifying relationships for $f_{U_1|V}(u_1|v)$ and $f_{U_0|V}(u_0|v)$ and resulting sample analog estimators can be constructed. Note that given estimators of PDF's, it is straightforward to obtain estimators of corresponding CDF's by integrating the estimated PDF's, and to obtain estimators of corresponding quantiles by inverting the estimated CDF's. In section 4.3 we describe in detail the corresponding nonparametric estimators.

Heckman and Vytlacil (2001,2005) show how we can construct a variety of treatment effect parameters as weighted averages of $E(Y_1 - Y_0|X=x,V=v)$, and develop weights for several parameters of interest. Drawing on their work, we can estimate $E[Y_j]$, $E[Y_j|S=1]$, and $E[Y_j|S=0]$ by integrating out our estimator of $E[Y_j|X=x,V=v]$ with some suitable weights for $j=0,1$. Specifically, we obtain estimators of $E[Y_j]$, $E[Y_j|S=1]$, and $E[Y_j|S=0]$ by the sample analogs of the following formulae:

$$\begin{aligned} E[Y_j] &= \int\int_0^1 E[Y_j|X=x,V=v] f_X(x)\, dv\, dx, \\ E[Y_j|S=1] &= \int\int_0^1 E[Y_j|X=x,V=v] \frac{1-F_{P|X}(v|x)}{\Pr(S=1)} f_X(x)\, dv\, dx, \\ &\text{and} \\ E[Y_j|S=0] &= \int\int_0^1 E[Y_j|X=x,V=v] \frac{F_{P|X}(v|x)}{\Pr(S=0)} f_X(x)\, dv\, dx. \end{aligned} \tag{3.4}$$

for $j=0,1$. See Appendix B.1 for details on implementing (3.4). Using estimates of these conditional

[10] Fan and Gijbels (1996) provide a detailed discussion of the properties of local polynomial estimators. The advantages of the local polynomial estimators are that (1) the form of bias is simpler than that of the standard kernel estimator, (2) it adapts to various types of distributions of explanatory variables, (3) it does not require boundary modifications to achieve the same convergence rate, and (4) it has very good minimax efficiency property.

expectations, standard treatment effect parameters can be estimated:

$$\begin{aligned}
\text{ATE (Average Treatment Effect)} &= E[Y_1] - E[Y_0], \\
\text{TT (Average Treatment Effect on the Treated)} &= E[Y_1|S=1] - E[Y_0|S=1], \\
\text{TUT (Average Treatment Effect on the Untreated)} &= E[Y_1|S=0] - E[Y_0|S=0], \\
\text{OLS (Ordinary Least Squares)} &= E[Y_1|S=1] - E[Y_0|S=0].
\end{aligned}$$

$E[Y_1|S=1]$ and $E[Y_0|S=0]$ can also be estimated directly by taking sample means of observed college and high school wages. Therefore, comparison between model-based and direct estimates of $E[Y_1|S=1]$ and $E[Y_0|S=0]$ provides a goodness-of-fit check of our model.[11] Similarly, integrating our estimators of $f_{Y_j|X,V}(y_j|x,v)$ for $j=0,1$ with the weights in (3.4), we can obtain estimators of $f_{Y_j}(\cdot)$, $f_{Y_j|S=1}(\cdot|S=1)$, and $f_{Y_j|S=0}(\cdot|S=0)$ for $j=0,1$. Note that $f_{Y_1|S=1}(\cdot|S=1)$ and $f_{Y_0|S=0}(\cdot|S=0)$ can also be estimated directly by taking sample analogs of observed college and high school wages, which again allows us to do a goodness-of-fit check of our model.

4 Details of Estimation Procedure

4.1 Estimating P, β_1 and β_0

This section provides a detailed description of our estimators of P, β_1 and β_0. Assume that the data consist of i.i.d. observations $\{(Y_i, S_i, X_i, Z_i) : i = 1, \ldots, n\}$. First, we consider series estimation of P. In Section 3, $P = F_{U_S}[\mu_S(Z)] = \Pr(S=1|Z)$. In order to avoid the curse of dimensionality, we model $\Pr(S=1|Z=z)$ as a partially linear additive regression model:

$$\Pr(S=1|Z=z) = \varphi_1(z_1) + \ldots + \varphi_d(z_d) + z_{pc}'\vartheta, \tag{4.1}$$

where $z = (z_c, z_{pc})$, $z_c = (z_1, \ldots, z_d)'$ is a d-dimensional vector of continuous random variables (non-parametric components), z_{pc} is a vector of parametric components, $\varphi_1, \ldots, \varphi_d$ are unknown functions, and ϑ is a vector of unknown parameters. The partially linear additive structure in (4.1) is adopted to have a good precision in our estimation procedure.

To describe the series estimator, let $\{p_k : k = 1, 2, \ldots\}$ denote a basis for real-valued smooth functions defined on $\mathbb{R}$ such that a linear combination of $\{p_k : k = 1, 2, \ldots\}$ can approximate $\varphi_j(\cdot)$ for each $j = 1, \ldots, d$ as the number of approximating functions increases to infinity. For any positive integer κ, define

$$P_\kappa(z) = [p_1(z_1), \ldots, p_\kappa(z_1), \ldots, p_1(z_d), \ldots, p_\kappa(z_d), z_{pc}]'.$$

[11] In this paper, we have not yet developed asymptotic distribution theory for average treatment effects, such as ATE, TT, and TUT, nor asymptotic properties of the proposed goodness-of-fit check. These are topics for future research.

Then for each i, the series estimator of $P(Z_i)$ is

$$\tilde{P}(Z_i) = P_\kappa(Z_i)'\hat{\theta}_{n\kappa},$$

where

$$\hat{\theta}_{n\kappa} = \left[\sum_{i=1}^{n} P_\kappa(Z_i)P_\kappa(Z_i)'\right]^{-1}\left[\sum_{i=1}^{n} P_\kappa(Z_i)S_i\right].$$

In finite samples, estimated $P(Z_i)$'s might be negative or larger than one. To solve this, our estimator is a trimmed version:

$$\hat{P}(Z_i) = \tilde{P}(Z_i) + (1-\delta-\tilde{P}(Z_i))1(\tilde{P}(Z_i) > 1) + (\delta - \tilde{P}(Z_i))1(\tilde{P}(Z_i) < 0) \tag{4.2}$$

for sufficiently small $\delta > 0$.[12]

We now consider estimation of β_1 and β_0. For convenience, we assume linear-in-parameters forms for μ_1 and μ_0, that is $\mu_j(x,\beta_j) = \mu_j(x)'\beta_j$ for each $j = 0, 1$. Then β_1 and β_0 are estimated as in Robinson (1988) (using the estimated rather than the true P) with the $S = 1$ subsample and the $S = 0$ subsample, respectively. Specifically, for $j = 0, 1$,

$$\begin{aligned}\hat{\beta}_j = &\left[\sum_{i=1}^{n} W_{ji}\left\{\mu_j(X_i) - \hat{E}_h\left[\mu_j(X_i)\middle|\hat{P}(Z_i), W_{ji}\right]\right\}\left\{\mu_j(X_i) - \hat{E}_h\left[\mu_j(X_i)\middle|\hat{P}(Z_i), W_{ji}\right]\right\}'\right]^{-1} \\ &\times\left[\sum_{i=1}^{n} W_{ji}\left\{\mu_j(X_i) - \hat{E}_h\left[\mu_j(X_i)\middle|\hat{P}(Z_i), W_{ji}\right]\right\}\left\{Y_i - \hat{E}_h\left[Y_i\middle|\hat{P}(Z_i), W_{ji}\right]\right\}\right],\end{aligned} \tag{4.3}$$

where $W_{ji} = 1(Z_i \in \mathcal{Z})(S_i)^{1(j=1)}(1-S_i)^{1(j=0)}$ and $\hat{E}_h[\cdot|\cdot]$ denotes the kernel mean regression estimator with a bandwidth h. Here, $\mathcal{Z}$ is a strict subset of the support of Z. A trimming function of the form $1(Z_i \in \mathcal{Z})$ is considered here to avoid unduly influences of outliers of Z. Alternatively, one could consider $W_{n,ji} = (S_i)^{1(j=1)}(1-S_i)^{1(j=0)}\omega_{n,ji}$, where $\omega_{n,ji}$ is some trimming function that may converge to one at a certain asymptotic rate.

4.2 Estimating $E[U_1|V=v]$ and $E[U_0|V=v]$

This section gives a detailed description of nonparametric estimators of $E[U_1|V=v]$ and $E[U_0|V=v]$. First consider local polynomial estimation of $E[U_1|V=v]$. In general, use of higher order polynomials may reduce the bias but increase the variance by introducing more parameters. Fan and Gijbels (1996) suggest that the order π of polynomial be equal to $\pi = \mu + 1$, where μ is the order of the derivative of the function of interest. That is, Fan and Gijbels (1996) recommend a local linear estimator for fitting a function and a local quadratic estimator for fitting a first-order derivative. Following their

[12] Alternatively, one may develop the series estimator of P based on a logit or probit model, so that the fitted probability always lies between 0 and 1.

suggestions, $E(U_1|P=v,S=1)$ is estimated by a local linear estimator using observations with $S=1$ and $\partial E(U_1|P=v,S=1)/\partial p$ is estimated by a local quadratic estimator.

To be more specific, let $\{(\hat{U}_{1i},\hat{P}_i,S_i):i=1,\ldots,n\}$ denote observations of estimated U_1 and P along with S, where $\hat{U}_{1i}=Y_i-\mu_1(X_i,\hat{\beta}_1)$ for $i=1,\ldots,n$. The local linear estimator $\hat{E}(U_1|P=v,S=1)$ is obtained by solving the problem

$$\min_{c_0,c_1}\sum_{i=1}^{n}S_i\left[\hat{U}_{1i}-c_0-c_1(\hat{P}_i-v)\right]^2K\left(\frac{\hat{P}_i-v}{h_{n1}}\right),$$

where $K(\cdot)$ is a kernel function and h_{n1} is a bandwidth. The resulting value of c_0 is the local linear estimator of $E(U_1|P=v,S=1)$. Similarly, the local quadratic estimator $\hat{\partial}E(U_1|P=v,S=1)/\partial p$ is obtained by solving the problem

$$\min_{c_0,c_1,c_2}\sum_{i=1}^{n}S_i\left[\hat{U}_{1i}-c_0-c_1(\hat{P}_i-v)-c_2(\hat{P}_i-v)^2\right]^2K\left(\frac{\hat{P}_i-v}{h_{n2}}\right),$$

where h_{n2} is a bandwidth that can be different from h_{n1}. The resulting value of c_1 is the local quadratic estimator of $\partial E(U_1|P=v,S=1)/\partial p$. Then the estimator of $E[U_1|V=v]$ is given by

$$\hat{E}[U_1|V=v]=v\frac{\hat{\partial}}{\partial p}E(U_1|P=v,S=1)+\hat{E}(U_1|P=v,S=1). \tag{4.4}$$

Similarly, the estimator of $E[U_0|V=v]$ can be obtained by replacing unknown functions in the right hand side of (3.3) with their nonparametric estimators.

4.3 Estimating $f(u_1|v)$ and $f(u_0|v)$

This section describes nonparametric estimators of $f(u_1|v)$ and $f(u_0|v)$. As in (3.2) and (3.3), an application of Theorem 1 yields the following relationships

$$f_{U_1|V}(u_1|v)=f_{U_1|P,S=1}(u_1|v,S=1)+v\frac{\partial}{\partial p}f_{U_1|P,S=1}(u_1|v,S=1) \text{ and} \tag{4.5}$$

$$f_{U_0|V}(u_0|v)=f_{U_0|P,S=0}(u_0|v,S=0)-(1-v)\frac{\partial}{\partial p}f_{U_0|P,S=0}(u_0|v,S=0). \tag{4.6}$$

Sample analogs of the right-hand sides of equations (4.5) and (4.6) can be obtained by some suitable nonparametric estimators.

We only discuss estimation of $f\left(u_1|v\right)$ in detail, since estimation of $f\left(u_0|v\right)$ is similar. To develop an estimator of $f\left(u_1|v\right)$ using the equation (4.5), it is necessary to estimate $f_{U_1|P,S=1}(u_1|p,S=1)$ and its derivative with respect to p. Specifically, the estimator of $f\left(u_1|v\right)$ can be obtained by

$$\hat{f}(u_1|v)=v\frac{\hat{\partial}}{\partial p}f_{U_1|P,S=1}(u_1|v,S=1)+\hat{f}_{U_1|P,S=1}(u_1|v,S=1), \tag{4.7}$$

where $\hat{f}_{U_1|P,S=1}(u_1|v,S=1)$ and $\hat{\partial}f_{U_1|P,S=1}(u_1|v,S=1)/\partial p$ are defined below.

