Separating risk in education from heterogeneity: a semiparametric approach

Mazza, J.; van Ophem, H.

Publication date
2010
Document Version
Submitted manuscript

Separating risk in education from heterogeneity: a semiparametric approach.

Jacopo Mazza *
and
Hans van Ophem
Department of General Economics
Universiteit van Amsterdam
Roetersstraat 11 1018WB Amsterdam, The Netherlands

November 15, 2010

Abstract

Returns to education are variable within the same educational group. If uncertain payoffs are a concern for individuals when selecting education, wage variance is the resultant of unobserved heterogeneity and pure uncertainty. The first element is known to the individual, but unknown to the researcher. If individuals exploit private information to select their level of education the variance observed in the data will overestimate the real magnitude of education uncertainty and the impact that risk has on educational decisions. In this paper we will apply a "semi-non parametric" estimator of an unknown density to tackle selectivity issues. This method does not rely on the joint normality of errors in the selection and primary equation and is robust to misspecification of the residual distribution. Results suggest that pure risk tend to increase with education. Private information accounts for a very minor share of total wage variance observed in the data.

Keywords: Wage inequality; Wage uncertainty; Unobserved heterogeneity; Variance differential; Selection bias, Return to education, Semiparametric estimation, Two-step estimation

JEL classification: C14; C24; D81; J31

*Corresponding author: j.mazza@uva.nl. The results presented in this paper are preliminary, please do not quote. All data and computer programs are available on request.

1 Introduction

Empirical evidence shows that earnings inequality has increased in the US in the second half of the past century (Joshua Angrist, 2006; Katz and Autor, 1999) the phenomenon involves both between and within educational group inequality (Autor et al., 2006; Acemoglu, 1999) and has attracted attention on the link between wage variance and schooling. If wage variance increases with schooling level and individuals are risk-averse, the increasing differences might reflect compensation for risk.

The identification of causal effect of risk on education attainments is complicated by selection biases (Acemoglu, 2002). Observed wages inequality are calculated upon truncated wages distributions. The truncation is an effect of private information: individuals possess information about their tastes and inclinations and might use these information to select the level of education assuring the best risk/pay-off profile. Researchers, ignoring these factors, have to rely solely on the revealed choices and outcomes confusing total observed variance with risk. In our terminology wage uncertainty or *risk* is the part of wage variability which is not foreseeable by the individual even with the superior knowledge that he possesses about himself. *Unobserved heterogeneity,* instead, is that part of wage variability that depends on factors known to the individual, but not observable by the econometrician with the available data. The inability to disentangle risk and heterogeneity will cause an overestimation of real risk and, in turn, an underestimation of risk premium offered in the labor market in the form of higher salaries.

The two main goals of this paper are: a) to establish a causal relation between education and wage inequality; b) to estimate a proper measure of *risk* that various educational categories entail disentangling it from *unobserved heterogeneity*. The literature on this issue is quite scarce. Chen (2008) tackles issues of selectivity and unobserved heterogeneity taking dispersion of wages as outcome of interests implementing a standard parametric selection model with instrumental variables as originally proposed by Heckman (1979) and extending it to the polychotomous case. We apply the same formalization of her theoretical model, but we depart from it in an essential issue: we do not impose any restriction on the distribution of disturbances in the primary equation.

Parametric methods have undergone increasing criticism for imposing excessive restriction over the model (Vella, 1998; Goldberger, 1983; Moretti, 2000) the main one being that incorrect specification of joint normality of residuals of wage and selection equation leads to inconsistent estimates. These criticisms spurred a growing literature (Ahn and Powell, 1993; Cosslett, 1991; Newey, 2009; Robinson, 1988) proposing a series of different semi-parametric estimators.

In this paper we apply a two stage semi nonparametric estimation method. Similarly to the two step Heckman method we first estimate a selection equation from which we build four selection correction terms, one for each category, and include them as additional regressors in the outcome equation to re-establish zero conditional mean on the error term. We do so exploiting a distribution-free semiparametric estimator developed by Gallant and Nychka (1987) in the first stage and a procedure proposed by Cosslett (1983)(1991) in the second. The estimates so obtained are robust to miss-specification of residuals distribution function and allow us to minimize the distribution assumptions required to obtain identification.

To our knowledge, this is the first paper dealing with both issues of self-selection and unobserved heterogeneity semi parametrically. Chen (2008) deals with both issues, but strictly parametrically, while the semiparametric methods proposed in the literature tackle either self-selection or unobserved

heterogeneity. Chen and Khan (2007) use kernel weighting schemes and symmetry conditions on the joint distribution of outcome and selection equation errors obtaining estimates for wage inequality among college graduates corrected for selection, but do not examine the impact of unobserved heterogeneity. Abadie (2002) proposes a method based on instrumental variables concerned with estimation of causal effects on the entire distribution and not only mean effects, while Abadie et al. (2002) propose a generalization of the quantile treatment effect estimator when selection into treatment is endogenous with the first step estimated non parametrically, but both these works are not interested in distinguishing between intrinsic heterogeneity and uncertain shocks.

Our empirical analysis will be based on the well known National Longitudinal Survey of Youth (NLSY). We first obtain estimates of potential wages inequality robust to selection and truncation biases with the aforementioned methodology, then we distinguish between the two components of inequality: heterogeneity and risk semi parametrically.

Results suggest that observed inequality, potential inequality and pure risk increase with educational attainments while unobserved heterogeneity is almost non-existent or at lest it is not acted upon when selecting the desired level of education.

2 Econometric specification

If our goal is that of identifying the magnitude of risk in each education and, eventually, its impact on individual choices and wages, two obstacles might intervene: a) observed wage inequality is not the correct quantification of real wage inequality due to self-selection; b) even if we are able to correct for self selection the so corrected wage inequality would pool real risk together with unobserved heterogeneity. The following section describes the model stemming, in its general structure, from Chen (2008). The procedure is divided in two separate parts. First wage inequality corrected for self-selection is identified, then unobserved heterogeneity is separated from risk. The model exploits the panel structure of NLSY to control for time invariants individual fixed effects which are not observable as, for example, taste for education. Being individual effects constant across time using a fixed effect model we can difference them out from the wage equation and obtain consistent estimation.

2.1 The model

The model presented in Chen (2008) is an extension of a classical Roy model (Roy, 1951) with four possible choices, in which the choice of "occupation" is substituted with a choice over which educational level to acquire. In this model individuals (i) have four possible schooling choices (s_i): no high school diploma ($s_i = 0$); high school diploma ($s_i = 1$); some college ($s_i = 2$); and four years of college or more ($s_i = 3$). Individuals are observed for T periods; each time period is indexed by subscript (t). The total number of individuals in the sample will be indicated by N.

For each individual we will observe one wage y_{it} for each time period t given his educational level s. Which of the four possible wage will be observed is determined by the relation:

$$y_{it} = y_{0it} I\{s_i = 0\} + y_{1it} I\{s_i = 1\} + y_{2it} I\{s_i = 2\} + y_{it} I\{s_i = 3\},$$

where $I\{.\}$ is the indicator function assuming value one if that particular schooling level is selected.

The potential wage (y_{sit}) is a latent variable and represent the wage that we would observe in each category if the subject would have chosen that particular level. In other words, the potential wage is the hypothetical wage that the subject would earn if, instead of the educational choice he actually took, he had to chose any of the other three counterfactuals and it is determined by the regression model:

$$y_{sit} = \alpha_s + x_{it}\beta_s + \sigma_s e_{si} + \psi_{st}\epsilon_{it} \text{ if } s_i = s, \tag{1}$$

α_s is a schooling specific constant; β_s is a vector of coefficients for the matrix of observable covariates x_{it}; the individual fixed effect is represented by the time invariant term ($\sigma_s e_{si}$) and it is allowed to correlate with x_{it}; the error term $\psi_{st}\epsilon_{it}$ denotes transitory shocks uncorrelated with personal characteristics; e_{si} and ϵ_{it} are random variables uncorrelated with each other. Inequality in potential wages within schooling levels is $\sigma_s^2 + \psi_{st}^2$: the sum of a permanent component created by variation in individual specific effect and a transitory component incorporating institutional or macroeconomic shocks uncorrelated with the individual effects.

