

Accounting for Heterogeneous Returns in Sequential Schooling Decisions

Gema Zamarro¹
CEMFI and Tilburg University
Preliminary version

November 2, 2005

¹Special thanks to Manuel Arellano for his advice, remarks and comments. I would like to thank Pedro Jesús Hernández for providing me with the data set and for his help when needed. I am grateful to James J. Heckman for accepting me as a visiting student at the University of Chicago from January to May 2004, for making me feel part of his workgroup and for discussions that were essential for the development of this paper. My thanks are also due to Samuel Bentolila, Pedro Mira and Salvador Navarro for their suggestions and time. I acknowledge the comments and support of Aitor Lacuesta, Jesús Carro, Stéphane Bonhomme, Viswa Velur and all my Ph.D. colleagues. Finally, this paper has also benefited from comments of participants at the following colloquia, meetings and workshops: CEMFI Ph.D. workshop; Heckman's workgroup seminar, Chicago 2004; COST conference, Amsterdam 2004; European Winter Meeting of the Econometric Society, Stockholm 2004; FEDEA seminar, Madrid 2004; XXIX Simposio de Análisis Económico, Pamplona 2004; PEW day conference, Madrid 2005; World Congress of the Econometric Society, London 2005. Errors, if any, are mine. Address for correspondence: Tilburg University, P.O. Box 90153, NL-5000 LE Tilburg, The Netherlands. Phone: (+31) 134663219. Email: G.Zamarro@uvt.nl

Abstract

This paper presents a method for estimating returns to schooling that takes into account that returns may be heterogeneous among agents and that educational decisions are made sequentially. A sequential decision model is interesting because it explicitly considers that the level of education of each individual is the result of previous schooling choices and so, the variation of supply-side instruments over time will emerge as a source of identification of the desired parameters. A test for heterogeneity in returns from sequential schooling decisions is developed and expressions for Marginal Treatment Effects are obtained in this context. Returns are estimated and tested from cross-sectional data from a Spanish household survey that contains rich family background information and useful instruments. This data is stratified by level of education and so estimators are adapted to take this feature into account. Finally, this methodology is used to analyze possible effects of the 1970 reform of the Spanish education system.

JEL Classification: I21, I28, C10, J31. Keywords: Schooling, Selection Models, Heterogeneity, Sequential decisions, Policy Evaluation.

1 Introduction

Education plays a central role in the modern labor literature and in particular the economic return to schooling is a fundamental parameter of interest in many areas of economics and public policy. Although there are many other ways to measure returns to schooling, most of the literature has focus on returns in term of wages and this is also the case in this paper.

The traditional approach to estimate returns to schooling in terms of wages consists in regressing the logarithm of observed wages on a variable about schooling (i.e. years of schooling or a dummy variable for a given level of education) controlling for variables that may affect wages (i.e. tenure, kind of job, kind of activity). The main concern when trying to draw inferences from previous equations is the possible endogeneity of the schooling variable used. That is, it is difficult to know whether individuals who have higher education also have higher earnings because of their schooling or because individuals with the ability to obtain greater earnings have chosen to acquire higher education (selection problem). As a result of this, a wide range of estimates, using different methods, have been presented. The traditional approach consisting in assuming common returns for all individuals and using instrumental variables for the schooling outcomes. Supply-side variables emerge in this context as natural candidates for instruments. Card (1999, 2001) surveys this literature with special emphasis on instrumental variable methods estimates.

However, different people may have different returns. If this is the case there is no single return of the schooling choice but rather a distribution of them and the assumptions that validate instrumental variables estimates break down. This is because estimates using traditional instrumental variable methods will recover the average return for those induced to change their decisions by changes in the instrument values.¹ This summary measure may not necessarily answer the policy questions we have in

¹Imbens and Angrist (1994) show that, under independence and monotonicity assumptions, linear instrumental variables estimates can be interpreted as a weighted average of treatment effects. In particular, they show that if the instrument used is a binary variable we can recover the average effect for those induced to change their decisions because of the instrument value. This summary measure is known as Local Average Treatment Effect (LATE). Moreover, they show that in the general case, where the instrument is not a binary variable, instrumental variables estimates identify a particular weighted average of previous terms.

mind. Then, there are two problems that arise when we try to analyze the distribution of returns. The first one is how to summarize the distribution to respond to policy relevant questions and the second one, is how to obtain estimates of these summary measures. This is the aim of an emerging modern literature on returns to schooling. Estimates of the expected return of those who are likely to be affected by a policy can be useful to construct parameters that give an answer to the first question. This summary measure is called Marginal Treatment Effect (MTE) and was first introduced in this context by Heckman (1997).² Concerning the second question, the traditional econometric approach for dealing with it consists in specifying structural models for wage and decision equations joint with distributional assumptions for the error terms in the model.³ However, in spite of its popularity and its computational convenience, this approach has been criticized for its reliance on arbitrary distributional assumptions. On the other hand, the Policy Evaluation approach, which has focused on the high school-college transition, seek to obtain robust estimates of different summary measures of the distribution of returns without invoking distributional assumptions. As an example of this approach we find Carneiro, Heckman and Vytalacil (2003) who proposed a semiparametric test for the homogeneity of the returns to college, focusing on the high school-college transition. In the empirical application of their test they found that selection in the returns to college was an empirically important phenomenon using different US data sets. They showed that in this case different estimators and instruments yield different measures of the distribution of returns which do not, in general, answer the economic question we are interested in. Finally, they used the MTE to construct estimates of different treatment parameters and to characterize what instrumental variables methods estimate in this case.

If returns to schooling are heterogeneous and individuals are sorted accordingly, it is not sufficient to focus the analysis of educational policies on the high school-college transition. This is so because high school graduates are a self-selected group of the population. So, the dynamics and timing of schooling decisions will be very important if we want to draw valid conclusions for the whole population. However,

²Bjorkland and Moffit (1987) introduced a first version of a marginal treatment effect in the context of the parametric normal Roy model.

³See, e.g., Willis and Rosen (1977), Heckman (1990).

less attention has been paid to this fact when estimating returns. An exception is Carneiro, Hansen and Heckman (2003), who used factor structure models and a generalized version of ordered choice models to identify the whole distribution of returns. However, the application of ordered choice models is justified only if there are no option values of continuing studying that affect the decisions. That is, it is not possible for individuals to decide to continue studying because of the only reason of accessing higher levels of education with greater earnings. Concerning this, Taber (2001) develops and estimates an structural dynamic programming model that allows for such option values of continuing studying. Heckman, Lochner and Todd (2003) considered also a model that allowed for the existence of these option values and analyzed the empirical role that they play in determining rates of return to schooling. They concluded that more research is needed to determine their empirical importance.

In this paper I also analyze the problem of heterogeneous returns from multiple educational choices but from a different perspective. I extend existing methods of estimating heterogeneous returns to college, considering multiple levels of education and that educational decisions are made in a sequential way. A sequential decision model is interesting because it explicitly considers that the level of schooling of each individual is the result of previous schooling choices. Then, at any stage, an alteration in the benefits or costs of a given level of education affects the proportion of people in any preceding stage. Therefore, the variation of supply-side instruments over time will emerge as a source of identification of the desired parameters. Moreover, the notion of the option value of education emerges naturally in this context. So, estimated returns will take into account the selection of individuals in previous levels of education in a more general fashion than ordered choice models. Alternative methods recently developed in the nonparametric statistics literature are applied to test for the homogeneity of sequential returns (Stute (1997) and Stute, González Manteiga and Presedo Quindimil (1998)).

In addition, this paper has an empirical contribution. Estimated returns are obtained and tested, under different assumptions, using an interesting cross-sectional Spanish data set (*“Encuesta de Conciencia y Biografía de Clase”* (ECBC) (1991)). Spain has experienced a rapid growth in their secondary and post-secondary school

enrollment rates at the same time as substantial transformations occurred in the educational system. Therefore, supply side variables will emerge as good candidates for being instruments. In particular, the variation in distance to college due to the creation of new universities that took place in Spain will be used as a source of identification of the desired parameters. The ECBC is a survey which contains rich family background information, including area of residence at the time of schooling decisions, but it is stratified by level of education. This is helpful because it provides more information about this group of people that can have sparse representation in a random sample. However, this is something that should be considered when estimating the model. Estimators that do not take this into account will be inconsistent. In order to solve this problem weighted estimators are used where oversampled categories are given a low weight while undersampled ones are given a high weight. The empirical application suggests that heterogeneity is an important feature of the data set used. Moreover, marginal returns are obtained to analyze possible effects of the 1970 reform of the Spanish educational system.

The rest of the paper is organized as follows: Section 2 shows the framework to estimate and test heterogeneous returns to schooling. Expressions for MTEs when returns are heterogeneous and schooling decisions are made sequentially are presented in Section 3. Section 4 describes some facts about the Spanish education system. The main characteristics of the data set used are shown in Section 5. Section 6 shows estimates of the returns to schooling assuming common returns, a test for the homogeneity of sequential returns as well as, estimated returns when they vary among individuals and schooling decisions are made sequentially. Moreover, latter methodology is used to analyze the 1970 reform of the Spanish educational system. Finally, Section 7 concludes.

2 Heterogeneity in the returns to schooling

Accounting for individual-level heterogeneity on the returns and individual decisions on education level taking returns into account is the main concern of modern literature on returns to schooling. If this is the case, the dynamics and timing of the decisions will be very important to be able to draw conclusions that are valid for the whole

population. From the point of view of policy designers, knowing the value added of attending different educational levels would be essential. In order to obtain a comprehensive evaluation of these policies it will not be sufficient to consider high-school graduates only. This is because these individuals would make up a self-selected group of the population.

In this section, I extend the methodology developed by Carneiro, Heckman and Vytlacil (2003), for testing and estimating heterogeneous returns to college, considering multiple levels of education and that decisions among them are made sequentially. A natural extension to the binary choice case is to consider three levels of education. Extensions to more complex finite horizon specifications are, however, straight forward. Let me define the following general levels of education and a dummy variable for each one:

$$D_0^i = 1 \text{ if the individual } i \text{ left school at a low level of education.}$$

$$D_1^i = 1 \text{ if the individual } i \text{ left school at a medium level.}$$

$$D_2^i = 1 \text{ if the individual } i \text{ left school at a high level of education.}$$

Moreover, consider that each individual can obtain the following latent log wages ($\ln(y_j^i) \forall j = 0, 1, 2$) according to the level of education achieved:

$$\ln(y_0^i) = \alpha(x^i) + u_0^i \quad \text{for the low level}$$

$$\ln(y_1^i) = \alpha(x^i) + \bar{\beta}_1 + u_1^i \quad \text{for the medium level}$$

$$\ln(y_2^i) = \alpha(x^i) + \bar{\beta}_2 + u_2^i \quad \text{for the high level}$$

where x are observable variables in the wage equations, whereas u_0 , u_1 and u_2 are unobservable variables that are assumed to be independent of x . $\bar{\beta}_1$ and $\bar{\beta}_2$ are constant. It is assumed that $\ln(y_j^i) \forall j = 0, 1, 2$ are defined for each individual and that these outcomes are independent across persons, i.e. that there are no interactions among agents. Then, the returns, in terms of wages, of achieving a certain level of education are:

$$\beta_{10}^i = \ln(y_1^i) - \ln(y_0^i) = \bar{\beta}_1 + u_1^i - u_0^i$$

$$\beta_{20}^i = \ln(y_2^i) - \ln(y_0^i) = \bar{\beta}_2 + u_2^i - u_0^i$$

$$\beta_{21}^i = \ln(y_2^i) - \ln(y_1^i) = \beta_{20}^i - \beta_{10}^i$$

As can be seen from previous equations, if there is a distribution of unobservable variables in the population there is also a distribution of returns across individuals. β_{10}^i and β_{20}^i denote the individual returns, in terms of wages, to achieve medium and high level of education as compared to the low level. β_{21}^i denotes the return to achieving the high level of education as compared to the medium level.⁴

However, for each individual we only observe the wages related to his/her schooling choice ($\ln(y_0^i)$ if $D_0^i = 1$, $\ln(y_1^i)$ if $D_1^i = 1$ and $\ln(y_2^i)$ if $D_2^i = 1$). Then, the observed log wages may be written as:

$$\begin{aligned}\ln(y^i) &= D_1^i \ln(y_1^i) + D_2^i \ln(y_2^i) + D_0^i \ln(y_0^i) = \\ &= D_1^i \ln(y_1^i) + D_2^i \ln(y_2^i) + (1 - D_1^i - D_2^i) \ln(y_0^i) = \\ &= \alpha(x^i) + (\bar{\beta}_1 + u_1^i - u_0^i) D_1^i + (\bar{\beta}_2 + u_2^i - u_0^i) D_2^i + u_0^i = \\ &= \alpha(x^i) + \beta_{10}^i D_1^i + \beta_{20}^i D_2^i + u_0^i\end{aligned}\tag{1}$$

where the coefficients of education variables are allowed to be different across individuals. For simplicity the individual superscript (i) will be dropped from now on.