In order to compute $\hat{f}_{U_1|P,S=1}(u_1|v, S=1)$ and $\hat{\partial} f_{U_1|P,S=1}(u_1|v, S=1)/\partial p$ in (4.7), we begin with estimated data $\{(\hat{U}_{1i}, \hat{P}_i) : i = 1, \ldots, n, S_i = 1\}$, where $\hat{U}_{1i} = Y_i - \mu_1(X_i, \hat{\beta}_1)$. One could estimate the conditional density of U_1 given P and its derivative by estimating the joint and marginal densities using the standard kernel density estimators, taking the ratio between them to estimate the conditional density, and finally computing a derivative of the conditional density. This procedure would yield consistent estimators but it is quite cumbersome. Instead we use a direct method of Fan, Yao, and Tong (1996), who develop local polynomial estimators of the conditional density function and its derivative. To motivate the estimators of Fan, Yao, and Tong (1996), notice that, as $\delta_n \to 0$,

$$E\left[\delta_n^{-1} K\left(\frac{U_1 - u_1}{\delta_n}\right) \middle| P = v, S = 1\right] \approx f_{U_1|P,S=1}(u_1|v, S=1)$$
$$\approx f_{U_1|P,S=1}(u_1|v_0, S=1) + \frac{\partial}{\partial p} f_{U_1|P,S=1}(u_1|v_0, S=1)(v - v_0)$$

for any v in a neighborhood of v_0, where K is a nonnegative density function and δ_n is a bandwidth. This suggests that the local linear estimator of $f_{U_1|P,S=1}(u_1|v, S=1)$ can be defined as $\hat{f}_{U_1|P,S=1}(u_1|v, S = 1) \equiv \hat{c}_0$, where $(\hat{c}_0, \hat{c}_1)$ solves the problem

$$\min_{c_0, c_1} \sum_{i=1}^{n} S_i \left[\delta_n^{-1} K\left(\frac{\hat{U}_{1i} - u_1}{\delta_n}\right) - c_0 - c_1(\hat{P}_i - v)\right]^2 K\left(\frac{\hat{P}_i - v}{h_{n1}}\right), \tag{4.8}$$

and the local quadratic estimator of $\partial f_{U_1|P,S=1}(u_1|v, S=1)/\partial p$ can be defined as $\hat{\partial} f_{U_1|P,S=1}(u_1|v, S = 1)/\partial p \equiv \hat{c}_1$, where $(\hat{c}_0, \hat{c}_1, \hat{c}_2)$ solves the problem

$$\min_{c_0, c_1, c_2} \sum_{i=1}^{n} S_i \left[\delta_n^{-1} K\left(\frac{\hat{U}_{1i} - u_1}{\delta_n}\right) - c_0 - c_1(\hat{P}_i - v) - c_2(\hat{P}_i - v)^2\right]^2 K\left(\frac{\hat{P}_i - v}{h_{n2}}\right). \tag{4.9}$$

The estimator defined in (4.7) is an unrestricted estimator. Thus, it can be negative for a given finite sample, although it is a consistent estimator of $f(u_1|v)$ under certain regularity conditions. To ensure that the estimator is positive in finite samples, we consider a trimmed version of (4.7):

$$\hat{f}_{pdf}(u_1|v) = \max[\varepsilon, \hat{f}(u_1|v)],$$

where ε is a fixed, very small positive number.

Now we describe estimators of $F(u_1|v)$ and $F(u_0|v)$. Again we only discuss estimation of $F(u_1|v)$.

To develop an estimator that is a distribution function for a given finite sample, note that

$$F(u_1|v) = F_{U_1|V}(\underline{u}_1|v) + \int_{\underline{u}_1}^{u_1} f_{U_1|V}(u|v)du, \tag{4.10}$$

for any fixed constant $\underline{u}_1 < u_1$. We estimate $F(u_1|v)$ by replacing $F_{U_1|V}(\underline{u}_1|v)$ and $f_{U_1|V}(u|v)$ in (4.10) with their sample analogs. More specifically, the estimator of $F_{U_1|V}(\underline{u}_1|v)$ is defined as

$$\hat{F}^{cdf}_{U_1|V}(\underline{u}_1|v) = \max[0, \hat{F}_{U_1|V}(\underline{u}_1|v)], \tag{4.11}$$

where

$$\hat{F}_{U_1|V}(\underline{u}_1|v) = v\frac{\hat{\partial}}{\partial p}F_{U_1|P,S=1}(\underline{u}_1|v, S=1) + \hat{F}_{U_1|P,S=1}(\underline{u}_1|v, S=1),$$

and $\hat{F}_{U_1|P,S=1}(\underline{u}_1|v, S=1)$ and $\hat{\partial}F_{U_1|P,S=1}(\underline{u}_1|v, S=1)/\partial p$, respectively, are local linear and quadratic estimators that solve the problems similar to those in (4.8) and (4.9) with $\delta_n^{-1}K\left((\hat{U}_{1i} - u_1)/\delta_n\right)$ replaced by $1(\hat{U}_{1i} \leq \underline{u}_1)$. Then our estimator of $F(u_1|v)$ is defined as

$$\hat{F}_{cdf}(u_1|v) = \min\left[1, \hat{F}^{cdf}_{U_1|V}(\underline{u}_1|v) + \int_{\underline{u}_1}^{u_1} \hat{f}_{pdf}(u|v)du\right].$$

The constant $\underline{u}_1$ can be chosen such that most of estimated values $\hat{U}_{1i}$ are greater than $\underline{u}_1$. Notice that by construction, our estimator is a strictly increasing, continuous function of u_1 (for $u_1 > \underline{u}_1$) and is restricted to be between 0 and 1. In other words, our estimator is a distribution function for a given finite sample. One could also use an unrestricted estimator (4.11), which is not necessarily a distribution function in finite samples.

Notice that as a by-product of estimating $\hat{F}_{cdf}(u_1|v)$, we obtain an estimator of the τ-th quantile of U_1 conditional on $V = v$ for any $\tau \in (0,1)$, which is denoted by $Q_{U_1|V}(\tau|v)$. Simply, the estimator is given by

$$\hat{Q}_{U_1|V}(\tau|v) = \hat{F}^{-1}_{cdf}(\tau|v),$$

where the right-hand side is unique for a given finite sample provided that $\underline{u}_1$ is sufficiently small, since $\hat{F}_{cdf}(u_1|v)$ is a strictly increasing function when $u_1 > \underline{u}_1$. Furthermore, under the assumption that U_1 and V are independent of X and Z, the τ-th quantile of Y_1 conditional on $X = x$ and $V = v$ can be estimated by

$$\hat{Q}_{Y_1|X,V}(\tau|x, v) = \mu_1(x, \hat{\beta}_1) + \hat{Q}_{U_1|V}(\tau|v).$$

Therefore, we can also obtain estimators of marginal quantile treatment effects, which are defined as

$$\hat{Q}_{Y_1|X,V}(\tau|x, v) - \hat{Q}_{Y_0|X,V}(\tau|x, v).$$

This is a quantile analog to the marginal treatment effect of Heckman and Vytlacil (1999, 2001, 2005).

5 Asymptotic Properties of the Estimators

This section provides asymptotic properties of the proposed estimators. The proof of theorems in this section are provided in the Appendix. Recall that Z_c denotes the components of Z which enter nonparametrically in the estimation of P. We consider regression splines as approximating functions $\{p_k : k = 1, \ldots\}$ since regression splines have a smaller bias than power series (Newey, 1997). The following assumptions are standard in the literature on series estimation (Newey, 1997).

Assumption 3. *The data* $\{(Y_i, S_i, X_i, Z_i) : i = 1, \ldots, n\}$ *are independent and identically distributed.*

This is a standard assumption in empirical microeconomics, but it has some limitations. One limitation that might be related with our application is that we do not allow for clustered data. The extension of the asymptotic results obtained in this section to clustered data is non-trivial, and we leave it for future research.

Assumption 4. *The support of Z_c is known and is a Cartesian product of compact connected intervals on which Z_c has a probability density function that is bounded away from zero.*

Assumption 5. *Each function φ_j in (4.1) is r_φ-times continuously differentiable on the support of Z_c for some $r_\varphi > 2$.*

Assumptions 4 and 5 are standard in the literature (Newey, 1997). In particular, Assumption 5 implies that the asymptotic bias (due to the series approximation by regression splines) converges to zero at a rate of κ^{-r_φ} as the number of approximation functions, κ, diverges to infinity.

Note that a finite-sample correction in (4.2) would not have any effect on the asymptotic properties of the estimator. Then the following theorem is a standard result in series estimation.

Theorem 2. *Let Assumptions 2, 3, 4, and 5 hold. Then with regression splines as approximating functions, we have*

$$\max_{i=1,\ldots,n} |\hat{P}(Z_i) - P(Z_i)| = O_p\left[\frac{\kappa}{n^{1/2}} + \kappa^{-(2r_\varphi - 1)/2}\right].$$

We now consider the asymptotic distribution of $n^{1/2}(\hat{\beta}_j - \beta_j)$ for $j = 0, 1$. Let $W_j = 1(Z \in \mathcal{Z})S^{1(j=1)}(1-S)^{1(j=0)}$, where $\mathcal{Z}$ is a strict subset of the support of Z. We make additional assumptions that are standard in semiparametric estimation.

Assumption 6. *Assume that $P(Z)$ is continuously distributed and its density is bounded away from zero on $\mathcal{Z}$.*

It might be too strong to assume that P is bounded away from zero on the whole support of Z.[13] Instead, we assume that it holds in an interior of the support of Z.

Assumption 7. *The conditional expectation $E[\mu_j(X)|P = p, W_j]$ is twice continuously differentiable with respect to p and its kernel estimator $\hat{E}_h[\mu_j(X)|P = p, W_j]$ is consistent uniformly over $p \in \mathcal{P}$, where $\mathcal{P}$ is an interior of the range of $P(Z)$ on $\mathcal{Z}$. Furthermore, assume that*

$$\sup_{p\in\mathcal{P}} \left|\hat{E}_h[\mu_j(X)|P = p, W_j] - E[\mu_j(X)|P = p, W_j]\right| = o_p\left(n^{-1/4}\right).$$

Assumption 8. *Assume that $\kappa^4/n \to 0$ and $\kappa^{2r_\varphi}/n \to \infty$.*

[13] We would like to thank an associate editor for pointing this out.

Note that Assumption 8 is satisfied, for example, if $\kappa \propto n^a$ with $1/(2r_\varphi) < a < 1/4$. For $j = 0, 1$, define

$$\Omega_j = E\left[W_j \left(\mu_j(X) - E[\mu_j(X)|P, W_j]\right)\left(\mu_j(X) - E[\mu_j(X)|P, W_j]\right)'\right],$$
$$\nu_j(z) = E\left[W_j(\mu_j(X) - E[\mu_j(X)|P, W_j])\frac{\partial\ \lambda_j(p)}{\partial p}\bigg|_{p=P}\bigg| Z = z\right],$$

and

$$\Sigma_j = E\left[W_j U_j^2 \left(\mu_j(X) - E[\mu_j(X)|P, W_j]\right)\left(\mu_j(X) - E[\mu_j(X)|P, W_j]\right)'\right] + E\left[P(1-P)\nu_j(Z)\nu_j(Z)'\right],$$

where $\lambda_j(p)$ was defined in (3.1) for $j = 1$ and can be defined similarly for $j = 0$.

Assumption 9. *For each $j = 0, 1$, Ω_j is positive definite, $\nu_j(z)$ is continuously differentiable with respect to z, $E[\nu_j(Z)\nu_j(Z)']$ is nonsingular, and Σ_j is finite.*

The following theorem gives the asymptotic distribution of the estimator of β_j for $j = 0, 1$.

Theorem 3. *Let Assumptions 2-9 hold. Then for each $j = 0, 1$, as $n \to \infty$,*

$$n^{1/2}(\hat{\beta}_j - \beta_j) \to_d \mathbf{N}(0, \Omega_j^{-1}\Sigma_j\Omega_j^{-1}),$$

Our estimation details are different from Das, Newey and Vella (2003); however, the asymptotic distribution of $n^{1/2}(\hat{\beta}_j - \beta_j)$ is comparable to that of Das, Newey and Vella (2003). It is straightforward to construct a sample analog of the asymptotic variance $\Omega_j^{-1}\Sigma_j\Omega_j^{-1}$. We now turn to estimation of $E[U_j|V = v]$ for $j = 0, 1$.

Assumption 10. *$E[U_j|V = v]$ is four times continuously differentiable for $j = 0, 1$.*

Assumption 11. *K is a second-order kernel function with compact support and is Lipschitz continuous.*

Assumption 12.

$$\max_{i:1\le i\le n} |\hat{P}_i - P_i| = o_p\left(h_{n1}^2\right) \quad \textit{and} \quad \max_{i:1\le i\le n} |\hat{P}_i - P_i| = o_p\left(h_{n2}^2\right).$$

In addition,

$$\frac{h_{n2}}{h_{n1}} \to \infty \quad \textit{and} \quad \frac{h_{n2}^3}{h_{n1}} \to 0.$$

The following theorem gives the asymptotic distribution of the estimators of $E[U_j|V = v]$ for $j = 0, 1$.

Theorem 4. *Let Assumptions 2-12 hold. Then for any point v that is in the interior of the range of $P(Z)$ on $\mathcal{Z}$, the asymptotic distributions of the estimators of $E[U_1|V = v]$ and $E[U_0|V = v]$ are*

normal with the same means and variances that they would be if U_{1i}, U_{0i}, and P_i were observed directly. Furthermore,

and

$$(nh_{n2}^3)^{1/2}\left\{\hat{E}[U_0|V=v]-E[U_0|V=v]-B_0(v)h_{n2}^2\right\}\to_d \mathbf{N}(0,V_0(v)),$$

where

$$\begin{aligned}
B_1(v) &= \frac{v}{3!}\frac{\int u^4K(u)du}{\int u^2K(t)du}\frac{\partial^3 E[U_1|P=v,S=1]}{\partial p^3},\\
V_1(v) &= v^2\frac{\int u^2K^2(u)du}{\left(\int u^2K(t)du\right)^2}\frac{E[\left(U_1-E[U_1|P=v,S=1]\right)^2|P=v,S=1]}{f_{P,S=1}(v)},\\
B_0(v) &= \frac{-(1-v)}{3!}\frac{\int u^4K(u)du}{\int u^2K(t)du}\frac{\partial^3 E[U_0|P=v,S=0]}{\partial p^3},\\
V_0(v) &= (1-v)^2\frac{\int u^2K^2(u)du}{\left(\int u^2K(t)du\right)^2}\frac{E[\left(U_0-E[U_0|P=v,S=0]\right)^2|P=v,S=0]}{f_{P,S=0}(v)}.
\end{aligned}$$

This theorem says that the asymptotic distribution of the estimators of $E[U_1|V=v]$ and $E[U_0|V=v]$ are driven by corresponding partial derivative estimators and that estimation errors from $\hat{\beta}_j$ and $\hat{P}_i$ are asymptotically negligible. The asymptotic bias is not easy to estimate because it involves nonparametric estimation of higher-order partial derivatives, but one can adopt undersmoothing to make the asymptotic bias negligible (at the expenses of slower rates of convergence in distribution). The asymptotic variance is relatively easy to estimate (e.g., see equations (4.8) and (4.9) of Fan and Gijbels, 1996). Combining theorems above gives the main result of this section.

