CentER

Discussion Paper

No. 2009–88

COLLEGE EDUCATION AND WAGES IN THE U.K.: ESTIMATING CONDITIONAL AVERAGE STRUCTURAL FUNCTIONS IN NONADDITIVE MODELS WITH BINARY ENDOGENOUS VARIABLES

By Tobias J. Klein

September 2009

ISSN 0924-7815

TILBURG UNIVERSITY

College Education and Wages in the U.K.: Estimating Conditional Average Structural Functions in Nonadditive Models with Binary Endogenous Variables

Tobias J. Klein[*]
Netspar, CentER, Tilburg University

September 4, 2009

Abstract

Recent studies debate how the unobserved dependence between the monetary return to college education and selection into college can be characterized. This paper examines this question using British data. We develop a semiparametric local instrumental variables estimator for identified features of a flexible correlated random coefficient model. These identified features are directly related to the marginal and average treatment effect in policy evaluation. Our results indicate that returns to college systematically differ between actual college graduates and actual college non-graduates. They are on average higher for college graduates and positively related to selection into college for 96 percent of the individuals. The dependence between selection into college and returns to college education is strongest for individuals with low math test scores at the age of 7, individuals with less educated mothers, and for working-class individuals.

JEL Classification: C14, C31, J31.

Keywords: Returns to college education, correlated random coefficient model, local instrumental variables estimation.

[*] *Address:* Tilburg University, Department of Econometrics and OR, PO Box 90153, 5000 LE Tilburg, The Netherlands. *E-Mail:* T.J.Klein@uvt.nl.

1. Introduction

In labor economics it is well understood that estimating the returns to college education is anything but straightforward.[1] One reason for this is that an individual's educational choice is likely to be based on information that is superior to what is recorded in the data. For example, if ability is unobserved and positively related to the return to a college degree, then we might expect individuals with high ability to be more likely to obtain a college degree. A direct consequence of this is that a simple regression of wages on an indicator for college education yields upward biased estimates of the average return to college education.

Neither matching techniques nor standard instrumental variables estimators (referred to as 2SLS from now on) are able to overcome this problem because they preclude any unobserved dependence between the return to a college degree and selection into college education (Heckman and Vytlacil, 1998; Heckman, Urzua, and Vytlacil, 2006; Wooldridge, 2007). However, following Imbens and Angrist (1994), Heckman and Vytlacil (1999, HV in the remainder) show that the average return for a well-defined subgroup can be estimated using a local instrumental variables estimator. It is local because it estimates the return using only observations for which the probability to obtain a college degree falls into a small neighborhood.[2]

In this paper we set up a flexible correlated random coefficient model that allows for binary endogenous variables. We develop a semiparametric local instrumental variables estimator for identified features of this model, among others the conditional average structural function (CASF), which is the expected wage if (no) college education is assigned to an individual. The CASF depends both on observed and unobserved characteristics that lead to selection into college. The identifiable features that are estimated are directly related to the marginal and average treatment effect in policy evaluation. Hence, the estimator is suitable for many other situations in which a binary endogenous variable is related to its effect. One example is the estimation of

[1] See Griliches (1977) and Card (2001) for surveys on the returns to schooling and Solmon and Taubman (1973) on the returns to college education.

[2] See Heckman and Vytlacil (2001, 2005) as well as Heckman, Urzua, and Vytlacil (2006) for a comprehensive discussion.

the effect of participation in a training program on unemployment duration, where the participation decision is made in light of the idiosyncratic expected effect of the program.

We implement this estimator using data from United Kingdom's National Child Development Survey (NCDS). Our results indicate that returns to college systematically differ between college graduates and college non-graduates. They are on average higher for college graduates. We find that returns are positively related to selection into college for 96 percent of the individuals. The difference in returns between those who actually attend college and those who do not is largest for individuals with low math test scores, less educated mothers, and individuals whose father's occupation is associated with a lower social class. Many of those individuals actually do not attend college, but would profit from doing so if they were to have high levels of unobserved ability. Thus, within this well-specified subgroup of individuals, those with high levels of unobserved ability should be encouraged to attend college.

The remainder of this paper is organized as follows. Section 2 presents and discusses the econometric approach. In Section 3 we describe the data set. Section 4 contains the empirical results. In Section 5 we assess the validity of the instruments. Section 6 concludes.

2. Econometric Approach

2.1. Econometric Model

Our point of departure is the correlated random coefficient model

$$Y = X'\varphi(D, U, V) \tag{1}$$

$$D = 1\{P(Z) \geq V\}. \tag{2}$$

Y is the log hourly wage at the age of 33, X is a K-vector of covariates that includes a constant term, a math test score, family background variables, and regional indicators.[3] D is an indicator for having received college education. Z contains X and variables that are excluded from the wage equation (instruments). In our analysis these excluded variables are indicators for the parents' interest in the education of the child. At least one variable in X or Z is continuous. U is unobserved and possibly vector-valued. V is an unobserved scalar random variable. $\varphi(\cdot, \cdot, \cdot)$ is a vector valued function and $P(\cdot)$ is a scalar-valued function. We will refer to (1) as the wage equation and (2) as the selection equation. This model allows for unobserved dependence between wages and selection into college because V enters both the wage equation and the selection equation. We will later interpret low values of V as representing high unobserved ability regarding formal schooling, and high values of V as representing low ability.

We impose the following stochastic restrictions.

Assumption 1 (Stochastic Restrictions): (i) (U, V) *are jointly independent of* (X, Z) *and* (ii) U *is independent of* V.

Assumption 1(i) prescribes that the distribution of V is unrelated to all variables in X and Z. Assumption 1(ii) restricts the randomness in Y through U to be unrelated to V.[4]

The approach taken here is inspired by the nonparametric identification result in HV. They show nonparametric identification of various parameters of interest under the assumption that (U, V) is jointly independent of Z conditional on X, and that X is is not affected by the choice of D.[5] In addition they require that there is a continuous variable in Z that is not in X. In practice, however, the typical situation is that researchers only have access to discrete instruments, e.g.

[3] We will denote (vectors of) random variables by uppercase letters and their respective typical elements by lowercase letters.

[4] One can show that this is not restrictive because there exists an observationally equivalent model with three unobservables, U_D in the selection equation, U_0 in the outcome equation for $D = 0$, and U_1 in the outcome equation for $D = 1$, that are not restricted to be independent of one another. Derivations are available upon request from the author.

[5] The requirement that X is is not affected by the choice of D is weaker than the assumption that X is exogenous. See the discussion in Heckman and Vytlacil (2005) and Section 2.4.

when exogenous variation in eligibility rules is used (Battistin and Rettore, 2008, e.g.). It is therefore worth noting that under Assumption 1 we can instead exploit continuous variation in X. Another difference is that HV require that Z shifts the probability to observe $D = 1$ from 0 to 1, given X, whereas we only require this for the unconditional probability to observe $D = 1$. To summarize, the stochastic restrictions in HV are weaker but the support conditions are stronger. In Section 2.4 we discuss how Assumption 1 restricts the set variables that can be included in X.

Apart from the stochastic restrictions, we assume that the following regularity conditions hold.

Assumption 2 (Regularity Conditions): *(i) All first moments exist and (ii) the distribution of V is absolutely continuous with respect to Lebesgue measure.*

Assumption 2(i) ensures that all parameters of interest defined below exist. Assumption 2(ii) implies that V is a continuous random variable. This allows us, without loss of generality (w.l.o.g.), to normalize V from now on to be uniformly distributed on the unit interval, see, e.g., Vytlacil (2002) for details. It then follows immediately from Assumption 1(i) that $P(Z)$ is identified since it is equal to $\Pr(D = 1|Z)$. For simplicity, we will write P for $P(Z)$ in the remainder, and denote its typical element by p.

The CASF is the average outcome when we assign $D = d$, $X = x$, and $V = v$, i.e.

$$G(d, x, v) \equiv x'\mathbb{E}[\varphi(d, U, v)]. \tag{3}$$

Here, we average over U. The terminology CASF is related to the one used by Blundell and Powell (2003). They suggest to focus on recovering the average structural function,

$$\bar{G}(d, x) \equiv x'\mathbb{E}[\varphi(d, U, V)] \tag{4}$$

in our case, where we also average over V.

We are further interested in

$$\frac{\partial G(d,x,v)}{\partial x} = \mathbb{E}[\varphi(d,U,v)], \tag{5}$$

the vector of conditional average *ceteris paribus* effects, understanding the notion of *ceteris paribus* as holding all other factors constant, including V, while again averaging over U. Finally,

$$\frac{\partial \overline{G}(d,x)}{\partial x} = \mathbb{E}[\varphi(d,U,V)] \tag{6}$$

is the vector of average *ceteris paribus* effects, where we again average over V.

