Comparative Advantage and Schooling*

Pedro Carneiro
University College London, Institute for Fiscal Studies and IZA

Sokbae Lee
University College London and Institute for Fiscal Studies

June 7, 2004

Abstract

In this paper we empirically study how individuals sort across different levels of schooling using a sample of white males from the 1979 cohort of the National Longitudinal Survey of Youth. We find that individuals select into the level of schooling in which they have an absolute advantage. Had high school graduates enrolled in college their wages would have been lower than the wages of college participants. Conversely, had college participants not enrolled in college their wages would have been lower than the wages of regular high school graduates. We illustrate how accounting for changes over time in these sorting patterns is important for the study of wage inequality.

1 Introduction

Individuals may sort into different labor market activities according to a variety of assignment rules (Sattinger, 1993). It is crucial to understand this sorting mechanism in order to be able to characterize labor market behavior and the determinants of inequality. One possible assignment rule is comparative advantage. There is comparative advantage when individuals sort into the activities where they perform relatively better when compared to other individuals in the economy. Roy (1951), Mandelbrot (1962) and Heckman and Honoré (1990) characterize how sorting based on comparative advantage influences the distribution of earnings in the labor market.

In this paper we study how individuals select into high school and college using a semiparametric selection model applied to a sample of white males in the National Longitudinal Survey of Youth (see Heckman and Vytlacil, 2000, 2004; Carneiro, Heckman and Vytlacil, 2003; Carneiro, 2003a). As in Carneiro, Heckman and Vytlacil (2003), we find that the returns to college are higher for individuals

*We would like to thank Sami Berlinski, Richard Blundell, Costas Meghir and seminar participants at the Timbergen Institute, Pompeu Fabra and the Institute for Fiscal Studies for valuable suggestions. We thank James Heckman for multiple discussions on the ideas presented in this paper.
who enroll in college, which means that comparative advantage governs the sorting of individuals across schooling levels. We also find that individuals who enroll in college have wages that are higher in college than the wages high school graduates would have if they had gone to college. However had college attendees not gone to college (and became high school graduates) they would have wages which are lower than the wages that high school individuals earn in the labor market.\footnote{As an example, suppose that all college attendees become teachers and all high school graduates become plumbers. Then we find that, if their roles were reversed, teachers would be bad plumbers (compared to actual plumbers) and plumbers would be bad teachers (compared to actual teachers).} Individuals choose the sector where they have an absolute advantage.\footnote{Sorting based on absolute advantage in the chosen sector implies comparative advantage but it is not a consequence of comparative advantage.} These results resemble the findings of Willis and Rosen (1979) who used a parametric method applied to a different dataset.

The idea that individuals sort into different levels of schooling (or different occupations) based on comparative advantage is very natural in economics. Sattinger (1993) shows that in different models of the labor market comparative advantage may or may not govern sorting patterns. Many assignment rules are possible. However, in the Roy (1951) model sorting is determined according to comparative advantage.

The standard Roy model has a simple structure. Suppose there are two sectors ($S$): 0 (high school) and 1 (college). $Y_0$ and $Y_1$ are the potential incomes of an individual in each sector. Individuals choose the sector where their incomes are the highest:

$$S = 1 \text{ if } Y_1 > Y_0,$$
$$S = 0 \text{ if } Y_1 \leq Y_0.$$  

(1)

Therefore $\frac{Y_1}{Y_0} > 1$ for individuals with $S = 1$ and $\frac{Y_1}{Y_0} \leq 1$ for individuals with $S = 0$. If we compare $\frac{Y_1}{Y_0}$ for any two individuals, one choosing sector 1 and the other choosing sector 0, we always have that:

$$\left(\frac{Y_{1i}}{Y_{0i}}\right)_{S=1} > \left(\frac{Y_{1j}}{Y_{0j}}\right)_{S=0}.$$  

(2)

where $\left(\frac{Y_{1i}}{Y_{0i}}\right)_{S=1}$ is $\frac{Y_1}{Y_0}$ for individual i in sector 1 and $\left(\frac{Y_{1j}}{Y_{0j}}\right)_{S=0}$ is $\frac{Y_1}{Y_0}$ for individual j in sector 0. Therefore individuals sort according to comparative advantage. In this model individuals may also sort based on absolute advantage in the sector they choose, which would mean that:

$$\left(\frac{Y_{1i}}{Y_{0i}}\right)_{S=1} > \left(\frac{Y_{1j}}{Y_{0j}}\right)_{S=0},$$

$$\left(\frac{Y_{0i}}{Y_{0j}}\right)_{S=1} < \left(\frac{Y_{0j}}{Y_{0j}}\right)_{S=0}.$$  

(3)

If the conditions in (3) are satisfied then the condition in (2) is automatically satisfied. The income maximization rule in equation (1) imposes sorting based on comparative advantage but not necessarily based on absolute advantage.
The pure version of the Roy Model is rarely used in practice. When using this framework in empirical work economists usually postulate a generalized version of the Roy Model where the selection rule is based on utility maximization instead of income maximization, and where there are costs of choosing each alternative. In our schooling example we could have:

\[ S = 1 \text{ if } W_1(Y_1) - C > W_0(Y_0) \]

where \( W_1 \) is the utility of income in college, \( W_0 \) is the utility of income in high school, and \( C \) is the cost (financial, psychic) of going to college. \( Y_1, Y_0 \) and \( C \) are allowed to be arbitrarily correlated. In such a model any pattern of selection is possible (depending on the correlation between \( Y_1, Y_0 \) and \( C \)) and comparative advantage is no longer imposed as the rule for sorting.

There are several papers that show that comparative advantage is an important feature of the labor market. A few examples are Sattinger (1978), Willis and Rosen (1979), Heckman and Sedlacek (1985, 1990), Heckman and Scheinkman (1986) and Carneiro, Heckman and Vytlacil (2003). Willis and Rosen (1979), Heckman and Sedlacek (1985, 1990) and Carneiro, Heckman and Vytlacil (2003) estimate generalized Roy models. Since comparative advantage has to do with the sorting of individuals across activities according to their performance in each activity, in principle one would need to know both the joint density of \( Y_1 \) and \( Y_0 \), say \( f(y_1, y_0) \), and how individuals sort across sectors in order to be able to have a complete picture of the patterns of comparative advantage. Of the papers cited, only Sattinger (1978) observes the performance of the same individuals in different activities. In a cross section \( f(y_1, y_0) \) is not nonparametrically identified since we never observe the same individual in both states.

Nevertheless, under some conditions it is possible to estimate marginal densities of \( Y_1 \) and \( Y_0 \), say \( f(y_1) \) and \( f(y_0) \). Furthermore, it is possible to estimate objects such as \( f(y_1|S = 1) \), \( f(y_1|S = 0) \), \( f(y_0|S = 1) \) and \( f(y_0|S = 0) \) which can provide suggestive evidence of the existence or absence of comparative and/or absolute advantage in the labor market. To illustrate this point, suppose we estimate the model using log wages. Finding that:

\[ E(\ln Y_1 - \ln Y_0|S = 1) > E(\ln Y_1 - \ln Y_0|S = 0) \]

suggests sorting based on comparative advantage. Finding that:

\[ E(\ln Y_1|S = 1) > E(\ln Y_1|S = 0) \]

\[ E(\ln Y_0|S = 1) < E(\ln Y_0|S = 0) \]

suggests sorting based on absolute advantage. This means that individuals who enroll in college have on average high \( Y_1 \) but low \( Y_0 \), while those who do not enroll in college have on average low \( Y_1 \) but high \( Y_0 \).
Our study is relevant for the analysis of schooling and wages, and for the study of wage inequality. We contribute to the literature on heterogeneity in the returns to education by characterizing selection in levels and in returns in a semiparametric framework. The return to college for an individual is the difference between his potential wages in college \((Y_1)\) and in high school \((Y_0)\), and heterogeneity in the returns to college is a consequence of heterogeneity in counterfactual wages in high school and college. By decomposing the returns to college in these two components we gain a better understanding of heterogeneity in returns. Willis and Rosen (1979) provide a similar characterization of the returns to college, but the model we employ is semiparametric and our characterization of heterogeneity is more complete than theirs. Our dataset is also more recent than the one used in their paper.\(^3\) Carneiro, Hansen and Heckman (2003) also analyze selection on levels and on returns but they use a substantially different framework.

We contribute to the literature on wage inequality by characterizing how individuals sort into different levels of schooling and what are the consequences of this sorting for the distribution of wages in the economy. Several papers such as Mandelbrot (1962) and Heckman and Honoré (1990) have shown how self-selection of individuals into different sectors can generate a skewed distribution of wages from a symmetric distribution of abilities. Heckman and Sédlacek (1985, 1990) provide some empirical evidence of the effect of self-selection on the distribution of wages. We add to that literature by examining empirical patterns of self-selection into schooling. In particular, it is interesting to analyze the impact on the wage distribution of education policies that affect the sorting of individuals across schooling levels, such as tuition subsidies. Assuming that the policy does not have substantial general equilibrium effects we can analyze the effect of the policy on the assignment of individuals across schooling levels and therefore understand its consequences for the overall wage distribution. For example, if the population of high school graduates increases over time (say, because of increased tuition subsidies) then the composition of that population is changing, probably in a non-random way. These compositional changes may affect inequality in an important way. Our framework and empirical results provide useful insights for the study of the impact of sorting on wage inequality within and between educational groups.

We also clarify what is meant by the contribution of the college wage premium for the study of wage inequality. In our framework, the college wage premium varies across individuals. Therefore, conventional (OLS) estimates of the wage premium may overestimate the true wage premium for some individuals and underestimate the wage premium for other individuals. For example, our empirical work shows that the OLS estimates are underestimates of the average returns to college for individuals that enroll in college and are overestimates of the average returns to college for individuals

---

\(^3\) Furthermore, our dataset is a random sample of the U.S. population of white males, whereas theirs is not.
who do not enroll in college. This finding is useful in itself to understand what conventional estimates of the returns to schooling mean (see also Carneiro, Heckman and Vytlacil, 2003). It is also important for understanding the relationship between the distribution of schooling and the distribution of income across time, within and across generations.

Although we use the same data and related methods and draw on ideas which are similar to the ones in Carneiro, Heckman and Vytlacil (2003), this paper is different. While Carneiro, Heckman and Vytlacil (2003) characterize heterogeneity in the returns to college, our emphasis is on heterogeneity in levels of wages in college and high school. Heterogeneity in the returns to college is a consequence of heterogeneity in levels. Therefore we provide a more complete characterization of sorting in the labor market than the one given in the previous paper. Furthermore, we also estimate counterfactual marginal distributions of wages in college and high school for different groups of individuals and illustrate the relevance of our findings for the study of wage inequality. Both of these are substantial differences from Carneiro, Heckman and Vytlacil (2003). This paper also draws on the ideas of Carneiro, Hansen and Heckman (2003), but the methodology that is used and the assumptions that are imposed on the data are distinct.