People first select into one education according to their personal tastes and inclinations, in a second stage their choice is revealed and they earn a wage influenced by their schooling choice. Specifically, we observe the outcome y_{it}.

The assignment to one of the four categories is governed by the rule:

$$s_i = s \text{ if } a_{si} \leq \nu_i < a_{s+1,i} \text{ for } s = 0, 1, 2, 3 \tag{2}$$

In this expression ν_i is the unobserved schooling factor known to the individual and includes tastes for education, motivation and all those factors influencing the choice of the individual, but unobservable by the researcher. $a_{si}$is the minimal level of ν_i for those individuals that chose schooling level s and it is determined by the relation:

$$a_{si} = \kappa_s - z_i\theta. \tag{3}$$

The vector z_i contains all observable characteristics x_{it} plus an instrument for education; θ is the vector of coefficients for z_i and κ_s is a constant with $\kappa_0 = -\infty$ and $\kappa_4 = \infty$, respectively. We assume ν_i to be uncorrelated with the transitory shocks ϵ_{it}, but to be correlated with the permanent component e_{si}. The correlation coefficient is indicated by ρ_s and it can assume both negative or positive values. In case of $\rho_s > 0$ we have positive correlation and high-skilled workers will obtain more education; the opposite occurs in case of negative selection ($\rho_s < 0$).

In order to be able to disentangle the share of wage variance due to real uncertainty from that caused by unobserved heterogeneity we will have to rely on some additional assumption regarding the disturbances in the primary and selection equation. Following Olsen (1980) and Maddala (1983) we assume linearity on the conditional expectations of e_{si} given ν_i so that:

$$\sigma_s e_{si} = \sigma_{e\nu s}\nu_i + \xi_s \tag{4}$$

with $E[\xi_s|\nu_i] = 0$, $Var[e_{si}|x_{it}, z_{it}] = \sigma^2_{es}$, $Var[\xi_s] = \sigma^2_\xi$ and $Cov[e_{si}, \nu_i] = \sigma_{\epsilon\nu s}$.

The assumption of linearity in the error term is needed in order to disentangle the two components of $\sigma_{e\nu s}$ which are the correlation coefficient ρ_s and the variance of the permanent component σ^2_s.

From the personal standpoint[1] the expected value of future wages is given by:

$$E[y_{sit}|z_i, x_{it}, \nu_i] = \alpha_s + x_{it}\beta_s + \gamma_s\nu_i \tag{5}$$

where $\gamma_s \equiv \sigma_s\rho_s$. This decomposition of expected wages introduce an important feature of this model. When selection is positive (i.e.: $\rho_s > 0$) the labor market rewards workers with high taste for education whilst the opposite occurs when selection is negative (i.e.: $\rho_s < 0$).

Since individual posses a more accurate assessment of their own abilities then researchers, private information (ν_i) has to be accounted for when we want to build a true measure of risk. Wage uncertainty is the variance of permanent and transitory component from the individual standpoint that has to say separated from unobserved heterogeneity. We indicate wage uncertainty or risk with the Greek letter τ. Using equation (4) and distributional assumptions over disturbances illustrated above we obtain[2] a formal expression for risk as the variance of wage uncertainty given observed and unobserved heterogeneity:

$$\tau^2_s = Var[\sigma_s e_{si} + \psi_{st}\epsilon_{it}|z_i, x_{it}, \nu_i] = \sigma^2_s(1 - \rho^2_s) + \psi^2_{st} \tag{6}$$

Remembering that the extent of predictability of wage uncertainty from the personal standpoint is expressed by the correlation coefficient ρ_s equation (6) makes explicit the formal link between uncertainty and private information. In fact, if the correlation between unobserved schooling factor (ν_i) and permanent component of wage inequality (e_{si}) is perfect (i.e.: $\rho_s = 1$) the subject can predict exactly the permanent part of his wage variability and uncertainty is only caused by transitory shocks (ψ^2_{st}). On the other hand, if correlation is absent (i.e.: $\rho_s = 0$) the subject does not posses any additional information compared to the researcher and wage uncertainty is observed in the data.

Rearranging equation (6) as $\sigma^2_s + \psi^2_{st} = \gamma^2_s + \tau^2_s$ helps us visualizing how potential wage inequality ($\sigma^2_s + \psi^2_{st}$) is the sum of two elements: variance of unobserved heterogeneity (γ^2_s) and wage uncertainty (τ^2_s). Note also that if correlation between schooling and unobserved tastes for education exists (i.e.: $\rho_s \neq 0$) potential wage inequality overstates the real degree of risk ($\tau^2_s < \sigma^2_s + \psi^2_{st}$).

3 Estimation

In the previous section we have amply discussed the possible source of self-selection. In presence of self-selection the zero conditional mean of the error term in the outcome equation is violated leading to inconsistent parameters estimates if an OLS regression is used[3]. The first solution[4] to correct for sample selection bias has been introduced by Heckman(1974; 1976; 1979). Heckman's approach restores the zero conditional mean of errors in the outcome equation via the inclusion of a selection correction

[1] See appendix for derivation.

[2] Details of derivation in appendix.

[3] For a textbook discussion of selection and self-selection and methods to correct for it see Cameron and Trivedi (2005) .

[4] For a survey on sample selection models estimations methods see Vella (1998).

term λ_i. Under normality, the selection correction term is proportional to the hazard rates and depends only on known parameters of the selection equation: $\lambda_i = \frac{\phi(z_i'\theta)}{\Phi(z_i'\theta)}$ with ϕ and Φ denoting the probability density and cumulative distribution functions of the standard normal distribution respectively. The so called two-step procedure prescribes to estimate the correction term by Probit or Logit and then include the estimated term into the outcome equation as an additional regressor.

The wide success that this estimator has encountered in the literature is explained with the readiness of application. Heckman's procedure provides consistent estimates given a valid exclusion restriction[5] of one variable in z_i from x_i; additionally, the error terms in the selection and outcome equation need to have a bivariate normal distribution. However, if the true joint distribution of the error terms is not correctly specified, the correction term can become inconsistent. This lead to a criticism of the Heckman model to be too restrictive to be able to eliminate the selection bias (Goldberger, 1983).

A fertile line of research (Ahn and Powell, 1993; Cosslett, 1983, 1991; Dahl, 2002; Powell, 1989; Robinson, 1988; Newey, 2009) offered new semiparametric methods to correct for self-selection with limited reliance on distributional assumptions. Generally all these methods imply a two-step approach, with a specified selection and structural equation and generic selection correction function and error term density. The assumption that those methods usually imply is:

$$E[\nu_i | a < \nu_i < b; x_i, z_i] = g(z_i'\theta)$$

with g an unknown function. If we compare the parametric to the semiparametric case two difficulties rise: a) it is not possible to invoke any distributional assumptions over ν_i to estimate θ; b) it is not possible to use distributional relationships to estimate $E[\sigma_s e_{si} + \psi_{st}\epsilon_{it} | a \leq \nu_i < b]$.