Firstly, consider that returns are the same among individuals ($u_1^i = u_0^i = u_2^i$), i.e. that β_{10}^i and β_{20}^i will be constant. Let z denote a set of observed variables that affect the probability of choosing each level of education and that are independent of the unobserved terms (u_0, u_1, u_2). Then, the following expression for the expected value of $\ln(y)$ conditional on z and x is valid in this case:

$$\begin{aligned}E(\ln(y)|z, x) &= \alpha(x) + E(\beta_{10} \cdot D_1|z, x) + E(\beta_{20} \cdot D_2|z, x) = \\ &= \alpha(x) + \beta_{10} \Pr(D_1 = 1|z, x) + \beta_{20} \Pr(D_2 = 1|z, x)\end{aligned}$$

This is a linear function of the propensity scores ($\Pr(D_1 = 1|z, x)$, $\Pr(D_2 = 1|z, x)$). The result would be the same in case of mean independence between D_1 , D_2 and the distribution of $(u_1 - u_0)$ and $(u_2 - u_0)$. That is, if there is heterogeneity in the

⁴In these expressions it is assumed that there is no observed heterogeneity in the returns. This is an assumption normally made in the returns to schooling literature. Carneiro, Heckman and Vytlačil (2003) considered that the only observed heterogeneity in the returns came from the scores of the individuals in different tests. I do not consider this case given that I do not have this information for the empirical application section.

returns to schooling but it does not affect the decisions of the individuals. In both cases, the only econometric problem that could arise when estimating (1) is due to the correlation between the decisions (D_1 and D_2) and the error term (u_0) (ability bias or measurement error bias). This problem can be solved using instrumental variables methods.

On the other hand, assume that there is a distribution of returns in the population, then:⁵

$$\begin{aligned} E(\ln(y)|z, x) &= \alpha(x) + E(\beta_{10}D_1|z, x) + E(\beta_{20}D_2|z, x) = \\ &= \alpha(x) + E((\bar{\beta}_1 + (u_1 - u_0))D_1|z, x) + E((\bar{\beta}_2 + (u_2 - u_0))D_2|z, x) = \\ &= \alpha(x) + \bar{\beta}_1 \Pr(D_1 = 1|z, x) + \bar{\beta}_2 \Pr(D_2 = 1|z, x) + K(\Pr(D_1 = 1|z, x), \Pr(D_2 = 1|z, x)) \end{aligned}$$

where

$$\begin{aligned} K(\Pr(D_1 = 1|z, x), \Pr(D_2 = 1|z, x)) &= \\ &= E((u_1 - u_0)|D_1 = 1, \Pr(D_1 = 1|z, x), \Pr(D_2 = 1|z, x)) \Pr(D_1 = 1|z, x) + \\ &+ E((u_2 - u_0)|D_2 = 1, \Pr(D_1 = 1|z, x), \Pr(D_2 = 1|z, x)) \Pr(D_2 = 1|z, x) \end{aligned}$$

This is an unknown function of $\Pr(D_1 = 1|z, x)$ and $\Pr(D_2 = 1|z, x)$. So, to test the heterogeneity of the returns to schooling a linearity test of $E(\ln(y)|z, x)$ with respect to $\Pr(D_1 = 1|z, x)$ and $\Pr(D_2 = 1|z, x)$ can be performed.⁶ Alternative methods recently developed in the nonparametric statistics literature are applied, in the empirical application to test for the homogeneity of sequential returns.

3 Marginal Treatment Effects

If returns vary with unobserved ability and thus there is a distribution of returns in the population, we need to find a manner in which to summarize this random variable.

⁵In obtaining this expression the following is assumed:

$$E(\ln(y)|x, \Pr(D_1 = 1|z, x), \Pr(D_2 = 1|z, x)) = E(\ln(y)|x, z)$$

This assumption can be obtained from a version of the index sufficiency restriction assumption frequently used in the selection models' literature (see, e.g., Powell (1994)).

⁶The same result for the binary choice case was presented in Carneiro, Heckman and Vytlačil (2003).

The traditional treatment parameters defined in the literature: the average return for a person randomly drawn from the population (Average Treatment Effect (ATE)), the average return for those that achieved a certain level of education (Treatment on the Treated (TT)), the average return for those that decided not to achieve a certain level of education (Treatment on the Untreated (TUT)), as well as, different instrumental variables parameters will give different summary measures of the distribution. The problem is that, these parameters will not in general answer the policy questions we may have in mind.⁷ It is in this context where the MTE grows in importance because it can be useful to obtain the desired summary measures. This is so because the MTE gives us the average return for those who are likely to be affected by policies. Moreover, under the appropriate assumptions, all the conventional treatment effect parameters can be estimated as weighted averages of the MTE.⁸

The MTE was defined in a binary choice framework in the following way. Consider two levels of education (high school (0) and college (1)) and the following outcomes that correspond to high school and college, respectively, $\ln(Y_0^i)$ and $\ln(Y_1^i)$. Define also the binary variable D such that:

$$D^i = \begin{cases} 0 & \text{if the individual left with the high school level} \\ 1 & \text{if the individual left with the college level} \end{cases}$$

and assume that the following latent variable model generates the variable D in this way:⁹

$$\begin{aligned} D_*^i &= \mu_D(z^i) - U_D^i \\ D^i &= 1 \text{ if } D_*^i \geq 0, = 0 \text{ otherwise.} \end{aligned}$$

⁷See Carneiro, Heckman and Vytlačil (2003).

⁸See, Heckman and Vytlačil (1999, 2000).

⁹The assumption of the existence of this kind of unobserved index that determines selection is equivalent to the assumptions needed in the LATE approach. In this approach we need to invoke independence and monotonicity assumptions to be able to interpret linear instrumental variables estimates as a weighted average of treatment effects. Vytlačil (2002) shows that, given the additive separability assumption, latent index models imply LATE assumptions. On the other hand, he shows that, given LATE assumptions, there always exists a selection model that rationalizes the observed and counterfactual data. Then, LATE assumptions are not weaker than the assumption of a latent index model, they impose the same restrictions on the counterfactual data, when nonparametric functional forms or distributional assumptions are invoked.

D_*^i denotes the net utility or gain to the decision-maker from choosing going to college. z^i is a set of covariates that includes valid instrumental variables and U_D^i is an unobserved random variable. Without loss of generality U_D will be assumed to be uniformly distributed between 0 and 1. As in previous section, the observed wages have the following form:

$$\ln(Y^i) = D^i \ln(Y_1^i) + (1 - D^i) \ln(Y_0^i)$$

Then, the MTE is defined in this context as:

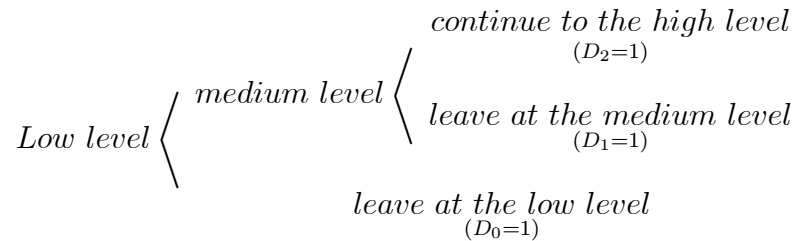
$$MTE(u) \equiv E(\ln(Y_1) - \ln(Y_0) | U_D = u)$$

If the instrument is externally set such that $p(z) = u$, where $p(z) = \Pr(D = 1 | z)$, this is the average return of going to college for those in the margin of indifference between going or not. In order to estimate the MTE, the Local Instrumental Variable Parameter (LIV) is used. This treatment parameter is defined as:

$$LIV(p(z)) = \frac{\partial E(\ln(Y) | P(z) = p(z))}{\partial p(z)}$$

LIV is the limit version of the Local Average Treatment Effect (LATE) parameter introduced by Imbens and Angrist (1994) and, given the previous index model, is equal to $MTE(p(z))$. One thing to notice is that under homogeneity or if returns are heterogeneous but people is not sorted in each educational level accordingly $MTE(p(z))$ will be the same than estimated returns using instrumental variables methods.

In this section, expressions for the MTE in the case of having multiple levels of education and decisions about enrolling in each one made in a sequential fashion are presented. Consider three levels of education and that people make their schooling decisions sequentially:



This model takes into account that decisions are made at different moments in time. Firstly ($t = 0$), everybody achieves the low level of education. Then, ($t = 1$), the individual has to decide whether to leave school at a low level or to continue to a medium level. Finally ($t = 2$), the individual decides to attend a high level of education or to leave at the medium level. Then, define as $\pi_0(z)$ the probability of continuing studying from the low level of education and as $\pi_1(z')$ the probability of continuing studying from the medium level for those who at least achieved this level.¹⁰ Given that the decision of continuing studying from the medium level of education takes place in a second moment of time the set of variables in z' can be different or can have different values compared to the variables in z . Then, the probabilities of achieving each level of education can be written as:

$$\begin{aligned}\Pr(D_0 = 1 \mid z, z') &= (1 - \pi_0(z)) \\ \Pr(D_1 = 1 \mid z, z') &= \pi_0(z)(1 - \pi_1(z')) \\ \Pr(D_2 = 1 \mid z, z') &= \pi_0(z)\pi_1(z')\end{aligned}$$

With this notation, consider the equation for observed log wages and rewrite it in the following way:

$$\begin{aligned}\ln(y) &= \alpha(x) + \beta_{10}D_1 + \beta_{20}D_2 + u_0 = \\ &= \alpha(x) + \beta_{10}(D_1 + D_2) + (\beta_{20} - \beta_{10})D_2 + u_0 = \\ &= \alpha(x) + \beta_{10}(\tilde{D}_1) + \beta_{21}D_2 + u_0\end{aligned}\tag{2}$$

where $\tilde{D}_1$ is equal to $D_1 + D_2$. In the context of our application $\tilde{D}_1$ takes the value one if the individual overcame the low level of education. Given the sequential model the probability of this even was defined as $\pi_0(z)$. Moreover, D_2 takes a value of one if the individual achieved the high level of education. According to the sequential decision model the probability of this occurring is equal to $\pi_0(z)\pi_1(z')$.