This paper proceeds as follows. In the next two sections we present our framework. Then we discuss our empirical results which we divide in two sections. In section 4 we discuss our main findings on comparative and absolute advantage. In section 5 we present our results on wage inequality. In section 6 we present a short summary of the paper and a brief description of work currently in progress and future work.

2 Models

Our point of departure is the binary treatment model that is standard in the programme evaluation literature. Let $Y_1$ and $Y_0$ be counterfactual outcomes in two states 1 and 0. In this paper $Y_1$ and $Y_0$ are the potential wages in states 1 and 0. We assume

\[
Y_1 = \mu_1 (X, U_1) \\
Y_0 = \mu_0 (X, U_0)
\]

where $X$ is a random vector influencing counterfactual outcomes, $\mu_1$ and $\mu_0$ are unknown functions, and $U_1$ and $U_0$ are unobserved random variables. Individuals choose to be in state 1 or 0 according to the following equation:

\[
S = 1 \text{ if } \mu_S (Z) - U_S > 0
\]
where $Z$ is a random vector influencing the decision equation, $\mu_S$ is an unknown function of $Z$, and $U_S$ is an unobserved random variable. 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 $Z$. In this paper we assume that there is at least one variable in $Z$ that is not in $X$ (exclusion restriction). Equation (5) should be interpreted as the reduced form of a well specified economic model. In our paper $S = 1$ is college attendance and $S = 0$ is high school graduation. Carneiro, Heckman and Vytlacil (2003) and Carneiro (2003a) use this model to examine heterogeneity in the returns to college and present an economic model that can justify the specification in (5). We can rewrite (5) as:

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

where $V = F_{U_S|X}[U_S|X]$, $P = F_{U_S[X}[\mu_S(Z)|X]$, and $F_{U_S[X|X}(u_s|x)$ is the c.d.f. of $U_S$ conditional on $X.$

By construction, $V \sim \text{Unif}[0,1]$ conditional on $X$.

It is well known that in general $f(y_1, y_0)$ cannot be nonparametrically identified in a cross section. Nevertheless, as shown in Heckman and Smith (1998), it is possible to identify marginal distributions of $Y_1$ and $Y_0$ conditional on $X$ and $V$.

In this paper we estimate the entire marginal distributions of $Y_1$ and $Y_0$ conditional on $X$ and $V$ as well as functionals of the marginal distributions, including the means and quantiles of $Y_1$ and $Y_0$ conditional on $X$ and $V$. Estimation of these conditional means, quantiles, p.d.f.s and c.d.f.s allows us to gain important insights into the sources of heterogeneity in the returns to college. Estimation of marginal distributions of potential outcomes have been previously considered by Imbens and Rubin (1997) and Abadie (2002). Imbens and Rubin (1997) develop histogram-type estimators of marginal distributions of counterfactual outcomes under the local average treatment effect (LATE) framework of Imbens and Angrist (1994). Abadie (2002) proposes bootstrap tests for distributional treatment effects under the same LATE framework.

Our estimation method builds on Heckman and Vytlacil (2000, 2004), who develop the Marginal Treatment Effect (MTE from now on, a parameter introduced in the literature by Bjorklund and Moffit, 1987) and show how using this parameter one can understand the relationship between

---

4Throughout the paper $f_X(x)$ denotes the p.d.f. of $X$ while $F_X(x)$ denotes the c.d.f. of $X$. Also, $f_{Y,X}(y,x)$, $f_{Y,X|X}(y|x)$, and $F_{Y|X}(y|x)$ denote the joint p.d.f. of $Y$ and $X$ and the conditional p.d.f. and c.d.f. of $Y$ on $X = x$, respectively. We suppress subscripts in the notation whenever this can be done without causing confusion.

5Heckman, Smith and Clemens (1996) compute bounds for $f(y_1, y_0)$ using data from a training program. Carneiro, Hansen and Heckman (2001,2003) estimate $f(y_1, y_0)$ assuming that $Y_1$ and $Y_0$ have a factor structure that governs all the dependence between the two variables. They fit models with two independent factors which have a flexible mixture of normals distribution. They develop identification results for models with an arbitrary number of factors. Cunha, Heckman and Navarro (2003) extend this work and estimate models with more than two factors.

6See also Ichimura and Taber (2000).
different treatment parameters, policy relevant parameters, standard least squares and linear instrumental variable estimates of treatment effects. Carneiro, Heckman and Vytlacil (2003) and Carneiro (2003a) apply this framework to the study of the returns to college.

The following lemma, which is an extension of Heckman and Vytlacil (2000, 2004a), provides identification of the nonparametric model given by (4) and (5).

Lemma 1 Consider a nonparametric selection model given by (4) and (5). Let \( V = F_{U_S|X}[U_S|X] \) and \( P = F_{U_1|X}[\mu_S(Z)|X] \). 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 distributions of \( U_S \) and \( \mu_S(Z) \) conditional on \( X \) are absolutely continuous with respect to Lebesgue measure; (4) For a measurable function \( G \), \( E [G(Y_1)|X = x, V = p] \) \( E [G(Y)|X = x, P = p, S = 1] \) \( E [G(Y)|X = x, P = p, S = 0] \), and (5) \( 0 < Pr(S = 1|X) < 1 \). Then

\[
E [G(Y_1)|X = x, V = p] = E [G(Y)|X = x, P = p, S = 1] + p \frac{\partial E [G(Y)|X = x, P = p, S = 1]}{\partial p} - (1 - p) \frac{\partial E [G(Y)|X = x, P = p, S = 0]}{\partial p}
\]

provided that \( E [G(Y)|X = x, P = p, S = 1] \) and \( E [G(Y)|X = x, P = p, S = 0] \) 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 (5) guarantees that \( P \) is strictly between 0 and 1 conditional on \( X \). Assumption (4) is needed to ensure that expectations considered below are finite. Notice that

\[
E [G(Y)|X = x, P = p, S = 1] = E [G(Y)|X = x, P = p, V < p] = \int_0^p E [G(Y_1)|X = x, V = v] \, dv / p,
\]

where the second equality follows from assumption (2) and the fact that \( V \) is uniformly distributed on \([0, 1]\) conditional on \( X \). Then 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.

The assumptions in Lemma 1 are basically identical to those of Heckman and Vytlacil (2000, 2004a). 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 MTE, defined as \( E (Y_1 - Y_0|X = x, V = v) \), 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.
The identification result in Lemma 1 is very general since it does not impose any restrictions on the functional forms of \( \mu_1 \) and \( \mu_0 \) in (4). 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 Lemma 1 decreases rapidly as the number of continuously distributed components of \( X \) increases (curse of dimensionality). To avoid this disadvantage, we specify and estimate a separable version of (4) under a more stringent assumption on unobservables. The following lemma, which is a special case of Lemma 1, gives identification of the model that will be estimated in sections 4 - 5.

**Lemma 2**  Consider a semiparametric selection model given by (5) and

\[
Y_1 = \mu_1(X) + U_1 \\
Y_0 = \mu_0(X) + U_0.
\]

Let \( V = F_{U_S|U_S} \) and \( P = F_{U_S|\mu_S(Z)} \). 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) \( 0 < \Pr(S = 1|X) < 1 \); and (6) \( E[U_1|P = p, S = 1], E[U_0|P = p, S = 0], \Pr[U_1|P = p, S = 1] \) and \( \Pr[U_0|P = p, S = 0] \) are continuously differentiable with respect to \( p \). Then the expectations, quantiles, and marginal distributions of \( Y_1 \) and \( Y_0 \) conditional on \( X = x \) and \( V = v \) are identified.

**Proof.** The lemma is proved in the appendix. ■

Separability in (6) and assumption (2) in Lemma 2 reduce the dimension of the estimation problem. We can test for selection on \( Y_1 \) by checking whether \( E(Y_1|X = x, V = p) \) changes with \( p \). Similarly, we can test for selection on \( Y_0 \) by checking if \( E(Y_0|X = x, V = p) \) changes with \( p \). Notice that we can only identify \( E(Y_1|X = x, V = p) \) over the support of \( P \) for individuals in \( S = 1 \), and \( E(Y_0|X = x, V = p) \) over the support of \( P \) for individuals in \( S = 0 \). As a consequence, we can only identify \( E(Y_1 - Y_0|X = x, V = p) \) over the common support of \( P \) for individuals in \( S = 1 \) and \( S = 0 \).

The methods used here are a version of the local instrumental variables method of Heckman and Vytlacil (2000), and therefore the intuition for our identification results is a simple instrumental variable intuition. When we use instrumental variables in a random coefficient model we interpret the estimates as a local average treatment effect or a weighted average of local average treatment

---

effects (see Imbens and Angrist, 1994; Angrist, Imbens and Graddy, 2000; Heckman and Vytlacil, 2000, 2004; Carneiro, Heckman and Vytlacil, 2003; and Carneiro, 2003a, 2003b). The idea is that the instrument shifts some individuals from high school to college and we identify an average return to college for individuals induced to go to college by the observed change in the instrument. Assume we have a continuous valid instrument $Z$. Say tuition is such an instrument. Suppose we pick two values of the instrument which are very close together: $z_1$ and $z_2$. If we estimate the return to college using standard IV using only individuals for whom the instrument takes these two values, we estimate the average return to college for individuals induced to switch from high school to college when the instrument switches from $z_1$ to $z_2$. Similarly, we can identify the average college wage and the average high school wage for these same individuals, just by examining how average college wages and average high school wages (and the proportion of individuals enrolled in college) change when $Z$ changes from $z_1$ to $z_2$.\footnote{Notice that if we have a binary instrument taking two values ($z_1$ and $z_2$) we can identify the average return to college ($Y_1 - Y_0$) for individuals induced to switch from high school to college when the instrument changes from $z_1$ to $z_2$ (Imbens and Angrist, 1994) by computing:

$$\text{LATE}_{z_1,z_2} = \frac{E(Y|Z = z_1) - E(Y|Z = z_2)}{E(S|Z = z_1) - E(S|Z = z_2)}.$$}

We can also estimate the marginal densities of college and high school wages (but not the joint density) for this same group of individuals. These individuals define a particular margin: these are individuals indifferent between going to college or not when tuition is between $z_1$ and $z_2$ (and therefore they switch when tuition varies over this range). By comparing pairs of $z$ values over the whole observed range of $Z$ we can identify all the objects above for individuals at different margins. This is exactly what we have shown in Lemmas 1 and 2.

3 Heterogeneity in Observables and Unobservables

It is interesting to examine heterogeneity both in observables and unobservables. Heterogeneity in observables is usually analyzed using the method of matching and our study benefits from important insights from that literature (for a recent exposition, see Heckman, Ichimura and Todd, 1997, 1998; and Heckman, Ichimura, Smith and Todd, 1998). However, we go beyond matching by analyzing selection on unobservables as well as selection on observables.
Notice that, under the assumptions in Lemma 2, the MTE is:

$$MTE = E(Y_1 - Y_0|X = x, V = v)$$

$$= E[\mu_1(X) - \mu_0(X) + U_1 - U_0|X = x, V = v]$$

$$= E[\mu_1(X) - \mu_0(X)|X = x] + E[U_1 - U_0|V = v].$$

Obviously, $E[\mu_1(X) - \mu_0(X)|X = x] = \mu_1(x) - \mu_0(x)$. Because $X$ can be multidimensional this difference can be hard to analyze and to link to the choice equation in a systematic way. On the other hand, notice that $E[U_1 - U_0|V = v]$ has a natural interpretation as the average return to college (in terms of unobservables) for individuals who have different propensity (in terms of unobservables) to attend college. How can we compute a similar function for observables?