To overcome the first complication we adopt the "semi-nonparametric" (SNP) estimation strategy proposed by Gallant and Nychka (1987). This estimator does not necessitate any imposition on the distribution of the error term ν_i in the selection equation to obtain estimates of θ. The basic underlying idea of this methodology is to approximate the true density by the product of an order K series of polynomials and a normal density. In this way, many different features of the unknown density - density itself, variance and higher moments, derivatives and integrals etc.- can be consistently estimated. The approximation is specified as:

$$f_K(\nu) = \frac{1}{\pi} \sum_{k=0}^{2K} \iota_k^* \nu^k \phi(\nu) \tag{7}$$

where:

$$\pi = \int_{-\infty}^{\infty} \left(\sum_{k=o}^{K} \iota_k \nu^k \right)^2 \phi(\nu) \tag{8}$$

and

[5] If the inverse Mills ratio are sufficiently nonlinear, identification can be achieved only relying on distributional assumptions allowing $x_i = z_i$. See Vella (1998) and Cameron and Trivedi (2005).

$$\iota_k^* = \sum_{i=a_k}^{b_k} \iota_i \iota_{k-i} \tag{9}$$

with $a_k = max(0, k-K)$ and $b_k = min(k, K)$. $\phi(\nu)$ is the standard normal density. In principle any moment generating density other than the normal could be used; the normal density is a convenient choice since this form nests the ordered Probit model which becomes a special case with $K = 1$.

The corresponding cumulative function is then given by:

$$F_K(u) \frac{1}{\theta} \sum_{k=0}^{2K} \iota_k^* \int_{-\infty}^{u} \nu^k \phi(\nu) d\nu \tag{10}$$

Gallant and Nychka show that estimates of θ are consistent provided that the order of polynomials K increases with sample size. The choice of the right K is then essential. This is a standard model selection problem that we tackle by applying two different model selection criteria: the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) and chose the preferred one according to these methods.

We overcome the second complication caused by the non reliance on normality applying the method proposed by Cosslett (1991). With this procedure, after having estimated $\widehat{\theta}$ via the SNP estimator in the first step, the selection correction term to be included in the primary equation are approximated by J indicator variables $\{1(z_i\hat{\theta} \in \hat{I}_j)\}$. The correction term in our case assumes the form:

$$g(z_i\theta) = \sum_{j=1}^{J} b_j I_{ij}(\widehat{z_i\theta}) \tag{11}$$

the unknown parameters b_j can be estimated by OLS. Consistency requires J to increase with sample size. Inclusion of the correction term in the primary equation re establishes the zero conditional mean on the error term. The estimated equation takes the form:

$$y_{it} = \alpha_s + x_{it}\beta_s + \gamma_s \sum_{j=1}^{J} b_j I_{ij}(\widehat{z_i\theta}) + \omega_{it} \tag{12}$$

In this equation, by construction, we have that $E[\omega_{is}|s_i = s; x_{it}, z_i] = 0$ and we can apply OLS to obtain consistent estimates of the vector β.

For comparison we adopt an alternative strategy in the second step. Gallant and Nycha's method allows us to produce an estimate for $E[\nu_i|a \leq \nu_i < b]$. The inclusion of these estimates in the wage equation is sufficient to reestablish the zero conditional mean on the error term and the wage equation can be safely estimated by OLS. The exact specification of the alternative estimation of the primary equation is:

$$y_{it} = \alpha_s + x_{it}\beta_s + \gamma_s E[\nu_i|a \leq \nu_i < b] + \omega_{it} \tag{13}$$

Remember that our main interest relies in estimation of variances of wages. The variance of

observed wage is expressed by[6]:

$$Var[\sigma_s e_{si} + \psi_{st}\epsilon_{it} | a \leq \nu_i < b] = \sigma_s^2(1 - \rho_s^2 \delta_{si}) + \psi_{st}^2 \tag{14}$$

δ_{si} is referred to as the truncation adjustment. Its expression is[7]: $\delta_{si} = Var[\nu_i | a \leq \nu_i < b] - \sigma_\nu^2$ with σ_ν^2 indicating the unconditional variance of ν_i. The sign of δ_{si} determines whether observed wage inequality will understates or overstates potential wage inequality. In fact, in case $\delta_{si} > 0$ **(i.e.:** $Var[\nu_i | a \leq \nu_i < b] > \sigma_\nu^2$**)** observed wage inequality is greater than potential wage inequality while the opposite will occur when $\delta_{si} < 0$.

After we have constructed the selection and truncation adjustment terms from the first stage, we apply a fixed-effect model to identify the transitory component. The fixed-effect model allow us to filter out the individual permanent component $\sigma_s e_{si}$. In this way we identify the transitory component ψ_{st}^2. Defining $\zeta_{sit} \equiv \psi_{st}\epsilon_{it}$, our model takes the form:

$$(y_{it} - \bar{y}_i) = (x_{it} - \bar{x}_i)\beta_s + (\zeta_{sit} - \bar{\zeta}_{si}) \text{ if } s_i = s, \tag{15}$$

$\bar{y}_i$, $\bar{x}_i$ and $\bar{\zeta}_{si}$ denote the average over time of the corresponding variables. The transitory component of wage inequality will be identified as the variance of the error term in equation (15).

The permanent component will be identified via a between individual model:

$$\overline{y}_i = \alpha_s + \overline{x}_i\beta_s + \gamma_s g(z_i\theta) + \overline{\omega}_i \tag{16}$$

Where $g(z_i\theta) = \sum_{j=1}^{J} b_j I_{ij}(\widehat{z_i\theta})$ in the Cosslett specification and $g(z_i\theta) = E[\nu_i | a \leq \nu_i < b]$ in the alternative specification. Consequently, the permanent component of wage inequality is identified as[8]:

$$\widehat{\sigma}_s^2 = \widehat{\gamma}_s^2 \overline{\widehat{\delta}}_s + \widehat{Var}[\omega_{si} | a \leq \nu_i < b] - \frac{\sum_t \widehat{\psi}_{st}^2}{\overline{T}} \tag{17}$$

The parameter $\widehat{\gamma}_s^2$ is estimated as the coefficient for the correction terms distinguished by schooling level in an OLS regression.; $\widehat{Var}[\omega_{si} | a \leq \nu_i < b]$ is estimated as the mean squared of the error term with the between individual estimator $\overline{T} \equiv (\sum_i T_i^{-1}/N)^{-1}$ and $\overline{\widehat{\delta}}_s$ is the sample average of the truncation adjustment. We now have all elements to identify wage uncertainty as defined in equation (6): $\widehat{\tau}_s^2 = \hat{\sigma}_s^2 - \widehat{\gamma}_s^2 + \hat{\psi}_{st}^2$. To obtain a separate identification for the two components of γ_s we need to substitute equation (17) in the expression for $Var[\sigma_s e_{si} | a \leq \nu_i < b]$ and we obtain: $\widehat{\rho}_s^2 = 1 - \frac{\widehat{\sigma}_{\xi s}^2}{\widehat{\sigma}_s^2}$.

All parameters of interest - σ_s, ψ_{st}, τ_s and ρ_s - are in this way identified.

4 Data

For our purposes, we will use the National Longitudinal Survey of Youth 1979 (NLSY79). The survey is a well known and widely exploited data set of 12,686 young American citizens who were 14 to 22 years old in 1979. The participant to the survey were interviewed annually from 1979 until 1994

[6] See appendix for derivation.
[7] See appendix for derivation.
[8] See appendix for derivation.

and biennially from then on. NLSY79 provides information on schooling, labor market experiences, training expenses, family income, health condition, household composition, geographical residence and environmental characteristics.

We will restrict our analysis to males between the survey years 1991 and 2000 (calendar years 1990 to 1999). By selecting men only we do not have to worry about issues of labor market participation decisions that would rise if also women were included, while the wave restriction will allow us to focus on individuals already out of school and into the labor market. Additionally, we will exclude respondents who do not provide any information about parental education, highest grade completed, exact work experience history, hourly rate of pay and ability index as defined below. After having selected the sample according to these guidelines, we remain with a balanced panel sample of 3,373 individuals.

Our dependent variables are two: schooling for the choice equation and earnings for the outcome equation. Schooling is measured as highest schooling level completed in 1990. From this information we construct four dummies for the highest educational achievement: no high school, high school, college drop outs and college graduates or beyond. Earnings are defined as the logarithm of hourly earnings in 1992 dollars.