Equation (2) prescribes how the decision to attend college depends on V. For a given $P = p$, those individuals with $V \leq p$ sort into college and those with $V > p$ do not. Hence, we can think of low values of V as representing high levels of ability, and high values representing low ability. The dependence of the CASF and the conditional average *ceteris paribus* effects on V is therefore informative about the relationship between wages and selection into college, and how this relationship depends on observed characteristics X.

The parameters of interest that were defined above are directly related to the treatment effect parameters in policy evaluation. The difference in the CASF between $D = 1$ and $D = 0$ is Björklund and Moffit's (1987) marginal treatment effect (MTE) for $X = x$,

$$G(1,x,v) - G(0,x,v) = x'(\mathbb{E}[\varphi(1,U,v)] - \mathbb{E}[\varphi(0,U,v)]). \tag{7}$$

This is the expected effect of a college degree on wages for a given level of unobserved ability and for a given vector of covariates. The average treatment effect (ATE) is the average MTE when we average over the population distribution of unobserved ability. For a given $X = x$ it is

$$x' \int_0^1 (\mathbb{E}[\varphi(1,U,v)] - \mathbb{E}[\varphi(0,U,v)])\, dv, \tag{8}$$

recalling that we have normalized V to be uniformly distributed. It follows from Assumption 1 that the unconditional ATE is equal to the expression in equation (8), evaluated at the mean of X. The average treatment effect on the treated (ATT) can be obtained by integrating over the distribution of V, conditional on $D = 1$, and evaluating the expression at the mean of X conditional on $D = 1$. For the average treatment effect on the untreated (ATU) we use the distribution of V conditional on $D = 0$ and the mean of X conditional on $D = 0$.[6]

2.2. Identification

In this subsection, we show that the CASF is identified. Because of the multiplicative structure of the wage equation, identification of the CASF at $D = d$, $V = v$ and $X = x$, equation (3), is equivalent to identification of the conditional average *ceteris paribus* effects, equation (5).[7] The average structural function, equation (4), and average *ceteris paribus* effects, equation (6), are identified at $D = d$ if the CASF is identified at all v in the open unit interval, recalling that we have normalized V to be uniformly distributed and that the endpoints have probability measure zero. Finally, if the (conditional) average structural function is identified at both $D = 0$ and $D = 1$, then so is the average (marginal) treatment effect.

From equation (1) it follows that

$$\mathbb{E}[Y|D = 1, P = p, X = x] = x'\mathbb{E}[\varphi(1, U, V)|D = 1, P = p, X = x] \tag{9}$$

which is equal to

$$x'\mathbb{E}[\varphi(1, U, V)|P \geq V, P = p, X = x]$$

by the selection model in equation (2). Assumption 1(i) implies that P is independent of (U, V)

[6] See Heckman and Vytlacil (1999, 2000) for the relationship between treatment parameters within a latent variable framework.

[7] Since X includes a constant, the intercept is identified once the conditional average *ceteris paribus* effects are identified.

so that this is equal to

$$x'\mathbb{E}[\varphi(1,U,V)|X=x,p\geq V].$$

By Assumption 1(i), we can reexpress this as

$$x'\mathbb{E}[\varphi(1,U,V)|p\geq V]=:x'\beta(1,p).$$

$\mathbb{E}[\varphi(1,U,V)|p \geq V]$ is a vector valued function of p which we will denote by $\beta(1,p)$ in the remainder. Since the left hand side of equation (9) is identified at points x and p of the support of X and P in the $D=1$ population, respectively, $\beta(1,p)$ is identified if we observe at least K linearly independent values of X for $D=1$ and $P=p$. $\beta(0,p)$ is defined in an analogous manner and a similar result holds for $D=0$ and $P=p$.

Starting from this, we show that the CASF is identified.[8] We call p a limit point of the support of P, if P has a continuous density in a neighborhood around p which is bounded away from zero. Notice that at $P=p$ derivatives of differentiable functions of P are identified.

Proposition 1 (Identification): *Assume that $\beta(0,p)$ and $\beta(1,p)$ are continuously differentiable with respect to p and that we observe at least K linearly independent values of X for $D=0, D=1$, and all values of P in a neighborhood around p (rank condition). Then, under Assumptions 1 and 2 the CASF is identified at $V=p$, where p is a limit point of the support of P, and given by*

$$G(0,x,v)=x'\left(\beta(0,p)-(1-p)\cdot\frac{\partial\beta(0,p)}{\partial p}\right)$$
$$G(1,x,v)=x'\left(\beta(1,p)+p\cdot\frac{\partial\beta(1,p)}{\partial p}\right).$$

[8] We state the result in a proposition which resembles Theorem 1 in Carneiro and Lee (2009). Following HV, they show nonparametric identification under weaker stochastic restrictions than the ones in Assumption 1, at the price of stronger support conditions that need to hold for their result. Only when they estimate the model they impose the restrictions in Assumption 1. We show the proof for two reasons. First, strictly speaking, our identification result is not implied by their Lemma 1, even though their proof is similar to ours. Second, our rank condition differs from theirs.

Proof. We show identification of $G(1, x, v)$. The proof for $G(0, x, v)$ is similar. Recall that we have normalized V to be uniformly distributed. By definition,

$$x'\mathbb{E}[\varphi(1, U, V)|p \geq V] = x'\beta(1, p).$$

From the normalization on V and Assumption 1(ii) it follows that

$$x' \int_0^p \int_{-\infty}^{\infty} \varphi(1, u, v)\ \mu(du)\ dv/p = x'\beta(1, p), \tag{10}$$

where $\mu(du)$ is the marginal probability measure of u. Multiplying both sides by p gives

$$x' \int_0^p \int_{-\infty}^{\infty} \varphi(1, u, v)\ \mu(du)\ dv = p \cdot x'\beta(1, p)$$

and differentiating both sides with respect to p using Leibniz' rule reveals that

$$x' \int_{-\infty}^{\infty} \varphi(1, u, p)\ \mu(du) = x'\beta(1, p) + p \cdot x' \frac{\partial\beta(1, p)}{\partial p}.$$

If p is a limit point of the support of P then the rank condition implies that $\beta(1, p)$ and $\partial\beta(1, p)/\partial p$ are identified at $P = p$. The left hand side is the object of interest. □

Finally, notice that identification relies on the monotonicity of D in P, which is implied by the selection model and allows us to formulate equation (10).[9]

2.3. Estimation

We have established that under the conditions of Proposition 1

$$\mathbb{E}[Y|D = d, P = p, X = x] = x'\beta(d, p)\ ,\ \ d \in \{0, 1\},$$

[9] See Klein (forthcoming) for a discussion and an analysis of the case in which monotonicity does not hold, but is wrongly assumed.

where $\beta(d,p)$ is a coefficient vector that is a function of the observable D, and P, which can be estimated.[10] We parametrically estimate P using a logit model. We assume that the coefficient functions are bounded and have bounded second derivatives. This allows us to estimate them by local linear smoothing.[11] This estimation procedure is usually motivated by a Taylor expansion of the coefficient function in $\tilde{p}$ about $\tilde{p} = p$ which yields

$$\beta_k(d,\tilde{p}) = \beta_k(d,p) + \frac{\partial\beta_k(d,p)}{\partial p} \cdot (\tilde{p} - p) + \frac{1}{2}\frac{\partial^2\beta_k(d,\bar{p})}{\partial p^2} \cdot (\tilde{p} - p)^2,$$

where $\bar{p}$ is a point between p and $\tilde{p}$. We select all observations with $D = d$ and index them by i, $i = 1, \ldots, n$. The estimates of $\beta(d,p)$ and $\partial\beta(d,p)/\partial p$ are given by

$$(11) \qquad \begin{pmatrix} \widehat{\beta(d,p)} \\ \partial\widehat{\beta(d,p)}/\partial p \end{pmatrix} = \arg\min_{a,b} \left\{ \sum_{i=1}^{n} K\left(\frac{p_i - p}{h}\right) \cdot \left(y_i - \begin{bmatrix} x_i \\ (p_i - p)\cdot x_i \end{bmatrix}' \begin{pmatrix} a \\ b \end{pmatrix} \right)^2 \right\},$$

where $K(\cdot)$ is a kernel function with the usual properties and h is the bandwidth.[12]

From these estimates of $\beta(d,p)$ and $\partial\beta(d,p)/\partial p$, which we provide with hats in the remainder, we calculate the vector of conditional average *ceteris paribus* effects, equation (5), and the

[10] This is a version of the varying coefficient model which was suggested by Cleveland, Grosse, and Shyu (1991) and Hastie and Tibshirani (1993).

[11] This assumption is stronger than what is required for identification. A sufficient condition for this to hold is that the second derivative of $\mathbb{E}[\varphi(D,U,V)|D = d, V = v]$ with respect to V is bounded for $d = 0$ and $d = 1$. Concerning the properties of the estimator see e.g. Fan and Zhang (1999) and Xia and Li (1999) for details as well as a proof of consistency and results on rates of convergence.