We motivate our approach with a simple example. For our purposes it is useful to partition $Z = [X, Z']$, where $Z'$ is the set of variables in $Z$ that is not in $X$ (exclusion restrictions), so that we can write $\mu_S(Z) = \mu_S(X, Z')$. Assume for the moment that $X$ is a scalar and that $\mu_S(X, Z') = \mu_{SX}(X) + \mu_{SZ'}(Z')$. Furthermore, assume that $\mu_{SX}(X) = X$. In this case $E[\mu_1(X) - \mu_0(X)|X = x] = \mu_1(x) - \mu_0(x)$ has exactly the same interpretation as $E[U_1 - U_0|V = v]$ as the average return to college (in terms of observables) for individuals who have different propensity (in terms of observables) to attend college. If $\mu_{SX}(X)$ is a general function of a vector $X$ then we may redefine $\mu_1(x)$ and $\mu_0(x)$ as:

$$\mu_1(x) \equiv \mu_1'[\mu_{SX}(x)]$$

$$\mu_0(x) \equiv \mu_0'[\mu_{SX}(x)],$$

where, in general, $\mu_1'$ and $\mu_0'$ are correspondences, not necessarily functions. If $\mu_{SX}(x)$ is a one-to-one function of $x$, and $\mu_1'$ and $\mu_0'$ are functions of $\mu_{SX}(x)$, we can interpret $\mu_1'[\mu_{SX}(x)] - \mu_0'[\mu_{SX}(x)]$ as exactly the same way as $E[U_1 - U_0|V = v]$. In this case we can reduce the possible multidimensionality of $X$. However, this may not be satisfied in practice. Therefore, to implement this method we use the following approximation. We first estimate $\mu_{SX}(X)$ (up to scale) from equation (5). Then we estimate $E[\mu_1(X)|\mu_{SX}(X) = \mu_{SX}(x)]$, $E[\mu_0(X)|\mu_{SX}(X) = \mu_{SX}(x)]$ (and consequently $E[\mu_1(X) - \mu_0(X)|\mu_{SX}(X) = \mu_{SX}(x)]$) by running nonparametric regressions of $\mu_1(X)$ and $\mu_0(X)$ on $\mu_{SX}(X)$, separately. Naturally, it is only possible to estimate $E[\mu_1(X)|\mu_{SX}(X)]$ over the support of $\mu_{SX}(X)$ for individuals choosing $S = 1$, $E[\mu_0(X)|\mu_{SX}(X)]$ over the support of $\mu_{SX}(X)$ for individuals choosing $S = 0$ and $E[\mu_1(X) - \mu_0(X)|\mu_{SX}(X)]$ over the common support of $\mu_{SX}(X)$ for these two groups.

In this paper we characterize heterogeneity in observables and unobservables. As will be shown in the following empirical sections, the analyses of observables and unobservables tell the same story.
about comparative and absolute advantage.

4 Estimating Patterns of Comparative Advantage

In this section we study the levels of counterfactual wages for high school graduates and college attendees and the returns to college using a sample of white males from the National Longitudinal Survey of Youth. The data are summarized in table 1.9 S = 1 denotes college attendance. In our data set there are 688 high school graduates who never attend college and 729 individuals who attend any type of college.10 Table 1 documents that individuals who attend college have on average a 31% higher wage than those who do not attend college.11 They also have one year less of work experience since they spend more time in school. Scores on a measure of cognitive ability, the Armed Forces Qualifying Test (AFQT), are much higher for individuals who attend college than for those who do not.12 Persons who only attend high school come from larger families and have less educated parents than individuals who attend college. They also live in counties where tuition is higher, and they live farther away from a college, two measures of direct costs of schooling. The county unemployment rate is a measure of the opportunity cost of schooling and is only marginally higher for those individuals who only attend high school.

The dependent variable in the potential wage equations is the logarithm of average wages between 1991 and 1993, and the X vector in the potential wage equations consists of experience, schooling-adjusted AFQT, number of siblings, mother’s years of schooling, and father’s years of schooling. The Z vector in the school choice equation consists of AFQT, number of siblings, mother’s education, father’s education, distance to college at age 14, average tuition in the county the individual lived in at age 17, and local unemployment rate in the county of residence in 1979. Our instruments Z’ are distance to college, tuition, and local unemployment rate.13 The AFQT and family background

9A detailed description of the data is provided in appendix B.
10These are white males, in 1992, with either a high school degree or above and with a valid wage observation, as described in appendix B.
11Wages are constructed as an average of all nonmissing wages between 1991 and 1993 for each individual. Using this procedure we are able to get nonmissing wages for close to 90% of the sample. Therefore, in this paper we ignore self selection into employment. Actual work experience (not potential experience) is measured in 1992. Since individuals in the NLSY are born between the years of 1957 and 1964, in 1992 they are 28 to 35 years of age.
12We use a measure of this score corrected for the effect of schooling attained by the participant at test date, since at the date the test was taken, in 1981, different individuals have different amounts of schooling and the effect of schooling on AFQT scores is important. We use a version of the nonparametric method developed in Hansen, Heckman and Mullen (2003). We standardize the AFQT to have mean 0 and variance 1 in the sample of white males.
13We realize that each of these instruments may have serious limitations, although they are usually used in the literature (for example, Card, 1993, uses the presence of a college in the county of residence; Kane and Rouse, 1995, use tuition; Cameron and Taber, 2004, use a local labor market variable, although in their case the variable is local wage). Carneiro and Heckman (2002) and Carneiro, Heckman and Vyllclicl (2003) show that these may not be valid instruments since they are correlated with AFQT. However, in this paper (and in Carneiro, Heckman and Vyltlacil, 2003) we control for AFQT and we hope that by doing so we greatly attenuate the importance of the problem. Tuition may also not be a valid instrument because it may be correlated with college quality. The literature on the effects of college quality on wages is not conclusive. Black and Smith (2004) find important effects of quality on labor market outcomes, but Dale and Krueger (1999) find that the effects of attending a selective college are strong only for children
variables enter the schooling choice equation but do not play the role of an instrument since they are included in the $X$ vector as well.

We use a logit model for schooling choice. That is, $S = 1$ if $P(Z) > V$, where $P(z) = L[\mu_S(z)]$ and $L(u) = \exp(u)/(1 + \exp(u))$. The function $\mu_S(z)$ is specified parametrically to be a sum of two parts, that is $\mu_S(z) = \mu_{SX}(x) + \mu_{SZ}'(z')$, where $\mu_{SX}(x)$ consists of linear, quadratic, and interaction terms of AFQT, number of siblings, mother’s education, and father’s education (excluding experience) and $\mu_{SZ}'(z')$ consists of a constant, tuition, distance to college, and local unemployment rate. The specification of $\mu_S(z)$ is an attempt to use a very flexible functional form. The propensity score $P(Z)$ could be estimated nonparametrically; however, dimension reduction is needed here to achieve reasonable precision of estimates since the dimension of $Z$ is large. The flexible logit model used here provides good estimation precision, allows for nonlinear effects of $Z$, and ensures that the estimated probability lies between 0 and 1.

Table 2 presents estimates of the average marginal derivatives for each variable in $Z$. The average derivatives are computed by taking an average of partial derivatives of $P(Z)$. That is,

$$n^{-1} \sum_{i=1}^{n} L'[\hat{\mu}_S(Z_i)] \frac{\partial \hat{\mu}_S(Z_i)}{\partial z},$$

where $L'(u) = dL(u)/du$, $\hat{\mu}_S(z)$ is the logit estimator of $\mu_S(z)$, and $Z_i$ is the $i$-th observation of $Z$, and $n$ is the sample size. AFQT, family background, and tuition are quite strong predictors of schooling, whereas distance to college and local unemployment rate are weak predictors of college attendance once we control for AFQT and family background. Our tuition effects (measured in hundreds of dollars) conform to the ones found in the literature that measures enrollment-tuition responses in the US: a $1000 reduction in (four year college) tuition leads to an increase in enrollment of 4-5% (see Kane, 1994 or Cameron and Heckman, 2001 for summaries of the literature).

The support of the estimated $P(Z)$ is shown in the top panels of figure 1 and it is almost the full unit interval for individuals both in college and in high school, although at the extremes of the interval the cells of data become quite thin. The sparseness of data in the tails results in a large amount of noise (variability) in the estimation of $E(Y|X, P, S = 1)$ and $E(Y|X, P, S = 0)$ for students from poor backgrounds. Nevertheless, a common finding across studies of college quality is that cognitive test scores of students may be a good proxy for college quality. Our wage equations include controls for family background and cognitive test scores which may also work as adequate controls for college quality. The use of tuition as an instrument also relies on lack of migration across counties. Although there is no available data on county migration for college attendance, it is estimated that about 75% of american students attend college in the state they live in (Digest of Education Statistics, 1993). We also estimated models where we included quarter of birth as an instrument (Angrist and Krueger, 2001) but with no change in our results.

These are partial equilibrium estimates of the effects of tuition. Heckman, Lochner and Taber (1998, 1999) show that partial and general equilibrium analyses of tuition policy can lead to very different conclusions.

Formally, for nonparametric analysis, we need to investigate the support of $P(Z)$ conditional on $X$. However, the partially linear structure that we will impose below implies that we only need to investigate the marginal support of $P(Z)$.
values of $P$ very close to 0 or very close to 1. The bottom panels of figure 1 give us an idea of the support of $X$ for individuals in high school and college. To reduce the dimensionality of $X$ we aggregate it into an index $\mu_{SX}(X)$, constructed as only $X$-components of $\mu_S(Z)$.\footnote{In the matching literature, this type of aggregation is common. The usual aggregator is the propensity score.} We graph the support of $\mu_{SX}(X)$ for individuals in college and in high school. Although these supports do overlap substantially there exist no individuals in high school with high values of $\mu_{SX}(X)$ and therefore we cannot estimate how the return to college varies with $X$ at high values of $\mu_{SX}(X)$.