The control variables added both in the schooling and wage equations and presented in table 1 are the highest education completed for both parents, the Armed Forces Qualification Test score (AFQT), the family income, the number of siblings and the ethnic origin. All these variables are meant to control for intrinsic ability and family background of the individual. To control for characteristics of the geographical area of origin we include a set of dummies for urban area and for the region of residence at 14 (Northeast, South or West in table 1).

The AFQT is a series of four tests in mathematics, science, vocabulary and automotive knowledge. The test was administered in 1980 to all subjects regardless their age and schooling level. For this reasons it can include age and schooling effects in the ability index that the test is meant to construct. To correct for this undesired effects we will follow Kane and Rouse (1995) and Neal and Johnson (1996) regressing the original score on age dummies and quarter of birth and by using the residuals obtained from the regression instead of the original test score.

For family income we intend family income at age 17. If no measure for family income at 17 is recorded, we will plug in the family income closest to 17 available.

Besides the aforementioned controls common to choice and wage equation, the latter is augmented by the inclusion of experience in the labor market. Work experience is here defined as the cumulative number of working weeks divided by 49: the amount of working weeks in a calendar year. In this way we transform work weeks in work years.

The instrument for schooling that we will exploit for identification is the average national unemployment rate stratified by sex, age group and ethnic origin in the economy during the years that each respondent spent in school after mandatory schooling age and the last year of mandatory schooling. The intuition behind this instrument is that the unemployment rate that an individual would have to face in the market influences his outside option making the possibility to drop out of school more or less attractive. To our knowledge, the only paper exploiting the same instrument is Arkes (2010), but with state level information over youth unemployment. Information about unemployment rates are taken

Table 1: Summary statistics

Time Invariant Variables			
(a) Schooling Variables		Number of siblings	3.63
Years of schooling	12.99		(2.52)
	(2.57)	Family income (1999 dollars)	23,320
Categorical education:			(16,941)
No high school	.20	Black	.25
	(.40)		(.42)
High school	.36	Hispanic	.14
	(.48)		(.35)
Some college	.21	*(c) Geographic Controls at age 14*	
	(.41)	Urban	.79
Four year college or beyond	.22		(.41)
	(.41)	Residence in Northeast	.19
(b) Ability and Family Background			(.39)
Armed Forces Qualifying Test score (adjusted)	43.30	Residence in South	.33
	(29.21)		(.47)
Highest grade mother	11.10	Residence in West	.19
	(3.20)		(.39)
Highest grade father	11.12	*(d) Instrument for Schooling*	
	(3.93)	Average unemployment rate (%)	25.33
			(5.62)

Time Variant Variables					
Calendar year	1990	1993	1995	1997	1999
Actual work experience	10.03	12.77	14.40	16.09	17.93
	(3.58)	(4.05)	(4.35)	(4.68)	(5.04)
Log hourly earnings	2.18	2.16	2.28	2.26	2.21
	(.98)	(1.06)	(1.06)	(1.14)	(1.25)
Unemployment rate (%)	5.81	8.70	6.01	5.59	4.12
	(2.13)	(2.51)	(1.74)	(1.87)	(1.05)

Note: Standard deviations in parentheses. Unemployment rates calculated from CPS data.

from the Current Population Survey (CPS). The CPS is freely accessible from Internet[9] and conducted by the American Bureau of Census for the Bureau of Labor Statistics on a sample of 50,000 American families each month for the last 50 years. Since it could be possible that past level of unemployment affect current levels and thus wages and that labor market conditions at entry might carry over in subsequent years, we will include the unemployment rate for the same years wages are observed directly in the wage equation. The assumption is then that conditional on current unemployment rates, past unemployment rates have no effect on current wages.

Means and standard deviations of dependent and independent variables are illustrated in Table 1. We see that distribution among the four educational category is quite equal, but a substantial share stopped after high school, black are overrepresented and the large majority of respondents was raised in a urban environment.

[9] HTTP://data.bls.gov/data/ accessed the 15/06/2010.

Table 2: Model comparison

K	log likelihood	LR-test of OP	p-value	LR-test of K-1	p-value	AIC	BIC
OP	-16,610.284					33,272.57	33,478.57
3	-15,917.239	1386.09	.000	1386.09	.000	31,890.48	32,112.33
4	-15,286.300	2647.96	.000	1261.87	.000	30,630.61	30,860.38
5	-15,346.265	2528.04	.000	-119.92	1.000	30,752.53	30,990.23
6	-15,334.075	2552.42	.000	24.38	.000	30,730.15	30,975.77
7	-15,325.313	2569.94	.000	17.52	.000	30,714.63	30,968.17
8	-15,259.880	2700.81	.000	130.86	.000	30,585.76	30,847.23

5 Empirical results

In the empirical section we illustrate the two stages of the selection model described in section 3 estimated on NLSY79 data. After we have obtained the mean wages corrected for self-selection, we will identify the key parameters of our model: *permanent component* (σ_s^2); *transitory component* (ψ_{st}^2); *unobserved heterogeneity* (ν_i) and *wage uncertainty* (τ_s^2).

5.1 Selection of the preferred model and first stage

In the first stage we estimate the choice equation via the Gallant and Nychka method discussed in section 3. In this way we obtain estimates for the density function of the unobserved heterogeneity component and we can finally substitute these estimates in the outcome equation reestablishing the zero conditional mean on the error term.

In this method it is essential for the degree of polynomial K to increase with sample size. To select the best approximation we apply two standard method for selection: AIC and BIC[10]. The two methods differ on how steeply they penalize model complexity. AIC tends to penalize complexity less than BIC, thus if parsimony is important BIC should be the preferred criteria. In table 2 we present the two test.

OP is the ordered Probit model. We start from the 3rd degree polynomial since this is the first model generalizing the ordered Probit to the semi non-parametric case. We can see that all three tests conducted - likelihood ratio, AIC and BIC criteria - select the 8th degree polynomial. We have not carried on our test on higher order polynomials since the maximization of the log-likelihood function with $K = 9$ does not converge[11].

Results of ordered Probit model and for the SNP at 3rd and 8th degree polynomial are presented in table 3. It must be noted that estimates of ι_s cannot be compared directly across model without adjustment that is because the fitted densities differ (Stewart, 2004). What we can compare are ratios of different coefficients from different models. Anyhow, what is more interesting in this estimates is the strong effect that our instrument has irrespective of the selected model.

Results show that our instrument has a significant and strong impact on schooling decisions. In the OP model, the *t-statistic* for the instrument is a reassuring -54.58 corresponding to an *F-statistic* of about 2,979. Even stronger is the impact that average unemployment rate has in the SNP(8) model,

[10] See Cameron and Trivedi (2005) for a textbook discussion of the two criteria.

[11] The process was stopped after the 80th iteration. Data and programs available on request.