As in the binary case, consider that individuals will decide to leave with the low level of education ($\tilde{D}_1 = 0$), in a first stage, if the following is true:

¹⁰See Appendix part D for an example of an economic model that can be behind these probabilities.

$$D_1^* = \gamma_1^*(z) - \varepsilon_1 < 0$$

that is, if the expected net utility of continuing studying from the low level of education is less than 0. z are observed variables that affect the decision of continuing studying in a first moment of time.

On the other hand, consider that individuals, that achieved the medium level of education, decide to acquire the high level according to the following latent variable model:

$$\begin{aligned} D_2^* &= \gamma_2^*(z') - \varepsilon_2 \\ D_2 &= 1 \text{ if } D_2^* \geq 0 \text{ and } \tilde{D}_1 = 1 \end{aligned}$$

D_2^* represents the expected net utility of achieving the high level of education. In this stage individuals make the decision with the information they have in this second moment of time and so, the value of the variables in z' can be different to the value of the variables in z .¹¹ Without loss of generality, we can assume that ε_1 and ε_2 are uniformly distributed between 0 and 1 in which case we have: $\gamma_1^*(z) = \pi_0(z)$ and $\gamma_2^*(z') = \pi_1(z')$.

Then, taking expectations in (2) conditional to $\pi_1(z')$ and $\pi_0(z)$:¹²

¹¹It is assumed that individuals make their decisions about continuing studying with the information that is common knowledge at each moment of time.

¹²In this result it is assumed that:

$$E(\ln(y)|x, z, z') = E(\ln(y)|x, \gamma_1^*(z), \gamma_2^*(z'))$$

and that $Supp(\{\varepsilon_j\}) = \Re \forall j = 1, 2$. If this is the case:

$$E(\ln(y)|x, z, z') = E(\ln(y)|x, \pi_0(z), \pi_1(z'))$$

This a version of the index sufficiency restriction usually exploited in the literature about semi-parametric estimation of selection models (See Powell (1994)).

$$\begin{aligned}
& E[\ln(y)|x, \pi_1(z'), \pi_0(z)] = \\
& = \alpha(x) + E[\beta_{10}\tilde{D}_1|x, \pi_1(z'), \pi_0(z)] + E[\beta_{21}D_2|x, \pi_1(z'), \pi_0(z)] = \\
& = \alpha(x) + E[(\bar{\beta}_1 + (u_1 - u_0))\tilde{D}_1|x, \pi_1(z'), \pi_0(z)] + \\
& \quad + E[(\bar{\beta}_{21} + (u_2 - u_1))D_2|x, \pi_1(z'), \pi_0(z)] = \\
& = \alpha(x) + \bar{\beta}_1\pi_0(z) + \bar{\beta}_{21}\pi_0(z)\pi_1(z') + E[(u_1 - u_0)\tilde{D}_1|\pi_1(z'), \pi_0(z)] + \\
& \quad + E[(u_2 - u_1)D_2|\pi_1(z'), \pi_0(z)]
\end{aligned} \tag{3}$$

Where:

$$\begin{aligned}
E[(u_1 - u_0)\tilde{D}_1|\pi_1(z'), \pi_0(z)] & = E_{\tilde{D}_1}(E[(u_1 - u_0)|\pi_1(z'), \pi_0(z), \gamma_1^*(z) - \varepsilon_1 \geq 0]) = \\
& = \int_0^{\gamma_1^*(z)} \int_{-\infty}^{\infty} (u_1 - u_0)f(u_1 - u_0|\varepsilon_1 = r)d(u_1 - u_0)dr \\
& = \int_0^{\pi_0(z)} \int_{-\infty}^{\infty} (u_1 - u_0).f(u_1 - u_0|\varepsilon_1 = r)d(u_1 - u_0)dr
\end{aligned}$$

$$\begin{aligned}
& E[(u_2 - u_1)D_2|\pi_1(z'), \pi_0(z)] = \\
& = E_{D_2}(E[(u_2 - u_1)|\pi_1(z'), \pi_0(z), \gamma_1^*(z) - \varepsilon_1 \geq 0, \gamma_2^*(z') - \varepsilon_2 \geq 0]) = \\
& = \int_0^{\gamma_2^*(z')} \int_0^{\gamma_1^*(z)} \int_{-\infty}^{\infty} (u_2 - u_1)f(u_2 - u_1|\varepsilon_1 = r_1, \varepsilon_2 = r_2)d(u_2 - u_1)dr_1dr_2 \\
& = \int_0^{\pi_1(z')} \int_0^{\pi_0(z)} \int_{-\infty}^{\infty} (u_2 - u_1)f(u_2 - u_1|\varepsilon_1 = r_1, \varepsilon_2 = r_2)d(u_2 - u_1)dr_1dr_2
\end{aligned}$$

So,

$$\begin{aligned}
E[\ln(y)|\pi_1(z'), \pi_0(z)] & = \alpha(x) + \bar{\beta}_1\pi_0(z) + \bar{\beta}_{21}\pi_0(z)\pi_1(z') + \\
& \quad + \varphi_0(\pi_0(z), \pi_1(z'))\pi_0(z) + \varphi_1(\pi_0(z), \pi_1(z'))\pi_0(z)\pi_1(z')
\end{aligned} \tag{4}$$

where $\varphi_0(\pi_0(z), \pi_1(z'))$ and $\varphi_1(\pi_0(z), \pi_1(z'))$ are general functions of the propensity scores $(\pi_0(z), \pi_1(z'))$.

In this context it is natural to define the following MTEs:

$$\begin{aligned} MTE_0(u, \pi_1(z')) &\equiv E(\ln(y_{0*}^i) - \ln(y_0^i) | \varepsilon_1 = u, \pi_1(z')) \\ MTE_1(\pi_0(z), u) &\equiv E(\ln(y_2^i) - \ln(y_1^i) | \varepsilon_2 = u, \pi_0(z), \tilde{D}_1 = 1) \end{aligned}$$

where $\ln(y_{0*}^i)$ denotes the logarithm of wages in the final best option in the set of all choices except the low level of education ($\ln(y_{0*}^i) = D_2^i \ln(y_2^i) + D_1^i \ln(y_1^i)$). If u is set equal to $\pi_0(z)$ $MTE_0(u, \pi_1)$ denotes the average return to continue studying from the low level of education for those who are in the margin of indifference between continuing or not. $MTE_1(\pi_0, u)$ denotes the average return of achieving the high level of education for those who achieved at least the medium level but that do not have a preference about continuing after this level when u equals $\pi_1(z')$.

As in the binary case, point-wise estimates of these MTEs can be obtained from the following Local Instrumental Variable parameters:

$$\begin{aligned} LIV_0(\pi_0(z), \pi_1(z')) &= \frac{\partial E[\ln(y) | \pi_1(z'), \pi_0(z)]}{\partial \pi_0(z)} = \bar{\beta}_1 + \bar{\beta}_{21} \cdot \pi_1(z') + \\ &\quad + \int_{-\infty}^{\infty} (u_1 - u_0) \cdot f(u_1 - u_0 | \varepsilon_1 = \pi_0(z)) \cdot d(u_1 - u_0) + \\ &\quad + \int_0^{\pi_1(z')} \int_{-\infty}^{\infty} (u_2 - u_1) \cdot f(u_2 - u_1 | \varepsilon_1 = \pi_0(z), \varepsilon_2 = r_2) \cdot d(u_2 - u_1) dr_2 = \\ &= MTE_0(\pi_0(z), \pi_1(z')) \end{aligned}$$

$$\begin{aligned} LIV_1(\pi_0(z), \pi_1(z')) &= \frac{\partial E[\ln(y) | \pi_1(z'), \pi_0(z)]}{\partial \pi_1(z')} = \bar{\beta}_{21} \cdot \pi_0(z) + \\ &\quad + \int_0^{\pi_0(z)} \int_{-\infty}^{\infty} (u_2 - u_1) \cdot f(u_2 - u_1 | \varepsilon_1 = r_1, \varepsilon_2 = \pi_1(z')) \cdot d(u_2 - u_1) \cdot dr_1 = \\ &= MTE_1(\pi_0(z), \pi_1(z')) \end{aligned}$$

For the validity of this result it is needed to be able to change $\pi_0(z)$ without changing $\pi_1(z')$ and *vice versa*. One way to obtain this variation in $\pi_0(z)$ without varying $\pi_1(z')$

is with exclusion restrictions. That is, having a variable in z that it is not in z' . On the other hand, we can have the variation needed in $\pi_1(z')$ given the sequential nature of the decisions. That is, given that the decision about achieving the high level or stopping at the medium level of education takes place in a second period and so it can be an unexpected shock that changes it without changing the decision made in a first stage.¹³ So, the variation of supply-side instruments over time can be used to obtain this result. The expansion of the number of universities that took place in Spain, specially in the years between 1970 and 1990, is an example of this kind of variation. Note that what we can obtain are point-wise estimates for the values of $\pi_0(z)$ and $\pi_1(z')$ in the support of the data set used. Finally, it should be pointed out that in this case, in contrast with the binary case, in spite of assuming homogeneity of the returns estimated returns using IV methods will not coincide with previous MTEs. If we apply instrumental variable methods using two valid instruments into equation (2) we will obtain estimates of: $\bar{\beta}_1$ and $(\bar{\beta}_2 - \bar{\beta}_1)$. Whereas MTEs parameters have the following form under homogeneity:

$$\begin{aligned} MTE_0(\pi_0(z), \pi_1(z')) &\equiv \bar{\beta}_1 + (\bar{\beta}_2 - \bar{\beta}_1)\pi_1(z') \\ MTE_1(\pi_0(z), \pi_1(z')) &\equiv (\bar{\beta}_2 - \bar{\beta}_1)\pi_0(z) \end{aligned}$$

4 Some facts concerning Spanish education.

Spain is the country of the OECD (Organization for Economic Co-operation and Development) that has undergone, in recent years, the most spectacular educational transformation. The average level of education in the population has increased generation by generation as shown in Table 1. This table shows the proportion of individuals in each finished level of education for five-year cohorts of those born between 1918 and 1967. As one can see in this table the level of illiteracy has been falling dramatically while the proportion of people with a college degree increased from 5.68 % for those born between 1918 and 1922 to 16.14 % for those who were born between

¹³Taber (2000) showed that the same kind of exclusion restrictions, along with a monotonic transformation and strong support conditions, were needed for the semiparametric identification of general sequential discrete choice models.

1953 to 1957. Then, this proportion fell down to 14.87%, for those born between 1958 and 1962, and to 13.13%, for those born between 1963 and 1967. This can be due to the fact that a great proportion of those who decided to go to college in these cohorts had not finished studying in the moment the survey was carried out. One thing to notice is that the greatest proportion of people left school when they were 10 years old for those who were born between 1918 and 1957, whereas those born between the years 1958 and 1967 left school at 14 years old. This phenomenon took place at the same time as the 1970 reform of the Spanish educational system.

The 1970 reform of the Spanish educational system raised the age of compulsory schooling from 10 to 14 years old. Yet another reform, in 1990, raised this age to 16 years. Figure 1 shows the main characteristics of the system previous to the 1970 reform and the one between 1970 and 1990.¹⁴ These are the relevant ones for the individuals in the data set used.