[12] Write $\tilde{y}_i = \sqrt{K((p_i - p)/h)} \cdot y_i$ and $\tilde{x}_i = \sqrt{K((p_i - p)/h)} \cdot (x_i', (p_i - p)x_i')'$. Then,

$$\begin{pmatrix} \widehat{\beta(d,p)} \\ \partial\widehat{\beta(d,p)}/\partial p \end{pmatrix} = \left(\sum_{i=1}^{n} \tilde{x}_i\tilde{x}_i'\right)^{-1} \left(\sum_{i=1}^{n} \tilde{x}_i\tilde{y}_i\right).$$

Following Fan (1992) we add a matrix with elements equal to 0.001 to the matrix $\sum_{i=1}^{n} \tilde{x}_i\tilde{x}_i'$ to ensure that it can be inverted. We use an Epanechnikov kernel and estimate the coefficient vectors at 101 grid points between 0 and 1. The bandwidths are chosen using a leave-one-out cross validation procedure. Figure 5 in the Appendix shows, separately for $D = 0$ and $D = 1$, the sample mean integrated squared error plotted against the bandwidth. It decreases until a value of the bandwidth of 1.2 for $D = 0$ and flat thereafter, so we use 1.2 as the bandwidth to allow for maximum flexibility. For $D = 1$ it is minimal at 0.8, so we use 0.8 as the bandwidth. In Proposition 1, we require the rank to be K. Here, we also use interaction terms between P and X for the estimator in (11), and therefore we require it to be $2K$. This rank condition holds in our data, i.e. the weighted $n \times 2K$ matrix of explanatory variables and interaction terms is of rank $2K$ at all evaluation points p.

CASF using the formulae in Proposition 1. From these we calculate values of other identifiable features of interest and simulate average effects.

Since fitted values p_i were parametrically estimated in a first step we do not expect them to have an impact on the distribution of $\widehat{\beta(d,p)}$ and $\widehat{\partial\beta(d,p)}/\partial p$ in a first order asymptotic sense. However, we obtain confidence intervals, accounting for the first step estimation error, from $1,000$ bootstrap replications. In our application they are wider than bootstrapped confidence intervals that do not account for the first step estimation error. We also account for simulation error if simulations are undertaken.

2.4. Discussion

In this section we briefly discuss the econometric approach taken here. There are two key advantages. First, functional form restrictions are mild. X could include approximating functions in such a way that the number of approximating functions grows with the sample size. Then, following Newey (1997), equation (1) could be interpreted as a series approximation to a general nonseparable structural equation $Y = g(X, D, U, V)$.

Second, the estimator that is proposed here allows for, and is able to recover, richer selection patterns than 2SLS, matching, and the local instrumental variables estimator for the additive model that is implemented e.g. by Carneiro and Lee (2009).[13] The additive model allows for selection based on the return, but imposes that the selection pattern does not depend on X. If we express X as $(1, X'_{-1})'$ and $\varphi(\cdot,\cdot,\cdot)$ as $(\varphi_1(\cdot,\cdot,\cdot), \varphi_{-1}(\cdot,\cdot,\cdot)')'$, equation (1) can be written as

$$Y = \varphi_1(D, U, V) + X'_{-1}\varphi_{-1}(D, U, V),$$

[13] Heckman and Vytlacil (1998), Heckman, Urzua, and Vytlacil (2006) and Wooldridge (2007) point out that 2SLS requires that the difference in the coefficients, $\varphi(1, U, V) - \varphi(0, U, V)$, is not correlated with D and Z. This precludes selection that is related to the return to a college degree. Matching estimators require that conditional on a set of observed variables, say $\overline{X}$, the difference in the coefficients is mean independent of D (Rosenbaum and Rubin, 1983), i.e.

$$\mathbb{E}[\varphi(1, U, V) - \varphi(0, U, V)|\overline{X}, D] = \mathbb{E}[\varphi(1, U, V) - \varphi(0, U, V)|\overline{X}]$$

if we maintain the functional form restrictions. Hence, matching estimators cannot be used if there is selection based on the return conditional on $\overline{X}$.

and the additive model as

$$Y = \mu(D, U, V) + X'_{-1}\gamma(D, U). \tag{12}$$

The sorting pattern is characterized by the dependence of the MTE on V, and this dependence is unrelated to X for the additive model.[14] In Section 4, we examine whether the additive model is consistent with our data.

The estimator that is proposed in this paper requires that there is a continuous variable in X or Z. 2SLS does not require continuous variation in X or Z because the assumption of uncorrelatedness between the effect and the endogenous variable can be exploited instead (Heckman and Vytlacil, 1998).

Interestingly, both the nonparametric identification result in HV and the matching estimator do not require the conditioning variables in X and $\bar{X}$, respectively, to be exogenous. For the nonparametric identification result in HV, the addition of variables to X increases the likelihood that the assumption of independence between Z and (U, V) conditional on X holds. At the same time, however, it becomes less likely that there is continuous variation in Z given X that shifts the probability to observe $D = 1$ from 0 to 1, which is necessary unless one directly estimates local average treatment effects instead (Frölich, 2007). Also for matching, adding more variables to $\bar{X}$ makes it more likely that the conditional mean independence assumption holds. In both cases the conditioning variables can be thought of as predictors of wage levels. However, also with these techniques we can only recover the causal effect of X on wages and on the returns to a college degree if we assume that X is exogenous. Besides, under exogeneity of X we can rely on weaker assumptions on the support of Z given X, which turns out to be of key importance in our application because the instruments are discrete.

[14] The MTE in the additive model is $\mathbb{E}[\mu(1, U, v) - \mu(0, U, v)] + x'_{-1}\mathbb{E}[\gamma(1, U) - \gamma(0, U)]$ and hence the effect of a change in v is not related to x.

3. Data

The estimator is implemented using NCDS data from the U.K. The NCDS is conducted by the Centre for Longitudinal Studies at the Institute of Education in London. It is a longitudinal data set and keeps detailed records for all those living in the U.K. who were born between March 3 and 9, 1958. The data were first collected at birth in 1958, in 1965 (age 7), in 1969 (age 11), in 1974 (age 16), in 1981 (age 23), in 1991 (age 33), in 1999-2000 (age 41-42), in 2004-2005 (age 46-47), and in 2008-2009 (age 50-51). The NCDS has gathered data on child development from birth to early adolescence, as well as on child care, medical care, health, physique, school readiness, home environment, educational progress, parental involvement, cognitive and social growth, family relationships, economic activity, income, training, and housing.

In a related application, Blundell, Dearden, and Sianesi (2005) study these data using 2SLS, a control function estimator, and matching techniques. We use the same procedures to prepare the data for analysis. For a more detailed data description and variable definitions the reader is referred to their paper.

For the analysis we select working men for whom information on their highest educational degree is available.[15] Our core sample thus consists of $3,609$ observations, of which 646 (17.9%) did not complete their O-levels, 986 (27.3%) completed their O-levels, 960 (26.6%) did so for the A-levels, and $1,017$ (28.2%) completed college education.[16] We distinguish between college graduates ($D = 1$), who have completed some kind of higher education, and the remaining individuals ($D = 0$). Test scores are rescaled so that each of them lies between zero and one.

Following Mincer (1974), the outcome of interest is the log hourly wage in 1991, at the age of 33. The NCDS contains information on a number of family background variables such as the respective parents' ages, their years of education, whether the mother was working when the

[15] Information on the education is not available for non-working individuals. Out of $3,945$ individuals 270 (6.84%) are not working. For 66 of the remaining individuals information on the education is missing.

[16] We say that an individual completes his A-levels if he completed at least one A-level, which is generally obtained at the end of secondary school, see Blundell, Dearden, and Sianesi (2005) for details.

child was 16, as well as the number of siblings. Furthermore, we observe the occupation of the father when the child was 16, in particular whether he was an intermediate employee. This can be interpreted as a proxy for the social class.

As discussed in Section 2.4, variables in X need to be selected such that they are unrelated to U and V. This precludes the inclusion of indicators for secondary school type because certain secondary schools are more likely to be chosen by individuals who plan to attend college thereafter. Also test scores might be related to unobserved ability. For that reason we include only the math test score at the age of 7 in the set of covariates and consider it a proxy for purely analytical skills, which are unrelated to other types of unobserved ability such as assertiveness and social intelligence that affect wages at the age of 33 as well as the decision to attend college. We do not include the number of siblings because it might be related to unobserved ability through interaction of the child with his siblings, again thinking of unobserved ability as being related to traits such as assertiveness and social intelligence. However, we include the mother's years of education into X and consider it a proxy for the kind of education the child receives at home, *irrespective* of his unobserved ability.[17]

We discarded some additional variables such as the respective age of the parents when the child was 16, because the corresponding coefficient estimates were insignificant in a first stage regression of an indicator for college education on the full set of covariates and indicators for the mother's interest in the child's education when the child was 16.[18] In addition, these discarded variables had insignificant coefficient estimates in the wage equation when the efficient GMM estimator (to be described in Section 5) was implemented. Most of the indicators for both the father's occupation and region were dropped as well, which did not have a big effect on the remaining significant coefficient estimates.