Theorem 5. *Let Assumptions 2-12 hold. Then for any x in the support of X and for any point v that is in the interior of the range of $P(Z)$ on $\mathcal{Z}$, the asymptotic distributions of the estimators of $E[Y_1|X=x,V=v]$, $E[Y_0|X=x,V=v]$, and $E[Y_1-Y_0|X=x,V=v]$ are as follows:*

$$\begin{aligned}
&(nh_{n2}^3)^{1/2}\left\{\hat{E}[Y_1|X=x,V=v]-E[Y_1|X=x,V=v]-B_1(v)h_{n2}^2\right\}\to_d \mathbf{N}(0,V_1(v)),\\
&(nh_{n2}^3)^{1/2}\left\{\hat{E}[Y_0|X=x,V=v]-E[Y_0|X=x,V=v]-B_0(v)h_{n2}^2\right\}\to_d \mathbf{N}(0,V_0(v)),
\end{aligned}$$

and

$$\begin{aligned}
&(nh_{n2}^3)^{1/2}\left\{\hat{E}[Y_1-Y_0|X=x,V=v]-E[Y_1-Y_0|X=x,V=v]-\{B_1(v)-B_0(v)\}h_{n2}^2\right\}\\
&\to_d \mathbf{N}(0,V_1(v)+V_0(v)),
\end{aligned}$$

where $B_j(v)$ and $V_j(v)$, $j=0,1$, are defined in Theorem 4.

It would be straightforward to establish similar results for estimators of distributions of Y_1 and Y_0 conditional on $X=x$ and $V=v$. In particular, the asymptotic distributions of the estimators

would be driven by corresponding nonparametric partial derivative estimators. We have not developed asymptotic theory for average treatment effects, such as ATE, TT, and TUT. Notice that Theorem 5 provides asymptotic normality only for interior points of V. To develop asymptotic results for objective defined in (3.4), it would be necessary to extend our asymptotic results for boundary points with careful treatment on tail conditions. This is a topic for future research.

6 Data

The dataset we use consists of a sample of white males surveyed in the NLSY. In the NLSY there exists detailed information on cognitive ability and family background, which are important determinants of both schooling and labor market outcomes. Furthermore we know the place of residence of most respondents in the NLSY during their adolescent years. As a result, we can construct school and labor market characteristics in different areas of residence of adolescent NLSY respondents and use them as instrumental variables for schooling, as is often done in the literature.

We estimate the model for 1992, 1994, 1996 and 1998. The reason we choose to start our analysis in the 1990s and not before is because NLSY respondents were very young in the 1980s. Our sample consists of white males born between 1957 and 1964. The hourly wage measure we use was created by the NLSY. In order to minimize measurement error and reduce concerns with selective unemployment, our wage measure for each year is a 5 year average of all non-missing wages reported in the five year interval centered in the year of interest.[14] The model of Section 2 only allows for selection into two levels of schooling, so we need to group some schooling categories into these two. The two groups we consider are: high school graduates plus high school dropouts; and some college plus college graduates and above.[15]

The instruments for schooling we use are standard in the literature: distance to college, tuition, and local unemployment rates, all measured in the place of residence of each individual during late adolescence.[16] We provide details on their construction in the Appendix.

[14] The percentage of individuals in our sample who have a missing observation for our measure of wages (due to unemployment or non-reporting, but not due to attrition in the panel) is the following for each year: 3.21% in 1992, 3.03% in 1994, 3.19% in 1996, and 2.78% in 1998. When we use different measures of wages such as yearly wages or averages over three years of wages, our results are qualitatively similar but they are more imprecise.

[15] The wage distribution of the NLSY roughly replicates that of the CPS during the 1990s for white males born between 1957 and 1964 (available on request from the authors).

[16] For example, see Cameron and Taber (2004), Card (1995, 1999), Carneiro, Heckman and Vytlacil (2007), Currie and Moretti (2003), Kane and Rouse (1995), Kling (2001), among others.

7 Selection Bias, Composition Effects and the Evolution of Inequality

7.1 Specification of the Model

We consider potential wage equations for the college and high school sectors, denoted by Y_1 and Y_0. Considering only two schooling levels is a limitation of our analysis, but nevertheless a common practice in the literature on inequality (and in studies of the returns to schooling using selection models, such as Willis and Rosen, 1979, and Carneiro, Heckman and Vytlacil, 2007).[17] Using the econometric framework described in previous sections, we can estimate the following objects: $E(Y_1|X=x,V=v)$, $E(Y_0|X=x,V=v)$, $F_{Y_1|X,V}(y_1|x,v)$ and $F_{Y_0|X,V}(y_0|x,v)$. These functions tell us how the distributions of counterfactual wages vary with observed (X) and unobserved (V) heterogeneity.

Our focus on $E(Y_1|X=x,V=v)$, $E(Y_0|X=x,V=v)$, $F_{Y_1|X,V}(y|_1x,v)$ and $F_{Y_0|X,V}(y_0|x,v)$ is useful for two reasons. First, they help us characterize how individuals sort across different levels of schooling according to observed and unobserved heterogeneity. Second, these objects are especially useful for simulating the effect of changes in composition on inequality. As college enrollment increases, there will be changes in the distribution of X and V at each level of schooling because some individuals will switch from high school to college (those who switch will probably be those at the margin). If we can characterize these changes, then we can use them together with our estimates of $F_{Y_1|X,V}(y_1|x,v)$ and $F_{Y_0|X,V}(y_0|x,v)$ to compute the implied effects on inequality (see Heckman and Vytlacil, 1999, 2001, 2005).

We now turn to the exact specification of the equations that we estimate. The X vector in the log wage equations includes years of actual experience, the Armed Forces Qualifying Test score (AFQT, a measure of cognitive ability), number of siblings, mother's years of schooling, father's years of schooling, cohort dummies, and the state unemployment rate in the current state of residence (five year average centered in the year of interest, mimicking our construction of wages). In order to use a flexible specification for the Xs each variable (except the dummy variables and the current state unemployment rate) enters with a linear and a quadratic term. We also interact number of siblings, mother's education and father's education. Finally, we include a dummy variable for being a high school dropout, another dummy variable for being a college attendee without a college degree, and interactions of these variables with quadratic polynomials in experience and AFQT (the most important observables in the wage equations). This is an attempt to allow for some selection on observables within each broad schooling category (fully interacted models where we interact all variables with one another produce qualitatively similar results

[17] Heckman and Vytlacil (2005) consider models with multiple levels of schooling, but which we have difficulty implementing with the data at hand. Allowing for multiple levels of schooling would in principle require different instruments for different transitions, although that may not be strictly necessary provided that different regressors have different effects in different transitions (see, e.g., Heckman and Navarro, 2007).

but more imprecision).

We use a partially linear additive model for schooling choice. In particular, regressors (the Z vector) include a constant, cohort dummies, distance to college (an indicator variable for whether a four-year college is in the county of residence at age 14), linear terms of family background variables (number of siblings, mother's schooling, and father's schooling), interactions between distance to college and family background variables, and cubic B-splines with equally spaced knots (based on quantiles of variables of interest) for corrected AFQT, unemployment at 17, and college tuition at 17. The number of interior knots as well as the inclusion of interaction terms were determined by the least squares cross-validation method. The variables that we exclude from the outcome equations are distance to college, tuition and local unemployment rate. Distance is a strong predictor of schooling, but it takes only two values. By interacting it with the remaining variables in the model we are able to expand the available variation in this variable. The proportion of observations that were trimmed in implementing (4.2) is between 0.07 and 0.08 across years. We set δ to be $1e-8$ in (4.2).

Sample statistics are presented in table 1. In each year of our data, individuals who attend college have on average higher wages than those who do not attend college. They also have less years of experience, higher levels of cognitive ability, fewer siblings, more educated parents, live nearer to colleges and in counties with lower tuition at age 17 than those individuals who never enrolled in college. College enrollment rates increase from 50% (1992) to 52% (1998), although we follow relatively mature cohorts for college enrollment (the NLSY respondents in year 1992 are between 28 and 35 years old).

In implementing our selection model we estimate the model for each year where the dependent variable is college attendance.[18] Average derivatives are presented in table 2. Ability and family background are strong predictors of college attendance. The presence of a college in the county of residence at 14 is also an important determinant of enrollment in college, as are local unemployment and tuition. We test the null hypothesis that three average derivatives for instruments are all zeros and reject this null hypothesis at any conventional level.

For each year we estimate $f_{Y_1|X,V}(y_1|x,v)$ and $f_{Y_0|X,V}(y_0|x,v)$ and then weight these objects with appropriate weights to construct the counterfactuals of interest, as described in Section 4. However, it is only possible to estimate these functions within the support of the data. In particular, we can only estimate them for values of X and P (accordingly V) for which we have individuals both in college and high school. Figure 1 shows the support of the data for 1992, a representative year in our sample (this figure varies very little across years). The top two figures refer to P and the bottom two figures refer

[18] Alternatively we could have estimated a single selection model for all the years of the sample. The reason we choose not to do it is that, even though these individuals are well into their adult years in the beginning of the 1990s, there are still changes in schooling attainment during the decade. In particular, the college enrollment rate in this sample increases from 50% to 52%. A similar pattern is found in the CPS. When we redo the analysis considering that schooling is fixed at a particular level for all the years the overall results do not change substantially.

to AFQT. AFQT is only one of the variables in the X vector on which we condition, but it is the most important one and is also most likely to have non-overlapping supports (see table 1). Notice that the support of P is almost the full unit interval which allows us to estimate our model over the full support of V. We are able to achieve large support for P because: (i) we combine multiple instruments into an index; (ii) if we assume that X is independent of (U_1, U_0, V) we can trade-off variation in X and Z to increase the support of P (since X is controlled for in the outcome equations in a very flexible way).[19] Most of our simulations are within the range of the data, since we only consider movements in P in regions well within the support.

7.2 Choosing Tuning Parameters

In our implementation of (4.2), we used $\delta = 1e-8$.[20] Also, in our application, we use cubic B-splines with equi-spaced knots (based on sample quantiles of variables) as $\{p_k : k = 1, 2, \ldots\}$. The number of approximating functions is chosen by least-squares cross validation. In our empirical work, β_1 and β_0 are estimated with $\omega_{n,ji} \equiv 1$ in (4.3) and a bandwidth of $h = 0.10$ (with the standard normal density as kernel function) for estimation of the kernel estimator. The main estimation results did not change as we used alternative bandwidths (0.05 and 0.20), or we trimmed the data by 5 or 10% of the observations with the smallest density estimates of the estimated P.

Estimating $E(U_1|P = v, S = 1)$ and its derivative requires choices of two bandwidths h_{n1} and h_{n2}. A reasonable data-driven bandwidth selection rule is important to carry out nonparametric estimation. We carry out some initial search for bandwidths using a method called residual squares criterion (RSC) proposed in Fan and Gijbels (1996, Section 4.5). After experimenting different bandwidths around RSC-chosen bandwidths, we finally choose $h_{n1} = 0.35$ and $h_{n2} = 1.25h_{n1}$ for estimating both $E(Y_1|X, V)$ and $E(Y_0|X, V)$ for all the years. The bandwidth h_{n2} is chosen to be larger than h_{n1} since h_{n2} has to go to zero at a rate slower than h_{n1} asymptotically. Varying the value of h_{n1} from 0.2 to 0.5 did not make any important changes in the shape of estimated functions. Throughout the paper, we use the standard normal density function as the kernel function K.

To estimate these conditional PDF's and CDF's, we adopt the same bandwidths h_{n1} and h_{n2} that are used to estimate the corresponding conditional means. The bandwidth δ_n is chosen by Silverman's

[19] When X is not independent of (U_1, U_0, V) our procedure is not valid and the identification condition is that P has full support at each value of X, which is a very demanding condition. For each X, variation in P identifies the objects of interest for small intervals of V. However, if X is independent of (U_1, U_0, V) we can put these intervals all together and identify the objects interest over the whole support of V. This is equivalent to using not only Z, but also interactions of X and Z as instruments for college attendance (controlling for X in the wage regressions). In such a case it is important to ensure that variation in P is not driven exclusively by variation in X. In order to assess the importance of this problem we performed the following exercise. Let $D = 1$ indicate the presence of a college in the county of residence at 14. We divided the sample in four groups according to S and D, and checked the support of P in each group: $S = 1$ and $D = 1$, $S = 0$ and $D = 1$, $S = 1$ and $D = 0$, $S = 0$ and $D = 0$. For each group, the support of P is close to the interval between 0 and 1. Conversely, if we look at the extremes of the support of P, there are individuals with both $D = 1$ and $D = 0$.

[20] This is an arbitrary choice; however, the result would not be very sensitive, provided that δ is sufficiently small.

normal reference rule (Silverman, 1986, p.45). These choices of bandwidths are arbitrary, but our estimation results were not very sensitive to the choices of the bandwidths.

7.3 Empirical Results

There are three components in our empirical analysis. First, we analyze how individuals sort into different levels of schooling and illustrate how sorting affects inequality. Second, we investigate the role of composition changes for the evolution of inequality. Third, we characterize selection bias and its evolution over time.[21]

7.3.1 Characterizing the Patterns of Sorting

We start by presenting estimates of $E(Y_1|X,V)$ and $E(Y_0|X,V)$ for 1992, a representative year in our sample. Figure 2 shows estimates of $E(Y_1|AFQT,X,V=0.5)$, $E(Y_0|AFQT,X,V=0.5)$, and the difference between these two objects, as functions of AFQT, along with 95% pointwise asymptotic confidence intervals for $E(Y_1-Y_0|AFQT,X,V=0.5)$.[22]

We fix years of experience at 10 to abstract from life-cycle effects, V at its median value, and the remaining variables in X at: 3 siblings, 12 years of mother's and father's education, cohort at 1964 and 7% for the local unemployment rate. This figure shows that, in 1992, on average AFQT is a strong determinant of college wages (Y_1) and of the return to college (Y_1-Y_0), but not of high school wages (Y_0).