Concerning the supply side, the number of universities has increased a lot specially in the years between 1970 and 1990. Before 1970 there were 19 universities in Spain. This figure increased to 33 in the 1970's and to 38 in 1980's. Finally, the number of universities increased to 68 in the 1990's. It should be pointed out, however, that Spanish universities offer two kind of degrees: the university disciplines, that are considered traditional in other countries and that are designed to last for between four and six years, and studies that are completed in a three year duration. The latter kind of study may be to train nurses, teachers, physiotherapists etc. and are considered vocational studies in many other countries. Most of the new universities were located in places where three-year courses were already instructed. As a result of this, the creation of universities affected mainly the decision of accessing four-year college degrees. This phenomenon was strengthened because the number of four-year degrees offered in some existing universities also increased during the 60's and 70's. In fact, there are only five old universities that traditionally have offered a great variety of four-year courses.

In addition, no authoritative comparative ranking has even been drawn up of Spanish universities. Furthermore, students used to go to the university closest to

¹⁴These pictures are based on Pons and Gonzalo (2001).

home. In fact, they were not able to go to a different one by government decree. This did not changed until 1991, when the government, trying to introduce some competition in the hope of raising standards, allowed prospective students to choose between any of the universities in the region where they lived.

5 Data

This paper estimates returns to schooling based on the ECBC (*“Encuesta de Conciencia y Biografía de Clase”*) survey (1991).¹⁵ One peculiarity of this survey is that it has information related to the educational level of the individuals, earnings and family background. It should be stressed that this information is not found in other Spanish surveys. In particular, this survey has the following data about these topics:

- Level of education completed by the individuals and number of years completed in that level.
- Information on whether the individual obtained education prior to or after the reform of 1970.
- Information about the kind of studies pursued by those in vocational studies or at college.
- Retrospective data about the head of household at the age of 16, like the level of education and labor situation. The same information is available for the mother if she is not the head of household.
- Information about where the individual lived for the most part of his/her childhood.
- Wages.

It has been found that in most surveys with a size similar to the ECBC and that have been done with the same aim as this one, there is scant representation of top professionals, such as entrepreneurs or managers. This is the reason for over-representing

¹⁵See Appendix part A for a detailed description of this survey.

these categories in the ECBC. It was decided to overrepresent these individuals, with the help of data that is available in the Spanish Electoral Census (“*Censo Electoral*”) (1986). Specifically, the survey over-represented those with a medium or high level of education. This bias should be considered when estimating the model. In particular, the following weights ($w(s)$) are constructed for each individual depending on the stratum (s) he/she belongs to:

$$\begin{aligned} w(s) &= \frac{Q(s)}{H(s)} \\ w(s) &> 0 \end{aligned}$$

where $Q(s)$ denotes the population density of stratum (s) and $H(s)$ is the arbitrary distribution used when sampling the data. $Q(s)$ is constructed with the help of data in the 1991 Census. These weights eliminate the bias giving a low weight to over-sampled categories and *vice versa*. Then, each observation is weighted properly according to the estimation method used.¹⁶

6 Empirical Application

In this section I present estimates of returns to different levels of education using a sample from the ECBC survey (1991). Firstly, I assume that returns are homogenous among individuals of a same level of education or that they are heterogeneous but people is not making their schooling decisions considering their returns. Under this assumption estimates assuming that each additional unit of schooling has the same return as well as, of models where different levels of schooling have different returns are obtained. Then, alternative methods recently developed in the nonparametric statistics literature are applied to test if returns are heterogeneous and individuals are sorted in each level of education considering them. This test shows that heterogeneity is an important feature of the sample used. Then, estimates of the MTEs are obtained, both assuming normality as well as without imposing distributional assumptions. Finally, estimates of MTEs are used to analyze some effects of the 1970 reform of the Spanish educational system.

¹⁶See Appendix part B for a description of how these weights are constructed and used.

In order to obtain these estimates different schooling attainments are classified in the following three levels of education:

1. Low level. Illiterate people, people without studies and people who left school at the primary level (“*EGB*”, “*Graduado Escolar*”, “*Primaria*”, “*Bachiller Elemental*”) (at age 14 at most) are in this group.
2. Medium Level. People in this group completed at least one year of post-compulsory secondary education (vocational studies or high school) (“*FPI*”, “*FPII*”, “*Bachiller Superior*”, “*BUP*”, “*COU*”, “*Diplomado*”).¹⁷
3. High Level. People in this group completed at least one year of college.¹⁸

Concerning the instrumental variables used in the analysis, a first group of instruments considered is related to institutional sources of variation in schooling. In particular the following variables are used:¹⁹

- Distance to the nearest university. Variables about distance to college have been used as instruments in the literature of returns to schooling as they are a good measure of the cost of higher education.²⁰ These variables have been criticized as being good instruments because of the possibility of endogenous mobility decisions. However, regional migrations in Spain, for the period of the sample used, were mostly motivated by families moving from the country-side to big cities searching for a job. This, joint with the fact that individuals were obliged to go to the nearest university, makes of these variables good candidates for being instruments in the case of Spain.
- Information about whether the individual studied before the 1970 reform or after it.

¹⁷Given the similarity between three-year courses imparted in Spanish universities and what are considered vocational studies in many other countries, they are included in the medium level of education. This makes the results comparable to similar works with data for other countries. One problem with this classification is that individuals in the medium level that achieved a three-year college degree have studied a greater number of years than some dropouts in the high level. However, this problem does not seem to be relevant in the sample used as only 4.58% of those in the high level are in this situation.

¹⁸Courses imparted in Spanish universities that last for more than three years.

¹⁹See appendix part C for a description of these variables.

²⁰See for example Card (1999, 2001) and Carneiro, Heckman and Vytalil (2003).

The second group of instruments consists on variables which reflect the labor situation in the moment of decision of the individuals. In particular, each individual is assigned the family income and employment in the province of residence when they were making schooling decisions. Data used to construct these variables was obtained from Alcaide (2003). Finally, background information is used as instrument. In particular, I use information about the level of education and kind of job of the head of household when the individual was a child as well as, a dummy that takes value one if he was a wage earner doing a supervisory job. Although this kind of information has been criticized for being good instruments it turned out to be an extra source of identification and so, it is used in some of the following estimates.

6.1 Estimates under homogeneity

The traditional literature on returns to schooling treated returns as if they were homogenous. Basically, two kind of models have been considered in this context: models where each additional unit of schooling has the same return and more flexible models, where different levels of schooling have different returns. As an example of the first kind of models we find the popular Mincer's model of earnings (Mincer, 1974). Mincer's specification has been widely used over the last fifty years and continues to be applied in recent work. Table 3 shows estimates of the Mincer's model using the ECBC survey.²¹ The sample considered consists on individuals between 19 and 65 years old, who were working, who stated have finished studying and who reported a valid wage in the moment of the survey. Wages are measured per hour and net of taxes. Variables in x are: 11 dummies for Spanish regions, age, age squared, tenure,

²¹In this Table OLS estimates with weights and without weights are reported. Estimated returns using OLS with weighs were very similar to the ones without weights. Note that if we define the weighted estimator as:

$$\hat{\beta}_w = (X'WX)^{-1}(X'WY)$$

we have that the difference between the weighted and unweighted estimator has the following form:

$$\hat{\beta}_w - \hat{\beta}_{ols} = (X'WX)^{-1}(X'W(I - X(X'X)^{-1}X'))Y$$

So, a test for the significance of the difference between the two coefficients can be performed by testing if the coefficient of the weighted regression of the unweighted regression errors on X are different from zero (See Deaton (2000)). In this case, this hypothesis could not be rejected at a confidence level of 99% ($F(31, 1622) = 0.68$; $p - value = 0.91$).

tenure squared, a gender dummy that equals one if the individual is a woman, 4 occupational dummies and 10 industry dummies. The effect of these covariates on log wages is assumed linear for comparability with the literature.

GMM estimates are performed using as instruments: two variables with the income in the province of residence when the individual was 14 and 18 years old, the employment to population ratio in the province of residence when the individual was 14 years old, a dummy that takes the value 1 if the head of household, when the individual was a child, was a wage earner doing a supervisory job and a variable about the distance to the nearest university when the individual was 18 years old.²² One thing to notice is that IV estimates of the returns to schooling are greater than OLS estimates. This is a result frequently found in studies that use supply-side variables as instruments.

Previous estimates considered that each additional year of education has the same return. Now this assumption is relaxed considering that the three levels of education (low, medium and high) have different returns. However, returns are still assumed to be the same among individuals in the same level of education. Table 4 shows estimates for this model using OLS and GMM.²³ The set of instruments for the GMM estimates consists on: three dummies for the level of education (illiterate, primary and high school) and three dummies for the kind of job (agriculture, mining, blue collar job) of the head of household when the individual was a child, a dummy that takes a value of one if this person is a wage earner doing a supervisory job, a variable about the income in the province of residence when the individual was 14 years old and a variable for the distance to traditional universities when the individual was 18 years old. It should be stressed that under homogeneity it will be concluded that individuals will benefit from educational policies that encourage them to acquire higher levels of schooling.

²²See Appendix part C for details of how these variables are constructed.

²³Estimated returns to schooling using OLS without weights were similar to the ones using weights. The hypothesis of being equal could not be rejected at a confidence level of 99% ($F(32, 1620) = 0.76$; $p - value = 0.83$).

6.2 Schooling choices estimates

The sample considered to estimate the sequential decision model consists of individuals who were between 19 and 65 years old and stated have finished studying. Those who stated a level of education lower than the level registered in the 1986 Census or a level of education that implies an increment of more than four academic years were not considered.

As it can be seen in Table 2,²⁴ those with the high and medium level of education are on average younger, have more educated parents in higher qualified jobs and lived in richer regions during their childhood than those with the low level of education. On the other hand, the proportion of women is higher in the low level. Finally, those with a high level used to live in regions that were closer to a province that traditionally offered a great variety of four-year college degrees.

As it was shown in Section 3, for the identification of MTEs using LIV parameters it is needed to have independent variation of the probability of continuing studying after the low level of education ($\pi_0(z)$) and the probability of achieving the high level for those that completed the low level ($\pi_1(z)$). One way of obtaining this variation is imposing exclusion restrictions. In the following estimates variables about the family income in the moment of decision will be explanatory variables in $\pi_0(z)$ but not in $\pi_1(z)$. On the other hand, variables about regional employment rate in the moment of decision will be considered for estimates of $\pi_1(z)$ but not for the ones of $\pi_0(z)$. These variables are excluded as explanatory variables in these probabilities because they turned out not to be significant. On the other hand, another source of variation comes from the sequential nature of schooling decisions. It can be that something changes $\pi_1(z)$, in the second moment of decision, without changing $\pi_0(z)$. For the sample considered, the increment of the number of four-year degrees offered in existing universities during the latter 60's and 70's can be used as a source of this kind of variation. However, it could not be possible to take advantage of the creation of new universities that took place in Spain among the 80's and 90's due to the limitation in the number of Spanish household surveys with the required information.