Fully nonparametric estimation of $E(Y|X, P, S = 1)$, $E(Y|X, P, S = 0)$, and their derivatives with respect to $P$ is not feasible due to the curse of dimensionality. We impose additional structure on the model that results in a feasible semiparametric estimation problem. More specifically, we assume

$$E[Y_1|X = x, V = v] = \mu_1(x, \beta^1) + E[U_1|V = v],$$

and

$$E[Y_0|X = x, V = v] = \mu_0(x, \beta^0) + E[U_0|V = v],$$

where the functional forms of $\mu_1$ and $\mu_0$ are specified up to finite dimensional parameters $\beta^1$ and $\beta^0$. Thus, estimates of $E[Y_1|X = x, V = v]$ and $E[Y_0|X = x, V = v]$ can be obtained by estimating $\beta^1$, $\beta^0$, $E[U_1|V = v]$, and $E[U_0|V = v]$. The exact functional form of $\mu_1(x)$ includes linear, quadratic, and interaction terms of the $X$ vector not including an intercept term, which is not identified without imposing further restrictions. The same functional form is also used for $\mu_0(x, \beta^0)$. Again, the specifications of $\mu_1(x, \beta^1)$ and $\mu_2(x, \beta^2)$ are motivated by the concern that it may be important to be flexible in the way we model the effects of observables.

In this paper, $\beta^1$ and $\beta^0$ are estimated using a semiparametric version of the sample selection estimator of Das, Newey, and Vella (2003). Note that under the assumption that $U_1$ and $V$ are independent of $X$ and $Z$, we have

$$E[Y|X = x, P = p, S = 1] = \mu_1(x, \beta^1) + \lambda_1(p),$$

where $\lambda_1(\cdot)$ is an unknown function of $P$. The equation (8) suggests that $\beta^1$ 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 is construction of the estimated $P$ and the second step is estimation of $\beta^1$ using the estimated $P$. In this paper, the first step is carried out by a logit regression of $S$ on $Z$ using $\mu_S(z)$ specified above, and the second step is accomplished via series estimation, which is convenient for imposing the partially linear restriction in (8).
To describe an estimator of $\beta^1$ more specifically, let $\{(Y_i, X_i, \hat{P}_i, S_i) : i = 1, \ldots, n\}$ denote a sample of observations, where $\hat{P}_i = L[\hat{\mu}_S(Z_i)]$. Also, let $\mathbf{q}_\kappa(p) = [q_{1\kappa}(p), \ldots, q_{\kappa\kappa}(p)]'$ be a $(\kappa \times 1)$ vector of approximating functions and let $\tau(p) = 1(\tau_0 \leq p \leq \tau_1)$ be a trimming function with predetermined constants $\tau_0$ and $\tau_1$ such that $0 \leq \tau_0 < \tau_1 \leq 1$. The estimator $\hat{\beta}^1$ of $\beta^1$ is obtained by solving the problem

$$\min_{b, \theta} \sum_{i=1}^{n} \tau(\hat{P}_i)S_i[Y_i - \mu_1(X_i, b) - \mathbf{q}_\kappa(\hat{P}_i)'/\theta]^2.$$  

The resulting value of $b$ is the estimator of $\beta^1$.\textsuperscript{17} Analogously, $\beta^0$ can be estimated by a partially linear regression of $Y$ on (polynomials in) $X$ and estimated $P$ using only observations with $S = 0$.

Estimation results are reported in table 3, which shows estimates of the average marginal derivatives for each variable in $X$. There is no trimming of the data, that is $\tau_0 = 0$ and $\tau_1 = 1$. To approximate unknown functions ($\lambda_1(p)$ and $\lambda_0(p)$), we have tried polynomial approximation from a second-order polynomial to a fourth-order polynomial. Our preferred specification is a second-order polynomial, since it has a nonlinear, parsimonious specification and estimation results do not vary much from a second-order polynomial to a fourth-order polynomial.\textsuperscript{18} Standard errors are computed using an asymptotic variance formula similar to that presented in Das, Newey, and Vella (2003), while accounting for estimation of $P$ using the logit model.

We start by discussing simple tests of selection on observables and unobservables (assuming that flexible specification we choose is correct). Table 3 reports that we cannot reject the null hypothesis of no selection on unobservables in college (equivalently, the null hypothesis that the coefficients on $P$ and $P^2$ are equal to zero) but we can reject this null hypothesis for high school wages at the 10% level. It is clear that there is selection on observables by looking at $t$-values for the average derivatives shown in table 3. In particular, AFQT (including interactions of AFQT with other variables) is an important determinant of schooling decisions, of counterfactual wages (especially in college), and consequently of the returns to college. In short, we have strong evidence against no selection on observables and unobservables.\textsuperscript{19}

Now we consider estimation of $E[U_1|V = v]$ and $E[U_0|V = v]$. Applying Lemma 1 with $G(u) = u$,

\textsuperscript{17}In practice, we regress $Y$ on polynomials in $X$ (allowing for cross terms between different $X$s) and polynomials in $P$.

\textsuperscript{18}Additional estimation results with third-order and fourth-order polynomials are available upon request.

\textsuperscript{19}It is interesting that we cannot reject the hypothesis of no selection on unobservables in college wages. It is possible that controlling for cognitive skills and family background is enough to account for selection in college, although it is not enough to account for selection in high school. The variables we use in the $X$ vector may be good measures of the types of (more academic) skills required in the college sector but not of the type of (less academic) skills required in the high school sector.
we have

\[
E(U_1|V = v) = E[U_1|P = v, S = 1] + v \frac{\partial E[U_1|P = v, S = 1]}{\partial p}, \tag{9}
\]

\[
E(U_0|V = v) = E[U_0|P = v, S = 0] - (1-v) \frac{\partial E[U_0|P = v, S = 0]}{\partial p}. \tag{10}
\]

Equations (9) and (10) are the basis for nonparametric estimators of \( E[U_1|V = v] \) and \( E[U_0|V = v] \) proposed in this paper. Local polynomial estimation is used here to estimate \( E(U_1|P = v, S = 1), E(U_0|P = v, S = 0) \) 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. Fan and Gijbels (1996) provide a detailed discussion of the properties of local polynomial estimation.

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 the order \( \pi \) of polynomial be equal to \( \pi = \mu + 1 \), where \( \mu \) is the order of the derivative of a function of object. 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.

First consider estimation of \( E[U_1|V = v] \). Following suggestions of Fan and Gijbels (1996), \( E(U_1|P = v, S = 1) \) 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 \( \{(\hat{U}_{i1}, \hat{P}_i, S_i) : i = 1, \ldots, n\} \) denote observations of estimated \( U_1 \) and \( P \) along with \( S \), where \( \hat{U}_{i1} = Y_i - \mu_1(X_i, \beta) \) for \( i = 1, \ldots, n \). The local linear estimator \( \hat{E}(U_1|P = v, S = 1) \) of \( E(U_1|P = v, S = 1) \) is obtained by solving the problem

\[
\min_{c_0, c_1} \sum_{i=1}^n S_i \left[ \hat{U}_{i1} - c_0 - c_1(\hat{P}_i-v) \right]^2 K \left( \frac{\hat{P}_i - v}{h_{n1}} \right),
\]

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 \( \hat{\partial E}(U_1|P = v, S = 1)/\partial p \) of \( \partial E(U_1|P = v, S = 1)/\partial p \) is obtained by solving the problem

\[
\min_{c_0, c_1, c_2} \sum_{i=1}^n S_i \left[ \hat{U}_{i1} - c_0 - c_1(\hat{P}_i-v) - c_2(\hat{P}_i-v)^2 \right]^2 K \left( \frac{\hat{P}_i - v}{h_{n2}} \right),
\]

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 \( \partial E(U_1|P = v, S = 1)/\partial p \). Therefore, the estimator of \( E[U_1|V = v] \) is given

\(^{20}\)One may estimate \( \hat{\beta}^1 \) and \( \hat{\beta}^0 \) using a local polynomial approximation in \( P \) as well. This is not done here because the series estimation procedure is simpler to implement and choices of nonparametric methods are relatively less important for the estimation of \( \beta^1 \) and \( \beta^0 \) than for the estimation of \( E(U_1|P = v, S = 1) \) and \( E(U_0|P = v, S = 0) \).

\(^{21}\)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.
by

\[ \hat{E}[U_1|V = v] = v \frac{\partial}{\partial p} E(U_1|P = v, S = 1) + \hat{E}(U_1|P = v, S = 1). \] (11)

Similarly, the estimator of \(E[U_0|V = v]\) can be obtained by replacing unknown functions in the right hand side of (10) with their nonparametric estimators.

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 use a method called residual squares criterion (RSC) proposed in Fan and Gijbels (1996, Section 4.5). The bandwidth chosen by RSC is an optimal bandwidth in a sense that it minimizes the integrated mean squared error of the estimator and is relatively easy to compute. Throughout the paper, we use the standard normal density function as the kernel function \(K\).

Figure 2 shows our estimates of \(E(Y_1|X, V)\) (dashed line), \(E(Y_0|X, V)\) (dotted line), and \(E(Y_1 - Y_0|X, V)\) (solid line), each as a function of \(V\), evaluated at \(X = \bar{X}\), where \(\bar{X}\) is a vector of the sample means of each variable in \(X\) (reported on the first panel of table 1). The estimated functions are shown in the domain of \([0.05, 0.95]\), which is well within the support of \(P\) in the data. The bandwidths are \(h_{n1} = 0.197\) and \(h_{n2} = 0.220\) for estimating \(E(Y_1|X, V)\), and \(h_{n1} = 0.194\) and \(h_{n2} = 0.198\) for estimating \(E(Y_0|X, V)\). \(E(Y_1|X, V)\) is a decreasing function of \(V\), which means that individuals who are likely to enroll in college (low levels of \(V\)) receive on average higher wages as college attendees than individuals who are not likely to enroll in college (high \(V\)). This difference is not quantitatively negligible. Individuals with \(V = 0.05\) have on average college wages that are about 65% higher than individuals who have \(V = 0.95\). On the contrary, \(E(Y_0|X, V)\) is an increasing function of \(V\), which means that individuals who are likely to enroll in college (low levels of \(V\)) receive on average lower wages as high school graduates than individuals who are not likely to enroll in college (high \(V\)). This difference is quantitatively quite large. Individuals with \(V = 0.05\) have on average high school wages that are about 210% lower than individuals who have \(V = 0.95\). As a result, the returns to college are higher for individuals with low levels of \(V\) and become negative as \(V\) gets larger than 0.6. Figure 2 also shows 95% pointwise confidence intervals for \(E(Y_1 - Y_0|X, V)\) that are obtained using asymptotic approximation without adjusting for the bias. Relatively narrow confidence intervals suggest strong evidence on selection on unobservables.