We use indicators for the mother's interest in the child's education when aged 16 as instru-

[17] Currie and Moretti (2003) show that maternal education is endogenous in a regression that explains birth outcomes. Carneiro, Meghir, and Parey (2007) extend this finding to outcomes such as ability test scores and other types of outcomes for children and adolescents. However, to our knowledge, it has not been shown that maternal education is also endogenous in a wage equation.

[18] Choices were made using both linear regression and logit estimates.

	no college		college	
	mean	std.	mean	std.
log hourly wage at the age of 33	1.929	0.400	2.329	0.369
math ability at 7	0.446	0.277	0.584	0.293
missing	0.113	-	0.116	-
FAMILY BACKGROUND VARIABLES WHEN THE CHILD WAS 16				
mother's years of education	7.133	4.455	7.857	4.945
missing	0.269	-	0.256	-
father is intermediate employee	0.019	-	0.108	-
missing	0.108	-	0.101	-
MOTHER'S INTEREST IN THE EDUCATION OF THE CHILD WHEN THE CHILD WAS 16				
expects too much	0.020	-	0.023	-
very interested	0.190	-	0.462	-
some interest	0.221	-	0.163	-
little interest	0.141	-	0.025	-
number of obs.	2,592		1,017	

Summary statistics for our core sample of individuals who are working at the age of 33 and for whom information on the highest educational degree is available. Standard deviations for indicator variables are not shown.

Table 1: Summary statistics.

ments for the decision to attend college. The interest in the child's education is assessed by the child's head teacher. It is an objective assessment of the parent's behavior, because the head teacher is not asked to evaluate the appropriateness of the parents' interest in the education, but rather to describe it. We expect this variable to be measured accurately because the teacher usually knows the parents from personal meetings.[19] It is plausibly unrelated to a child's ability, as long as the importance parents attach to the child's education does not depend on the child's unobserved characteristics. This is an assumption that will be formally tested by conducting tests of overidentifying restrictions in Section 5.

Table 1 shows summary statistics for our sample. The statistics are shown separately for college graduates and college non-graduates. On average, college graduates have a higher math ability test score, more highly educated mothers, and they are more likely to have a father who

[19] It would be problematic if the assessment was made by an interviewer because he would have to make it based on the impression he gained in the interview. It would be even more troublesome if the parents were asked to answer this question themselves, because their answer would probably be related to their child's ability.

is an intermediate employee. In addition, individuals who went to college are more likely to have parents who were interested in their education.

There are missing values for some variables in the data. In the analysis we assume that they are missing at random, set the value of the respective variable to zero, and include indicators for missing information in the set of covariates.[20]

4. College Education and Wages in the U.K.

First stage estimates were obtained using a logit model and are not reported here.[21] Since most explanatory variables are indicator variables our specification is very close to the series logit specification implemented by Hirano, Imbens, and Ridder (2003).[22]

Figure 1 shows the sample distributions of the fitted values of P. For both $D = 0$ and $D = 1$ the support is almost equal to the full unit interval. Note that the distributions differ between $D = 0$ and $D = 1$. This illustrates that the variables in Z have explanatory power.

The identification result in Proposition 1 implies that in principle, instead of using the mother's interest in the child's education as an excluded variable, we can exploit the non-linearity of P in X for identification. To check whether this is possible here we obtained logit estimates, with and without the excluded variables in Z, obtained fitted values of P, and then regressed Y on X and P. With the excluded variables in Z the coefficient on P is 0.715 with a standard error of 0.065.[23] Without the excluded variables the coefficient estimate changes sign and is equal to -0.305 with a standard error of 0.340. This shows that identification off the

[20] The table shows that the probability a value is missing is about equal for college graduates and college non-graduates.

[21] They are reported in the Online Appendix to this paper. The coefficient estimates for our final specification confirm to expectations and are in line with the literature which takes a closer look at the channels through which parents' education is transmitted to the children, see Goldberger (1989) and Haveman and Wolfe (1995) for an overview and discussion.

[22] We tried several specifications with interaction terms but they were generally not significant. There are two variables, math ability at the age of 7 and the mother's years of education, which we treat as continuous. In the second stage, we will estimate coefficients on those variables, which are functions of V. Therefore, to keep the results nicely interpretable, we have not included higher order terms for those variables.

[23] Again standard errors are obtained from $1,000$ bootstrap replications and correct for the first stage estimation error. For comparison, the two stage least squares estimate is 0.781 with a standard error of 0.073.

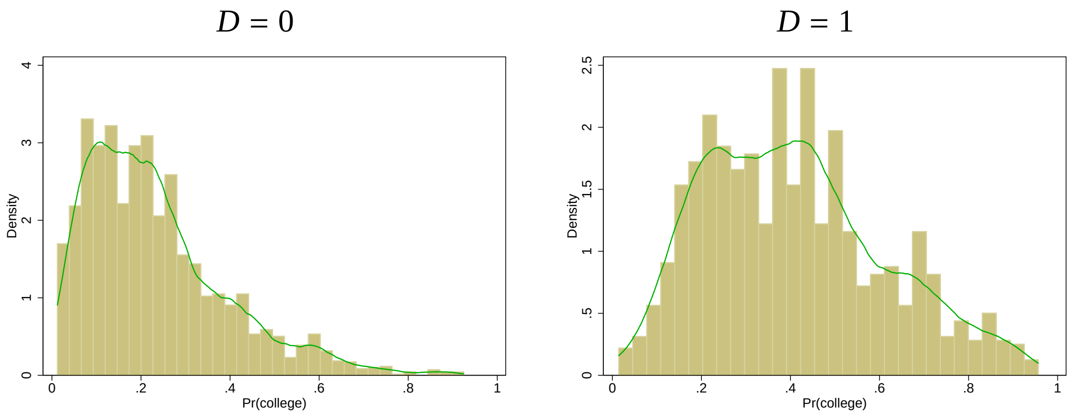


Figure 1: Sample distribution of the propensity score.

functional form is not a successful strategy here. The reason for this is that P is very close to being linear in X.[24] We now present the main results.

4.1. Wage Levels and Effects of Covariates

Figure 6 in the Appendix contains estimates of the conditional average *ceteris paribus* effects with respect to the variables in the vector X, (5), that are plotted against V, separately for $D = 0$ and $D = 1$. Figure 2 is an example and shows that for $D = 0$ the effect of the math test score at age 7 is positive and significant for low values of V. Recall that, according to the selection model, low values of V induce individuals to attend college. Thus, we should think of low values of V as representing high unobservable ability. Figure 6 shows that for individuals with high levels of unobserved ability and $D = 0$ wages increase in the math test score and the mother's years of education, and are higher when the father is an intermediate employee. When $D = 1$ the effect of those variables does not depend on V.

Table 5 (in the Appendix) presents estimates of the impacts of covariates on wages for $D = 0$ (left panel) and for $D = 1$ (right panel).[25] The fourth column contains the result of a test for a

[24] The R^2 of a linear regression of the fitted value of P on Z is 0.967 with the excluded variables in Z and 0.9787 without them.