In figure 3 we graph $E(Y_1|AFQT=0,X,V)$, $E(Y_0|AFQT=0,X,V)$, $E(Y_1-Y_0|AFQT=0,X,V)$ (the Marginal Treatment Effect, or MTE, of Heckman and Vytlacil, 2001, 2005), as functions of V, along with 95% pointwise confidence intervals for $E(Y_1-Y_0|AFQT=0,X,V)$. Again we fix years of experience at 10 and the remaining X variables at the values described above apart from AFQT, which we fix at its mean value 0. As V increases, college wages decrease and so does the return to college, while high school wages increase (recall that the higher the V is, the smaller the likelihood that an individual enrolls in college).

These figures show that those individuals most likely to attend college (the ones with high levels of AFQT and low levels of V) have high wages in the college sector (since college wages increase with AFQT and decrease with V) but have low wages in the high school sector (since high school wages do

[21] We have carried out an informal goodness-of-fit check of our model specification by comparing estimates of $E(Y_1|S=1)$, $E(Y_0|S=0)$, Quantile$(Y_1|S=1)$ and Quantile$(Y_0|S=0)$ from the model with the corresponding quantities in the data, for all the years of our analysis. Overall, our model fits the data relatively well, giving us confidence in the specification of the model.

[22] The pointwise asymptotic confidence intervals were constructed by normal approximations with estimated pointwise asymptotic variances, while ignoring asymptotic biases (undersmoothing). The asymptotic variance is estimated based on equations (4.8) and (4.9) of Fan and Gijbels (1996). Alternatively, one could consider bootstrap confidence intervals (for example, a percentile method for each point). See Horowitz (2001, Section 4.2) for general exposition on bootstrapping kernel-type estimators and Chen, Linton, and van Keilegom (2003, Theorem B) for the asymptotic validity of the bootstrap inference for the GMM-type semiparametric estimators. It is expected that the bootstrap provides an asymptotically valid inference for our case, but it is beyond the scope of this paper to prove its validity.

not move substantially with AFQT and increase with V). Conversely, individuals less likely to attend college have low college wages and high high school wages. In summary, individuals sort into the sector where they have both comparative and absolute advantage.

These results confirm the findings in Willis and Rosen (1979), Carneiro, Heckman and Vytlacil (2007), and Deschenes (2007). Single skill models of the labor market implicit in standard specifications of earnings equations with no heterogeneity predict college goers to have higher earnings both in the high school and college sectors than high school graduates. Our findings (and those of the recent literature cited above) are inconsistent with such a model (see also Heckman and Scheinkman, 1987, Heckman and Sedlacek, 1985, and Gould, 2002, 2005). Similar patterns are found for other years, as shown in figure 4 (which shows these graphs for 1992, 1994, 1996 and 1998).

Figure 5 presents estimates of the 25th, 50th and 75th percentiles of $f(u_1|v)$ and $f(u_0|v)$ for 1992 (which, under our assumptions, correspond up to location to $f(y_1|x,v)$ and $f(y_0|x,v)$). U_1 and U_0 are normalized to have mean zero. The way these three quantiles vary with V parallels the patterns we observed for means. While the dispersion in Y_1 is flat over a large range of V, the dispersion in Y_0 increases more visibly with V. The latter indicates that the components of heterogeneity that do not determine selection are more disperse for individuals with a higher level of V (indicating more uncertainty in high school wages; see Carneiro, Hansen and Heckman, 2003), or that the prices of these components of heterogeneity (skills) are higher for individuals with a high V.[23]

7.3.2 Composition Effects and Education Policy

In this section we examine the importance of changes in the educational composition of the population for wage inequality. Since we follow a single cohort of individuals over time, there are no significant composition changes which we can examine in the raw data. Therefore, instead of looking directly to the data for evidence of composition effects, we use our estimates of the selection and outcome equations to simulate what would happen to inequality if college enrollment rates were different than the ones we observe, keeping prices fixed (partial equilibrium framework; see also Ferreira and Leite, 2005).

The main difficulty of this exercise is to determine which individuals shift across schooling levels when the college enrollment rate changes. This is why a model is needed. Even though our data is only representative of a fixed set of cohorts working in the 1990s, our model can be useful for studying other time periods. The restriction we face is that we can only simulate changes in composition for skill prices fixed at their 1990s levels, and skill prices are probably higher in the 1990s than ever before. Therefore

[23] When estimating the model using data for the 1980s the patterns of selection we obtained were quite unstable, unlike what we observe in the 1990s when, across different years, all the curves have similar shapes. In terms of selection on AFQT, we observe roughly similar patterns for the 1980s and 1990s (at least in qualitative terms). However, the patterns of selection on unobservables are more erratic. Therefore we believe our estimates for the 1990s to be much more reliable and we ignore the 1980s data in the rest of our analysis. We also note that the 1980s data is not very adequate for our analysis since most individuals are in their 20s, and their wages have not yet stabilized.

composition effects will be larger when we evaluate them using 1990s prices than they would be if we evaluated them instead using 1980s prices or 1970s prices.

The mechanics of the simulation are simple: first we change the intercept of the schooling equation and we identify the distribution of (X, V) for individuals who are induced to enroll in college; second we generate the distribution of high school and college wages for this set of individuals; third, we compute how their exit from the high school sector affects the high school wage distribution and how their entry into the college sector affects the college wage distribution. We use the estimates using the 1992 data. The details of the simulation procedure are presented in Appendix B.2. When conducting the simulations, our aim is to mimic the change in college participation among working-age (25-65) white males that is observed between the 1980 and 1990 Censuses, which is an increase from 41% to 55%.

Table 3 (columns (1) and (2)) shows the result of an experiment where we increase the fraction of individuals in college from 41% to 55% (which, according to the Census, is roughly the same change that is observed from 1980 to 1990 for white males aged 25 to 65). The consequence is a decrease in average college wages by 5%, and an increase in average high school wages by 7%. The reason is that the marginal individuals induced to attend college are of below average college quality and they are also of below average high school quality. As a result, the OLS estimate of the return to schooling decreases from 54% to 42%.

We simulate much smaller changes in within group inequality and overall inequality as a result of changes in composition. The ratio of the 90th to the 10th percentile of college wages increases from 1.29 to 1.31, an increase of 2%. In high school, the 90-10 percentile wage differential does not change. Finally, our simulations show a very small decrease in the 90-10 differential in the overall wage distribution, from 1.30 to 1.29.

Our simulation shows that in the absence of composition effects the college premium in the 1980s would have grown by 12 percentage points more than it did in the data. Even if we exaggerate the magnitude of these effects by using 1990s prices, we conjecture that they would be still large if evaluated at 1980s prices. Ignoring them would lead us to severely underpredict the increase in the college premium in the 1980s. As a flip side, if we were to estimate the elasticity of substitution between college and high school labor in this data, it would be overstated.

The consequences of our simulated changes in composition for within group inequality are smaller, although they are still sizeable in the college sector. At first glance it is surprising to find large effects of composition on between group inequality but small effects on within group and overall inequalities. However, it is possible to reconcile these facts. This will happen if the amount of heterogeneity on which individuals select does not explain a lot of the dispersion in wages. As emphasized in Carneiro, Hansen and Heckman (2003) and Cunha, Heckman and Navarro (2005), even if the returns to schooling are very

heterogeneous across individuals, individuals only select on returns to the extent that this heterogeneity is known at the time they make the schooling decision. In our data individuals do select into schooling based on their returns. However, their ex-ante expectation of returns is quite imperfect, and it only accounts for a small portion of the total dispersion in the returns to schooling.[24] In such a context, it is possible to have a large impact of changes in composition on average parameters such as the college premium, but only a small effect on dispersion parameters, such as the 90-10 percentile ratio.

In table 4 we compute the variance of counterfactual wages in each sector ($\text{Var}(Y_1)$ and $\text{Var}(Y_0)$) for all the years and decompose it into components due to X and V (on which agents select), and a residual component (on which agents do not select).[25] Carneiro, Hansen and Heckman (2003) and Cunha, Heckman and Navarro (2005) interpret the former as (ex-ante) heterogeneity and the latter as uncertainty. We could also relate the former to the permanent component and the latter to the transitory component of an earnings variance decomposition. The latter accounts for 66-74% of the variance of college wages and 37-66% of the variance of high school wages (the literature on earnings dynamics usually finds that the permanent component represents 50% of the total variance of earnings and that this number is higher in college than high school; see, e.g., Meghir and Pistaferri, 2004).

7.3.3 The Importance of Selection Bias

The fact that individuals sort into different levels of schooling implies that selection bias affects both within and between group inequality and their evolution over time. Selection bias is always defined relatively to a specific parameter of interest. Here we illustrate the role of selection bias by comparing inequality in the observed economy with inequality in a simulated counterfactual economy where individuals are randomly assigned to different schooling levels, as in Heckman and Sedlacek (1985, 1990). Therefore, we assess the effect of selection bias on inequality parameters under random assignment. We are able to approximate random assignment fairly well because we have close to full support on P, although, as mentioned above, this relies on the assumption of full independence between (X, Z) and (U_1, U_0, V).

The first column of Table 5 characterizes the observed distribution of log wages in 1992, and the second column of the table corresponds to the counterfactual distribution of log wages in the same year

[24] This interpretation will change under different assumptions about the agents' access to insurance markets. As argued in Cunha, Heckman and Navarro (2005), it is not possible to estimate the information set of the agents without first specifying the market structure they face.

[25] In particular, since X and U_1 are assumed to be independent:

$$\begin{aligned}\text{Var}(Y_1) &= \text{Var}[\mu_1(X)] + \text{Var}(U_1) \\ &= \text{Var}[\mu_1(X)] + \text{Var}[E(U_1|V)] + E[\text{Var}(U_1|V)],\end{aligned}$$

where $\text{Var}[E(U_1|V)]$ is the component of variance due to V and $E[\text{Var}(U_1|V)]$ is the remainder. $\text{Var}(Y_0)$ can be decomposed in the same way. The first row of table 4 corresponds to $E[\text{Var}(U_1|V)]$, the part of the variance that cannot be associated with selection, the second corresponds to $\text{Var}[\mu_1(X)]$, the third corresponds to $\text{Var}[E(U_1|V)]$ and the fourth corresponds to $E[\text{Var}(U_1|V)]/\text{Var}(Y_1)$, the fraction of the variance in Y_1 that is not due to any variable related to selection into schooling.

for the random assignment economy. Columns (3) and (4) of the table show the actual and counterfactual distributions of log wages in 1998, and columns (5) and (6) concern the evolution of these distributions between 1992 and 1998. Composition changes do not contaminate this exercise because most individuals in the sample are out of school during the 1990s.

In both 1992 and 1998, average wages in college (panel A) and high school (panel B) are higher in the observed economy than in the random assignment economy due to self selection. The college premium (panel C) is lower in the observed than in the random assignment economy in 1992, although in theory this was not guaranteed to happen (the college premium in the random assignment economy corresponds to the average treatment effect, or ATE). The observed OLS estimate increased by 9%, but there is no increase in the average return to college in the random assignment economy. Thus, the commonly used measure of college premium cannot reveal the true evolution of skill prices.

Table 5 also shows that selection leads to lower within group wage dispersion in college (measured by differences in percentiles of the log wage distribution). In high school the effect of selection is negligible in 1998. Finally, selection bias leads us to underestimate the growth in within group inequality in college relative to the random assignment economy (with the exception of the P90-P50 differential in college) and to have a different trend in within group inequality in high school.

8 Conclusion

Much of the literature on inequality considers simple models without heterogeneity and self-selection into schooling. Our paper examines the importance of accounting for selection into schooling in the empirical study of inequality. We estimate a semiparametric selection model with two levels of schooling (high school and college) using four years of data (1992, 1994, 1996 and 1998) from the NLSY, and use it for three different exercises. First, we have used it to understand the main patterns of sorting of individuals into different levels of schooling. We find that individuals sort into the level of schooling where they have absolute and comparative advantage. Second, we have used the model to simulate a change in college enrollment and examine its effect on the wage structure. In our data increases in educational attainment lead to reductions in between group inequality and increases in within group inequality in college. Third, we have used it to analyze the evolution of inequality and of its determinants in our sample during the 1990s, purging our estimates of selection bias. We find that the trends in the commonly measured college wage premium and within group inequality cannot reveal the trends of prices of skills.

We have also made two new contributions to the econometrics literature. First, we have extended the method of local instrumental variables of Heckman and Vytlacil (1999, 2001, 2005) to identify the distributions of potential outcomes. Second, we have developed a semiparametric method for estimating the entire marginal distributions of potential outcomes.

Some of the effects we emphasize in this paper are also present in some analyses of larger samples. Using the Census, Juhn, Kim and Vella (2005) find evidence of cohort quality effects systematically related to the educational attainment of different cohorts, but argue these can only explain a small fraction of recent fluctuations in the college premium. Carneiro and Lee (2007), using almost the same data but an alternative approach, find larger effects. However, more work remains to be done.