In particular, two variables about the distance to universities are used. The first

²⁴See the Appendix part C for a description of the variables in this table.

one is a measure of the distance to universities which offer a great variety of four-year degrees. Madrid, Barcelona, Salamanca, Valladolid, and Granada are considered the provinces that have had such a university for those who were born before 1961. This group was augmented considering also Seville, Zaragoza, A Coruña, Asturias and Valencia for those who were born after 1961. This variable is used as an explanatory variable for obtaining estimates of $\pi_1(z)$. Within the explanatory variables in $\pi_0(z)$ I consider a dummy variable that takes value one if the individual was living far away from one of the provinces that traditionally had a university. In this case, Madrid, Barcelona, Salamanca, Valladolid and Granada are the provinces in this group for those who were born before 1966. The augmented group considering also Seville, Zaragoza, A Coruña, Asturias and Valencia was the relevant for those who were born after 1966. When using this definition of the variables I am considering that those who were born between 1961 and 1966 are the ones that experienced the augmentation of the number of degrees offered in the period in between the two educational decisions.

Table 5A shows estimates of $\pi_0(z)$, using a probit model with and without weighting. Table 5B allows different variables to have different effects depending on the individual studying before the 1970 reform or after it. Estimates of $\pi_1(z)$, using a probit model, are presented in Tables 6A and 6B. Again Table 6B allows different effects of the variables according to the different educational systems. As we can see in these tables, those who lived in a region with higher family income and who studied before the 1970 reform have a higher probability of continuing studying from the low level of education. Moreover, if the head of household when the individual was a child was illiterate, had primary studies or worked in the agriculture or manufacture sector this probability is smaller. Being a woman or leaving far away from one of the traditional universities has a negative effect on this probability especially for those studying before the 1970 reform. Concerning $\pi_1(z)$, if the head of household had a low level of education the probability of achieving the high level of education is lower. In addition, this probability is also smaller if the head of household had a job in the agriculture, mining or manufacture sector. Furthermore, living in a province nearer to one of the traditional universities has a positive effect on this probability. Finally, being a woman and studying before the 1970 reform has a negative effect on

the probability of achieving the high level of education. The sign of the estimated coefficients was the one expected for all variables except for the coefficient of age, for those who studied after the 1970 reform, in the estimation of $\pi_1(z)$. Age has a positive and significant effect in this case. This can be due to the fact that those in younger cohorts who achieved the high level of education would be still studying in the moment of the survey. It can be seen from a comparison of estimates with and without weights that weighting changes the estimated coefficients of the variables as well as their level of significance.

The support of the estimated probabilities $\pi_0(z)$ and $\pi_1(z)$, using estimates of Tables 5B and 6B, is shown in Figures 2 and 3. As it can be seen in these figures it is almost the full unit interval for $\pi_0(z)$ whereas it is concentrated in small values for $\pi_1(z)$. Moreover, at the high extreme of the cell of data it become very thin, especially for $\pi_1(z)$. This will make it hard to estimate some of the parameters defined over the full support of $\pi_0(z)$ and $\pi_1(z)$. However, it will still be possible to obtain point-wise estimates of interesting parameters for a wide range of evaluation points; these points being representative of the sample used. Estimated probabilities using estimated coefficients of Tables 5B and 6B are used when needed to obtain estimates presented in next sections.

6.3 Testing homogeneity

With the framework of Section 2 in hand, one way of testing the homogeneity of sequential returns to schooling consists in testing if $E(\log(y)|x, P_1, P_2)$ is linear with respect to the propensity scores ($P_1 = \Pr(D_1 = 1|z, x)$, $P_2 = \Pr(D_2 = 1|z, x)$).²⁵ P_1 and P_2 are considered to be the result of sequential education decisions made by individuals. In this section, the method developed by Stute (1997) and Stute, González Manteiga and Presedo Quindimil (1998), for testing the goodness of fit of a parametric regression model, is applied to test the homogeneity of returns.

Under the null hypothesis of homogeneous returns to schooling:

²⁵ Alternatively, one could consider equation (4) and test for the linearity of $E[\ln(y)|\pi_1(z), \pi_0(z)]$ with respect to $\pi_0(z)$ and $\pi_1(z)$.

$$\begin{aligned} E(\log(y)|x, P_1, P_2) &= \\ &= \alpha(x) + \bar{\beta}_1 P_1 + \bar{\beta}_2 P_2 \end{aligned}$$

If this is the case, we can write:

$$\log(y) = \alpha(x) + \bar{\beta}_1 P_1 + \bar{\beta}_2 P_2 + \zeta$$

where ζ is an error term such that:

$$E(\zeta|x, P_1, P_2) = 0$$

So, The null hypothesis we want to test is:

$$H_0 : E(\zeta|x, P_1, P_2) = 0$$

against the alternative hypothesis:

$$H_1 : E(\zeta|x, P_1, P_2) = K(P_1, P_2)$$

where $K(\cdot)$ is a general function on the propensity scores.²⁶

If H_0 is true then we will have that:

$$E(K(P_1, P_2)\zeta) = 0$$

That is, the error term ζ is uncorrelated with any function of the propensity scores. Different methods have been used to test for this kind of hypothesis. For example, estimates of the conditional expectation of ζ , in a completely non parametric framework, can be obtained and tests for its dependence on the propensity scores could be performed. However, the main disadvantage of this method is that it requires smoothing of the data and leads to less precise fits in general.²⁷ Other methods used approximate $K(\cdot)$ using polynomials or assume different distributional forms for it. A problem using this approach is the lack of consistency on the whole H_1 . Stute

²⁶Here it is assumed that all the dependence on x is captured in a general fashion by $\alpha(x)$.

²⁷See Stute (1997).

(1997) and Stute, González Manteiga and Presedo Quindimil (1998) developed a test that overcomes these problems. The idea underlying this test is to capture all the dependence between the error terms and the propensity scores using indicator functions and then summarize it taking an average measure.

The test proposed by Stute (1997) is based on the integrated regression function that, in this context, has the following form:

$$I(\tilde{P}_1, \tilde{P}_2) = \int_0^{\tilde{P}_2} \int_0^{\tilde{P}_1} (\alpha(x) + m(P_1, P_2)) dF(P_1, P_2)$$

Where $\tilde{P}_1$ and $\tilde{P}_2$ are given values of the propensity scores, F is the distribution function of these probabilities and $m(P_1, P_2)$ is:

$$m(P_1, P_2) = \bar{\beta}_1 P_1 + \bar{\beta}_2 P_2$$

Tests of H_0 can be based on $I(\tilde{P}_1, \tilde{P}_2)$ because this function uniquely determines $\alpha(x) + m(P_1, P_2)$. A consistent estimator of $I(\tilde{P}_1, \tilde{P}_2)$ for a sample of size n is:

$$\hat{I}(\tilde{P}_1, \tilde{P}_2) = \frac{1}{n} \sum_{i=1}^n 1(P_1^i \leq \tilde{P}_1, P_2^i \leq \tilde{P}_2) (\log(y^i))$$

with this estimator the following function is constructed:

$$\hat{R}(\tilde{P}_1, \tilde{P}_2) = n^{-1/2} \sum_{i=1}^n 1(P_1^i \leq \tilde{P}_1, P_2^i \leq \tilde{P}_2) [\log(y^i) - \hat{\alpha}(x^i) - \hat{m}(P_1^i, P_2^i)]$$

Where $1(A)$ is the indicator function for the event A occurring. Then, mainly two statistic tests are used in order to test for H_0 : The Kolmogorov-Smirnov and Cramér-Von Mises statistics which has, respectively, the following form in this context:

$$\hat{D} = \sup_x |\hat{R}(\tilde{P}_1, \tilde{P}_2)|$$

and

$$\hat{W}^2 = \frac{1}{\hat{\sigma}^2 n^2} \sum_{i=1}^n \hat{R}(\tilde{P}_1, \tilde{P}_2)^2$$

where

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n [\log(y^i) - \hat{\alpha}(x^i) - \hat{m}(P_1^i, P_2^i)]^2$$

Finally, the null hypothesis is rejected if these statistics exceed a critical value. Given the complexity of the asymptotic distribution of $\hat{R}(\tilde{P}_1, \tilde{P}_2)$ bootstrap technics are required to obtain it. Stute, González Manteiga and Presedo Quindimil (1998) found that the wild bootstrap scheme yield a valid approximation for the distribution of $\hat{R}(\tilde{P}_1, \tilde{P}_2)$, whereas, classical bootstrap, smooth bootstrap and residual based bootstrap do not, in general.

Table 7 shows the results for the homogeneity test with the sample used in Section 6.1. 500 replications using wild bootstrap were performed to obtain the distribution of Crámer Von-Mises and Kolmogorov statistics.²⁸ The hypothesis of linearity on the propensity scores is rejected at a confidence level of 95% for the Crámer Von-Mises and at a confidence level of 90% for the Kolmogorov statistic.²⁹ Given this result, the hypothesis of heterogeneity in the returns to schooling and individuals deciding to attend different levels of education with some knowledge of them cannot be rejected. That is, selection bias is relevant in the sample considered.

6.4 Marginal Treatment Effects

6.4.1 Control function Approach

From latter estimates we saw that selection bias was an important feature of the sample used. One popular approach for dealing with selection bias consists on assuming a normal distribution for the error terms of the model. Although this approach has been criticized for its reliance on distributional assumptions and lack of robustness when departures from normality it is a natural point of reference for the estimates of the marginal returns in this context.

²⁸Wild bootstrap is performed obtaining draws from the following distribution for the residuals:

$$\zeta_i = \begin{cases} \frac{(1-\sqrt{5})}{2} \hat{\zeta}_i & \text{with probability } \frac{(1+\sqrt{5})}{2\sqrt{5}} \\ \frac{(1+\sqrt{5})}{2} \hat{\zeta}_i & \text{with probability } 1 - \frac{(1+\sqrt{5})}{2\sqrt{5}} \end{cases}$$

where $\hat{\zeta}_i = (\log(y^i) - \hat{\alpha}(x^i) - \hat{m}(P_1^i, P_2^i))$

²⁹In order to perform this test, $\alpha(x)$ is assumed to be linear on x . Then, $\log(y) - \alpha(x) - m(P_1, P_2)$ is estimated by the residuals of a linear regression of log wages on the explanatory variables and the propensity scores.