Figure 3 shows our estimates of \(E(Y_1|\mu_{SX}(X), V)\) (dashed line), \(E(Y_0|\mu_{SX}(X), V)\) (dotted line), and \(E(Y_1 - Y_0|\mu_{SX}(X), V)\) (solid line), each as a function of \(\mu_{SX}\), evaluated at \(V = 0.5\). As explained above, in order to reduce the dimensionality of our graphs we present an estimate of \(E(Y_1|\mu_{SX}(X), V)\) as a function of \(\mu_{SX}(X)\) instead of an estimate of \(E(Y_1|X, V)\) as a function of
X. To estimate $E[Y_1|\mu_{SX}(X), V]$, notice that from the assumptions in Lemma 2,

$$E[Y_1|\mu_{SX}(X), V] = E[\mu_1(X, \beta^1)|\mu_{SX}(X)] + E[U_1|V].$$

Thus, we only need to estimate $E[\mu_1(X, \beta^1)|\mu_{SX}(X)]$ in addition to $E(U_1|V)$ in order to obtain an estimator of $E[Y_1|\mu_{SX}(X), V]$. Specifically, we estimate $E[\mu_1(X, \beta^1)|\mu_{SX}(X)]$ by carrying out a local linear mean regression of $\mu_1(X, \beta^1)$ on $\mu_{SX}(X)$ using all observations. Here, as before, bandwidth selection is carried out by RSC. The selected bandwidth is $h_n = 1.493$ for estimating $E[\mu_1(X, \beta^1)|\mu_{SX}(X)]$, and $h_n = 1.020$ for estimating $E[\mu_0(X, \beta^0)|\mu_{SX}(X)]$.

Figure 3 shows that $E[Y_1|\mu_{SX}(X), V]$ is an increasing function of $\mu_{SX}(X)$ illustrating again that individuals who are highly likely to enroll in college (high values of $\mu_{SX}(X)$) have on average higher college wages than individuals less likely to enroll in college. The patterns of selection on observables and unobservables go in the same direction. Figure 3 also shows that $E[Y_0|\mu_{SX}(X), V]$ is a decreasing function of $\mu_{SX}(X)$, which means that individuals who are highly likely to enroll in college (high values of $\mu_{SX}(X)$) have on average lower high school wages than individuals less likely to enroll in college. Once again the patterns of selection on observables and unobservables go in the same direction. As a consequence, individuals with higher values of $\mu_{SX}(X)$ also have higher returns to college than individuals with lower values of $\mu_{SX}(X)$.

These results show that sorting based on absolute advantage in the selected sector is an important feature of the labor market. Those individuals who enroll in college are very good college attendees but would be very poor high school graduates if they did not go to college and chose instead to be high school graduates. Inversely, individuals who choose not to enroll in college and to become high school graduates instead, are the best high school graduates in the population and would be the worst college attendees in the population if they became college attendees. Individuals select into the activities in which they are absolutely better. Even though these results may seem striking, notice that we observe the same patterns of selection both in terms of observables and unobservables. Twenty-five years ago Willis and Rosen (1979) reached the same conclusion using a parametric framework and a different dataset.

Figure 4 graphs an annualized version of $E(Y_1 - Y_0|\mu_{SX}(X), V)$ as a function of both $\mu_{SX}(X)$ and $V$. The return to one year of college may be obtained by dividing $E(Y_1 - Y_0|\mu_{SX}(X), V)$ by 3.5, which (in this dataset) is the difference between the average years of schooling of high school graduates and individuals who attend at least some college. The MTE is negative for a substantial

---

22We do not interpret our results as saying that a high level of AFQT and high parental education cause individuals to have low wages as high school graduates. However, take the example where all college attendees become teachers and all high school graduates become plumbers. It is not clear that the child of a teacher will become a better plumber than the child of a plumber. Maybe the opposite is more likely. Similarly, individuals that lack cognitive academic skills may specialize early in non-academic skills. Specialization may generate a negative correlation between AFQT and other types of skills, more related to high school jobs.
set of values of \( \mu_{SX}(X), V \). The quantitative magnitude of heterogeneity is substantial. Returns can be as high as 90% per year of college for individuals with low \( V \) and high \( \mu_{SX}(X) \), and as low as -30% per year of college for individuals with high \( V \) and low \( \mu_{SX}(X) \). These results suggest that comparative advantage is an important feature of the labor market.\(^{23}\)

In summary, our estimates show that absolute and comparative advantage govern the sorting of individuals across these two schooling levels. There is strong selection on observable and unobservable determinants of wages and returns, and the story is the same whether we look at observables only, unobservables only, or simultaneously at both. In consonance with papers by Willis and Rosen (1979), Heckman and Sedlacek (1985, 1990) and Heckman and Scheinkman (1986) this evidence also challenges a single skill (efficiency units) view of the labor market. Individuals who are the best in the sector they choose are the worst in the sector from which they opt out. Carneiro, Heckman and Vytlacil (2003) also conclude that the pattern of absolute advantage uncovered in this paper is likely to be present in the data. They argue that the only way that we can simultaneously have that the IV estimate estimate can be above the OLS estimate of the return to college and that the MTE is downward sloping is if \( E(Y_0|S = 1) < E(Y_0|S = 0). \)\(^{24}\) In this section we restate their conclusion and present direct evidence for it by directly estimating the relevant quantities.

## 5 Estimating Potential Wage Distributions

Using the framework presented in section 2 we can not only estimate average counterfactual wages in each sector (high school and college) as a function of \( X \) and \( V \), but we can also estimate the entire marginal distribution of counterfactual wages in each sector as a function of \( X \) and \( V \).

The distribution of counterfactual wages may be useful for policy makers who are concerned with distributional effects of a policy (see, for example, Imbens and Rubin (1997) and Abadie (2002)). For example, using this framework we can examine the effects of different policies on the distribution of outcomes, provided the policy affects the distribution of \( P \) but not \( f(y_1|x, v) \) nor \( f(y_0|x, v) \). This is equivalent to a partial equilibrium assumption which may be restrictive, especially when we are analyzing the distribution of income for the whole economy (see Heckman, Lochner and Taber,\(^{23}\)

\(^{23}\)In this paper we lump together individuals with some college and college graduates. In results available from the authors we implement a version of this model where, for the college wage regression, we include a dummy variable in the \( X \)s indicating whether the individual graduated from college or not (Willis and Rosen, 1979, control for the years of college in the college wage regression). This procedure does not alter our results significantly, and the shapes of the functions we display in this section (and therefore, the patterns of sorting of individuals across schooling levels) remain roughly the same. Heckman and Vytlacil (2004) and Carneiro and Heckman (2004) present extensions this framework that account for multiple (ordered) levels of schooling.\(^{24}\)However, they do not show the patterns of selection in levels displayed in this paper and their conclusion is indirect.
The density of observed outcome $Y$ conditional on $X = x$ can be written as:

$$f(y|x) = \int_0^1 f_{Y|X,V}(y|x,v) \left[1 - F_{P|X}(v|x)\right] dv + \int_0^1 f_{Y|X,V}(y|x,v) F_{P|X}(v|x) dv.$$ 

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 $f(y_0|x, v)$ or $f(y_1|x, v)$ (no general equilibrium effects). Then the post-policy distribution of wages is:

$$f^*(y|x) = \int_0^1 f_{Y|X,V}(y|x,v) \left[1 - F_{p'|X}(v|x)\right] dv + \int_0^1 f_{Y|X,V}(y|x,v) F_{p'|X}(v|x) dv.$$ 

If we can simulate this policy and compute $F_{p'|X}(p|x)$ we can easily study the effect of the policy on the distribution of income just by comparing $f(y|x)$ and $f^*(y|x)$. For doing welfare analysis of this policy (compare the baseline state with no policy with the policy state) under the commonly invoked veil of ignorance all we need is to compare $f(y|x)$ and $f^*(y|x)$.

If the goal of the exercise were just policy analysis we could avoid the two step procedure used in this paper (first we estimate $f(y_1|x, v)$ and $f(y_0|x, v)$, and then we estimate $F_{p|X}(v)$ and $F_{p'|X}(v)$) and perform a similar analysis in a more direct way using a version of the method developed by Ichimura and Taber (2000) (who also assume there are no general equilibrium effects). However, in this paper our aim is to provide a complete characterization of heterogeneity and therefore estimation of all the components is useful for our analysis.

In the remainder of this section, we first describe our estimators and then present empirical results. Since $f(y_1|x, v)$ and $f(y_0|x, v)$ can be obtained by location shifts from $f(u_1|v)$ and $f(u_0|v)$, that is

$$f_{Y_1|X,V}(y_1|x,v) = f_{U_1|V}(y_1 - \mu_1(x)|v)$$

$$f_{Y_0|X,V}(y_0|x,v) = f_{U_0|V}(y_0 - \mu_0(x)|v),$$

we focus on estimation of $f(u_1|v)$ and $f(u_0|v)$. We only discuss estimation of $f(u_1|v)$ in detail, since estimation of $f(u_0|v)$ is similar. To develop an estimator of $f(u_1|v)$ using the following relationships

$$\frac{\partial}{\partial p} [f_{U_1|P,S=1}(u_1|p)p] = f_{U_1|V}(u_1|p)$$

$$\frac{\partial}{\partial p} [f_{U_0|P,S=0}(u_0|p)(1 - p)] = -f_{U_0|V}(u_0|p),$$

25This is equivalent to a policy that changes the density of $Z$ in the population. One example is a tuition subsidy such as the one simulated in Heckman and Vytlacil (2001) and Carneiro, Heckman and Vytlacil (2003).

26However, invoking the veil of ignorance may not be the best way to do policy analysis in many circumstances. In particular, one would think individuals have some knowledge of $f_{Y|Y^*}(y, y^*)$, where $Y$ and $Y^*$ are individual outcomes under the two policy states. Even though $f_{Y|Y^*}(y, y^*)$ is not identified nonparametrically, Carneiro, Hansen and Heckman (2003) develop a method that allows us to compute $f(y, y^*)$ under some assumptions. It is still to be fully worked out what can be done using the framework we developed in this paper. Notice also that even though the estimation of $f(y_1|v)$ and $f(y_0|v)$ requires a monotonicity assumption (see Imbens and Angrist, 1994; Heckman and Vytlacil, 2000, 2004; Vytlacil, 2002), the policy analysis does not require a monotonicity assumption. For example, it is possible to have a policy that induces some individuals to enroll in college at the same time that induces other individuals to drop out of college (increase tuition for the rich and subsidize tuition for the poor).
it is necessary to estimate \( f_{U_1|P,S=1}(u_1|p) \) and its derivative with respect to \( p \). Specifically, the estimator of \( f(u_1|v) \) can be obtained by

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

(14)

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

In order to compute \( \hat{f}_{U_1|P,S=1}(u_1|v) \) and \( \frac{\partial}{\partial p} f_{U_1|P,S=1}(u_1|v) \) in (14), we begin with estimated data \( \{(\hat{U}_i, \hat{P}_i) : i = 1, \ldots, n, S_i = 1\} \), where \( \hat{U}_i = Y_i - \mu_1(X_i, \beta^1) \). 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, and then taking the ratio between them to estimate the conditional density, and finally computing a derivative of the conditional density. This indirect method 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 \( \delta_n \rightarrow 0 \),

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

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

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

(15)

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

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

(16)

The estimator defined in (14) 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 (14):