Appendix

A Description of the Data

We restrict the NLSY sample to white males. We define four schooling categories: high school dropouts, high school graduates, some college and college graduates. Because there are multiple useful reports of schooling in the NLSY we construct the educational categories as follows: individuals without a high school degree are high school dropouts; individuals with a high school degree but with less than 13 years of schooling are high school graduates; those reporting 13 to 15 years of schooling and without a four year college degree go into the some college group; finally, those reporting a four year college degree or 16 or more years of schooling are considered to be four year college graduates. GED recipients who never attend college are included in the group of high school graduates. GED recipients that do not have a high school degree, who have less than 12 years of schooling completed and who never reported college attendance are excluded from the sample. The wage variables we use are deflated (to 1983) non-missing hourly wages from 1990 to 2000. We use these to construct 5 year averages which we use in the analysis. We delete all wage observations that are below 1 or above 100. Experience is actual work experience in weeks accumulated from 1979 to the year of interest (annual weeks worked are imputed to be zero if they are missing in any given year). The remaining variables that we include in the X and Z vectors are number of siblings, father's years of schooling, mother's years of schooling, schooling corrected AFQT, average tuition in four year colleges in the county of residence at age 17 deflated (to 1993), distance to four-year colleges at age 14 and local unemployment rate in state of residence at age 17. The distance variable, which is from Kling (2001), is an indicator variable whether a four-year college is in the county of residence at age 14. The state unemployment rate data comes from the BLS website. However, from the BLS website it is not possible to get state unemployment data for all states for all the 1970s (data is available for all states from 1976 on, and it is available for 29 states for 1973, 1974 and 1975), and therefore for some of the individuals we have to assign them the unemployment rate in the state of residence in 1976 (which will correspond to age 19 for those born in 1957 and age 18 for those born in 1958). Annual records on tuition, enrollment, and location of all public four year colleges in the United States were constructed from the Department of Education's annual Higher Education General Information Survey and Integrated Postsecondary Education Data System "Institutional Characteristics" surveys. By matching location with county of residence, we determined the presence of four-year colleges. Tuition measures are taken as enrollment weighted averages of all public four-year colleges in a person's county of residence (if available) or at the state level if no college is available. County and state of residence at 17 are not available for everyone in the NLSY, but only for the cohorts born in 1962, 1963 and 1964 (age 17 in 1979, 1980 and 1981). However, county and state of residence at age 14 is available for most

respondents. Therefore, we impute location at 17 to be equal to location at 14 for cohorts born between 1957 and 1962 unless location at 14 is missing, in which case we use location in 1979 for the imputation. For a description of the NLSY sample see BLS (2001). The NLSY79 has an oversample of poor whites which we exclude from this analysis. We also exclude the military sample. To remove the effect of schooling on AFQT we implement the same procedure as in Carneiro, Heckman and Vytlacil (2007) (based on Hansen, Heckman and Mullen, 2004).

B More Details of Estimation Procedure

B.1 Obtaining Sample Analogs of (3.4)

Estimators of $E[Y_j]$, $E[Y_j|S=1]$, and $E[Y_j|S=0]$ are obtained by

$$
\begin{aligned}
E[Y_j] &= n^{-1}\sum_{i=1}^{n}\int_0^1 \hat{E}[Y_j|X=X_i,V=v]\,dv,\\
E[Y_j|S=1] &= n^{-1}\sum_{i=1}^{n}\int_0^1 \hat{E}[Y_j|X=X_i,V=v]\frac{1-\hat{F}_{P|X}(v|X_i)}{\hat{\Pr}(S=1)}\,dv,\\
&\text{and}\\
E[Y_j|S=0] &= n^{-1}\sum_{i=1}^{n}\int_0^1 \hat{E}[Y_j|X=x,V=v]\frac{\hat{F}_{P|X}(v|X_i)}{\hat{\Pr}(S=0)}\,dv,
\end{aligned}
\tag{B.1}
$$

where $\hat{E}[Y_j|X=x,V=v]=\mu_j(x,\hat{\beta}_j)+\hat{E}[U_j|V=v]$ is defined in section 4.2, $\hat{F}_{P|X}(v|x)$ is a nonparametric kernel estimator of $F_{P|X}(v|x)$, and $\hat{\Pr}(S=j)$ is the sample proportion of $S=j$ for $j=0,1$. The integration with respect to v can be evaluated numerically.

B.2 Simulating Wage distributions of Different Compositions of Education Groups

This subsection describes how we carry out simulations whose results are shown in table 3. Suppose there is a policy that shifts the distribution of P in the population from $F_{P|X}(p|x)$ to $F^*_{P|X}(p|x)$, but has no effect on $f(y_0|x,v)$ nor $f(y_1|x,v)$. In view of (3.4), the post-policy distributions of college and high school wages are

$$
\begin{aligned}
f(y_1|S=1) &= \int\int_0^1 f(y_1|x,v)\frac{1-F^*_{P|X}(v|x)}{\Pr^*(S=1)}f_X(x)\,dv\,dx,\\
&\text{and}\\
f(y_0|S=0) &= \int\int_0^1 f(y_0|x,v)\frac{F^*_{P|X}(v|x)}{\Pr^*(S=0)}f_X(x)\,dv\,dx.
\end{aligned}
\tag{B.2}
$$

Thus, in order to simulate wage distributions of different compositions of education groups, we need to compute only $F^*_{P|X}(p|x)$ and $\Pr^*(S=1)$. Recall that in our empirical work, $\hat{P}$ is a series estimator. We simulate changes in college enrollment rates simply by varying the intercept of $\hat{P}$. Then $F^*_{P|X}(p|x)$ can be estimated by a nonparametric kernel regression of $1\{(\hat{P}+c_*)\leq p\}$ on X and $\Pr^*(S=1)$ can be estimated by a sample average of $(\hat{P}+c_*)$, where we choose the values of c_* to match college enrollment rates in 1980 and 1990. Finally the results shown in table 3 can be obtained by sample analogs of (B.2).

C Mathematical Proofs

Proof of Theorem 2. Given the partially linear additive structure in the modelling of P, this lemma is a direct consequence of Theorem 7 of Newey (1997). ∎

Proof of Theorem 3. This can be proved using general results for two-step semiparametric estimators. In particular, we verify regularity conditions of Ichimura and Lee (2006, hereafter IL) and apply their general theorem to our case. We consider only the case that $j = 1$. The other case is very similar. To simplify the notation, we make our derivation below implicit in the trimming function $1(z \in \mathcal{Z})$. Now we view the estimator $\hat{\beta}_1$ as an M-estimator with

$$m\left[(y, s, x, z), b, f(\cdot)\right] = \frac{1}{2} s \left[y - f_1(f_3(z)) - \{x - f_2(f_3(z))\}' b\right]^2,$$

where $f = (f_1, f_2, f_3)$ are the nonparametric components of the model. In particular, the true function $f_0 = (f_{10}, f_{20}, f_{30})$ satisfies $f_{10}(\cdot) = E[Y|P = \cdot, S = 1]$, $f_{20}(\cdot) = E[X|P = \cdot, S = 1]$, and $f_{30}(\cdot) = E[S|Z = \cdot]$.

First, we check their regularity conditions. Assumption 3.1(a) of IL is not needed in our case because we have an estimator that minimizes a convex objective function. Assumptions 3.1(b) is guaranteed by the assumption that Ω_j is positive definite (see Section 4 of Robinson, 1988). The consistency of the estimator can be easily obtained in view of Assumption 7, so that Assumption 3.1(c) is satisfied. Assumptions 3.2 and 3.3 of IL are trivially satisfied given the form of the objective function m. In view of Theorem 2, Assumption 8 implies that

$$\max_{i:1 \le i \le n} |\hat{P}(Z_i) - P(Z_i)| = o_p\left(n^{-1/4}\right).$$

Then using this and Assumption 7, Assumption 3.4 of IL is easily verified. Given the form of m, it is trivial to verify Assumption 3.5 of IL (see, Proposition 3.1 of IL and discussions on Examples 2.2 and 2.3 of IL). Assumption 3.6 of IL is a key assumption that characterizes the effect of the first stage estimation of P. Using the notation that is same as in IL, it is straightforward to calculate

$$\begin{aligned}
D_{f_1} m^*(b, f_0(\cdot))[h_1(\cdot)] &= -E\left[S\left\{(Y - E[Y|P, S = 1]) - (\mu_1(X) - E[\mu_1(X)|P, S = 1])' b\right\} h_1(\cdot)\right], \\
D_{f_2} m^*(b, f_0(\cdot))[h_2(\cdot)] &= E\left[S\left\{(Y - E[Y|P, S = 1]) - (\mu_1(X) - E[\mu_1(X)|P, S = 1])' b\right\} h_2(\cdot)' b\right], \\
D_{f_3} m^*(b, f_0(\cdot))[h_3(\cdot)] &= E\Big[S\left\{(Y - E[Y|P, S = 1]) - (\mu_1(X) - E[\mu_1(X)|P, S = 1])' b\right\} \\
&\qquad \times \left\{-\partial f_{10}(p)/\partial p\big|_{p=P} + \partial f_{20}(p)/\partial p'\big|_{p=P} b\right\} h_3(\cdot)\Big].
\end{aligned}$$

Then it easy to see that

$$\begin{aligned}
&\frac{\partial}{\partial b} D_{f_1} m^*(b, f_0(\cdot))[h_1(\cdot)]\big|_{b=\beta_1} = 0, \\
&\frac{\partial}{\partial b} D_{f_2} m^*(b, f_0(\cdot))[h_2(\cdot)]\big|_{b=\beta_1} = 0, \\
&\frac{\partial}{\partial b} D_{f_3} m^*(b, f_0(\cdot))[h_3(\cdot)]\big|_{b=\beta_1} = E\left[S(\mu_1(X) - E[\mu_1(X)|P, S = 1]) \partial \lambda_1(p)/\partial p\big|_{p=P} h_3(\cdot)\right].
\end{aligned}$$

Thus, only the third term above affects the asymptotic distribution. Its limiting behavior evaluated at $\hat{P} - P$ is easy to describe, because it is a linear functional of $\hat{P} - P$. For example, see Section 4 of Newey (1997). Assumption 7 of Newey (1997) is satisfied with $\nu(z)$ that is defined above (see Assumption 9). Then the desired result follows from Theorem 3.3 of IL with the restriction that $n^{1/2}\kappa^{-r_\varphi} \to 0$. ∎

Proof of Theorem 4. Define $\hat{\mathbf{U}} = \left[\hat{U}_{11}, \ldots, \hat{U}_{1n}\right]'$ and $\hat{\mathbf{W}}_h = \text{diag}\left\{S_1 K_h(\hat{P}_1 - v), \ldots, S_n K_h(\hat{P}_n - v)\right\}'$, where $K_h(\cdot) = h^{-1}K(\cdot/h)$. In addition, define

$$\hat{\mathbf{X}}_L = \begin{bmatrix} 1 & (\hat{P}_1 - v) \\ \vdots & \vdots \\ 1 & (\hat{P}_n - v) \end{bmatrix} \text{ and } \hat{\mathbf{X}}_Q = \begin{bmatrix} 1 & (\hat{P}_1 - v) & (\hat{P}_1 - v)^2 \\ \vdots & \vdots & \vdots \\ 1 & (\hat{P}_n - v) & (\hat{P}_n - v)^2 \end{bmatrix}.$$

Then using this notation, it follows from (4.4) that

$$\text{(C.1)} \quad \hat{E}[U_1|V = v] = v \times e_2' \left(\hat{\mathbf{X}}_Q' \hat{\mathbf{W}}_{h_{n2}} \hat{\mathbf{X}}_Q\right)^{-1} \left(\hat{\mathbf{X}}_Q' \hat{\mathbf{W}}_{h_{n2}} \hat{\mathbf{U}}\right) + e_1' \left(\hat{\mathbf{X}}_L' \hat{\mathbf{W}}_{h_{n1}} \hat{\mathbf{X}}_L\right)^{-1} \left(\hat{\mathbf{X}}_L' \hat{\mathbf{W}}_{h_{n1}} \hat{\mathbf{U}}\right),$$

where $e_1 = (1, 0)'$ and $e_2 = (0, 1, 0)'$.

Let $\mathbf{U} = [U_{11}, \ldots, U_{1n}]'$ and $\mu_1(\mathbf{X}) = [\mu_1(X_1), \ldots, \mu_1(X_n)]$. Since $\hat{U}_{1i} = U_{1i} - \mu_1(X_i)'\left(\hat{\beta}_1 - \beta_1\right)$, we have

$$\text{(C.2)} \quad \hat{E}[U_1|V = v] = T_n(v) - R_n(v)\left(\hat{\beta}_1 - \beta_1\right),$$

where

$$T_n(v) = v \times e_2' \left(\hat{\mathbf{X}}_Q' \hat{\mathbf{W}}_{h_{n2}} \hat{\mathbf{X}}_Q\right)^{-1} \left(\hat{\mathbf{X}}_Q' \hat{\mathbf{W}}_{h_{n2}} \mathbf{U}\right) + e_1' \left(\hat{\mathbf{X}}_L' \hat{\mathbf{W}}_{h_{n1}} \hat{\mathbf{X}}_L\right)^{-1} \left(\hat{\mathbf{X}}_L' \hat{\mathbf{W}}_{h_{n1}} \mathbf{U}\right)$$
$$R_n(v) = v \times e_2' \left(\hat{\mathbf{X}}_Q' \hat{\mathbf{W}}_{h_{n2}} \hat{\mathbf{X}}_Q\right)^{-1} \left(\hat{\mathbf{X}}_Q' \hat{\mathbf{W}}_{h_{n2}} \mu_1(\mathbf{X})\right) + e_1' \left(\hat{\mathbf{X}}_L' \hat{\mathbf{W}}_{h_{n1}} \hat{\mathbf{X}}_L\right)^{-1} \left(\hat{\mathbf{X}}_L' \hat{\mathbf{W}}_{h_{n1}} \mu_1(\mathbf{X})\right).$$

Then since $R_n(v) = O_p(1)$,

$$\text{(C.3)} \quad \hat{E}[U_1|V = v] = T_n(v) + O_p\left(n^{-1/2}\right),$$

which implies that the error from estimating β_1 is asymptotically negligible.