Consider the equation for $E[\ln(y)|x, \pi_1(z'), \pi_0(z)]$ (equation (3)) and assume that the error terms in the model have the following joint distribution:³⁰

$$\begin{pmatrix} u_i \\ \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{i,\varepsilon_1} & \rho_{i,\varepsilon_2} \\ \rho_{i,\varepsilon_1} & 1 & 0 \\ \rho_{i,\varepsilon_2} & 0 & 1 \end{pmatrix} \right) \quad i = 0, 1, 2$$

Given this assumption and the latent variable models assumed for the schooling decisions we have that:

$$\begin{aligned} E((u_1 - u_0)|\tilde{D}_1 = 1, D_2 = 0) &= E((u_1 - u_0)|\gamma_1^*(z) \geq \varepsilon_1, \gamma_2^*(z') < \varepsilon_2) = \\ &= (\rho_{1,\varepsilon_1} - \rho_{0,\varepsilon_1}) \left(\frac{-\phi(\gamma_1^*(z))\Phi(-\gamma_2^*(z'))}{\Phi(\gamma_1^*(z))\Phi(-\gamma_2^*(z'))} \right) + (\rho_{1,\varepsilon_2} - \rho_{0,\varepsilon_2}) \left(\frac{\phi(\gamma_2^*(z'))\Phi(\gamma_1^*(z))}{\Phi(\gamma_1^*(z))\Phi(-\gamma_2^*(z'))} \right) = \\ &= (\rho_{1,\varepsilon_1} - \rho_{0,\varepsilon_1}) \left(\frac{-\phi(\gamma_1^*(z))}{\Phi(\gamma_1^*(z))} \right) + (\rho_{1,\varepsilon_2} - \rho_{0,\varepsilon_2}) \left(\frac{\phi(\gamma_2^*(z'))}{1 - \Phi(\gamma_2^*(z'))} \right) \end{aligned}$$

and

$$\begin{aligned} E((u_1 - u_0)|\tilde{D}_1 = 1, D_2 = 1) &= E((u_2 - u_0)|\gamma_1^*(z) \geq \varepsilon_1, \gamma_2^*(z') \geq \varepsilon_2) = \\ &= (\rho_{2,\varepsilon_1} - \rho_{0,\varepsilon_1}) \left(\frac{-\phi(\gamma_1^*(z))}{\Phi(\gamma_1^*(z))} \right) + (\rho_{2,\varepsilon_2} - \rho_{0,\varepsilon_2}) \left(\frac{-\phi(\gamma_2^*(z'))}{\Phi(\gamma_2^*(z'))} \right) \end{aligned}$$

Then:

$$\begin{aligned} \ln(y) &= \alpha(x) + \bar{\beta}_1 \tilde{D}_1 + (\bar{\beta}_2 - \bar{\beta}_1) D_2 + (\rho_{1,\varepsilon_1} - \rho_{0,\varepsilon_1}) \left(\frac{-\phi(\gamma_1^*(z))}{\Phi(\gamma_1^*(z))} \right) D_1 + \\ &+ (\rho_{1,\varepsilon_2} - \rho_{0,\varepsilon_2}) \left(\frac{\phi(\gamma_2^*(z'))}{1 - \Phi(\gamma_2^*(z'))} \right) D_1 + (\rho_{2,\varepsilon_1} - \rho_{0,\varepsilon_1}) \left(\frac{-\phi(\gamma_1^*(z))}{\Phi(\gamma_1^*(z))} \right) D_2 + \\ &(\rho_{2,\varepsilon_2} - \rho_{0,\varepsilon_2}) \left(\frac{-\phi(\gamma_2^*(z'))}{\Phi(\gamma_2^*(z'))} \right) D_2 + \zeta \end{aligned}$$

where $E(\zeta|x, z, z') = 0$. The second column of Table 8 shows estimates of this equation using instrumental variables methods. These estimates are valid in presence of

³⁰ As it was assumed that:

$$f(\varepsilon_2|\tilde{D}_1 = 1) \sim N(0, 1)$$

when estimating the sequential decision model, implicitly we are assuming that ε_2 and ε_1 are independent.

selection bias as well as ability bias or measurement error bias. On the other hand, if the only important bias comes because of the heterogeneity of the returns we do not need to use instrumental variables methods. Estimates for this case are presented in the first column of the same table.

The marginal treatment effects have the following form given the normality distribution assumption:³¹

$$\begin{aligned} MTE_0(\pi_0(z), \pi_1(z')) &= \bar{\beta}_1 + (\bar{\beta}_2 - \bar{\beta}_1)\pi_1(z') + (\rho_{1,\varepsilon_1} - \rho_{0,\varepsilon_1})\Phi^{-1}(\pi_0(z))(1 - \pi_1(z')) + \\ &+ (\rho_{2,\varepsilon_1} - \rho_{0,\varepsilon_1})\Phi^{-1}(\pi_0(z))\pi_1(z') + (\rho_{1,\varepsilon_2} - \rho_{0,\varepsilon_2}) \left(\frac{\phi(\Phi^{-1}(\pi_1(z')))}{1 - \pi_1(z')} \right) (1 - \pi_1(z')) \\ &+ (\rho_{2,\varepsilon_2} - \rho_{0,\varepsilon_2}) \left(\frac{-\phi(\Phi^{-1}(\pi_1(z')))}{\pi_1(z')} \right) \pi_1(z') \end{aligned}$$

$$\begin{aligned} MTE_1(\pi_0(z), \pi_1(z')) &= (\bar{\beta}_2 - \bar{\beta}_1)\pi_0(z) + ((\rho_{2,\varepsilon_1} - \rho_{0,\varepsilon_1}) - (\rho_{1,\varepsilon_1} - \rho_{0,\varepsilon_1})) \\ &\left(\frac{-\phi(\Phi^{-1}(\pi_0(z)))}{\pi_0(z)} \right) \pi_0(z) + ((\rho_{2,\varepsilon_2} - \rho_{0,\varepsilon_2}) - (\rho_{1,\varepsilon_2} - \rho_{0,\varepsilon_2}))\Phi^{-1}(\pi_1(z'))\pi_0(z) \end{aligned}$$

If we compare latter expression with the one that would be derived in case of having only two levels of education the main differences are in the presence of the second term and of $\pi_0(z)$ in the rest of terms.³² This is the way the selection in the first level of education is being considered. Figure 4 and Figure 5 show the MTEs: $MTE_0(\pi_0(z), \pi_1(z') = 0.10)$ and $MTE_1(\pi_0(z) = 0.43, \pi_1(z'))$. Their values are based on the GMM estimated coefficients reported in Table 8.

6.4.2 Semiparametric Estimation Approach

The objective of this section is to perform estimates of the Marginal Treatment Effects without imposing arbitrary distributional assumptions. In order to do this, estimates of $\pi_0(z)$ and $\pi_1(z')$ that arise from the sequential model are used but non assumptions on the distribution of $(u_i, \varepsilon_1, \varepsilon_2)$ ($i = 0, 1, 2$) are invoked.

Then,

³¹To obtain these expressions we need to use the fact that: $\gamma_1^*(z) = \Phi^{-1}(\pi_0(z))$ and $\gamma_2^*(z) = \Phi^{-1}(\pi_1(z'))$.

³²See, Heckman, Tobias and Vytlačil (2001) for the binary case expressions.

$$\begin{aligned}
K(\pi_0(z), \pi_1(z')) &= \bar{\beta}_1 \pi_0(z)(1 - \pi_1(z')) + \bar{\beta}_2 \pi_0(z) \pi_1(z') + \\
&+ E((u_1 - u_0) | \tilde{D}_1 = 1, (1 - D_2) = 1, \pi_0(z), \pi_1(z')) \pi_0(z)(1 - \pi_1(z')) + \\
&+ E((u_2 - u_0) | \tilde{D}_1 = 1, D_2 = 1, \pi_0(z), \pi_1(z')) \pi_0(z) \pi_1(z')
\end{aligned}$$

is an unknown function that will be estimated nonparametrically. In general, unless $\pi_0(z)$ and $\pi_1(z')$ have full support we will not be able to identify separately the constant term in $\alpha(x)$, $\bar{\beta}_1$ and $\bar{\beta}_2$.³³ As this is not the case in the sample used, it is not possible to separately identify these parameters but we still will be able to obtain the Marginal Treatment Effects as estimates of $LIV_0(\pi_0(z), \pi_1(z'))$ and $LIV_1(\pi_0(z), \pi_1(z'))$, for the points in the support of $\pi_0(z)$ and $\pi_1(z')$.

Estimates are performed following next steps that are a version of the method used by Carneiro, Heckman and Vytlacil (2003) for the binary case:

1. Firstly, each continuous variable in x and $\ln(y)$ is regressed on $\pi_0(z)$ and $\pi_1(z')$, using Local Linear Regression,³⁴ and residuals of each of these regressions are obtained. A probit model is estimated for each binary variable.
2. Then, the residuals of the regression of $\ln(y)$ are regressed on the residuals of the regressions of x and the probit models, using OLS.³⁵ Residuals of this regression are obtained again.

³³Following Heckman (1990) we can identify these parameters separately using an “identification at infinity” argument:

1. For those with $\pi_0(z)(1 - \pi_1(z')) \simeq 1$ we will identify $\bar{\beta}_1$ plus the constant term in $\alpha(x)$.
2. For those with $\pi_0(z)\pi_1(z') \simeq 1$ we will identify $\bar{\beta}_2$ plus the constant term in $\alpha(x)$.

Then assuming that $K(\pi_0(z), \pi_1(z')) = 0$ for previous groups of individuals we manage to identify separately the constant terms in $\alpha(x)$ and in $K(\pi_0(z), \pi_1(z'))$, as well as, $\bar{\beta}_1$ and $\bar{\beta}_2$. Estimators that use this argument are presented, for example, by Heckman (1990) and by Andrews and Schafgans (1998).

³⁴All the Local Linear Regression estimates are performed using a bivariate normal kernel and a common bandwidth equal to 0.24 (rule of thumb bandwidth). The propensity scores are divided by the standard deviation to avoid problems because of their disparate variation.

³⁵This regression does not include a constant variable given that it is not separately identified from $K(\pi_0, \pi_1)$.

3. Finally, previous residuals are regressed on $\pi_0(z)$ and $\pi_1(z')$, using Local Linear Regression. This leads to an estimate of $K(\pi_0(z), \pi_1(z'))$.³⁶

Figure 6. shows estimated $K(\pi_0, \pi_1)$ as a function of π_0 for those individuals with π_1 smaller than 5%, whereas, Figure 7 shows it as a function of π_1 for those with an estimated value of π_0 between 20% and 30%. As it is shown in these figures, there are signs of departure from linearity. This result agreed with the result of the homogeneity test of previous section.

Once the estimate of $K(\pi_0(z), \pi_1(z'))$ is obtained its derivatives are approximated taking discrete differences:

$$\begin{aligned}\frac{\partial K(\pi_0(z), \pi_1(z'))}{\partial \pi_0(z)} &\simeq \frac{K(\pi_0(z) + h, \pi_1(z')) - K(\pi_0(z), \pi_1(z'))}{h} \\ \frac{\partial K(\pi_0(z), \pi_1(z'))}{\partial \pi_1(z')} &\simeq \frac{K(\pi_0(z), \pi_1(z') + h) - K(\pi_0(z), \pi_1(z'))}{h}\end{aligned}$$

where h is set to be equal to 0.01. These pseudo derivatives are estimates of the MTEs: $MTE_0(\pi_0(z), \pi_1(z'))$ and $MTE_1(\pi_0(z), \pi_1(z'))$ respectively. Figure 8 and Figure 9 show estimated MTE_0 as a function of π_0 and MTE_1 as a function of π_1 . One thing to remark from these figures is that there are individuals for whom estimated *MTEs* are negative. If this is the case, they will not benefit from policies that obliged them to continue studying. This is something that should be taken into account when evaluating this kind of policies and that cannot be inferred from the traditional treatment effect parameters in the literature. Moreover, estimated *MTEs* differ from the ones obtained under normality assumptions. This suggests that these assumptions are not valid in the sample used.