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

where \( \varepsilon \) 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 of \( F(u_1|v) \), note that

\[
F(u_1|v) = F_{U_1|V}(u_1|v) + \int_{u_1}^{u_2} f_{U_1|V}(u|v)du,
\]

for any fixed constant \( u_1 \). We estimate \( F(u_1|v) \) by replacing \( F_{U_1|V}(u_1|v) \) and \( f_{U_1|V}(u|v) \) in (17) with their sample analogs. More specifically, the estimator of \( F_{U_1|V}(u_1|v) \) is defined as

\[
\hat{F}_{U_1|V}(u_1|v) = \max[0, \hat{F}_{U_1|V}(u_1|v)],
\]

where

\[
\hat{F}_{U_1|V}(u_1|v) = v\frac{\hat{\partial}}{\partial p} F_{U_1|P,S=1}(u_1|v) + \hat{F}_{U_1|P,S=1}(u_1|v),
\]

and \( \hat{F}_{U_1|P,S=1}(u_1|v) \) and \( \hat{\partial} F_{U_1|P,S=1}(u_1|v)/\partial p \), respectively, are local linear and quadratic estimators that solve the problems similar to those in (15) and (16) with \( \delta_n^{-1}K((\hat{U}_{11} - u_1)/\delta_n) \) replaced by 1(\( \hat{U}_{11} \leq u_1 \)). Then our estimator of \( F(u_1|v) \) is defined as

\[
\hat{F}_{cdf}(u_1|v) = \min\left[ 1, \hat{F}_{U_1|V}(u_1|v) + \int_{u_1}^{u_2} \hat{f}_{pdf}(u|v)du \right].
\]

Notice that by construction, our estimator is a strictly increasing, continuous function of \( u_1 (u_1 > u_1) \) and is restricted to be between 0 and 1. In other words, our estimator is a distribution function for a given finite sample.\(^{27}\)

We are now ready to present our estimation results.\(^{28}\) Since we identify \( f_{U_1|V}(u_1|v) \) from \( f_{U_1|P,S=1}(u_1|p) \) we can only identify this function for pairs \((U_1, P)\) where we have positive support in the data. Similarly, it is only possible to identify \( f_{U_0|V}(u_0|v) \) on pairs of \((U_0, P)\) over which we observe positive support in the data. After checking scatter plots of \((\hat{U}_1, \hat{P})\) and \((\hat{U}_0, \hat{P})\), we restrict the domains of \( V \) accordingly and estimate \( f(u_1|v), f(u_0|v), F(u_1|v), \) and \( F(u_0|v) \) on the

\(^{27}\)One could use an unrestricted estimator (18), which is not necessarily a distribution function in finite samples.

\(^{28}\)Notice that as a by-product of estimating \( \hat{F}_{cdf}(u_1|v) \), we obtain an estimator of the \( \tau \)-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

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

where the right-hand side is unique for a given finite sample provided that \( u_1 \) is sufficiently small, since \( \hat{F}_{cdf}(u_1|v) \) is a strictly increasing function when \( u_1 > u_1 \). Furthermore, under the assumptions in Lemma 2, the \( \tau \)-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 quantile treatment effects, which are defined as

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

In the recent econometrics literature, quantile treatment effects have been considered by Abadie, Angrist, and Imbens (2002), Chernozhukov and Hansen (2004), Chesher (2003), and Imbens and Newey (2003) among others.
restricted domains.\textsuperscript{29}

Figure 5 graphs $f(u_1|v), f(u_0|v), F(u_1|v)$, and $F(u_0|v)$ evaluated at three different values of $v = 0.25, 0.50, 0.75$. By looking at estimated $f(u_1|v)$ (top-left panel), notice that even though $E(U_1|V)$ is decreasing in $V$ there is considerable overlap in the density of $U_1$ for individuals at different values of $V$. Nevertheless, estimated $F(u_1|v)$ (bottom-left panel) shows that there is a relationship of almost stochastic dominance between different densities of $U_1$ for different values of $V$. Specifically, the conditional distribution of $U_1$ given $V = 0.25$ stochastically dominates that of $U_1$ given $V = 0.50$, which in turns stochastically dominates that of $U_1$ given $V = 0.75$.\textsuperscript{30} Notice also that the shape of these densities varies considerably with $V$. The top-right and top-bottom panels display $f(u_0|v)$ and $F(u_0|v)$, separately. Again there is a clear ordering of the three densities: the conditional distribution of $U_0$ given $V = 0.25$ is stochastically dominated by that of $U_0$ given $V = 0.50$, which is in turns stochastically dominated by that of $U_0$ given $V = 0.75$.

Evidence based on figure 5 provides further support of the importance of absolute advantage in the labor market. It also shows that there is considerable higher dispersion in college wages than in high school wages, once we condition on $V$ and $X$. However, across values of $V$ and $X$ there is much higher variability in high school wages than in college wages ($E(Y_0|X,V)$ is much steeper in $V$ than $E(Y_1|X,V)$).\textsuperscript{31} If we interpret variability in wages across values of $V$ and $X$ as heterogeneity known at the time individuals are making schooling decisions and variability in wages within values of $X$ and $V$ as uncertainty (for a similar approach see Carneiro, Hansen and Heckman, 2003; and Cunha, Heckman and Navarro, 2003), we conclude that heterogeneity is relatively more important for potential high school wages while uncertainty is relatively larger for potential college wages.

Estimation of these objects also may help us gain some insights on the relationship between inequality and schooling and how this relationship has evolved over time. To start with, notice that the OLS estimate of the college-high school wage premium (a standard measure of between group inequality, where each group is defined according to education) is not the college wage premium for anyone in the economy. The college wage premium for individuals who do not attend college is

$$E(Y_1 - Y_0 | S = 0) = E(Y_1 | S = 0) - E(Y_0 | S = 0)$$

\textsuperscript{29}To estimate these conditional p.d.f.s and c.d.f.s, we adopt the same bandwidths $h_{n1}$ and $h_{n2}$ that are used to estimate the corresponding conditional means. The bandwidth $\delta_n$ is chosen by Silverman’s normal reference rule (Silverman, 1986, p.45). The selected bandwidth is $\delta_n = 0.137$ for estimating $f(u_1|v)$ and $\delta_0 = 0.122$ for estimating $f(u_0|v)$. These choices of bandwidths are arbitrary, but the estimation results are not sensitive to the choices of the bandwidths.

\textsuperscript{30}One could test stochastic dominance formally. See, for example, McFadden (1989), Abadie, (2002), and Linton, Maasoumi, and Whang (2003).

\textsuperscript{31}Therefore it is not clear whether $f(y_0|x)$ exhibits more or less dispersion than $f(y_1|x)$. 

22
while the return to college for those in college is
\[ E(Y_1 - Y_0 | S = 1) = E(Y_1 | S = 1) - E(Y_0 | S = 1) \]
and the OLS estimate is
\[ OLS = E(Y_1 | S = 1) - E(Y_0 | S = 0). \]

Then:
\[ OLS - E(Y_1 - Y_0 | S = 0) = E(Y_1 | S = 1) - E(Y_1 | S = 0) \]
\[ OLS - E(Y_1 - Y_0 | S = 1) = E(Y_0 | S = 0) - E(Y_0 | S = 1). \]

Using the results from our previous section it is simple to see that in our data the first expression is positive while the second is negative. The OLS estimate of the college wage premium is an overestimate of the wage premium for the average individual in high school and an underestimate for the average individual in college.\footnote{The usual intuition is that OLS estimates of the return to schooling are “too high” because of ability bias. In our data, this intuition only holds for the individuals in high school.} Suppose that we introduce a tuition subsidy in this economy and vary it from very low (negative) values (so that nobody attends college) to very high values (so that everyone attends college). As we increase the tuition subsidy and more and more individuals are induced to enroll in college the OLS estimate of the return to college should go down as individuals with higher \( V \) leave high school to enroll in college in increasingly higher numbers. As we expand education individuals of lower college quality enter college, and individuals of worse high school quality leave high school, decreasing average college wages and increasing average high school wages. As a result the OLS estimate of the college premium decreases.

Similarly, as individuals move from high school to college dispersion in wages may tend to decrease because heterogeneity in terms of \( V \) and \( X \) is more important in high school than in college. However, dispersion in wages within groups of individuals with the same values of \( X \) and \( V \) is more important in college than in high school, and therefore overall dispersion in wages may tend to rise.

We now simulate a model with a tuition subsidy where we vary net tuition from very low negative levels (so that almost everyone is induced to attend college) to very high levels (so that almost everyone is prevented from going to college by high tuition costs) and examine the change in overall inequality and in inequality between and within groups.\footnote{Recall that our choice model is the following: \( S = 1 \) if \( P(Z) > V \), where \( P(z) = L[\mu_S(z)] \) and \( L(u) = \exp(u)/[1 + \exp(u)] \). The function \( \mu_S(z) \) is specified parametrically to be a sum of two parts, that is \( \mu_S(z) = \mu_{S,X}(x) + \mu_{SZ}(z') \), where \( \mu_{S,X}(x) \) consists of linear, quadratic, and interaction terms of all the variables in \( X \) and \( \mu_{SZ}(z') \) consists of a linear function of variables in \( Z' \), including tuition. To simulate the effect of a particular tuition subsidy we change tuition for every individual by an amount equal to the subsidy, and then we recompute the new density of \( P \) in the economy. Since the density of \( P \) defines the distribution of \( V \) in each sector we can examine how different densities of \( P \) (corresponding to different subsidies) change the sorting of individuals across schooling levels in the economy. In practice, we compute weighted averages of \( E(Y_1 | X, V) \), \( E(Y_0 | X, V) \), \( F_{Y_1 | X, V}(y_1 | x, v) \) and \( F_{Y_0 | X, V}(y_0 | x, v) \) using the procedures proposed in Heckman and Vytlacil (2000, 2001, 2004) and Carneiro, Heckman and Vytlacil (2003).}
This will help us understand the effects of changes in the composition of the high school and college population on inequality. Figure 6 has four panels. In the top-left panel we show how probability of going to college changes when tuition increases. In the top-right panel we show how the OLS parameter (the standard measure of between-skill-group inequality) changes. In the bottom-left panel we look at trends in overall wages and within-skill-group wages by tracing the changes in the median of $Y, Y_1$ conditional on $S = 1$, and $Y_0$ conditional on $S = 0$, respectively. Finally, in the bottom-right panel we examine changes in overall inequality (changes in the distribution of $Y$) as well as changes in within-skill-group inequality (changes in the distributions of college and high school wages, shown separately) by computing the interquartile range (the difference between the 25th and 75th percentile of the wage distribution) for $Y, Y_1$ conditional on $S = 1$, and $Y_0$ conditional on $S = 0$, respectively.

The OLS parameter rises sharply as the proportion of individuals attending college decreases. The reason behind this is that the population of college graduates becomes more select and of better quality as tuition increases, and therefore their average wages increase. Simultaneously, the population of high school graduates becomes of lower quality and average wages decrease. As a result, the OLS estimate increases with the rise in the level of tuition. Similarly, median college wages increase, median high school wages decrease, and interestingly the trends in overall wages are not monotone as tuition changes.