To analyze $T_n(v)$, define

$$\mathbf{U}_* = (E[U_1|P = P_1, S = 1], \ldots, E[U_1|P = P_n, S = 1])'$$

and

$$\mathbf{X}_L = \begin{bmatrix} 1 & (P_1 - v) \\ \vdots & \vdots \\ 1 & (P_n - v) \end{bmatrix} \text{ and } \mathbf{X}_Q = \begin{bmatrix} 1 & (P_1 - v) & (P_1 - v)^2 \\ \vdots & \vdots & \vdots \\ 1 & (P_n - v) & (P_n - v)^2 \end{bmatrix}.$$

It follows from Assumptions 11 and 12 that

$$\text{(C.4)} \quad \max_i \left|K_h(\hat{P}_i - v) - K_h(P_i - v)\right| = o_p(1),$$

which implies that $\left\|\hat{\mathbf{W}}_{h_{n1}} - \mathbf{W}_{h_{n1}}\right\| = o_p(1)$. By Taylor series expansion,

$$E[U_1|P = P_i, S = 1] = E[U_1|P = v, S = 1] + \frac{\partial E[U_1|P = v, S = 1]}{\partial p}(P_i - v) + \frac{1}{2}\frac{\partial^2 E[U_1|P = v, S = 1]}{\partial p^2}(P_i - v)^2 + \frac{1}{3!}\frac{\partial^3 E[U_1|P = v, S = 1]}{\partial p^3}(P_i - v)^3 + R_p(v),$$

where $R_p(v)$ is a Taylor remainder term. Further, expand the equation above as

$$
\begin{aligned}
E[U_1|P=P_i,S=1] &= E[U_1|P=v,S=1] \\
&\quad + \frac{\partial E[U_1|P=v,S=1]}{\partial p}(\hat{P}_i - v) - \frac{\partial E[U_1|P=v,S=1]}{\partial p}(\hat{P}_i - P_i) \\
&\quad + \frac{1}{2}\frac{\partial^2 E[U_1|P=v,S=1]}{\partial p^2}(P_i - v)^2 \\
&\quad + \frac{1}{3!}\frac{\partial^3 E[U_1|P=v,S=1]}{\partial p^3}(P_i - v)^3 + R_p(v).
\end{aligned} \tag{C.5}
$$

Using (C.4) and (C.5), we have

$$
\begin{aligned}
& e_1'\left(\hat{\mathbf{X}}_L'\hat{\mathbf{W}}_{h_{n1}}\hat{\mathbf{X}}_L\right)^{-1}\left(\hat{\mathbf{X}}_L'\hat{\mathbf{W}}_{h_{n1}}\mathbf{U}\right) - E[U_1|P=v,S=1] \\
&= e_1'\left(\hat{\mathbf{X}}_L'\hat{\mathbf{W}}_{h_{n1}}\hat{\mathbf{X}}_L\right)^{-1}\left(\hat{\mathbf{X}}_L'\hat{\mathbf{W}}_{h_{n1}}[\mathbf{U}-\mathbf{U}_*]\right) + O_p\left(\max_{i:1\le i\le n}|\hat{P}_i - P_i| + h_{n1}^2\right) \\
&= e_1'\left(\mathbf{X}_L'\mathbf{W}_{h_{n1}}\mathbf{X}_L\right)^{-1}\left(\mathbf{X}_L'\mathbf{W}_{h_{n1}}[\mathbf{U}-\mathbf{U}_*]\right)[1+o_p(1)] + O_p\left(\max_{i:1\le i\le n}|\hat{P}_i - P_i| + h_{n1}^2\right) \\
&= O_p\left((nh_{n1})^{-1/2} + \max_{i:1\le i\le n}|\hat{P}_i - P_i| + h_{n1}^2\right).
\end{aligned}
$$

Similar arguments also give

$$
\begin{aligned}
& e_2'\left(\hat{\mathbf{X}}_Q'\hat{\mathbf{W}}_{h_{n2}}\hat{\mathbf{X}}_Q\right)^{-1}\left(\hat{\mathbf{X}}_Q'\hat{\mathbf{W}}_{h_{n2}}\mathbf{U}\right) - \frac{\partial E[U_1|P=v,S=1]}{\partial p} \\
&= e_2'\left(\hat{\mathbf{X}}_Q'\hat{\mathbf{W}}_{h_{n2}}\hat{\mathbf{X}}_Q\right)^{-1}\left(\hat{\mathbf{X}}_Q'\hat{\mathbf{W}}_{h_{n2}}[\mathbf{U}-\mathbf{U}_*]\right) + O_p\left(\max_{i:1\le i\le n}|\hat{P}_i - P_i| + h_{n2}^2\right) \\
&= e_2'\left(\mathbf{X}_Q'\mathbf{W}_{h_{n2}}\mathbf{X}_Q\right)^{-1}\left(\mathbf{X}_Q'\mathbf{W}_{h_{n2}}[\mathbf{U}-\mathbf{U}_*]\right)[1+o_p(1)] + O_p\left(\max_{i:1\le i\le n}|\hat{P}_i - P_i| + h_{n2}^2\right) \\
&= O_p\left((nh_{n2}^3)^{-1/2} + \max_{i:1\le i\le n}|\hat{P}_i - P_i| + h_{n2}^2\right).
\end{aligned}
$$

Then the theorem follows from standard results on local polynomial regression (for example, see Chapter 3 of Fan and Gijbels, 1996). ∎

Proof of Theorem 5. This theorem follows easily by combining Theorems 2, 3, and 4 with the fact that the $S=1$ and $S=0$ samples are independent of each other. ∎

References

Abadie, A. (2002), "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models," *Journal of the American Statistical Association*, 97: 284-292.

Abadie, A. (2003), Semiparametric instrumental variable estimation of treatment response models, *Journal of Econometrics*, 113:2, 231-263.

Abadie, A., J. D. Angrist, and G. W. Imbens (2002), "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70: 91-117.

Aakvik, A., Heckman, J. and E. Vytlacil (2005), Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs, *Journal of Econometrics*, 125:1-2, 15-51.

Cameron, S. and C. Taber (2004), "Estimation of Educational Borrowing Constraints Using Returns to Schooling", *Journal of Political Economy*, part 1, 112(1): 132-82.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Return to Schooling ", *Aspects of Labour Economics: Essays in Honour of John Vanderkamp*, edited by Louis Christofides, E. Kenneth Grant and Robert Swindinsky. University of Toronto Press.

Card, D. (1999), "The Causal Effect of Education on Earnings," Orley Ashenfelter and David Card, (editors), Vol. 3A, *Handbook of Labor Economics,* Amsterdam: North-Holland.

Card, D. (2001), "Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems," *Econometrica*, 69(5): 1127-60.

Card, D. and T. Lemieux (2001), "Can Falling Supply Explain the Rising Return to College For Younger Men? A Cohort Based Analysis," *Quarterly Journal of Economics* 116: 705-46.

Carneiro, P. and J. Heckman (2002), "The Evidence on Credit Constraints in Post-secondary Schooling," *Economic Journal* 112(482): 705-34.

Carneiro, P., K. Hansen and J. Heckman (2003), "Estimating Distributions of Counterfactuals with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on Schooling Choice," *International Economic Review*, 44(2): 361-422.

Carneiro, P., J. Heckman and E. Vytlacil (2007), "Understanding what Instrumental Variables estimate: estimating average and marginal returns to schooling", working paper, The University of Chicago.

Carneiro, P. and S. Lee (2007), "Trends in Quality-Adjusted Skill Premia in the United States, 1960-2000", working paper, University College London.

Chen, S. H. and S. Khan (2007), Estimating the Causal Effects of Education on Wage Inequality Using IV Methods and Sample Selection Models, working paper, SUNY at Albany.

Chen, X, O. Linton, and I. van Keilegom (2003), Estimation of semiparametric models when the criterion function is not smooth, *Econometrica*, 71:5, 1591-1608.

Chernozhukov, V. and C. Hansen (2005), "An IV model of Quantile Treatment Effects," *Econometrica*, 73(1): 245-61.

Chernozhukov, V. and C. Hansen (2006): Instrumental quantile regression inference for structural and treatment effect models, *Journal of Econometrics*, 132, 491-525.

Chernozhukov, V., G. W. Imbens, and W. K. Newey (2007): Instrumental variable estimation of nonseparable models, *Journal of Econometrics*, 139, 4-14.

Chesher, A. D. (2003), "Identification in Nonseparable Models," *Econometrica*, 71: 1405-1441.

Cunha, F., J. Heckman and S. Navarro (2005), "Separating uncertainty from heterogeneity in life cycle earnings", Oxford Economic Papers, 57: 191 - 261.

Currie, J. and E. Moretti (2003), Mother's Education and the Intergenerational Transimission of Human Capital: Evidence from College Openings, *Quartely Journal of Economics*, 118:4.

Das, M., W. K. Newey and F. Vella (2003), "Nonparametric Estimation of Sample Selection Models," *Review of Economics Studies*, 70: 33-58.

Deschenes, O. (2007), "Estimating the Impact of Family Background on the Returns to Education", *Journal of Business and Economic Statistics*, 25(3): 265-277.

Fan, J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications* (London: Chapman & Hall).

Fan, J., Q. Yao, and H. Tong (1996), "Estimation of Conditional Densities and Sensitivity Measures in Nonlinear Dynamical Systems," *Biometrika*: 83:189-206.

Ferreira, F. and P. Leite (2005), Educational Expansion and Income Distribution: a Micro-Simulation for Ceara, in A. Shorrocks and R. Hoeven, *Growth, Inequality and Poverty: Prospects for Pro-poor Economic Development*, Oxford University Press.

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): 105-47.

Gould, E. D. (2005), "Inequality and Ability", *Labour Economics*, 12, 169-189.

Hansen, K., K. Mullen and J. Heckman (2004), "The Effect of Schooling and Ability on Achievement Test Scores," *Journal of Econometrics*. 121(1-2): 39-98.

Heckman, J., H. Ichimura, J. Smith, and P. Todd (1998), "Characterizing Selection Bias Using Experimental Data," *Econometrica*, 66, 1017-1098.

Heckman J. and S. Navarro (2007), "Dynamic discrete choice and dynamic treatment effects", *Journal of Econometrics*, Volume 136, 341-396.

Heckman, J. and J. Scheinkman (1987), "The Importance of Bundling in a Gorman-Lancaster Model of Earnings", *Review of Economic Studies*, Vol. 54, No. 2., pp. 243-255.

Heckman, J. and G. Sedlacek (1985), "Heterogeneity, Aggregation and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market", *Journal of Political Economy*, Vol. 93, No. 6., pp. 1077-1125.

Heckman, J. and G. Sedlacek (1990), Self-Selection and the Distribution of Hourly Wages, *Journal of Labor Economics*, Vol. 8, No. 1, Part 2, pp. S329-S363.

Heckman, J., S. Urzua, and E. Vytlacil (2006), Understanding Instrumental Variables in Models with Essential Heterogeneity, *Review of Economics and Statistics August*, 88:3, 389-432.

Heckman, J. and E. Vytlacil (1999), Local Instrumental Variable and Latent Variable Models for Identifying and Bounding Treatment Effects, *Proceedings of the National Academy of Sciences*, 96, 4730-4734.

Heckman, J. and E. Vytlacil (2001), "Local Instrumental Variables," in C. Hsiao, K. Morimune, and J. Powells, (eds.), *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, (Cambridge: Cambridge University Press, 2000), 1-46.

Heckman, J. and E. Vytlacil (2005), "Structural Equations, Treatment, Effects and Econometric Policy Evaluation," *Econometrica*, 73(3):669-738.

Horowitz, J. L. (2001), The Bootstrap, Edited by: J.J. Heckman and E. Leamer, *Handbook of Econometrics* Volume 5 3159-3228.

Horowitz, J. L. and S. Lee (2007): Nonparametric instrumental variables estimation of a quantile regression model, *Econometrica*, 75: 1191-1208.

Ichimura, H., and C. Taber (2000), Direct Estimation of Policy Impacts, Northwestern University, unpublished manuscript.

Ichimura, H. and S. Lee (2006): Characterization of the asymptotic distribution of semiparametric M-estimators, Cemmap Working Papers, CWP15/06, available at: `http://cemmap.ifs.org.uk`.

Imbens, G. and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2):467-475.

Imbens, G. and W. Newey (2003), "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity", working paper, MIT.

Imbens, G. and D. Rubin (1997), "Estimating Outcome Distributions for Compliers in Instrumental Variable Models," *Review of Economics Studies*, 64: 555-574.

Juhn, C., D. -L. Kim, and F. Vella (2005), "The Expansion of College Education in the United States: Is There Evidence of Declining Cohort Quality?," *Economic Inquiry*, 43(2): 303-15.

Kane, T. and C. Rouse (1995), "Labor-Market Returns to Two- and Four-Year College", *American Economic Review*, 85(3):600-614.

Kling, J. (2001), "Interpreting Instrumental Variables Estimates of the Returns to Schooling", *Journal of Business and Economic Statistics*, 19(3), 358-364.

Lee, S. (2007): Endogeneity in quantile regression models: A control function approach, *Journal of Econometrics*, 141, 1131-1158.

Ma, L. and R. Koenker (2006): Quantile regression methods for recursive structural equation models, *Journal of Econometrics*, 134, 471-506.

Meghir, C. and L. Pistaferri (2004), "Income Variance Dynamics and Heterogeneity", *Econometrica*, 72(1): 1-32.

Moffitt, R. (2008), Estimating Marginal Treatment Effects in Heterogeneous Populations, Annales d'Economie et de Statistique, Special Issue on Econometrics of Evaluation, forthcoming.

W. K. Newey (1997): Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics*, 79, 147-168.

Robinson, P. M. (1988), "Root- N-Consistent Semiparametric Regression," *Econometrica*, 56(4): 931-54.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Vytlacil, E. (2002): Independence, Monotonicity, and Latent Index Models: An Equivalence Result *Econometrica*, 70, 331-341.

Vytlacil, E. and N. Yildiz (2007), "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75(3): 757-779.

Willis, R. and S. Rosen (1979), "Education and Self-Selection," *Journal of Political Economy*, 87(5):Pt2:S7-36.