Table 3: First stage estimates for different values of K

	OP	SNP(3)	SNP(8)
Avg. unemp. rate	-.063***	-.382***	-.344***
	(.005)	(.002)	(.002)
Mother attended college	.071***	-.010***	-.018***
	(.009)	(.001)	(.002)
Father attended college	.055***	-.008***	-.014***
	(.008)	(.001)	(.002)
Highest grade mother	.040***	.011**	.015***
	(.005)	(.003)	(.003)
Highest grade father	.048***	.025***	.021***
	(.004)	(.003)	(.002)
Number of siblings	-.027***	-.022***	-.015***
	(.005)	(.003)	(.003)
Family income bottom quartile	.010	-.134***	-.072**
	(.035)	(.026)	(.022)
Family income second quartile	-.058	-.036	-.007
	(.031)	(.023)	(.020)
Family income third quartile	.014	.031	.049**
	(.028)	(.022)	(.018)
Family income top quartile	.238***	.103***	.098***
	(.028)	(.022)	(.018)
AFQT score (adjusted)	.028***	.013***	.010***
	(.000)	(.000)	(.000)
Black	.713***	3.555***	2.931***
	(.026)	(.026)	(.025)
Hispanic	.587***	1.264***	1.085***
	(.037)	(.029)	(.025)
Constant	-11.955***	-10.101	-10.101
	(.254)		
Cut point (κ_2)	-9.793***	-8.283***	-8.563***
	(.241)	(.019)	(.018)
Cut point (κ_3)	-8.352***	-7.016***	-7.464***
	(.229)	(.024)	(.024)
Polynomial:			
1		1.971***	.323***
		(.140)	(.059)
2		.954***	-.619***
		(.108)	(.023)
3		-.068***	-.135***
		(.012)	(.030)
4			.185***
			(.011)
5			.024***
			(.004)
6			-.021***
			(.001)
7			-.001***
			(.000)
8			.001***
			(.000)
Wald χ^2	8,693.37		99,255.10

Note: Geographic and cohort controls added. Geographic controls include the urban dummy and three regional dummies for residence at 14. Cohort controls include a full set of birth cohort dummies and age in the initial survey year. */**/*** indicate confidence levels of 10/5/1 percent respectively. Standard errors in parentheses.

which is our favorite model. In this model the *t-statistic* is -204.13 leading to an *F-statistic* of 41,669. Remember that as a rule of thumb an *F-statistic* bigger than 10 is deemed to be sufficient to rule out concerns of weak instruments. We can safely affirm that average unemployment rate in the labor market influences schooling decision. The sign of the effect of our instrument over schooling length is puzzling. We would have expected our instrument to show a positive correlation with schooling level (i.e.: high unemployment encourages pupils to stay in school), but our elaboration shows the exact opposite. Since our identification strategy requires a valid and relevant instrument to achieve identification and since both features are not jeopardized by the odd sign of correlation we will not investigate the cause of this surprising result.

The other covariates all show the expected signs. Parents education, ability and family income are positively correlated with educational achievements while number of siblings negatively so. The only surprising results regards the coefficients for African-American and Hispanic students. Belonging to a ethnic minority encourages education at least in our sample. The surprising result is also encountered by Chen (2008) and Cameron and Taber (2004) on the same data.

5.2 Identification of selection correction term

Separate identification of wage uncertainty (τ_s^2) and unobserved heterogeneity (ν_i) requires four selection correction terms, one for each schooling level, to be estimated[12]. The Cosslett procedure that we apply in this paper produces as many 'correction' dummies as intervals for which the distribution is partitioned into. In our case we have set each interval to contain 500 observations ordered according to the estimated score in the first stage. Dividing the entire sample in groups of 500 observations generates 44 intervals: 7 for high school drop outs; 16 for high school graduates; 9 for college drop outs and 9 for college graduates.

In order to reduce 44 correction coefficients down to four we fit a series of polynomials on the estimated conditional correction dummies in the first stage. In a second moment we include the estimated coefficients of the polynomials in the second stage as additional regressors. These are the estimated γ_s in table 4 for the Cosslett model.

In figure 1 we show the conditional selection dummies for the four educational categories and the polynomials approximating their location over the distribution.

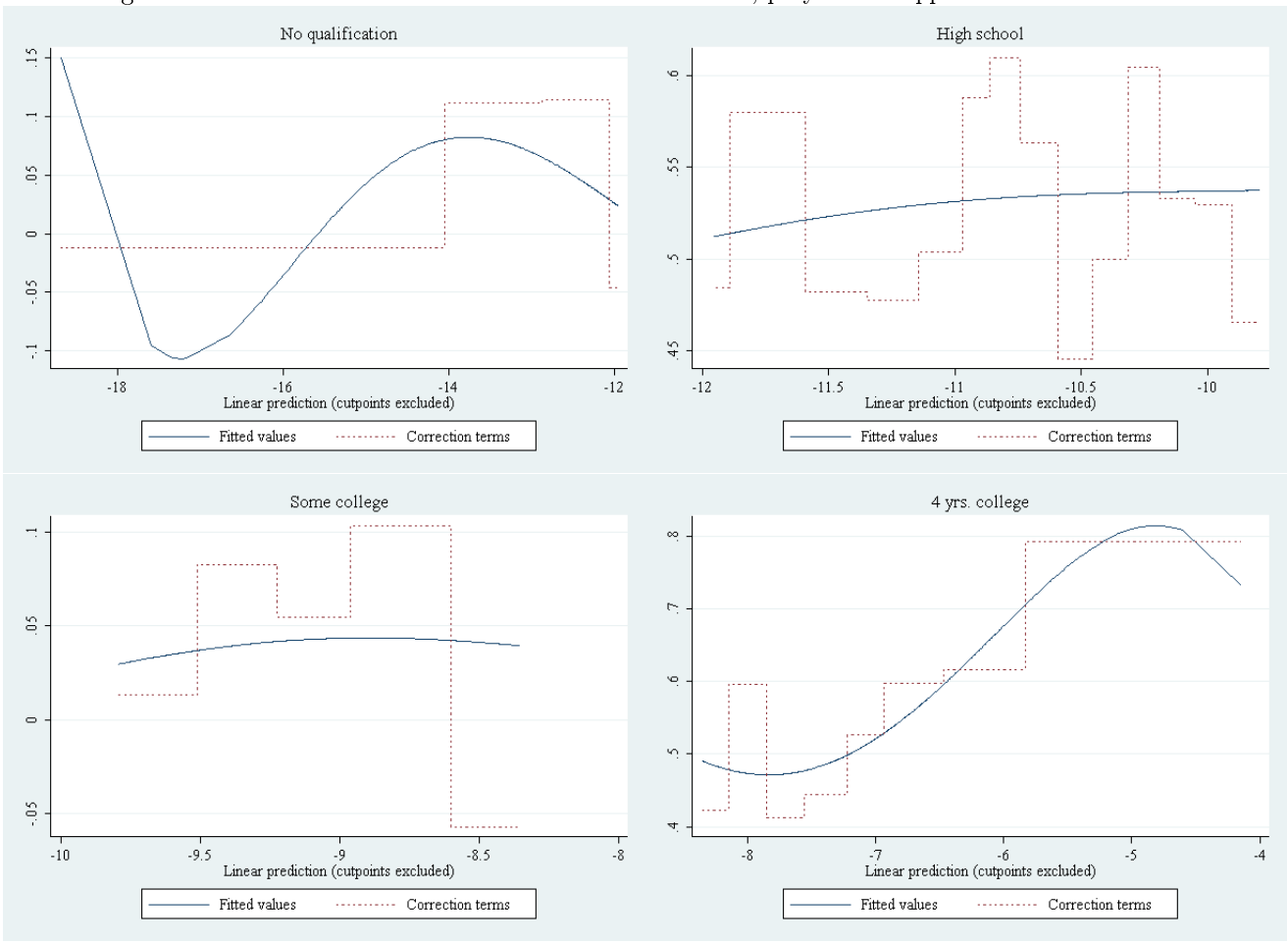
5.3 Wage equation

Here we report estimates of equation (16) with the two alternative specification for the function $g(.)$. Both estimates are based on a between-individual effect model and determine the causal impact of education on wages.

Estimates of the schooling coefficients based on the Cosslett procedure are presented in the first column. They show the expected effect that wages increase with schooling and all coefficients are estimated with extreme precision up to 1% confidence level. High school completion implies a 5% wage premium and college graduates earn about 32% more than high school dropouts with the same observable characteristics.

[12] See section (2.1) and appendix.

Figure 1: Estimated conditional selection correction term; polynomial approximation



Three out of four correction coefficients point towards an underestimation of education on wages the only exception being high school graduates. However, it has to be noted that the estimates are quite imprecise.

In the second column outcome of the alternative specification for correction terms are reported. Clearly, results are not robust to specification. Schooling coefficients invert their sign for high school graduates and college drop outs while it remains almost unaltered for college graduates. Estimates are insignificant for the lowest category and significant at a 5% confidence level for the middle one. Also the correction terms are consistently different. Two of them are statistically significant and positive pointing towards an overestimation of the impact of education on wages. The selection correction term for college drop outs is some order of magnitude bigger than all other correction terms estimated with the same method or with the Cosslett one.