[25] The additive model was estimated using a regression of the log hourly wage on a polynomial that is linear in X and quadratic in the first stage estimate of P, separately for $D = 0$ and $D = 1$. The order of the polynomial in P was chosen according to a leave-one-out cross-validation procedure. The average *ceteris paribus* effect in this

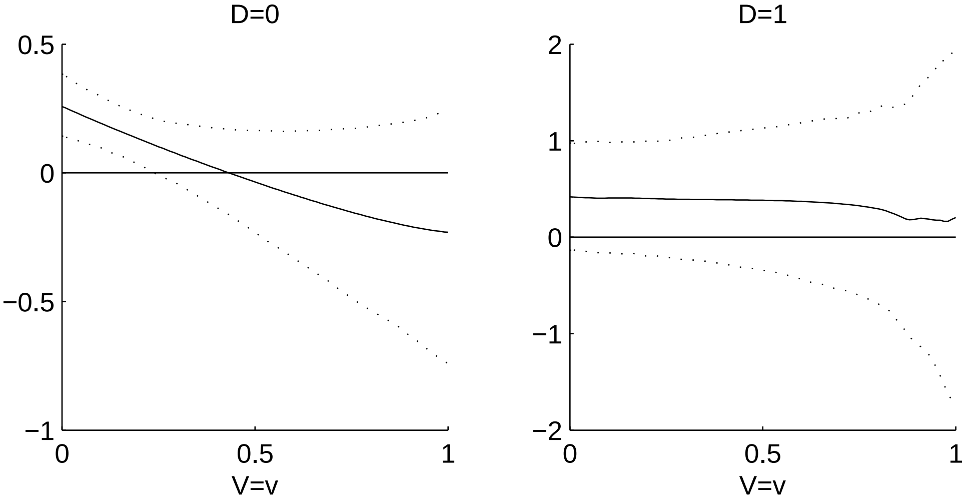


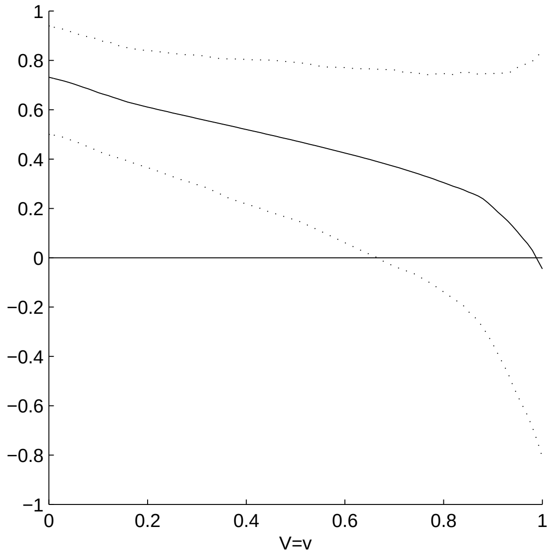
Figure 2: Average *ceteris paribus* effects of change in math score at the age of 7, estimates and 95% confidence intervals.

particular kind of unobserved heterogeneity. We say that unobserved heterogeneity is present whenever the impact of a component of X, including the constant, depends on V. Therefore, we test whether the linear approximation to the slope of the conditional average *ceteris paribus* effect with respect to V is zero.[26] We find evidence for a non-zero slope for the impact of the math ability test score, the indicator for Scotland and the constant for $D = 0$. Not surprisingly there is a bias in the coefficient estimate of the additive model whenever the test indicates that the type of heterogeneity we test for here is present. This shows that the additive model is too restrictive for our data.

Figure 7 in the Appendix contains an estimate of the CASF for a representative individual with median characteristics. Interestingly, the CASF increases in V for $D = 0$. This means that the kind of ability measured by V (for which low levels are associated with a higher likelihood of obtaining a college degree) is negatively related to wages if no college degree is obtained. Conversely, high V types do better when not obtaining a college degree than low V types do. This is compatible with the view that college graduates would in fact not do better on the labor

model is the coefficient on the respective variable in X.

[26] Here we face two sources of estimation error. First, the error that stems from estimating the conditional average *ceteris paribus* effect itself and second, the error from estimating the linear approximation to its slope.



Plotted for a man aged 33 who has a math ability test score of 0.5, whose mother has 9 years of education, whose father is not an intermediate employee, and who does not live in London, Scotland or Wales.

Figure 3: Marginal treatment effect for individual with median characteristics.

market than college non-graduates if one would have prevented them from attending college.

4.2. Selection into College

Figure 3 shows the MTE for the same representative individual with median characteristics. A negative slope implies that individuals with low values of V, i.e. high ability types, have a higher expected return to obtaining a college degree. Carneiro and Lee (2009, p. 201) point out that individuals base their selection into college education on their *comparative advantage* with respect to monetary benefits if the MTE is higher for those individuals who go to college, i.e. if the MTE decreases in V conditional on observables X.

Figures 3 and 7 were plotted for an individual with median characteristics. Since X varies across individuals, it is interesting to take a closer look at the dependence of the MTE on V when X varies across individuals. Variation in covariates induces variation in the slope of the MTE. Therefore, we estimated a linear approximation to the slope of the CASF and the MTE for every individual to investigate for how many individuals the comparative advantage hypothesis

	fraction	95% conf. int.	
level, no college	0.973	0.894	0.996
level, college	0.664	0.158	0.862
MTE	0.042	0.026	0.049

Table 2: Prevalence of positive dependence of CASF and MTE on V.

	estimate	ste.
ATE	0.366	0.129
ATT	0.538	0.079
ATU	0.303	0.154
OLS	0.282	0.016
ATT, matching estimate	0.274	0.030
GMM	0.773	0.107

For the matching estimate of the ATT as well as the OLS and GMM estimate the set of covariates is the same as for the estimates presented in Table 4. Standard errors for the first three estimates as well as the matching estimate were obtained from 1,000 bootstrap replications. Both OLS and GMM standard errors are analytic and robust to heteroskedasticity.

Table 3: Average treatment effects.

holds.

Table 2 contains the fractions of the population for which, respectively, the slope of the CASF and the MTE are positive. In order to obtain those numbers, linear approximations to the slope were estimated. The slope of the level is positive for 97% of the individuals if we assign $D = 0$ to all of them. If we assign $D = 1$ to everybody, it is positive for 66% of the individuals. Interestingly, the slope of the MTE is positive for only 4.2% of the individuals, indicating that the comparative advantage hypothesis holds for the remaining 95.8% of the individuals.[27]

If selection is based on V, it is interesting to calculate average returns for different subpop-

[27] This is in line with findings in previous studies including Willis and Rosen (1979) and Carneiro and Lee (2009). Those two studies impose additivity and therefore the comparative advantage hypothesis either holds for all individuals or for none. In Table 5, we observed that additivity does not hold and that coefficient estimates could therefore be biased. In light of this, our results are reassuring as they rule out this source of bias and nevertheless show that the comparative advantage hypothesis holds for almost all individuals. At this point, it is worth noting that this way of thinking about the comparative advantage hypothesis ignores nonmonetary costs and benefits. For example, it could well be that a college degree is additionally associated with nonpecuniary benefits such as the pleasure of being educated. Such *additional* returns are not addressed in this paper.

ulations. We first calculate ATE using equation (8), replacing x with the population mean of X. For ATT and ATU, we use the respective sample mean of X for $D = 1$ and $D = 0$ and simulate the distribution of V conditional on D, exploiting the structure of the selection model.[28] Respective confidence intervals account for the simulation error. In Table 3, our estimates are compared to those obtained from an ordinary least squares (OLS) regression, a matching estimate, and the efficient GMM estimate. Notably, and consistent with the finding that the MTE is negatively sloped for almost all individuals, we find that the ATT is higher than the ATU. The ATE is between the ATT and the ATU.[29]

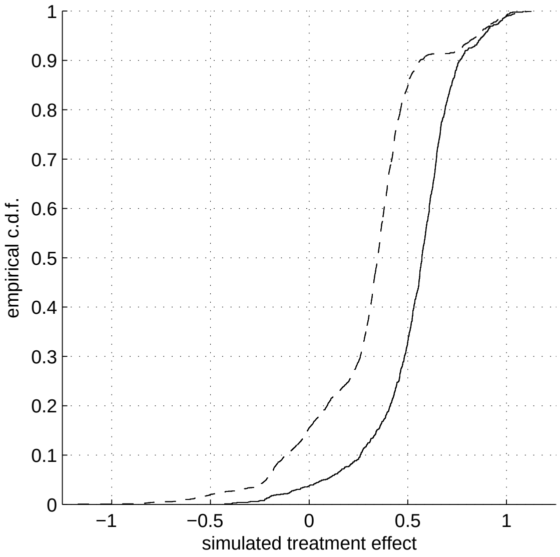
Not surprisingly, the OLS estimate is very close to the matching estimate since matching is built on the assumption that conditional on observables, D is independent of the error term in the outcome equation. The similarity between the OLS and the matching estimate indicates that the functional form assumption imposed by the OLS estimator is innocuous for our data. This could be because X contains many indicator variables. However, the fact that there is a difference between the matching estimate and our estimate of ATT (reported in Table 3) shows that the conditional independence assumption that matching is based on might be violated.

We find that both the OLS and the matching estimate are downward biased, indicating that the selection bias is negative. Controlling for covariates, the estimates are both roughly equal to the difference in the average wage for those individuals who are observed to have $D = 1$, averaging over low values of V according to equation (2), and the average wage for those individuals who are observed to have $D = 0$, averaging over high values of V. Figure 7 in the Appendix shows that for the representative individual, the wage depends positively on V when we assign $D = 0$. However, the wage is flat in V if we assign $D = 1$. Hence the downward bias.

Commonly, the GMM estimate is interpreted as estimating the average treatment effect for those individuals who are induced to attend college by the variables that are excluded from the

[28] For example, if we observe an individual with $D = 0$ and $P = p$, we draw values of V from a uniform distribution on $(p, 1]$.