Table 1: Summary Statistics of Data

Variable	Year 1992		Year 1994		Year 1996		Year 1998	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
College ($S = 1$)	$n = 907$		$n = 903$		$n = 893$		$n = 891$	
Log Wage	2.70	0.55	2.74	0.55	2.81	0.55	2.88	0.57
Years of Experience	7.65	3.36	9.46	3.57	11.14	3.70	12.93	3.91
Current Unemployment	7.46	1.37	5.93	1.29	5.25	1.05	4.27	0.93
Born in 57	0.10	0.30	0.10	0.30	0.10	0.30	0.11	0.31
Born in 58	0.12	0.32	0.11	0.32	0.11	0.32	0.11	0.32
Born in 59	0.09	0.29	0.09	0.29	0.10	0.29	0.10	0.30
Born in 60	0.12	0.33	0.12	0.33	0.12	0.33	0.13	0.34
Born in 61	0.14	0.34	0.14	0.34	0.13	0.34	0.13	0.34
Born in 62	0.17	0.37	0.17	0.37	0.16	0.37	0.15	0.36
Born in 63	0.14	0.34	0.14	0.34	0.14	0.34	0.14	0.35
Corrected AFQT	0.56	0.77	0.56	0.77	0.56	0.76	0.53	0.77
Number of Siblings	2.61	1.66	2.61	1.66	2.65	1.72	2.66	1.73
Mother's Schooling	12.91	2.25	12.90	2.21	12.87	2.24	12.82	2.25
Father's Schooling	13.70	3.15	13.67	3.14	13.61	3.17	13.58	3.19
Distance to College	0.58	0.49	0.58	0.49	0.57	0.50	0.57	0.50
Unemployment at 17	7.09	1.86	7.08	1.87	7.08	1.87	7.09	1.86
College Tuition at 17	2.05	0.78	2.05	0.79	2.05	0.78	2.06	0.79
Some College	0.46	0.50	0.45	0.50	0.45	0.50	0.45	0.50
High School ($S = 0$)	$n = 898$		$n = 872$		$n = 844$		$n = 821$	
Log Wage	2.36	0.58	2.37	0.55	2.40	0.50	2.45	0.54
Years of Experience	10.94	3.24	12.69	3.46	14.40	3.63	16.02	3.96
Current Unemployment	7.34	1.48	5.80	1.24	5.12	1.02	4.26	0.94
Born in 57	0.10	0.30	0.10	0.30	0.10	0.30	0.10	0.30
Born in 58	0.08	0.28	0.08	0.27	0.08	0.27	0.08	0.27
Born in 59	0.12	0.33	0.12	0.33	0.12	0.32	0.12	0.32
Born in 60	0.14	0.35	0.14	0.34	0.14	0.35	0.14	0.35
Born in 61	0.13	0.34	0.13	0.33	0.13	0.33	0.13	0.33
Born in 62	0.17	0.37	0.17	0.37	0.17	0.37	0.17	0.38
Born in 63	0.14	0.34	0.14	0.34	0.13	0.34	0.13	0.34
Corrected AFQT	-0.46	0.90	-0.47	0.89	-0.49	0.88	-0.50	0.89
Number of Siblings	3.25	2.06	3.26	2.08	3.24	2.03	3.28	2.05
Mother's Schooling	11.31	2.15	11.27	2.16	11.25	2.16	11.24	2.15
Father's Schooling	11.06	2.98	11.00	2.96	10.98	2.96	10.94	2.93
Distance to College	0.47	0.50	0.46	0.50	0.46	0.50	0.46	0.50
Unemployment at 17	7.10	1.81	7.09	1.81	7.07	1.81	7.09	1.82
College Tuition at 17	2.12	0.85	2.11	0.84	2.11	0.83	2.09	0.83
High School Dropout	0.19	0.39	0.18	0.38	0.17	0.37	0.17	0.37

Note: Entries in this table are means and standard deviations of variables. For each year, n denotes the sample size of each schooling group. The log wages are 5 year averages of non-missing hourly wages. Years of experience are actual work experience from 1979. Current unemployment is 5 year averages of the state unemployment in percentage in the current state of residence. The omitted variable for the birth-year is 1964. Correct AFQT is schooling-adjusted and normalized to have mean zero in the NLSY population. Parental schooling is measured in years of education. Distance to college is an indicator variable that has value one when there is a four-year college in the county of residence at age 14. Unemployment at 17 is the unemployment rate in percentage in the state of residence at 17. College tuition at 17 is the average tuition in thousand dollars of four year public colleges in the county of residence at 17. Finally, some college and high school dropout are indicator variables that have value one when an individual belongs to corresponding education groups.

Table 2: Average Derivatives for the College Attendance Logit Model

Variable	Year 1992	Year 1994	Year 1996	Year 1998
Corrected AFQT	0.2171	0.2194	0.2283	0.2231
	(0.0116)	(0.0115)	(0.0110)	(0.0111)
Number of Siblings	-0.0514	-0.0465	-0.0377	-0.0446
	(0.0158)	(0.0157)	(0.0167)	(0.0167)
Mother's Schooling	0.2110	0.2079	0.1231	0.1303
	(0.0840)	(0.0842)	(0.0847)	(0.0851)
Father's Schooling	0.3585	0.3839	0.4168	0.4093
	(0.0601)	(0.0597)	(0.0600)	(0.0608)
Unemployment at 17	0.0119	0.0140	0.0148	0.0140
	(0.0067)	(0.0067)	(0.0069)	(0.0070)
College Tuition at 17	-0.0270	-0.0335	-0.0365	-0.0270
	(0.0142)	(0.0144)	(0.0147)	(0.0147)
Distance to College	0.0403	0.0459	0.0402	0.0506
	(0.0200)	(0.0200)	(0.0202)	(0.0204)
Test for Instruments				
P-value	0.0073	0.0008	0.0012	0.0018

Note: For each year, the average derivatives are obtained from a partially linear additive regression of college attendance on explanatory variables using B-splines. In particular, regressors include a constant, cohort dummies, distance to college (a dummy variable), linear terms of family background variables (number of siblings, mother's schooling, and father's schooling), interactions between distance to college and family background variables, and cubic B-splines with equally spaced knots (based on quantiles of variables of interest) for corrected AFQT, unemployment at 17, and college tuition at 17. The number of interior knots as well as the inclusion of interaction terms were determined by the least squares cross-validation method. Standard errors are in parentheses. The last row shows p-values for the null hypothesis that three average derivatives for instruments are all zeros.

Table 3: Results of Simulating 1980's

Column Variable	(1) Simulation 1980	(2) Simulation 1990	(3) Census 1980	(4) Census 1990
	Panel A: College			
College Enrollment Rates	0.41	0.55	0.41	0.55
Average College Wages	2.78	2.73	2.72	2.76
90-10 College Wages	1.29	1.31	1.46	1.57
	Panel B: High School			
Average High School Wages	2.23	2.30	2.50	2.43
90-10 High School Wages	1.09	1.09	1.34	1.43
	Panel C: All Individuals			
Average Overall Wages	2.45	2.54	2.60	2.62
90-10 Overall Wages	1.30	1.29	1.41	1.56
	Panel D: Return to College			
College Premium (OLS)	0.54	0.42	0.22	0.33

Note: The first two columns present measures of average schooling and characteristics of the wage distribution using Census data from 1980 and 1990 for white males. The second two columns present characteristics of simulated wage distributions from our model under the assumption that the college participation rate is 41% (the third column) and 55% (the fourth column).

Table 4: Analysis of Counterfactual Variances of Y_1 and Y_0

Component	Year 1992	Year 1994	Year 1996	Year 1998
Panel A: Variance Decomposition of Y_1				
$E[\text{Var}(U_1\|V)]$	0.223	0.206	0.188	0.218
$\text{Var}[\mu_1(X)]$	0.068	0.078	0.070	0.099
$\text{Var}[E(U_1\|V)]$	0.013	0.023	0.015	0.015
$E[\text{Var}(U_1\|V)]/\text{Var}(Y_1)$	0.736	0.670	0.688	0.656
Panel B: Variance Decomposition of Y_0				
$E[\text{Var}(U_0\|V)]$	0.156	0.126	0.138	0.174
$\text{Var}[\mu_0(X)]$	0.071	0.085	0.033	0.036
$\text{Var}[E(U_0\|V)]$	0.116	0.132	0.039	0.085
$E[\text{Var}(U_0\|V)]/\text{Var}(Y_0)$	0.455	0.367	0.656	0.589

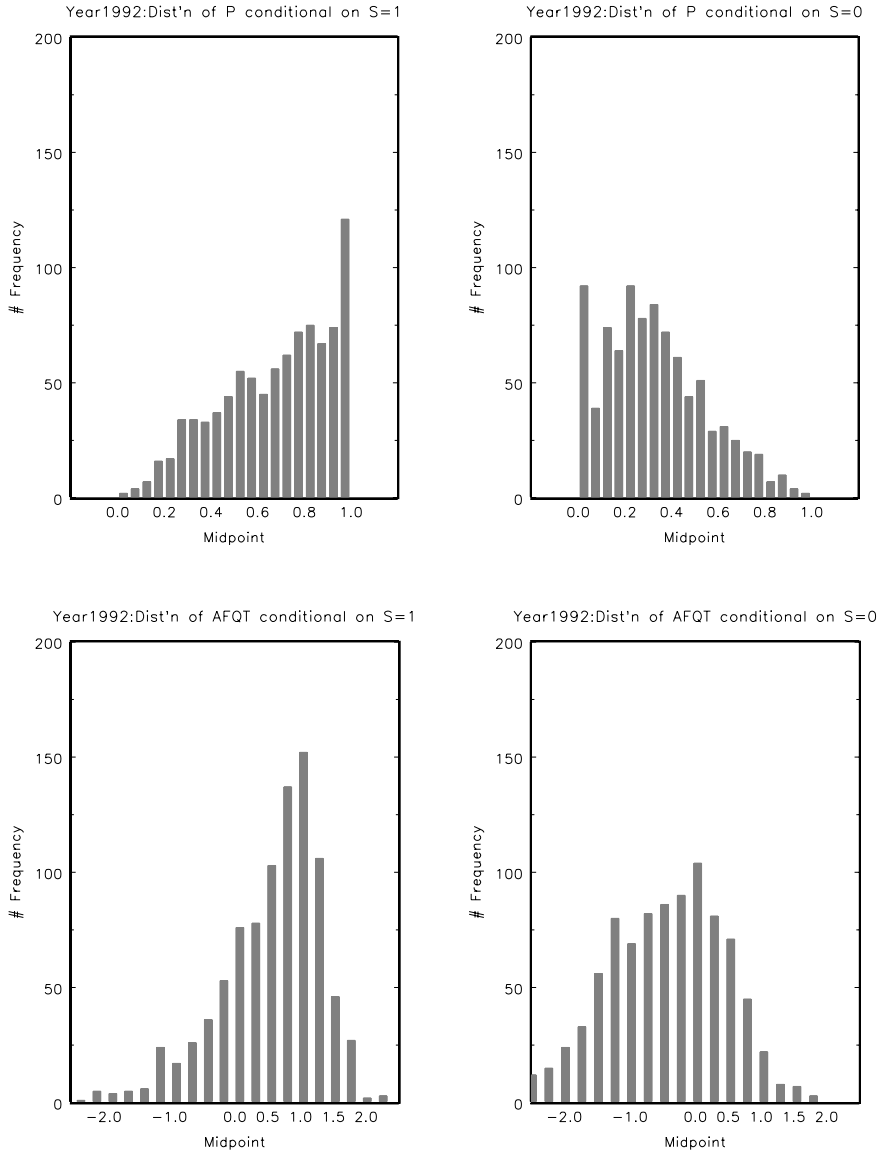
Note: The first panel of this table decomposes the variance of Y_1 in a component due to X (second line), another due to V (third line) and a third one due neither to X nor V (first line). The latter represents the variance in Y_1 that is not related with selection and in the fourth line of the panel we report the percentage of the total variance accounted for this component. The second panel presents a similar decomposition for Y_0.

Table 5: The Impact of Self-Selection on the Distribution of Log Wages, 1992 and 1998

Column	(1) NLSY 92	(2) Random 92	(3) NLSY 98	(4) Random 98	(5) NLSY 98-92	(6) Random 98-92
	Panel A: College					
Mean	2.70	2.58	2.88	2.79	0.18	0.21
P90-P10	1.25	1.39	1.31	1.50	0.06	0.11
P90-P50	0.62	0.63	0.70	0.72	0.08	0.09
P50-P10	0.63	0.76	0.61	0.78	-0.02	0.02
	Panel B: High School					
Mean	2.36	2.00	2.45	2.21	0.09	0.21
P90-P10	1.10	1.24	1.17	1.16	0.07	-0.08
P90-P50	0.54	0.76	0.59	0.58	0.05	-0.18
P50-P10	0.56	0.48	0.58	0.58	0.02	0.10
	Panel C: Return to College					
OLS	0.34	0.58	0.43	0.58	0.09	0.00