As tuition increases from very low to very high levels, the proportion of individuals attending college decreases from close to 100% to close to 0%, the 75-25% college wage differential (defined as the interquartile range of $Y_1$ conditional on $S = 1$) decreases by 13% and the 75-25% high school wage differential (defined as the interquartile range of $Y_0$ conditional on $S = 0$) increases by 6%. Therefore, as the proportion of individuals attending college decreases, between-group-inequality increases, one within-group-inequality (high school) rises but the other inequality (college) falls. As a result, overall inequality could increase or decrease. In our sample, changes in the overall 75-25% wage differential are not monotone with the increase in tuition.\footnote{For many policy purposes it is especially interesting to characterize the objects studied in this paper for the marginal students in the economy, the set of individuals just indifferent between enrolling or not in college. The reason is that these are the individuals more likely enroll in college in response to policy intervention. Carneiro, Heckman and Vytlacil (2003) estimate that the return to college for the marginal student is considerably below the return to college for the average student in the economy. Applying their framework to our paper we would find that the marginal student is a worse potential college participant than the average college participant and a worse potential high school graduate than the average high school graduate. Using this framework we would also be able to estimate the dispersion of wages of the marginal set of individuals and the uncertainty in wages they face at the time of the college decision, and compare these estimates with estimates for the average person in the economy.}

There are several recent studies that analyze the impact of changes in the demand and supply of high school and college graduates on the college wage premium.\footnote{Katz and Murphy (1992) and Card and Lemieux (2001) are two important examples.} Keeping the demand for college
and high school workers fixed, we should expect that an increase in the supply of college graduates
and a decrease in the supply of high school graduates should lead to a decrease in the wage of college
graduates and an increase in the wage of high school graduates. In such analysis we usually ignore
heterogeneity and assume that all college graduates (in the same cohort) have similar amounts of
skill. In this paper we show in an environment where there is heterogeneity in skill endowments,
even if the price of skill is kept fixed (instead of responding to changes in the relative scarcity of
different types of skill in the economy), an increase in the supply of college graduates would lead to a
fall in their average wage because the average quality of college graduates decreases, while a decrease
in the supply of high school graduates would lead to an increase in their average wage because the
average quality of the pool of high school graduates increases. Changes in sorting can account for
large changes in standard measures of the college wage premium, even when there is no change in the
price of skill. Therefore, studies of changes in the price of skill and of the way supply and demand
factors that contribute to such changes should take such sorting mechanism into account.

We end this section by mentioning some results of simple goodness-of-fit check. The OLS param-
eter $E[Y_1|S = 1] - E[Y_0|S = 0]$ can be estimated directly using the difference between two sample
means (reduced-form estimator) and can also be computed by integrating $E[Y_1|x, v]$ and $E[Y_0|x, v]$ with appropriate weights (given in Heckman and Vytlacil, 2004), respectively, and then by averaging
over $X$ (model-based estimator). The reduced-form and model-based annualized estimates are 0.088
and 0.103, respectively. The difference 0.015 is small, compared to the estimated values. Similarly,
the quantiles of $Y$, $Y_1$ conditional on $S = 1$, and $Y_0$ conditional on $S = 0$ can be estimated directly
from the data without using our model and indirectly using our model. Comparison between esti-
mates of quantiles (details are available upon request) suggest that our model fits the data pretty
well, although it fits the data worse at the tails of the distributions.

6 Conclusions

In this paper we estimate patterns of sorting in the labor market using a semiparametric selection
model applied to a sample of white males in the NLSY. We find that heterogeneity and selection
are important in this data and should be seriously considered in studies of schooling and wages, and
in studies of wage inequality. Comparative advantage and absolute advantage govern the sorting
of individuals across schooling levels. We are able to replicate the qualitative patterns that Willis
and Rosen (1979) found twenty-five years ago using a parametric model and a different dataset: i) returns to college are higher for individuals who go to college than for those who do not go to college;
ii) if he or she went to college, the average high school graduate would be a worse college attendee
than the average college attendee; iii) if he or she did not go to college, the average college attendee
would be a worse high school graduate than the average high school graduate.

Notice that even though we cannot estimate \( f(y_1, y_0) \), we can estimate \( f(y_1) \) and \( f(y_0) \). In such a setting, Heckman, Smith and Clemens (1997) argue that a variety of different assumptions about the correlation between \( Y_1 \) and \( Y_0 \) generate different estimates of \( f(y_1, y_0) \). One example is a perfect ranking assumption, where the best individual in terms of \( Y_1 \) is also the best in terms of \( Y_0 \). Our results show that this assumption is rejected in the data since we find that, on average, individuals with low levels of \( Y_0 \) have high levels of \( Y_1 \). Several other restrictions on \( f(y_1, y_0) \) may be possible to rule out based on our empirical results. In future work we plan to analyze what further restrictions our estimates impose on the joint distribution of counterfactual wages in college and high school.

Standard analyses of the evolution of inequality across time such as Juhn, Murphy and Pierce (1993) decompose overall inequality into within and between group inequality and present the main patterns in the evolution of these two components. However, changes in the sorting of individuals across different groups (in this case, schooling levels) are usually neglected, although they can be potentially important determinants of inequality within and between groups. Over the last twenty years there has been some growth in college attendance rates in the US. Even if the underlying distribution of skills in the economy is stable over time, the allocation of these skills across sectors is not stable over time. For example, the additional college graduates that have been entering the college sector over time may be of lower ability than the college graduates that were there before, which can be a force towards lower average college wages and higher wage dispersion across college graduates. Blundell, Reed and Stoker (2003) illustrate the importance of a similar problem for the study of wage growth and growth in the returns to schooling (inequality between schooling groups) over time. In their model, they have two sectors: the labor market sector and the home sector. We do not observe wages for individuals in the home sector and they are a self selected sample of the population. Furthermore, the sorting of individuals across these two sectors changes dramatically over time which has important consequences for the study of wage growth and wage inequality.

This problem is also widely recognized in the study of racial and gender wage differentials. For example, a recent paper by Chandra (2003) illustrates how different assumptions about the sorting of blacks across the working and non-working sectors influence dramatically inferences about black white wage convergence over time. Neal (2004) shows that what we assume about the selection of black and white women across the working and home sectors has important consequences for the measurement of the wage gap between these two groups. Even though it is hard for us to study the evolution of sorting over time (given that we use data on a limited set of cohorts which have already made their educational decisions), we can provide some insights about the impact of sorting into

---

36This is one of many other papers in this literature that are relevant to our study.
education levels on the distribution of wages and how the impact of such sorting evolved over the 1990s. In future work we plan to study part of the evolution of wage inequality during the 1990s using the NLSY over the years between 1990 and 2000. Such an analysis will help us understand the change in inequality in the 1990s keeping the composition of the high school and college population fixed (subject to assumptions on age, time and cohort effects). By simulating a tuition subsidy with different values, as in the previous section, we can analyze the effects of changes in composition. We also plan to expand our analysis to different demographic groups.

Finally, using the ideas first developed in Flavin (1981) and later revisited in Carneiro, Hansen and Heckman (2003), Cunha, Heckman and Navarro (2004) and Blundell, Pistaferri and Preston (2002), we plan to contribute to the study of uncertainty in wages and schooling choices. Since we can estimate how the observable and unobservable variables that determine schooling choice correlate with the observable and unobservable variables that determine potential wages in each sector we can estimate the amount of uncertainty that an individual faces by choosing each sector. We allow this uncertainty to differ across individuals with characteristics (observable and unobservable) which make them more or less likely to enroll in college. we plan to expand this idea in future research.

Appendix A: Proof of Lemma 2

Under the conditions assumed in Lemma 2, notice that

\[
E[Y|X = x, P = p, S = 1] = \mu_1(x) + \int_0^p E[U_1|V = v] \, dv / p \\
\equiv \mu_1(x) + \lambda_1(p)
\]

\[
E[Y|X = x, P = p, S = 0] = \mu_0(x) + \int_p^1 E[U_0|V = v] \, dv / (1 - p) \\
\equiv \mu_0(x) + \lambda_0(p),
\]

where \( \lambda_1 \) and \( \lambda_0 \) are only functions of \( p \). Therefore, \( \mu_1(x) \) and \( \mu_0(x) \) are identified up to location. This implies that \( U_1 \) and \( U_0 \) can be treated as observed variables for the purpose of identification. In addition, arguments identical to those used in the proof of Lemma 1 give

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

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

Then \( E(Y_1|X = x, V = p) \) and \( E(Y_0|X = x, V = p) \) are identified since

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

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

(19)
Similarly, we can identify $F_{U_1|V}(u_1|v)$ and $F_{U_0|V}(u_0|v)$. Then $F_{Y_1|X,V}(y_1|x,v)$ and $F_{Y_0|X,V}(y_0|x,v)$ are identified from the simple relationships

\[ F_{Y_1|X,V}(y_1|x,v) = F_{U_1|V}(y_1 - \mu_1(x)|v) \]
\[ F_{Y_0|X,V}(y_0|x,v) = F_{U_0|V}(y_0 - \mu_0(x)|v). \]

Thus, the conditional quantiles of $Y_1$ and $Y_0$ given $X = x$ and $V = v$ are identified by inverting $F_{Y_1|X,V}(y_1|x,v)$ and $F_{Y_0|X,V}(y_0|x,v)$, respectively.

Using a very similar reasoning and method to the one used for the MTE in Heckman and Vytlacil (2000, 2004a), we can also identify $E(Y_1|X = x, V = p)$ and $E(Y_0|X = x, V = p)$. Notice that:

\[ E(SY|X = x, P = p) = p\mu_1(x) + E(U_1|S = 1, P = p)p \]
\[ = p\mu_1(x) + \int_0^p E(U_1|V = v)dv \]
\[ E((1 - S)Y|X = x, P = p) = (1 - p)\mu_0(x) + E(U_0|S = 0, P = p)(1 - p) \]
\[ = (1 - p)\mu_0(x) + \int_0^p E(U_0|V = v)dv. \]

Then:

\[ \frac{\partial E(SY|X = x, P = p)}{\partial p} = E(Y_1|X = x, V = p) \]
\[ \frac{\partial E((1 - S)Y|X = x, P = p)}{\partial p} = -E(Y_0|X = x, V = p). \]

The identifying relationships described in (19) are used in our empirical work because it turns out that the estimation results based on (19) are more reliable than those based on (20).

Appendix B: Description of the Data

We restrict the NLSY sample to white males with a high school degree or above. We define high school graduates as individuals having high school degree, or having completed 12 grades and never reporting college attendance. We define participation in college as having ever gone to college or having complete more than 12 grades in school. 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 variable that is used is an average of all deflated (to 1983) no missing hourly wages from 1991 to 1993. We delete all wage observations that are below $1 and above $100. Experience is actual work experience in weeks accumulated from 1979 to 1992 (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 deflated (to 1993) tuition of the
colleges in the county the individual lives in at 17, distance to the nearest college at age 14 and local unemployment rate in county of residence in 1979. For the construction of the tuition variable see Cameron and Heckman (2001). Distance to college is constructed by matching college location data in HEGIS with county of residence in NLSY. 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. We are left with 2439 white males, and 1916 of these have a high school degree (or equivalent) or above. After deleting observations for which one or more of the variables we use are missing we obtain a sample of 1417 individuals. Many individuals report having a bachelors degree or more and at the same time having only 15 years of schooling (or less). We recode years of schooling for these individuals to be 16. This variable is only used to annualize the returns to schooling. If we did not perform this recoding, when computing returns to one year of college we would divide the returns to schooling by 3.2 instead of dividing by 3.5. This corresponds to multiplying all the estimated returns in the paper by 3.5/3.2 = 1.09. To remove the effect of schooling on AFQT we implement the same procedure as in Carneiro, Heckman and Vytlacil (2003) (based on Hansen, Heckman and Mullen, 2003).