[29] We also find this when we estimate the additive model, equation (12), and implement the matching estimator for ATE and ATU.



The left curve is the empirical c.d.f. of simulated treatment effects for college non-graduates and the right curve is the empirical c.d.f. for college graduates.

Figure 4: Distribution of simulated treatment effects.

outcome equation.[30] The GMM estimate is higher than those obtained from other estimation strategies.[31]

Next, we simulate the treatment effect for every individual. For this we draw 100 values of V, calculate the corresponding MTE, and then calculate the average MTE. If we observe $D = 0$, we draw values of V from a uniform distribution on $[p, 1]$, and if $D = 1$ we draw them from a uniform distribution on $[0, p]$, where, respectively, p is the fitted value of the probability to obtain a college degree. The idea behind this is that by observing D and P we can infer in which range V must lie. Since V is independent of P, it is uniformly distributed. Figure 4 shows that the distribution of simulated treatment effects for individuals who actually graduated from college first order dominates the distribution of simulated treatment effects for those who did not do so. However, the support overlaps, indicating that there are individuals who did not graduate from college and that would have benefitted more (in monetary terms) from obtaining

[30] See, e.g., the discussion in Blundell, Dearden, and Sianesi (2005) and Imbens and Angrist (1994) as well as Card (2001).

[31] In Section 5 we provide additional discussion and interpretation.

a college degree than others who are college graduates.

Finally, we examine the dependence of the selection pattern on observed characteristics more closely. It follows from equation (7) that the effect of covariates on the MTE is given by the difference in conditional average *ceteris paribus* effects between $D = 1$ and $D = 0$. Hence, it follows from Figure 6 that the dependence of the returns to college education on V is larger for individuals with low math test scores, less educated mothers, and fathers that are not intermediate employee. Within this well-defined group, a policy maker should target those individuals with high unobserved ability and strongly encourage them to attend college, assuming that the policy maker's objective is to allocate college education to those individuals with the highest expected returns.[32] Conversely, the screening effort within other groups (e.g. individuals with high math scores whose father is an intermediate employee) could be lower because those individuals' return to a college degree depends less on unobserved ability.[33]

5. Validity of Instruments

In this paper, indicators for the mother's interest in the education of the child (when the child was 16) serve as instruments for college education. These are valid instruments if the interest is unrelated to unobserved ability and at the same time related to the decision to attend college. In this section we assess the validity of the instruments by estimating standard Mincer (1974) type wage equations using the generalized method of moments (GMM). We carry out overidentifying restrictions tests and tests for joint significance of the excluded instruments. We do so for our preferred and potential alternative sets of instruments. The candidate instruments are indicators for the mother's and father's interest in the education of the child when the child was 7, 11, and 16, respectively.

To test whether the instruments are not weak or just mask variation in other individual

[32] See, e.g., Berger, Black, and Smith (2001) on profiling in the context of unemployment.

[33] Notice, however, that the test for heterogeneity in Table 5 shows that we cannot reject the null hypothesis that the effect of the mother's years of education and the father being an intermediate employee does not depend on V.

characteristics we additionally include in X a race indicator, math ability at age 11, verbal ability at age 7 and 11, indicators for secondary school type, additional family background variables, indicators for the occupation of the father, and a full set of region indicators.[34]

The sixth column of Table 4 presents efficient two-step GMM estimates of the effect of attending college on the log hourly wage at the age of 33. The set of excluded variables varies across the ten specifications. The first six specifications use indicators for all values of the interest in the child's education (and the ones for the respective missing value).[35] Specifications 7 through 10 use only the indicators for one category, respectively, as instruments.

F-statistics for the test of joint significance of the excluded instruments are presented in the fourth column of Table 4 and are calculated to detect the presence of weak instruments. Values above 10 are considered acceptable (Bound, Jaeger, and Baker, 1995; Staiger and Stock, 1997). The fifth column contains the p-values. Comparing specifications across Table 4 shows that adding more instruments never decreases the partial R^2 of the excluded instruments but results in a decreased value of the F statistic. Including all potential instrumental variables, specification 6, may result in biased estimates of the returns to college education because the corresponding F statistic is equal to 6.39, which is well below 10. By contrast, indicators for the mother's interest in the child's education when the child was 16 could be a good set of candidate instruments. The F statistic of 23.4 shows that these indicators are strongly related to D.

The last two columns of Table 4 display Hansen's J statistic and the corresponding p-value for the test of overidentifying restrictions. The null for this test is that the difference between observed and predicted wages is not correlated with the instrument.[36] It is rejected if the instruments are related to the error term or if different instruments identify different local average treatment effects, see Imbens and Angrist (1994) and Heckman, Urzua, and Vytlacil (2006, p. 392). In our case the null is rejected, at the 5% level, for specifications 2 and 3, suggesting

[34] Summary statistics for all variables but the region indicators can be found in the Online Appendix to this paper.

[35] There are four indicators for each parent and indicators for missing values. There were no missing values when the child was 16.

[36] It is often imposed that D has a non-random coefficient. Then, the null hypothesis is valid only if the usual exclusion restriction holds.

	instruments	num. iv	part. R^2	F	p	college	ste.	J	p
	INTEREST IN EDUCATION, ALL VALUES								
1	mother at 16	4	0.026	23.400	0.000	0.773	0.107	5.103	0.164
2	mother at 11 and 16	9	0.036	14.690	0.000	0.629	0.085	18.921	0.015
3	mother at 7, 11 and 16	14	0.037	9.820	0.000	0.625	0.083	26.197	0.016
4	mother and father at 16	8	0.028	12.600	0.000	0.810	0.105	6.243	0.512
5	mother and father at 11 and 16	17	0.042	9.140	0.000	0.658	0.079	24.078	0.088
6	mother and father at 7, 11 and 16	26	0.046	6.390	0.000	0.639	0.077	33.436	0.121
	INTEREST BY MOTHER AND FATHER AT 7, 11 AND 16								
7	only indicator for little interest	6	0.009	9.350	0.000	1.138	0.188	2.216	0.819
8	only indicator for some interest	6	0.008	5.320	0.000	0.434	0.165	2.845	0.724
9	only indicator for very interested	6	0.038	20.130	0.000	0.586	0.083	8.512	0.130
10	only indicator for overly concerned	6	0.005	3.130	0.000	0.537	0.227	2.950	0.708

All test statistics and standard errors are robust to heteroskedasticity.

Table 4: Performance of alternative sets of instruments.

that returns might be heterogeneous or that the exclusion restriction is violated for the instruments.

Specifications 7 through 10 provide evidence that returns are heterogeneous because they show that the estimate of the returns to college education depends on the indicators that are employed as instruments. Specification 9, e.g., estimates the return for those individuals whose decision to attend college changes if the parents become very interested in their education. Here, all indicators should identify the average effect for the same group. This is for instance because the group of individuals whose decision to attend college is changed by their father's taking an interest in their education when the individual concerned was 16 should be (roughly) the same group as the group of individuals whose decision to attend college is affected by the mother's interest in their education when aged 11. Therefore, the overidentifying restrictions test should only reject the null if the exclusion restriction is violated. This is not the case in specifications 7 though 10, so there is no evidence that the exclusion restriction does not hold.

To summarize, we conclude from Table 4 that the overidentifying restrictions tests yield evidence in favor of effect heterogeneity that is related to the decision to attend college, and in favor of the assumption that the instruments can be excluded, i.e. that they are unrelated to unobserved ability.[37] Results show that the instruments are strongly correlated with college attendance if not all indicators are selected. In Section 4, we therefore use indicator variables for the mother's interest in the child's education when aged 16 as excluded variables because for this set of variables the F statistic is particularly high.

[37] Heckman, Urzua, and Vytlacil (2006, p. 397) point out that one can directly test for unobserved dependence between the return to college and the decision to attend college by checking whether the expected wage conditional on X and P is linear in P. So we first fit P using a logit model and then regress Y on P, P^2 and X. The coefficient on P^2 is -1.034 with a standard error of 0.209. This standard error is obtained from $1,000$ bootstrap replications and accounts for the first stage estimation error. We conclude that the GMM estimates presented in Table 4 are indeed not estimates of the population average return to college, but rather estimates of local average returns for different subgroups, supporting our conclusion that the overidentifying restrictions tests yield evidence in favor of effect heterogeneity, rather than against the validity of the exclusion restriction.

6. Concluding Remarks

In this paper, we propose and implement a semiparametric local instrumental variables estimator and use it to characterize the unobserved dependence between the monetary return to college education and selection into college in the U.K. We relate this dependence to observable characteristics of the individuals. To accomplish this, the estimator requires that an exclusion restriction holds unconditionally, that covariates are exogenous, and that there is continuous variation in an instrument or a covariate.