5.4 Main results

The inequality measures for variance of wage residuals are: i) the observed wage inequality given the choice of schooling ($Var[y_{sit}|s_i = s, x_{it}, z_i]$); ii) the potential wage inequality purged of selection and truncation biases ($\sigma_s^2 + \psi_{st}^2$); and iii) the uncertainty in potential wages, after removing truncation and selection biases and incorporating unobserved heterogeneity factors (τ_s^2).

In table 5 panel A we report the observed wage inequality. The observed wage inequality is the resultant of the sum of two factors. The first is the permanent component, identified by the mean squared residuals in the between-individuals model not corrected for selectivity. The second is the transitory component identified by exploiting the mean-squared errors of the fixed-effects model.

From table 5 we can see that high school graduates are those showing a lower variance both in the permanent (panel A) and transitory (panel B) component. On the other hand, college graduates are those showing the highest. Observed wage inequality monotonically increases after high school and the results holds for both of its components this result is also encountered by Chen (2008) on the same data. College enrollment causes a 15% increase in inequality and college completion an additional 23%.

In panel C we show our estimates of permanent component corrected for selection and truncation biases. The hierarchical order of the four educational categories is changed. If we control for self-selection college drop outs show the highest variance in the permanent component. Marginal contribution of college entry is 33% increase in wage variability. College completion, on the other hand, diminishes variability by 13%. However, if we add the transitory component and we look at the total potential inequality we see that the hierarchical order is reestablished do to the considerably higher transitory shocks that college graduates have to face with respect to college drop outs. It is worth nothing that the transitory component it is not influenced by self selection since it is modeled as exogenous shocks that individuals can not act upon by construction.

An important and novel outcome of our estimation is that potential wage inequality is actually superior to the observed inequality; only for college drop outs the two are almost equal with a slight predominance of the latter. Our results suggest that if we were to assign education randomly intra educational wage variability would be smaller than the observed variability. At first sight this result is puzzling and former parametric estimates (Chen, 2008) find the exact opposite since in the parametric

Table 4: Wage equation

	Cosslett	GLS
High school	.051***	-.055
	(.016)	(.048)
Some college	.096***	-.171**
	(.017)	(.063)
4 yr. college or beyond	.317***	.340***
	(.019)	(.056)
Experience	.131***	.131***
	(.005)	(.005)
Experience^2	-.001***	-.001***
	(.000)	(.000)
AFQT score (adjusted)	.005***	.005***
	(.000)	(.000)
Highest grade mother	-.003	-.002
	(.002)	(.002)
Highest grade father	-.001	-.000
	(.002)	(.001)
Number of siblings	.003	.002
	(.002)	.002
Family income bottom quartile	.015	.014
	(.015)	(.015)
Family income second quartile	-.042**	-.041**
	(.014)	(.014)
Family income third quartile	-.008	-.006
	(.014)	(.013)
Family income top quartile	.060***	.066***
	(.013)	(.013)
Unemployment rate	-.001	-.001
	(.002)	(.002)
Black	.036*	.050**
	(.015)	(.016)
Hispanic	.037*	.050**
	(.018)	(.017)
γ_0	-.106	-.007
	(.088)	(.010)
γ_1	.202	.757*
	(.031)	(.332)
γ_2	-.273	3.229***
	(.193)	(.745)
γ_3	-.012	-.009
	(.011)	(.078)
Geographic controls	yes	yes
Cohort controls	yes	yes
R^2	.395	.393
N	20,398	20,398

Note: Geographic controls include the urban dummy and three regional dummies for residence at 14. Cohort controls include a full set of birth cohort dummies and age in the initial survey year. */**/*** indicate confidence levels of 10/5/1 percent respectively. Standard errors in parentheses. Bootstrapped standard errors on 200 replications. Unemployment rate calculated on CPS data.

Table 5: Parameters of interest

	Less than high school	High school	Some college	4 yr. college and beyond
I. Observed wage inequality				
A. Permanent component	.309	.307	.355	.419
B.Transitory component (ψ_{st}^2)	.156	.131	.142	.207
Age 25-30	-.055	-.052	-.043	-.105
Age 31-36	-.053	-.054	-.047	-.110
Age 37-42	-.023	-.023	-.020	-.055
Observed inequality (A+B)	.465	.438	.497	.626
II. Potential wage inequality				
C. Permanent component (σ_s^2)	.205	.263	.350	.304
D. Transitory component (same as B)				
Potential wage inequality (C+D)	.361	.394	.492	.511
III. Wage uncertainty				
E. Correlation coefficient (ρ_s)	.207	.326	.429	.017
F. Permanent component (C-CxE^2)	.197	.235	.285	.304
G. Transitory component (same as B)				
Degree of wage uncertainty (τ_s^2)	.357	.358	.407	.511
Unobserved heterogeneity (ν_i)	.005	.026	.085	.000

case by construction observed wage variability always understates potential wage variability. We believe this to be an unnecessary constraint of the parametric case. In fact, in the parametric case is implied that individuals exploit their private information to minimize their future wage risk when deciding upon educational levels. In reality we have no prior knowledge to suggest that this is the case. Even if individuals do posses superior knowledge about themselves, we cannot assume that they will try to minimize risk. It is reasonable to assume that other more compelling factors enter their utility function other than risk minimization and these other factors might offset considerations of future wage variability. If this is the case we cannot exclude *a priori* the apparently odd result that we encounter, result that is impossible if we assume normality of the error terms.

The difference between potential and observed wage inequality is particularly evident in the case of college graduates and high school drop outs. For the first category observed wage inequality exceeds the potential by 18%, for the latter the difference is about 23%. For the two middle categories the difference is less pronounced: 10% for high school graduates and only 1% for college drop outs. It is worth nothing how potential wage inequality increases monotonically with education. Wage inequality for the most educated individuals is about 45% higher than for the least educated ones.

The permanent component presented in panel C is corrected for self-selection and truncation, but it does not account for unobserved schooling factor ν_i which is included in estimates presented in panel F. It is interesting to compare the two estimates of panel C and panel F since from this comparison we can already understand the importance of unobserved heterogeneity for wage variability. Controlling for unobserved heterogeneity particularly affects high school graduates and college drop-outs; the estimates for high school drop-outs are less affected while those for college graduates are completely unaffected. These results are due to the strength of correlation between wages and schooling factor. The parameter ρ_s in panel E describes a positive correlation between the two for all four categories, but a very weak one for the last one. A positive ρ_s also tell us that the labor market rewards people with a high unobserved schooling factor. This result is at odd with previous parametric estimates (Chen, 2008).

Estimates for unobserved heterogeneity in panel G confirm the intuition. Unobserved heterogeneity accounts for only 1% of potential wage variability in the case of high school drop-outs and no unobserved heterogeneity for college graduates is traceable. The only two categories for which unobserved schooling factors affect wage variability are high school graduates and particularly college drop outs. For these two categories the unobservable parameter ν_i is responsible of 6.5% and 17% of total potential wage variability respectively.

The decomposition between the two determinants of inequality - uncertainty and heterogeneity - tell us that private information has no impact on wage inequality for two out of four categories, a minor impact for one and only in one case it shows to be an important factor to account for in identifying a causal effect of risk on wages. The other side of the medal is that almost the entirety of inequality can be referenced to uncertainty or risk. A possible explanation of the absence of heterogeneity in our estimates is that individuals have an imperfect knowledge about future earning streams and cannot correctly project their earning potential *ex-ante* on the *ex-post* realization. Dominitz and Manski (1996) report a low degree of awareness of real labor market pay-offs on a sample of American high school and college students. Thus, students might not be able to use private information to select the

appropriate level of education or they do not include wage variability as one of the decisive factor in their choice process. These explanation would also be consistent with the apparently odd finding of an overestimation of potential wage variability by observed variability encountered in our estimation.