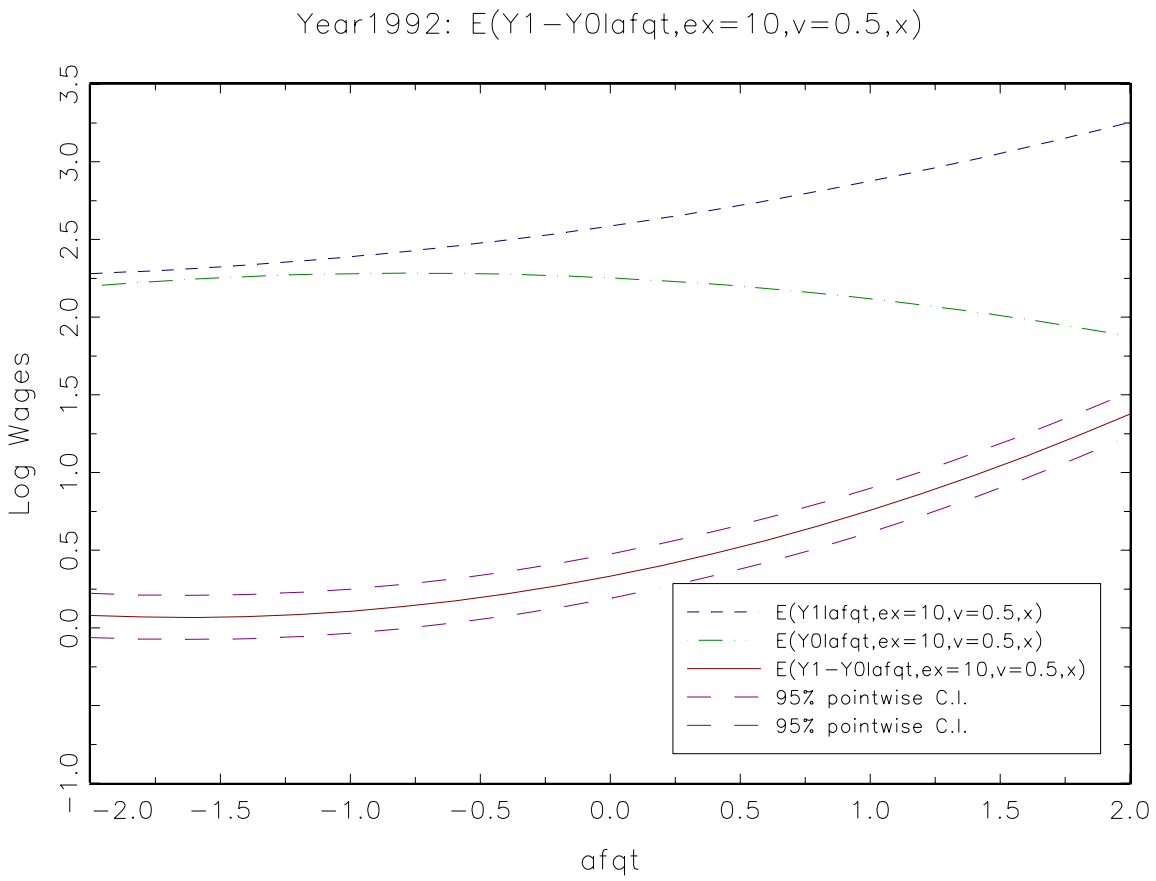
Note: The first column of the table reports actual values for the distribution of log wages in 1992, whereas the second column of the table provides counterfactual values for the distribution of log wages in 1992 that would be observed if individuals were randomly assigned to college and high-school sectors. Columns (3) and (4) of the table show the actual and counterfactual distributions of log wages in 1998 and columns (5) and (6) give the differences between the 1992 and 1998 values. OLS in the last row denotes the Ordinary Least Squares estimate of return to college.

Figure 1: Support of P and AFQT (Year 1992)



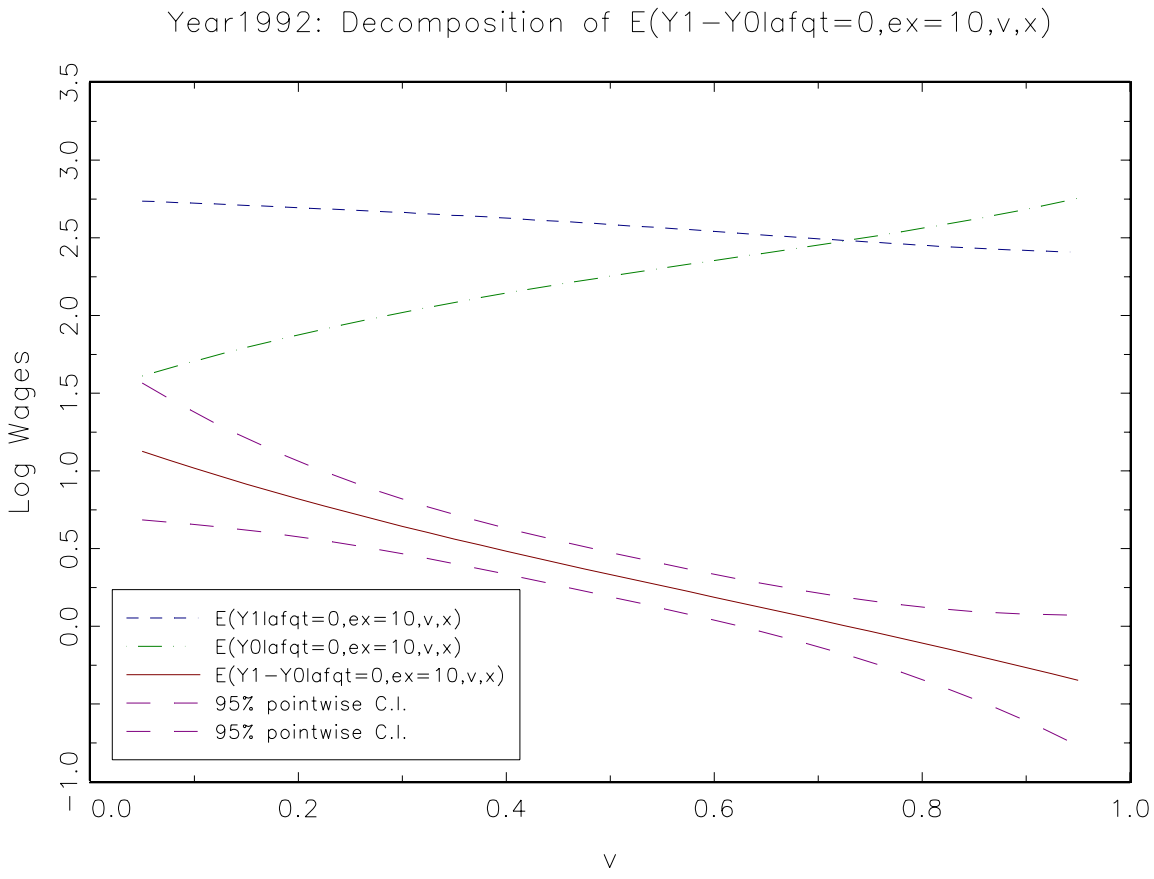
Note: This figure shows the support of the data for 1992. The top two figures refer to P and the bottom two figures refer to AFQT.

Figure 2: MTE as a Function of AFQT (Year 1992)



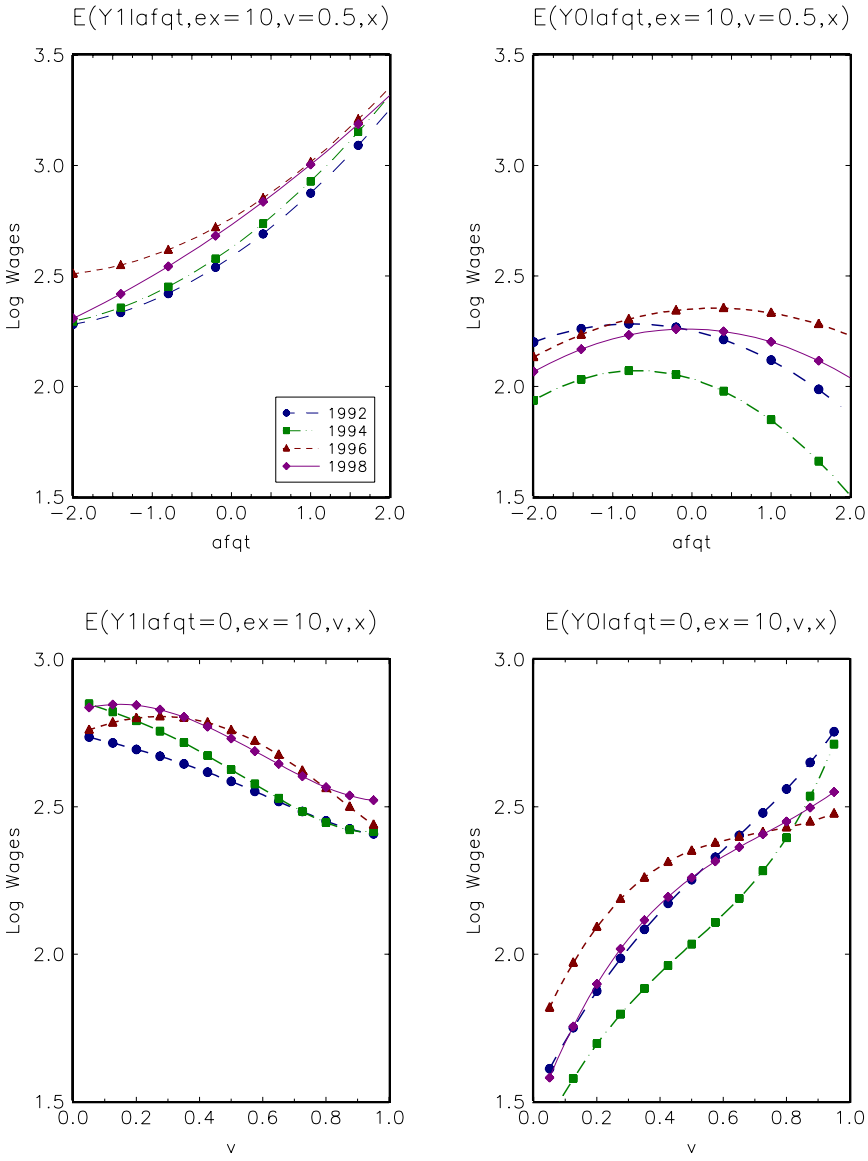
Note: This figure shows estimates of $E(Y_1|AFQT, X, V = 0.5)$, $E(Y_0|AFQT, X, V = 0.5)$, and $E(Y_1 - Y_0|AFQT, X, V = 0.5)$, as functions of AFQT, along with 95% pointwise asymptotic confidence intervals for $E(Y_1 - Y_0|AFQT, X, V = 0.5)$. The remaining X variables are fixed at 10 years of experience, 3 siblings, 12 years of mother's and father's education, cohort at 1964 and 7% for the local unemployment rate.

Figure 3: MTE as a Function of V (Year 1992)



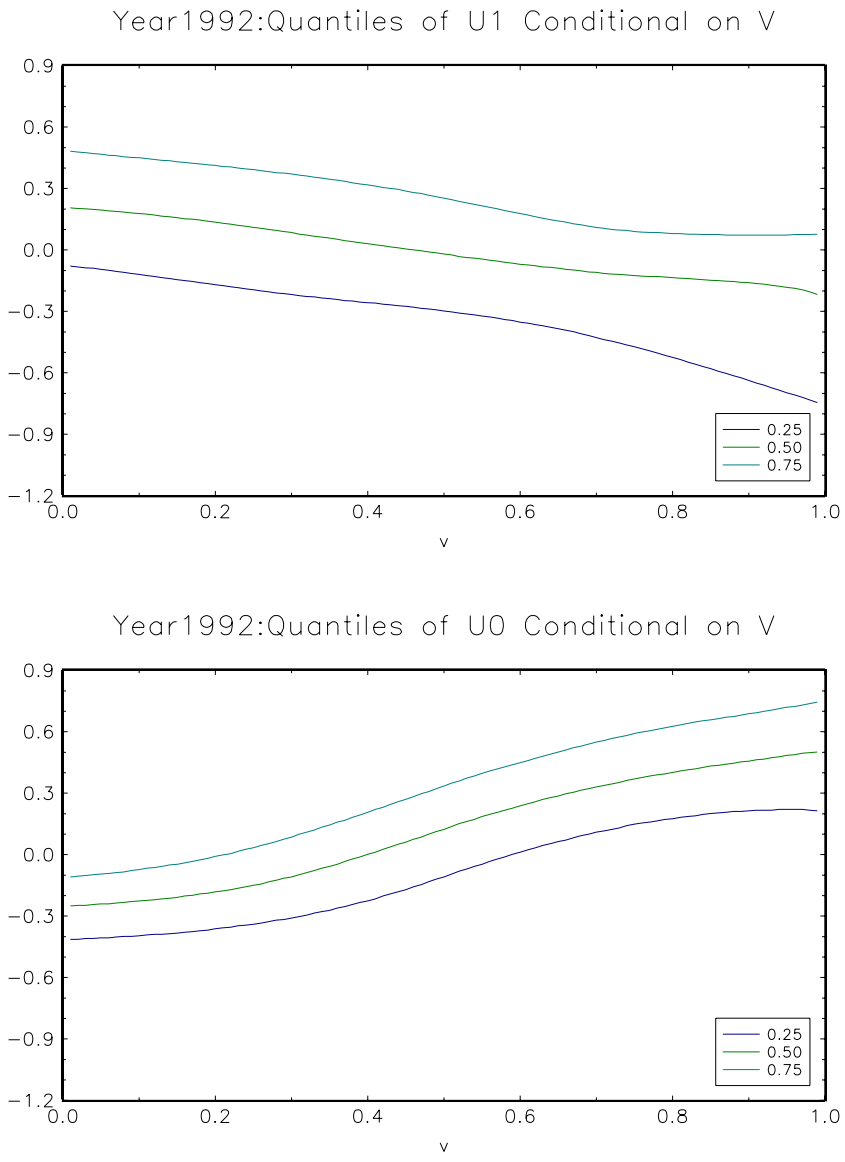
Note: This figure shows estimates of $E(Y_1|AFQT=0,X,V)$, $E(Y_0|AFQT=0,X,V)$, and $E(Y_1-Y_0|AFQT=0,X,V)$, as functions of V, along with 95% pointwise asymptotic confidence intervals for $E(Y_1-Y_0|AFQT=0,X,V)$. The remaining X variables are fixed at 10 years of experience, 3 siblings, 12 years of mother's and father's education, cohort at 1964 and 7% for the local unemployment rate.

Figure 4: $E[Y_1|X,V]$ and $E[Y_0|X,V]$ (All Years)



Note: This figure shows estimates of $E[Y_1|X,V]$ and $E[Y_0|X,V]$ for years 1992, 1994, 1996 and 1998, as functions of AFQT and V. The remaining X variables are fixed at 10 years of experience, 3 siblings, 12 years of mother's and father's education, cohort at 1964 and 7% for the local unemployment rate.

Figure 5: $Q[U_1|V]$ and $Q[U_0|V]$ (Year 1992)



Note: This figure shows estimates of the 25th, 50th and 75th percentiles of $f(u_1|v)$ and $f(u_0|v)$ for 1992. U_1 and U_0 are normalized to have mean zero.