Our empirical results indicate that sorting into college is based on the comparative advantage for almost all individuals. *Therefore*, the average return to college education for college graduates is larger than the average return for college non-graduates. We find that the dependence of selection into college and returns to college education is strongest for individuals with low math test scores at age 7, individuals with less educated mothers, and for working-class individuals.

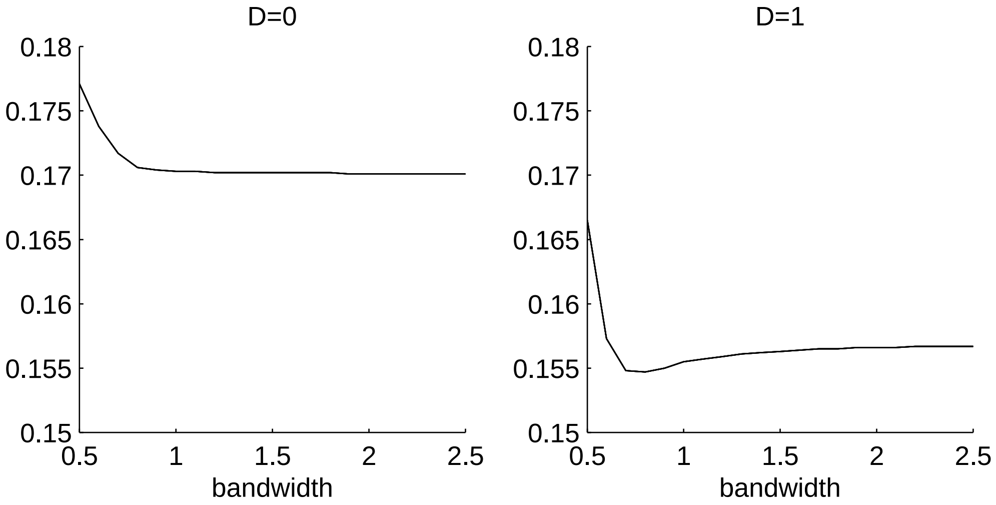
This knowledge is likely to be of value to policy makers who often design institutions in such a way that eligibility, e.g. for a subsidy, is related to observed individual characteristics. We find that returns to college education highly depend on unobserved ability for working-class individuals in the U.K. Based on this knowledge a policy maker may want to encourage more high ability working-class individuals to attend college. He could do so by offering a subsidy to all working-class individuals.[38] Our results imply that among the individuals who would not attend college without the subsidy, this subsidy will change the decision for the ones with the highest return, which is exactly the target group.

The methods developed in this paper could be applied in various other contexts. For example, they could be used to study how selection into a labor market program is related to the effect of the program, possibly as a function of observable characteristics and labor market history. Results could then be used to re-design eligibility rules in a such a way that the participation is

[38] In practice, it might not be possible to define eligibility according to social class. However, a good proxy for this could be family income.

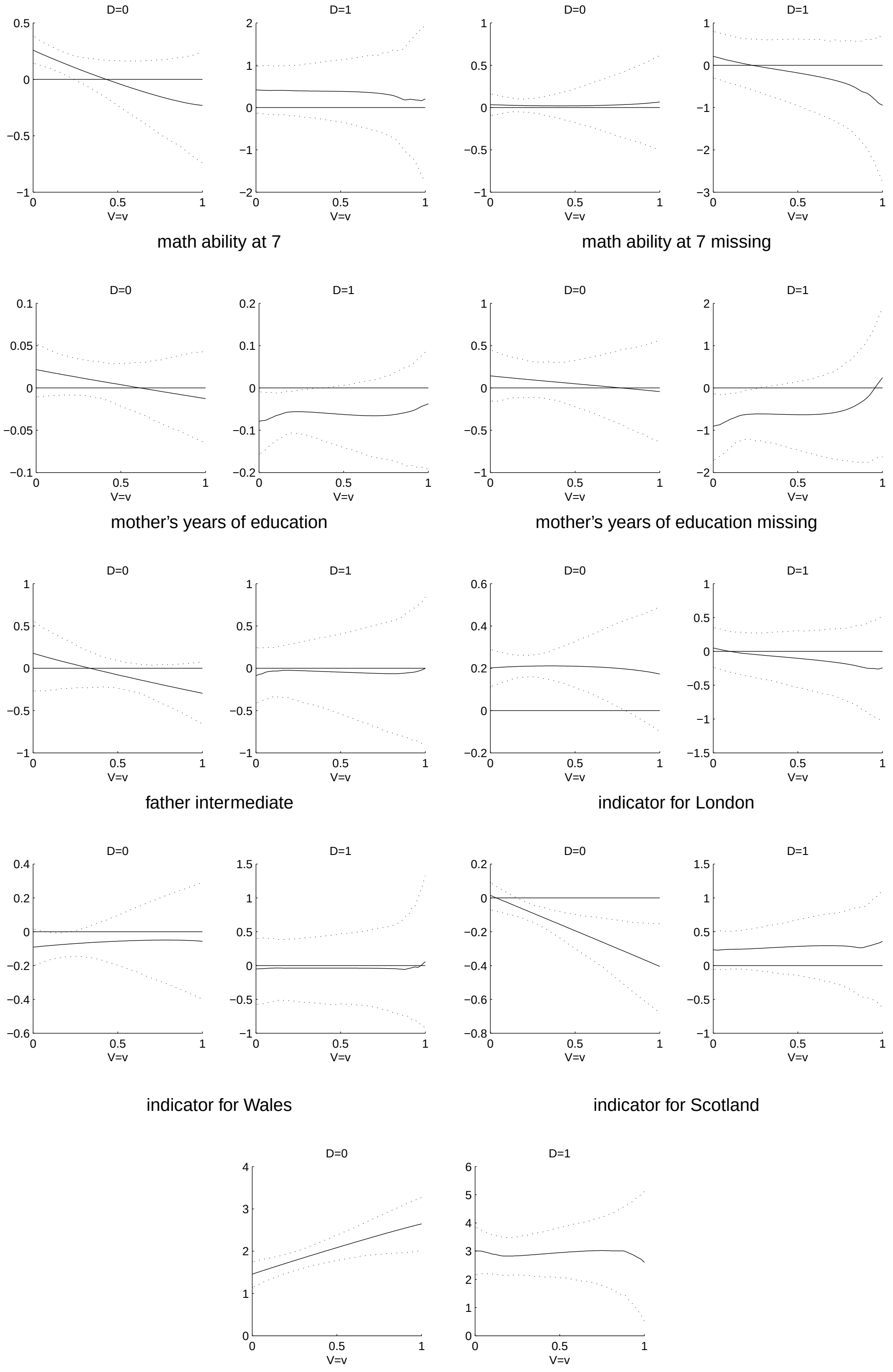
allocated more efficiently. Since our estimation procedure does not require the excluded variables to be continuous, eligibility rules that were in place when the data were collected could possibly be used as a source of exogenous variation.

Appendix: Additional Tables and Figures



These figures show, separately for $D = 0$ and $D = 1$, estimates of the mean integrated squared error as a function of the bandwidth. Obtained using a leave-one-out cross validation procedure.

Figure 5: Cross validation.



math ability at 7

math ability at 7 missing

mother's years of education

mother's years of education missing

father intermediate

indicator for London

indicator for Wales

indicator for Scotland

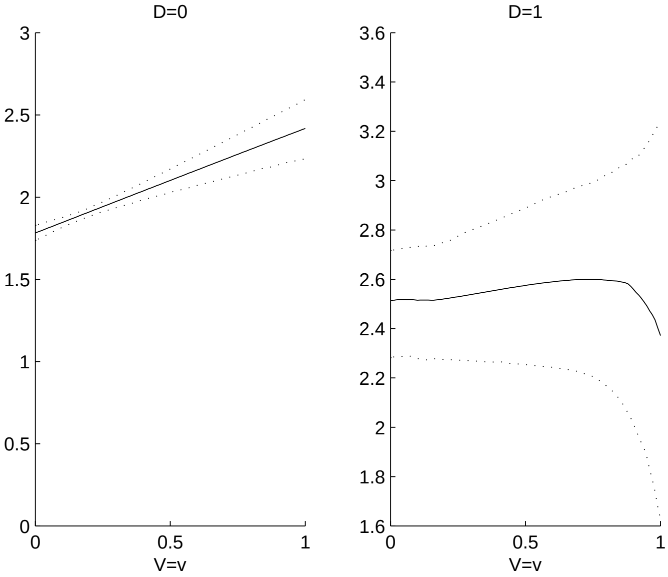
constant

Figure 6: Conditional average *ceteris paribus* effects, estimates and 95% confidence intervals.