Almost all the wage variation that we encounter in the data is then wage uncertainty. As for the other parameters of interest also risk monotonically increases with schooling. For risk-averse individuals this particular feature of wage variability might discourage investment into further education in absence of compensating mechanisms. The most immediate compensation that comes to mind is via higher salary, thus, the increasing between category wage inequality that has characterized the US economy in the past decades might be a compensation to superior risk. On the other hand, college enrollment and even more, college completion open up more opportunities both in term of further education (i.e.: M.Sc. or PhD etc.) and in terms of possible careers. The increased wage variability might simply reflect this enrichment in the choice set that college drop-outs and college graduates have at their disposal. Controlling for number of occupational choices that each education gives access too would shed some light on the exact mechanism, but that exceeds the scope of the present paper.

Our results clearly show that potential and observed wage variability increases with schooling and particularly with college entry. Almost the entirety of the encountered variability is explained by risk and not unobserved heterogeneity. It is also interesting to note that college graduates for whom variability is the highest are also those for whom risk completely accounts for it. Low impact of unobserved schooling factor and overestimation of potential by observed wage inequality suggest us that either the quality of information or the use that individuals make of it when selecting schooling might not be at odd with assumptions of risk minimizing agents.

6 Conclusion

In this paper we propose a new application of two semiparametric techniques developed by Cosslett (1983) and Gallant and Nychka (1987) to education. We extend the original case from the dichotomous to the polychotomous case providing consistent estimates of: within education potential wage variation, accounting for selection and truncation biases; degree of private information owned by the individuals and used to select their favorite level of education and the magnitude of pure risk that every education level entails.

Our results show that all decompositions of wage uncertainty (observed, potential and the transitory component) increase with schooling and particularly so at college entry. College graduates show a 44% increase of potential wage inequality compared to high school graduates. This result is in accordance to what already found by Cunha et al. (2005), but opposite to Chen (2008). Both estimates are conducted on the same sample of American young males.

Other important and novel results are the almost complete absence of unobserved heterogeneity and the fact that if education was randomly assigned to individuals withing educational groups wage variability would be reduced compared to the *status quo*. We offer a single interpretation to both findings: Individuals do not use their private information to minimize their future wage variability or simply they do not posses sufficient pieces of evidence on how their personal tastes and inclination for schooling match with possible educational levels. What it is clear from our analysis is that investing in

education has significant impact on uncertainty of future wages and this is especially true for college education. If compensation for risk exists in the labor market and if this compensation works via higher wages for riskier occupation, our findings might contribute to explain the increase between educational level inequality observed in the U.S. and other advanced economies in the past 30 years.

Uncertainty might as well be more complex than mere wage variability across educational categories. An obvious source of uncertainty is risk of unemployment. If education reduces the likelihood of unemployment spells, the increased uncertainty in pay-offs that we encounter in our estimates might be offset by the prospective of a continuous work career. A complete study of education risk have to account for both sources of uncertainty: wage variations and risk of unemployment. We let this task to future research.

References

Abadie, A.: 2002, 'Bootstrap Test for Distributional Treatment Effects in Instrumental Variable Models'. *Journal of the American Statistical Association* **97**(457), 284–292.

Abadie, A., J. Angrist, and G. Imbens: 2002, 'Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings'. *Econometrica* **70**(1), 91–117.

Acemoglu, D.: 1999, 'Changes 'in Unemployment and Wage Inequality:An Alternative Theory and Some Evidence'. *The American Economic Review* **89**(5), 1259–1278.

Acemoglu, D.: 2002, 'Technical Change, Inequality, and the Labor Market'. *Journal of Economic Literature* **40**(1), 7–72.

Ahn, H. and J. Powell: 1993, 'Semiparametric Estimation of Censored Selection Models with a Nonparametric Selection Mechanism'. *Journal of Econometrics* **58**(1/2), 3–29.

Arkes, J.: 2010, 'Using unemployment rates as instruments to estimate returns to schooling'. *Southern Economic Journal* **76**(3), 711–722.

Autor, D. H., L. F. Katz, and M. S. Kearney: 2006, 'The Polarization of the U.S. Labor Market'. *The American Economic Review* **96**(2), 189–194.

Cameron, C. and P. Trivedi: 2005, *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Cameron, S. V. and C. Taber: 2004, 'Estimation of Educational Borrowing Constraints Using Returns to Schooling'. *Journal of Political Economy* **112**(1), 132–182.

Chen, S.: 2008, 'Estimating the Variance of Wages in the Presence of Selection and Unobserved Heterogeneity'. *Review of Economics and Statistics* **90**(2), 275–289.

Chen, S. and S. Khan: 2007, 'Estimating the Casual Effect of Education on Wage Inequality Using IV Methods and Sample Selection Models'.

Cosslett, S.: 1983, 'Distribution Free Maximum Likelihood Estimator of the Binary Choice Model'. *Econometrica* **51**(3), 765–782.

Cosslett, S.: 1991, 'Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics'. In: W. Barnett, J. Powell, and G. Tauchen (eds.): *Semiparametric Estimation of a Regression Model with Sample Selectivity.* Cambridge, UK: Cambridge University Press, pp. 175–197.

Cunha, F., J. Heckman, and S. Navarro: 2005, 'Separating Uncertainty from Heterogeneity in Life Cycle Earnings'. *Oxford Economic Papers* **57**(2), 191–261.

Dahl, G. B.: 2002, 'Mobility and the Return to Education: Testing a Roy Model with Multiple Markets'. *Econometrica* **70**(6), 2367–2420.

Dominitz, J. and C. F. Manski: 1996, 'Eliciting Student Expectations of the Returns to Schoolingl'. *Journal of Human Resources* **31**(1), 1–26.

Gallant, R. A. and D. W. Nychka: 1987, 'Semi-Nonparametric Maximum Likelihood Estimation'. *Econometrica* **55**(2), 363–390.

Goldberger, A.: 1983, 'Abnormal Selection Bias'. In: S. Karlin, T. Amemiya, and L. Goodman (eds.): *Studies in Econometrics, Time Series, and Multivariate Statistics.* New York: Academic Press.

Heckman, J.: 1974, 'Shadow Prices, Market Wages and Labour Supply'. *Econometrica* **42**(4), 679–694.

Heckman, J.: 1976, 'The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models'. *Annals of Economics and Social Measurement* **5**(4), 475–492.

Heckman, J.: 1979, 'Sample Selection Bias as a Specification Error'. *Econometrica* **1**(74), 153–162.

Joshua Angrist, Victor Chernozhukov, I. F.-V.: 2006, 'Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure'. *Econometrica* **74**(2), 539–563.

Kane, T. J. and C. E. Rouse: 1995, 'Labor-Market Returns to Two and Four Years College'. *American Economic Review* **85**(3), 600–614.

Katz, L. F. and D. H. Autor: 1999, 'Changes in the Wage Structure and Earnings Inequality'. In: O. Ashenfelter and D. Card (eds.): *Handbook of labor economics.* Amsterdam, North-Holland: Elsevier Science, pp. 1463–1555.

Maddala, G. S.: 1983, *Limited-Dependent and Qualitative Variables in Econometrics.* New York: Cambridge University Press.

Moretti, E.: 2000, 'Do Wages Compensate for Risk of Unemployment? Parametric and Semiparametric Evidence from Seasonal Jobs'. *Journal of Risk and Uncertainty* **20**(1), 45–66.

Neal, D. A. and W. R. Johnson: 1996, 'The Role of Premarket Factors in Black-White Wage Differences'. *Journal of Political Economy* **104**(5), 869–895.

Newey, W.: 2009, 'Two-Step Series Estimation of Sample Selection Models'. *Econometrics Journal* **12**(1).