6.5 The 1970 reform

The 1970 reform of the Spanish education system raised the age of compulsory schooling to 14 years old. So, all the individuals after the reform were obliged to complete

³⁶An alternative method consists on approximate $K(\pi_0, \pi_1)$ using series based on orthogonal polynomials in which the number of terms increases with the sample size (Newey (1988)). See, Newey, Powell and Walker (1990) for an empirical application of this kind of estimators. These estimates were also performed using a bivariate normal kernel and a common bandwidth equal to 0.50.

the low level of education. Then, those individuals who had such big costs to continue studying that would have decided not to complete this level of education were obliged to do it. On the other hand, if costs of schooling are endogenous, as individuals were obliged to complete more years of education, their costs could have been reduced because of the reform. In the sample used, 63% of the individuals studied before this reform. Moreover, the proportion of individuals who decided to continue studying after the low level rose from 59% for those who studied before the reform to 72% for those who studied after it. Then, a natural question is if those who studied before the 1970 reform would have benefited in terms of wages from it. Figure 10 shows the Marginal Treatment Effect of continuing studying after the low level of education (MTE_0) as a function of the probability of being affected by the reform. This probability is defined as the difference between the probability of continuing studying from the low level that would have been obtained if the individual had studied after the reform and the probability that is obtained given that the individual studied before it. As it can be seen in this figure those individuals who would have benefited are those with a lower probability of changing their decisions. Individuals in this group had more educated parents, in supervisory jobs and with a lower proportion of them working in agriculture. On the other hand, those individuals with a higher change in the probability because of the reform did not benefit from continuing studying. Figure 11. shows the ratio of the probability that would be obtained if the individual would have studied after the reform and his probability *a priori* as a function of this latter probability. As it can be seen in this figure those who would have changed more their probability of continuing studying from the low level are those with a small probability *a priori*. Finally, the expected return for those who would change their decision because of the reform, given $\pi_0(z)$, $\pi_1(z)$ and a certain change in $\pi_0(z)$ has this form:

$$PE(\pi_0(z), \pi'_0(z), \pi_1(z)) = \int_{\pi_0(z)}^{\pi'_0(z)} MTE_0(u, \pi_1(z)) \frac{1}{(\pi'_0(z) - \pi_0(z))} du$$

Table 9 shows estimates of this parameter for different values of $\pi_0(z)$ and different increments in this probability ($\Delta\pi_0$). As it can be seen in this table those who would

benefit are those with high values of $\pi_0(z)$ and small increments in this probability. These individuals have also higher values of the probability, *a priori*, of continuing studying from the medium level of education $\pi_1(z)$.

7 Conclusions

This paper presents a framework to estimate sequential returns to schooling when they vary among the population and individuals decide their level of education taking into account their returns. A test for homogeneity in returns from sequential schooling decisions, based on methods recently developed in the nonparametric statistics literature, is presented. Using this framework, returns are estimated and tested using Spanish data. The data set used contains rich family background information and useful instruments but it is stratified by level of education. Then, estimators were adapted in order to obtain consistent estimates.

Firstly, homogeneity in the returns to schooling was assumed and estimated returns were obtained using instrumental variables methods. These estimates suggest that those individuals who normally would be affected by a supply side policy may benefit from it.

However, the data suggest that there is a distribution of returns in the population and people is sorted in each level of education taking into account their returns. If this is the case, different instruments, as well as different treatment effects, will give us different summary measures of this distribution. It is in this context where the average return for those individuals in the margin of indifference (MTE) gains importance. Estimated returns are obtained in this case both imposing normal distribution assumptions as well as nonparametrically. A comparison of the estimated marginal effects under each of these frameworks suggests that normality assumptions do not hold. Moreover, the empirical application shows that there could be groups of the population for whom marginal effects are negative. In this case, these individuals will not benefit, in term of wages, from an increment of their schooling levels. This is something that should be considered when evaluating policies that oblige individuals to achieve a certain level of education. An analysis of the 1970 reform of the Spanish educational system suggests that those individuals, who studied before the reform,

who would have experience a bigger change in the probability of continuing from the low level of education due to the reform, have on average negative marginal returns.

Finally, as an extension of this work, it would be useful to develop methods that take into account self-selection because of unemployment or non participation in the labor market, when estimating the returns. This is something that can be important especially for women and for high unemployment countries like Spain. It should also be noted that this methodology do not consider explicitly the existence of general equilibrium effects. It should be also good to the develop methods to assess the importance of such effects.

APPENDIX

A THE ECBC

The ECBC (*“Encuesta de Conciencia y Biografía de Clase”*) was carried out in 1991 by the Spanish National Statistical Office (INE) at the initiative of the regional government of Madrid and *“Instituto de la Mujer”*. The main purpose of this survey was the collection of data to carry out a sociological study of different regions in Spain, as well as, to compare it with similar works in other countries. 6632 people from all regions were interviewed and data about personal information, current and previous labor status, family background, information about the spouse and opinions about different issues were collected.

The process to collect the data was the following: In each region and each size of district, except for Madrid, the number of interviews was chosen in order to be proportional to their population. A total of 5000 interviews were first planned for these regions. On the other hand, in Madrid area they planned to do 1600 interviews. Then, in a second step, they decided on the number of interviews that should be done in each region, size of district and level of education for the overrepresenting of people with a medium or high level of education. They aggregated the levels of education reported in the Spanish electoral census (1986) in the next categories: low level of education (illiterate people, those without any degree but who have ever gone to school and people with an elementary degree (*“Bachiller Elemental”* or *“EGB”*)), medium level (vocational studies (*“Formación Profesional”*) and people with a high school degree (*“Bachiller Superior”* or *“BUP”*)) and high level (people with at least a college degree (*“Escuelas de Grado Medio”* included)). In the regions Cataluña and Madrid three times as many interviews were conducted for people with high level of education. For the rest of regions of Spain, the number of interviews for people with medium level of education was doubled and four times more interviews were conducted for those with high level of education. The interviews needed for the overrepresenting of these levels of education were subtracted of the ones assigned to the low level.

Once the number of interviews for each stratum (region, size of district and level

of education) was decided, people who should be interviewed were chosen, with a list of possible people to be interviewed as replacements. The substitutes were ordered according to the similarities they presented with the individuals chosen in the first place. Those similarities being: first gender, then level of schooling and finally, age. As a result of this process there were more interviews of young people with high levels of education than planned.

B Estimates from stratified samples

It is quite common to find that samples used in econometric estimates are not randomly drawn from the population; stratified samples are often obtained. In this case, the population is divided into different strata, based on the value of an observed vector and a random sample of a pre-assigned size is then drawn. This is useful in order to sample different subsets of the population with different frequencies. From the point of view of modelling stratification may be based on exogenous variables, endogenous variables or a mixture of both. We won't know in which case we are until the specification of the model and the assumptions in it.

Let $s = 1, 2, \dots, S$ denote the strata and i denotes elements of the population. Then, let $\beta = (\beta_0, \beta_1, \dots, \beta_k)$ be the fixed finite population parameters that we wish to estimate using a stratified sample of size n . Consider the following definitions:

Definition 1 Q_s is the population frequency of stratum s . That is, it is the probability that a randomly drawn observation falls into stratum s .

Definition 2 H_s is the arbitrary probability distribution over the strata that is used when sampling the data.

Then, for each element in the sample, define the function $w(s)$ in the following way:

$$\begin{aligned} w(s) &= \frac{Q(s)}{H(s)} \\ w(s) &> 0 \end{aligned}$$

where it is assumed that $Q(s)$, that denotes the population density of the stratum s , is known and we are able to recover $H(s)$. As the strata are supposed to be exhaustive and mutually exclusive:

$$Q_1 + Q_2 + \dots + Q_S = 1$$

Oversampled categories are given a low weight, and *vice versa*, while using these weights.

B.0.1 OLS with weights

Assume that we want to estimate the population parameters β from the following model:

$$\begin{aligned} Y &= X\beta + u \\ E(X'u) &= 0 \end{aligned} \tag{5}$$

where Y and X are observed variables, whereas, u is an unobserved variable.

Then,

$$\begin{aligned} E(X'(Y - X\beta)) &= 0 \Rightarrow E(X'Y) - E(X'X)\beta = 0 \\ \hat{\beta} &= \left(\frac{1}{n} \sum_{i=1}^n x_i^2\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n x_i y_i\right) \end{aligned}$$

In order to obtain a consistent estimate in the case of having a stratified sample we need the population totals of x_i^2 and $x_i y_i$ and so, we obtain the following weighted estimator:³⁷

$$\hat{\beta}_w = \left(\frac{1}{n} \sum_{i=1}^n w_i x_i^2\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n w_i x_i y_i\right)$$

³⁷If we assume $E((Y - X\beta)|X) = 0$ and the stratification is based on exogenous variables in the model then, the unweighted estimator is consistent and more efficient than the weighted estimator (See Deaton (2000) for a discussion on this topic).

B.0.2 GMM with weights

Consider that we want to estimate equation (5) but now:

$$E(X'u) \neq 0$$

Moreover, assume that there is a set of valid instruments Z such that:

$$E(Z'u) = 0$$

and that we want to estimate the set of parameters β using the Generalized Method of Moments (GMM).

In this context, estimates are based on a group of moments conditions $\Psi(X, Z; \beta)$ that are satisfied in the population.

$$E_Q\{\Psi(X, Z; \beta)\} = 0 \Rightarrow \int \Psi(X, Z; \beta) f_Q(X, Z) d(X, Z) = 0$$

$\Psi(X, Z; \beta) = Z'u$ in our case. E_Q denotes the Expectation with respect to the true population distribution of the variables X and Z ($f_Q(X, Z)$).

However, if we have a stratified sample of the population we will have:

$$E_H\{\Psi(X, Z; \beta)\} \neq 0 \Rightarrow \int \Psi(X, Z; \beta) f_H(X, Z) d(X, Z) \neq 0$$

where E_H denotes the Expectation with respect to the arbitrary distribution of the variables X and Z ($f_H(X, Z)$), used when sampling the data.

However,

$$\int \Psi^*(X, Z; \beta) f_H(X, Z) d(X, Z) = 0$$

where $\Psi^*(X, Z; \beta) = \frac{f_Q(X, Z)}{f_H(X, Z)} \Psi(X, Z; \beta)$. Then we can use GMM in this modified moments' conditions in order to obtain consistent estimates in case of having a stratified sample. If the number of moments' conditions is greater than the number of parameters we can compute a test for overidentifying moments' restrictions using the GMM based J-statistic.

B.0.3 Discrete choice models with weights

Consider that y is a binary variable and that we are interested on estimating a discrete choice model which gives us the following probability of choosing the alternative y_i :

$$P(y_i|z; \beta)$$

which is assumed to be known except for the values of a set of parameters (β) that we want to estimate from the choices of a sample of individuals.

Furthermore, assume that the sample is divided into subsets each consisting on individuals who choose a particular alternative and that we sample at different rates from different subsets. When the stratified sample is only based on the decisions of the individuals it is usually referred to as a choice based sample.

In this case, the log-likelihood of the stratified sample will have the following form:³⁸

$$\begin{aligned} \text{Log}L(\beta) &= \sum_{s=1}^S \sum_{i \in N(s)} \log f(z_i|y_i) = \\ &= \sum_{s=1}^S \sum_{i \in N(s)} \log \frac{P(y_i|z_i; \beta) \cdot f(z_i)}{Q(s|\beta, f)} \end{aligned} \quad (6)$$

where $N(s)$ denotes the number of observations in stratum s . In order to obtain the estimator that maximizes the likelihood of the sample we should maximize the latter expression subject to the following restrictions:

$$\begin{aligned} Q(s|\beta, f) &= \int P(y_i|z_i; \beta) \cdot f(z_i) dz \\ s &= 1, 2, \dots, S \end{aligned}$$

The problem that we face if we want to use this estimator is that $f(z_i)$ is an unknown function.

If we had an exogenous sample we would maximize the following log-likelihood function:

³⁸See Cosslett (1993).

$$LogL(\theta) = \sum_{i=1}^N (\log P(y_i|z_i; \beta) + \log(f(z_i))) \quad (7)$$

In this case, when we maximize, with respect to β , $f(z_i)$ is a constant and it can be ignored.