	no college				college			
	average *ceteris paribus* effect			test for	average *ceteris paribus* effect			test for
	$\mathbb{E}[\varphi(0, U, V)]$	additive	bias	het.	$\mathbb{E}[\varphi(1, U, V)]$	additive	bias	het.
math ability at 7	-0.020	0.114***	0.135	-0.500*	0.343	0.095**	-0.248	-0.227
	(0.104)	(0.042)	(0.084)	(0.275)	(0.390)	(0.042)	(0.386)	(0.755)
math ability missing	0.030	0.035	0.005	0.024	-0.226	0.122***	0.348	-0.949
	(0.110)	(0.029)	(0.097)	(0.323)	(0.396)	(0.029)	(0.391)	(0.685)
mother's years of education	0.004	0.010	0.006	-0.034	-0.060	0.007	0.067*	0.013
	(0.012)	(0.022)	(0.006)	(0.038)	(0.037)	(0.022)	(0.040)	(0.071)
mother's years of education missing	0.049	0.063**	0.014	-0.183	-0.551	0.107***	0.658	0.607
	(0.136)	(0.025)	(0.078)	(0.407)	(0.427)	(0.025)	(0.453)	(0.817)
father intermediate	-0.071	-0.061*	0.010	-0.467	-0.044	-0.003	0.042	-0.014
	(0.079)	(0.032)	(0.043)	(0.354)	(0.239)	(0.032)	(0.223)	(0.387)
indicator for London	0.201***	0.211***	0.011	-0.024	-0.109	0.063	0.171	-0.287
	(0.057)	(0.045)	(0.047)	(0.182)	(0.210)	(0.045)	(0.205)	(0.310)
indicator for Wales	-0.061	-0.074	-0.013	0.038	-0.036	-0.038	-0.002	0.013
	(0.074)	(0.052)	(0.060)	(0.213)	(0.260)	(0.052)	(0.257)	(0.440)
indicator for Scotland	-0.193***	-0.078	0.114**	-0.418***	0.270	0.006	-0.265	0.076
	(0.051)	(0.068)	(0.045)	(0.158)	(0.213)	(0.068)	(0.210)	(0.373)
constant	2.054***	1.595***	-0.459***	1.193***	2.898***	2.034***	-0.864*	0.036
	(0.150)	(0.130)	(0.108)	(0.407)	(0.470)	(0.130)	(0.512)	(0.989)

Standard errors in parenthesis were calculated from $1,000$ bootstrap replications and take the first stage estimation error into account. 1 through 3 stars indicates significance at the 10%, 5% and 1% level, respectively.

Table 5: Average *ceteris paribus* effects and test for heterogeneity.



Plotted for man aged 33 who has a math ability test score of 0.5, whose mother has 9 years of education, whose father is not an intermediate employee, and who does not live in London, Scotland or Wales.

Figure 7: Conditional average structural function for individual with median characteristics.

Acknowledgements

This paper is based on the second chapter of my Ph.D. thesis which I defended in July 2006 at the University of Mannheim. I am especially grateful to Erich Battistin, Richard Blundell, Pedro Carneiro, Andrew Chesher, Johannes Gernandt, Stefan Hoderlein, Enno Mammen, Melanie Schienle, and Edward Vytlacil for stimulating discussions and helpful comments on earlier versions of this paper. Moreover, I would like to thank Richard Blundell, Lorraine Dearden, as well as Leslie McGranahan for sharing their data. I would like to thank UCL for its hospitality during the academic year 2003/4, the European Commission for financial support through the Marie Curie program, and the Deutsche Forschungsgemeinschaft for financial support through SFB/TR 15. Furthermore, I would like to thank the audiences of various conference and seminar presentations for valuable comments. Finally, I would like to thank the three referees and the editor for their detailed comments and suggestions.

References

Battistin, E., and E. Rettore (2008): “Ineligible and Eligible Non-Participants as a Double Comparison Group in Regression-Discontinuity Designs,” *Journal of Econometrics*, 142(2), 715–730.

Berger, M. C., D. Black, and J. Smith (2001): “Evaluating Profiling as a Means of Allocating Government Services,” in *Econometric Evaluation of Labour Market Policies*, ed. by M. Lechner, and F. Pfeiffer, pp. 59–84. Physica, Heidelberg, Germany.

Björklund, A., and R. Moffitt (1987): “The Estimation of Wage Gains and Welfare Gains in Self-Selection Models,” *Review of Economics and Statistics*, 69(1), 42–49.

Blundell, R., L. Dearden, and B. Sianesi (2005): “Evaluating the Impact of Education on Earnings in the UK: Models, Methods and Results from the NCDS,” *Journal of the Royal Statistical Society, Series A*, 168(3), 473–512.

Blundell, R., and J. L. Powell (2003): “Endogeneity in Nonparametric and Semiparametric Regression Models,” in *Advances in Economics and Econometrics*, ed. by L. Hansen, Amsterdam. North Holland.

Bound, J., D. A. Jaeger, and R. M. Baker (1995): “Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak,” *Journal of the American Statistical Association*, 90(430), 443–450.

Card, D. (2001): “Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems,” *Econometrica*, 69(5), 1127–1160.

Carneiro, P., and S. Lee (2009): “Estimating Distributions of Potential Outcomes using Local Instrumental Variables with an Application to Changes in College Enrollment and Wage Inequality,” *Journal of Econometrics*, 149(2), 191–208.

Carneiro, P., C. Meghir, and M. Parey (2007): “Maternal Education, Home Environments and the Development of Children and Adolescents,” CEPR Discussion Paper No. 6505, CEPR, London, U.K.

Cleveland, W. S., E. Grosse, and W. M. Shyu (1991): “Local Regression Models,” in *Statistical Models in S*, ed. by J. M. Chambers, and T. J. Hastie, pp. 309–376. Wadsworth/Brooks-Cole, Pacific Grove, CA.

Currie, J., and E. Moretti (2003): “Mothers Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings,” *Quarterly Journal of Economics*, 118(4), 1495–1532.

Fan, J. (1992): “Design-adaptive Nonparametric Regression,” *Journal of the American Statistical Association*, 87(420), 998–1004.

Fan, J., and W. Zhang (1999): “Statistical Estimation in Varying Coefficient Models,” *Annals of Statistics*, 27(5), 1491–1518.

Frölich, M. (2007): “Nonparametric IV Estimation of Local Average Treatment Effects with Covariates,” *Journal of Econometrics*, 139(1), 35–75.

Goldberger, A. S. (1989): “Economic and Mechanical Models of Intergenerational Transmission,” *American Economic Review*, 79(3), 504–513.

Griliches, Z. (1977): “Estimating the Returns to Schooling: Some Econometric Problems,” *Econometrica*, 45(1), 1–22.

Hastie, T., and R. Tibshirani (1993): “Varying-Coefficient Models,” *Journal of the Royal Statistical Society, Series B (Methodological)*, 55(4), 757–796.

Haveman, R., and B. Wolfe (1995): “The Determinants of Children’s Attainments: A Review of Methods and Findings,” *Journal of Economic Literature*, 33(4), 1829–1878.

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

Heckman, J. J., and E. J. Vytlacil (1998): “Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling," *Journal of Human Resources*, 33(4), 974–987.

——— (1999): "Local Instrumental Variables and Latent Variables Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730–4734.

——— (2000): "The Relationship between Treatment Parameters within a Latent Variable Framework," *Economics Letters*, 66(1), 33–39.

——— (2001): "Local Instrumental Variables," in *Nonlinear Statistical Modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya*, ed. by C. Hsiao, K. Morimune, and J. Powell, pp. 1–46. Cambridge University Press, Cambridge.

——— (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," *Econometrica*, 73(3), 669–738.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

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

Klein, T. J. (forthcoming): "Heterogeneous Treatment Effects: Instrumental Variables without Monotonicity?," *Journal of Econometrics*.

Mincer, J. (1974): *Schooling, Experience, and Earnings*. Columbia University Press, New York.

Newey, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimates," *Journal of Econometrics*, 79, 147–168.

Rosenbaum, P. R., and D. B. Rubin (1983): "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70(1), 41–55.

Solmon, L. C., and P. J. Taubman (1973): *Does College Matter? Some Evidence of the Impacts of Higher Education*. Academic Press, New York.

Staiger, D., and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65(3), 557–586.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70(1), 331–341.

Willis, R. J., and S. Rosen (1979): "Education and Self-Selection," *Journal of Political Economy*, 87(5, Part 2: Education and Income Distribution), S7–S36.

Wooldridge, J. M. (2007): "Instrumental Variables Estimation of the Average Treatment Effect in Correlated Random Coefficient Models," in *Advances in Econometrics: Modeling and Evaluating Treatment Effects in Econometrics*, ed. by D. Millimet, J. Smith, and E. Vytlacil, vol. 21. Elsevier, Amsterdam.

Xia, Y., and W. K. Li (1999): "On the Estimation and Testing of Functional-Coefficient Linear Models," *Statistica Sinica*, 9, 735–757.