Olsen, R. J.: 1980, 'A Least Squares Correction for Selectivity Bias'. *Econometrica* **48**(7), 1815–1820.

Powell, J.: 1989, 'Semiparametric estimation of Censored selection models'.

Robinson, P. M.: 1988, 'Root-N-Consistent Semiparametric Regression'. *Econometrica* **56**(4), 931–954.

Roy, A. D.: 1951, 'Some Thoughts on the Distribution of Earnings'. *Oxford Economic Papers* **3**(2), 135–146.

Stewart, M. B.: 2004, 'Semi-nonparametric estimation of extended ordered probit models'. *The Stata Journal* **4**(1), 27–39.

Vella, F.: 1998, 'Estimating Models with Sample Selection Bias: A Survey'. *Journal of Human Resources* **33**(1), 127–169.

A Appendix: identification of ρ_s and σ_s

To see how the two parameters ρ_s and σ_s are identified consider the model as formalized by Chen (2008):

$$y_{si} = x_{si}\beta_s + \sigma_s e_{si} + \psi_{st}\epsilon_{it}, \tag{18}$$

$$s_i^* = z_{si}\theta_s + \nu_i \tag{19}$$

where $s_i^* = s$ if $a \leq \nu_i \leq b$. This is the usual sample selection model with ordered censoring rules. Following Olsen (1980) we make the following assumption on a part of the error term:

$$\sigma_s e_{si} = \sigma_{e\nu s}\nu_i + \xi_s \tag{20}$$

where $\sigma_{e\nu s} \equiv \sigma_s\rho_s$ is the covariance coefficient between the error term in the outcome equation and the error term in the choice equation. As Chen (2008), we assume that the error term ν_i in the choice equation is correlated with e_{si}, but not with ϵ_{it}. We also assume that ξ_s is independent of ν_i (Olsen, 1980) and that the ξ_s are uncorrelated across schooling levels. Additionally, $Var[\xi_s] = \sigma_{\xi s}^2$. From these assumptions we obtain that:

$$E[\sigma_s e_{si} + \psi_{st}\epsilon_{it} | a \leq \nu_i < b] = E[\sigma_s e_{si} | a \leq \nu_i < b] = \sigma_s\rho_s E[\nu_i | a \leq \nu_i < b] \tag{21}$$

$$Var[\sigma_s e_{si} | a \leq \nu_i < b] = \sigma_{e\nu s}^2 Var[\nu_i | a \leq \nu_i < b] + \sigma_{\xi s}^2 \tag{22}$$

To re-establish the zero conditional mean of the error term in the outcome equation (16) in presence of self-selection, we need a correction term accounting for $E[\nu_i | a \leq \nu_i < b]$ and an estimate for $\lambda_s \equiv \sigma_s\rho_s$. In fact, the equation:

$$y_{is} = \alpha_s + x_{it}\beta_s + \lambda_s g(z_i\theta) + \omega_{is} \tag{23}$$

can be consistently estimated with OLS since $E[\omega_{is}|x_{it}, z_i] = 0$ by construction. $g(z_i\theta) = E[\nu_i|a \leq \nu_i < b]$ is an unknown function entered using the method of Gallant and Nychka (1987). We can obtain estimates for both $E[\nu_i|a \leq \nu_i < b]$ and $Var[\nu_i|a \leq \nu_i < b]$ by approximating the unknown density function by a Hermite series of K degrees polynomials.

$$\widehat{E}[\nu_i|a \leq \nu_i < b] = \frac{\int_a^b \nu_i f_K(\nu_i)d\nu}{Pr[a \leq \nu_i < b]} \tag{24}$$

$$\widehat{Var}[\nu_i|a \leq \nu_i < b] = \widehat{E}[\nu_i^2|a \leq \nu_i < b] - \widehat{E}[\nu_i|a \leq \nu_i < b]^2 \tag{25}$$

where K represent the degree of polynomial used and $f_K(\nu_i)$ is a density function at ν_i assuming the form :

$$f_K(\nu_i) = \frac{1}{\pi}\sum_{k=0}^{2K} \iota_k^* \nu_i^k \phi(\nu_i) \tag{26}$$

$\pi = \int \sum_{k=0}^{2K} \iota_k^* \nu_i^k \phi(\nu_i)d\nu = \sum_{k=0}^{2K} \iota_k^* \mu_k$ is a scaling factor ensuring a proper approximation for the density function (i.e.: a function integrating to 1), μ_k is the kth moment of the standard normal distribution and ϕ the standard normal density.

According to our distributional assumptions the variance of the permanent component from an individual standpoint is given by: $Var[\sigma_s e_{si}|x_{it}] = Var[\sigma_s\rho_s\nu_i + \xi_s|x_{it}] = \sigma_s^2\sigma_\nu^2\rho_s^2 + \sigma_{\xi s}^2$. Remembering that $Var[\sigma_s e_{si}|x_{it}] = \sigma_s^2$ we obtain:

$$\sigma_{\xi s}^2 = \sigma_s^2(1 - \rho_s^2\sigma_\nu^2) \tag{27}$$

Substituting (27) into (22) and rearranging we obtain an equation for variance in observed wages corrected for truncation:

$$Var[\sigma_s e_{si}|a \leq \nu_i < b] = \sigma_s^2(1 - \rho_s^2\delta_{si}) \tag{28}$$

where we can estimate δ_{si} by: $\widehat{\delta}_{si} = \widehat{Var}[\nu_i|a \leq \nu_i < b] - \widehat{\sigma}_\nu^2$. The parameter $\hat{\gamma}_s = \widehat{\rho_s\sigma_s}$ is estimated as the coefficients for the correction terms distinguished by schooling level in an OLS regression.

In this model the error term is composed by two elements, the permanent component ($\sigma_s e_{si}$), for which we have explicated the variance in (22), and the transitory shocks $\psi_{st}\epsilon_{it}$. The expression for the variance of the complete error term is:

$$Var[\sigma_s e_{si} + \psi_{st}\epsilon_{it}|a \leq \nu_i < b] = Var[\sigma_s e_{si}|a \leq \nu_i < b] + \psi_{st}^2 = \sigma_s^2(1 - \rho_s^2\delta_{si}) + \psi_{st}^2 \tag{29}$$

$\hat{\gamma}_s = \widehat{\rho_s\sigma_{es}}$; $Var[\sigma_s e_{si} + \psi_{st}\epsilon_{it}|a \leq \nu_i < b] = Var[\omega_{si}|a \leq \nu_i < b]$ can be consistently estimated as the mean squared errors of the residuals in the between individual effects model of expression (23). ψ_{st}^2, as explained in Chen (2008), is identified from the fixed-effect model as the variance of residuals in equation (15). Substituting these elements into (29) and rearranging we identify the permanent

component of wage inequality corrected for truncation as:

$$\widehat{\sigma}_s^2 = \widehat{Var}[\omega_{si}|a \leq \nu_i < b] - (\widehat{\sigma_s \rho_s})^2 \overline{\widehat{\delta}}_s - \frac{\sum_t \widehat{\psi}_{st}^2}{\overline{T}} \tag{30}$$

With $\overline{T} \equiv (\sum_i T_i^{-1}/N)^{-1}$ and $\overline{\widehat{\delta}}_s$ is the sample average of the truncation adjustment. Substituting (30) into (27) we obtain estimates for $\sigma_{\xi s}^2$ and $\widehat{\rho}_s^2 = \frac{\widehat{\sigma}_s^2 - \widehat{\sigma}_\xi^2}{\widehat{\sigma}_s^2 \widehat{\sigma}_\nu^2}$.

Note that we have identified ρ_s^2 and σ_s^2 without assuming joint normality of the error terms in the wage and choice equations. The only two assumptions that we need to establish identification are:

1. linearity in the equation of the error term ($\sigma_s e_s = \omega_{\epsilon \nu s} \nu_i + \xi_s$);

2. the distribution of ξ_s is independent of ν_i.