References


Table 1. Summary Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev.</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Observations (n = 1417)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log Wage</td>
<td>2.434</td>
<td>2.433</td>
<td>0.486</td>
<td>0.063</td>
<td>4.311</td>
</tr>
<tr>
<td>College Attendance</td>
<td>0.514</td>
<td>1.000</td>
<td>0.500</td>
<td>0.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Years of Experience</td>
<td>9.896</td>
<td>10.192</td>
<td>2.924</td>
<td>0.731</td>
<td>16.000</td>
</tr>
<tr>
<td>Corrected AFQT</td>
<td>0.196</td>
<td>0.334</td>
<td>0.916</td>
<td>-3.119</td>
<td>2.381</td>
</tr>
<tr>
<td>Number of Siblings</td>
<td>2.795</td>
<td>2.000</td>
<td>1.761</td>
<td>0.000</td>
<td>15.000</td>
</tr>
<tr>
<td>Mother’s Years of Schooling</td>
<td>12.414</td>
<td>12.000</td>
<td>2.185</td>
<td>3.000</td>
<td>20.000</td>
</tr>
<tr>
<td>Father’s Years of Schooling</td>
<td>12.841</td>
<td>12.000</td>
<td>3.142</td>
<td>0.000</td>
<td>20.000</td>
</tr>
<tr>
<td>Average County Tuition</td>
<td>20.082</td>
<td>20.616</td>
<td>7.979</td>
<td>0.000</td>
<td>53.592</td>
</tr>
<tr>
<td>Distance to College</td>
<td>3.740</td>
<td>0.000</td>
<td>10.262</td>
<td>0.000</td>
<td>77.802</td>
</tr>
<tr>
<td>County Unemployment Rate</td>
<td>6.315</td>
<td>6.200</td>
<td>1.679</td>
<td>2.800</td>
<td>12.400</td>
</tr>
<tr>
<td>S = 1 (n = 729)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log Wage</td>
<td>2.583</td>
<td>2.596</td>
<td>0.495</td>
<td>0.063</td>
<td>4.311</td>
</tr>
<tr>
<td>Years of Experience</td>
<td>9.444</td>
<td>9.635</td>
<td>2.952</td>
<td>0.750</td>
<td>15.904</td>
</tr>
<tr>
<td>Corrected AFQT</td>
<td>0.637</td>
<td>0.810</td>
<td>0.718</td>
<td>-2.120</td>
<td>2.381</td>
</tr>
<tr>
<td>Number of Siblings</td>
<td>2.593</td>
<td>2.000</td>
<td>1.654</td>
<td>0.000</td>
<td>11.000</td>
</tr>
<tr>
<td>Mother’s Years of Schooling</td>
<td>13.103</td>
<td>12.000</td>
<td>2.260</td>
<td>4.000</td>
<td>20.000</td>
</tr>
<tr>
<td>Father’s Years of Schooling</td>
<td>13.951</td>
<td>14.000</td>
<td>3.080</td>
<td>4.000</td>
<td>20.000</td>
</tr>
<tr>
<td>Average County Tuition</td>
<td>19.300</td>
<td>19.656</td>
<td>7.867</td>
<td>0.000</td>
<td>53.592</td>
</tr>
<tr>
<td>Distance to College</td>
<td>2.890</td>
<td>0.000</td>
<td>9.655</td>
<td>0.000</td>
<td>76.026</td>
</tr>
<tr>
<td>County Unemployment Rate</td>
<td>6.305</td>
<td>6.200</td>
<td>1.674</td>
<td>2.800</td>
<td>12.400</td>
</tr>
<tr>
<td>S = 0 (n = 688)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Log Wage</td>
<td>2.277</td>
<td>2.290</td>
<td>0.424</td>
<td>0.564</td>
<td>3.711</td>
</tr>
<tr>
<td>Years of Experience</td>
<td>10.375</td>
<td>10.827</td>
<td>2.817</td>
<td>0.731</td>
<td>16.000</td>
</tr>
<tr>
<td>Corrected AFQT</td>
<td>-0.272</td>
<td>-0.184</td>
<td>0.870</td>
<td>-3.119</td>
<td>1.926</td>
</tr>
<tr>
<td>Number of Siblings</td>
<td>3.010</td>
<td>3.000</td>
<td>1.845</td>
<td>0.000</td>
<td>15.000</td>
</tr>
<tr>
<td>Mother’s Years of Schooling</td>
<td>11.683</td>
<td>12.000</td>
<td>1.840</td>
<td>3.000</td>
<td>18.000</td>
</tr>
<tr>
<td>Father’s Years of Schooling</td>
<td>11.664</td>
<td>12.000</td>
<td>2.757</td>
<td>0.000</td>
<td>20.000</td>
</tr>
<tr>
<td>Average County Tuition</td>
<td>20.910</td>
<td>21.350</td>
<td>8.018</td>
<td>0.000</td>
<td>47.056</td>
</tr>
<tr>
<td>Distance to College</td>
<td>4.641</td>
<td>0.000</td>
<td>10.803</td>
<td>0.000</td>
<td>77.802</td>
</tr>
<tr>
<td>County Unemployment Rate</td>
<td>6.326</td>
<td>6.200</td>
<td>1.686</td>
<td>3.200</td>
<td>12.400</td>
</tr>
</tbody>
</table>

Note: High school dropouts are excluded from the sample.
Table 2. Average Derivatives for the College Attendance Model Using Logit

<table>
<thead>
<tr>
<th>Variable</th>
<th>Average Derivative</th>
<th>Standard Error</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corrected AFQT</td>
<td>0.231</td>
<td>0.016</td>
<td>14.284</td>
</tr>
<tr>
<td>Number of Siblings</td>
<td>-0.015</td>
<td>0.008</td>
<td>-1.979</td>
</tr>
<tr>
<td>Mother’s Years of Schooling</td>
<td>0.025</td>
<td>0.007</td>
<td>3.707</td>
</tr>
<tr>
<td>Father’s Years of Schooling</td>
<td>0.024</td>
<td>0.005</td>
<td>5.206</td>
</tr>
<tr>
<td>Average County Tuition</td>
<td>-0.004</td>
<td>0.001</td>
<td>-2.734</td>
</tr>
<tr>
<td>Distance to College</td>
<td>-0.001</td>
<td>0.001</td>
<td>-0.681</td>
</tr>
<tr>
<td>County Unemployment Rate</td>
<td>0.008</td>
<td>0.007</td>
<td>1.135</td>
</tr>
</tbody>
</table>

Note: The average derivatives are obtained from a logit regression of college attendance on a constant, linear, quadratic, and interaction terms of AFQT, number of siblings, mother’s education, and father’s education, and linear terms of tuition, distance to college, and local unemployment rate, separately. The standard errors are computed using asymptotic approximation and the delta method.
Table 3. Average Derivatives for the Log Wages

(S = 1) Econometric Model: \( E[Y_1|X,V] = \mu_1(X,\beta^1) + E[U_1|V] \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Average Derivative</th>
<th>Standard Error</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of Experience</td>
<td>0.030</td>
<td>0.008</td>
<td>3.584</td>
</tr>
<tr>
<td>Corrected AFQT</td>
<td>0.222</td>
<td>0.081</td>
<td>2.756</td>
</tr>
<tr>
<td>Number of Siblings</td>
<td>0.006</td>
<td>0.015</td>
<td>0.401</td>
</tr>
<tr>
<td>Mother’s Years of Schooling</td>
<td>0.027</td>
<td>0.014</td>
<td>1.851</td>
</tr>
<tr>
<td>Father’s Years of Schooling</td>
<td>0.019</td>
<td>0.013</td>
<td>1.473</td>
</tr>
</tbody>
</table>

Inference
\( H_0: \) All coefficients of \( \hat{P} \) are zero.

<table>
<thead>
<tr>
<th>Wald Test Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.247</td>
<td>0.325</td>
</tr>
</tbody>
</table>

(S = 0) Econometric Model: \( E[Y_0|X,V] = \mu_0(X,\beta^0) + E[U_0|V] \)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Average Derivative</th>
<th>Standard Error</th>
<th>t-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of Experience</td>
<td>0.047</td>
<td>0.007</td>
<td>7.026</td>
</tr>
<tr>
<td>Corrected AFQT</td>
<td>-0.230</td>
<td>0.127</td>
<td>-1.811</td>
</tr>
<tr>
<td>Number of Siblings</td>
<td>0.021</td>
<td>0.017</td>
<td>1.181</td>
</tr>
<tr>
<td>Mother’s Years of Schooling</td>
<td>-0.014</td>
<td>0.021</td>
<td>-0.678</td>
</tr>
<tr>
<td>Father’s Years of Schooling</td>
<td>-0.018</td>
<td>0.015</td>
<td>-1.218</td>
</tr>
</tbody>
</table>

Inference
\( H_0: \) All coefficients of \( \hat{P} \) are zero.

<table>
<thead>
<tr>
<th>Wald Test Statistic</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.477</td>
<td>0.065</td>
</tr>
</tbody>
</table>

Note: The average derivatives are obtained, using observations with \( S = 1 \) and those with \( S = 0 \) separately, from partially linear regressions of log wages on linear, quadratic, and interaction terms of experience, AFQT, number of siblings, mother’s years of schooling, and father’s years of schooling (parametric part), and a series approximation of an unknown function of the predicted probability of attending college (nonparametric part). The predicted probability is estimated by a logit regression reported in Table 2. A quadratic polynomial is used for the series approximation. The standard errors are computed using asymptotic approximation and the delta method, while accounting for estimation of the predicted probability.
Figure 1. Support of $P(Z)$ and $\mu_{SX}(X)$. 
Figure 2. Decomposition of $E(Y_1 - Y_0|X = x, V = v)$ as a function of $v$, given at $X = \bar{X}$. 
Figure 3. Decomposition of $E[Y_1 - Y_0|\mu_{S_X}(X) = \mu, V = v]$ as a function of $\mu$, given at $V = 0.5$. 
Figure 4. Estimate of $E[Y_1 - Y_0|\mu_{SX}(X) = \mu, V = v]/3.5$ as a function of $\mu$ and $v$. 
Figure 5. Estimates of $f[u_1|v]$, $f[u_0|v]$, $F[u_1|v]$ and $F[u_0|v]$ given at $v = 0.25, 0.50, 0.75$. 
Figure 6. Simulation Results of Tuition Experiment