As we can see, (7) and (6) are different and so maximization of (7) gives us estimators that are generally inconsistent in case of having a sample that is stratified based on the decisions of the individuals. This is because we would be maximizing the wrong likelihood. However, there is a straightforward modification of (7) that gives us consistent and asymptotically normal estimators for the population parameters. This estimator is called weighted maximum likelihood estimator (WESML) and was first presented by Mansky and Lerman (1977). Although, in the case of endogenous sampling, the standard estimators are generally inconsistent once the problem of consistent estimation is recognized and solved endogenous stratification can improve the precision of estimates, usually by the overrepresenting of those individuals who choose an alternative that is uncommonly chosen. Please, note that if the stratification is based on exogenous variables it is not a problem, as the only thing that we change is $f(z_i)$; estimators that ignore it are consistent and asymptotically normal.

The WESML estimator has the following form:

$$\hat{\beta}_{WESML}^* = \arg \max_{\beta} \frac{1}{N} \sum_{i=1}^N \frac{Q_i}{H_i} \log P(y_i|z_i; \beta) \quad (8)$$

Under regularity conditions similar to those required in maximum likelihood estimation, estimates obtained by the maximization of (8) are strongly consistent and asymptotically normal, but generally not asymptotically efficient.³⁹ The idea underlying consistency, is the following:

Consider the log-likelihood for exogenous sampling of N observations evaluated at any β :

$$LogL(\beta) = \sum_{i=1}^N (\log P(y_i|z_i; \beta) + \log(f(z_i))) \quad (9)$$

³⁹See Wooldrige (2001).

When we maximize (9) we obtain the estimator $\hat{\beta}$:

$$\hat{\beta} = \arg \max_{\beta} \frac{1}{N} \sum_{i=1}^N \log P(y_i | z_i; \beta)$$

Then the true value of the parameter β is the following:

$$\beta_0 = \arg \max_{\theta} E_Q[\log P(y_i | z_i; \beta)]$$

However, the pseudo true value that we obtain in the case of having a stratified sample is:

$$\beta^* = \arg \max_{\beta} E_H[\log P(y_i | z_i; \beta)]$$

and so unless H equals Q we will not obtain consistent estimators without weighting.

On the other hand, the WESML estimator will be consistent:

$$\beta_{WESML}^* = E_H\left[\frac{Q_n}{H_n} \log P(y_i | z_i; \beta)\right] = E_Q[\log P(y_i | z_i; \beta)]$$

Given this, we should weight each observation with the weights:

$$w(s) = \frac{Q(s)}{H(s)} = \frac{Q(s)}{\frac{N_s}{N}}$$

in order to obtain consistent estimators.

B.1 ECBC sampling

In the case of the ECBC we have different subsamples in which they modify the probabilities of the educational levels. So, in this sample we have both exogenous and endogenous stratification. If the stratification were on T exogenous strata, the stratification could be ignored. Nevertheless, if each of those strata is then split into G endogenous strata, we have to work with $T \times G$ endogenous strata; the exogenous stratification can no longer be ignored. In this case, the population probabilities $Q(s)$ will be different for each exogenous stratum.⁴⁰

In particular, in the case of the ECBC the sampling is based on the following strata: province, size of district and level of education reported in the 1986 Census.

⁴⁰See Cosslet (1993).

This is a case of endogenous stratified sampling because the level of education reported by the 1986 Census is highly correlated with the level of education reported in the moment of the interview (December 1990/ March 1991). The number of strata that should be considered is given by the number of combinations in the categories of province, size of district and level of education.

The sample density of stratum j is obtained in the following way:

$$H(s) = \frac{N(s)}{N}$$

$$N = \sum_{s=1}^S N(s)$$

whereas the population densities $Q(s)$ are estimated as:

$$\begin{aligned} Q(s) &= \Pr(\text{Level of education} | \text{Province, Size of district}) \cdot \Pr(\text{Province, Size of district}) \\ &= \Pr(\text{Level of education} | \text{Province, Size of district}) \cdot \Pr(\text{Size of district} | \text{Province}) \\ &\quad \Pr(\text{Province}) \end{aligned}$$

$\Pr(\text{Level of education} | \text{Province, Size of district})$ and $\Pr(\text{Size of district} | \text{Province})$ are reported in the ECBC and $\Pr(\text{Province})$ is taken from the 1991 Census.

Then each individual observation is assigned the following weight depending on the stratum they belong to:

$$w(s) = \frac{Q(s)}{H(s)}$$

As an example of how this weights work, consider the highest level of education completed for the people who were born between 1928 and 1932. The ECBC sample shows the following shares for each level of education:

Level of Education	Percent	Percent (with weights)
compulsory level	55.75	80.43
high school	9.20	7.34
college	35.06	12.23

As it can be seen in previous table, weighting changes the share of individuals in each level of education. Using data from the surveys EPF80, EPF90 and Encuesta Continua de Presupuestos familiares 95 we have that these proportions are: 81.43% for the compulsory level, 8.05% for high school level and 10.52% for college.

C Definition of variables

- **Distance (In 100 km).** This variable measures the distance from the province where the individual used to live during his/her childhood to the nearest province in the group of provinces that traditionally had a university. Madrid, Barcelona, Salamanca, Valladolid, and Granada are the provinces considered in this group for those who were born before 1961. Given the increment of the number of four-year college degrees offered by some existing small universities during the latter 60's and 70's, this group was augmented for those who were born after 1961. In addition to previous provinces, Seville, Zaragoza, A Coruña, Asturias and Valencia were also included in this case.
- **Distance to the nearest university at the age of 18.** This variable takes value 0 if the individual lived in a province which had a university when he was 18 years old. If there was no university it is equal to the number of kilometers to the nearest capital of province which had a university in that moment.
- **High distance.** Dummy variable that takes value one if the individual was living far away from one of the provinces that traditionally had a university (more than 352 km) in the moment of decision. Madrid, Barcelona, Salamanca, Valladolid, and Granada are the provinces considered in this group for those who were born before 1966. This group was augmented for those who were born after 1966 considering also Seville, Zaragoza, A Coruña, Asturias and Valencia.
- **Family income 14 and family Income 18.**⁴¹ *Per capita* gross family income in prices of 1995. This variable measures the total income of the families, net of taxes and Social Security payoffs. It includes transfers of the government. It is

⁴¹This data was taken from Alcaide (2003).

measured *per capita* dividing by the total population of each province as at the 1st of July of each year. Each individual in the sample is assigned the family income at the age of 14 and 18, respectively, of the province where he/she used to live during his/her childhood.

- **Employment 14 and employment 18.**⁴² Number of occupations in the province, where the individual used to live during his/her childhood, at the age of 14 and 18, divided by the population of the province in that moment.
- **Illiterate, primary, high school.** Dummy variables for the level of education of the head of household when the individual was a child.
- **Blue collar, agriculture, mining, manufacture.** Dummy variables for the kind of job of the head of household when the individual was a child (blue collar job, agriculture sector's job, mining sector's job and manufacturing sector's job).
- **Unemployed.** Dummy variable that takes value one if the head of household when the individual was a child was unemployed.
- **Supervisory.** Dummy variable for the head of household, when the individual was a child, doing a supervisory job.
- **Female.** Dummy variable that takes value one if the individual is a woman.
- **Islands.** Dummy variable that takes value one if the individual lived in Canary Islands, Balearic Islands, Ceuta or Melilla when he/she was a child.
- **Medium and high.** Dummy variables that take value one if the individual achieved the medium and high level of education, respectively.

D A Sequential Decision Model for Schooling Choices

This section presents an economic model for sequential schooling decisions. Models of this kind were presented, for example, by Heckman, Lochner and Todd (2003) and Comay, Melnik and Pollatscheck (1973).

⁴²This data was taken from Alcaide (2003).

A sequential decision model takes into account that the level of schooling of each individual is the result of previous schooling choices. Thus, the probability that a person enters to university depends on high school graduation which also depends on deciding to start high school and so forth until the earliest schooling decision. Given this, something that alters the benefits or costs of a given level of education, at any stage, will affect the proportion of people in any preceding stage. This is important because of self-selection of individuals in each level of education that should be considered when estimating returns. Moreover, a sequential model allows the consideration of the existence of an option value of each education level which will affect the decisions made by the individuals. As explained by Heckman, Lochner and Todd (2003) there are mainly two reasons that can explain the fact that the option value of a given level of schooling is different from zero:

1. Returns to schooling may be not constant across different levels of education. In this case, an individual can be interested in finishing a certain level of schooling in order to access a higher level with greater earnings.
2. If there is uncertainty that is resolved as additional years of schooling are finished, the returns to schooling should include the expected value of the new information that is revealed. This uncertainty can be related, for example, to costs of schooling, earnings or the ability of the individuals.

In both cases, it will not be sufficient to analyze these decisions in a static binary choice framework like it has been usually done in the literature (see, e.g., Willis and Rosen (1979) or more recently Carneiro, Heckman and Vytlačil (2003)).

Consider the case in which there are no dropouts and if individuals want to achieve a level they are allowed to do it. People decide their level of education maximizing the expected present discounted value of lifetime earnings, given the available information, that would be obtained with each choice. An individual has two options in each moment of decision: continue studying or start working. Then, the present discounted value of earnings net of costs as of the schooling completion date are:

$$\begin{aligned}
w_0^i &= \sum_{t=0}^{T_r} (1+r)^{-t} y_0^i(t) - C_0 \\
w_1^i &= \sum_{t=0}^{T_r-T_1} (1+r)^{-t} y_1^i(t) - C_1 \\
w_2^i &= \sum_{t=0}^{T_r-T_2} (1+r)^{-t} y_2^i(t) - C_2
\end{aligned}$$

where, T_1 and T_2 are the number of years that the individual (i) has to study, from the low level, to achieve the medium and high level, respectively. T_r denotes the age of retirement and r is the discount rate in the moment of decision which is assumed to be externally specified. C_0 , C_1 and C_2 are the direct costs of achieving the low, medium and high level of education.

Define as $E_0(V_1^i)$ the total expected value of achieving the medium level of education given the information in the moment of decision ($t = 1$) :

$$E_0(V_1^i) = E_0 \max \left\{ w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i) \right\}$$

where V_j^i for $j = 1, 2$ denote the benefits of studying up to the low and medium level of education, respectively. E_j for $j = 0, 1$ denotes the expected value given the information the individual has in the moment of deciding to continue studying up to each level of education ($t = 1, t = 2$).

When making their decisions, individuals consider not only the direct benefits from completing the next level of education but also the value of the offered option of entry into even higher levels. The option value of continuing studying from the low level, as defined in Heckman, Lochner and Todd (2003), is:

$$O_0^i = E_0(V_1^i - w_1^i) = E_0 \max \left\{ 0, \left(\frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i) \right) - w_1^i \right\}$$

An individual will decide in ($t = 1$) to continue studying from the low level of education if the following is true:

$$\begin{aligned}
w_0^i &< \frac{1}{(1+r)^{T_1}} E_0(V_1^i) \implies w_0^i < \frac{1}{(1+r)^{T_1}} E_0 \max\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)\} \implies \\
&\implies w_0^i < \frac{1}{(1+r)^{T_1}} (E_0(w_1^i) + O_0^i)
\end{aligned}$$

This is assuming that an individual knows ones earning potential after completion of each level of education, but cannot foresee his potential after completion of higher levels of education.

On the other hand, given this model, an individual decides to continue studying from the medium level of education if the following is true:

$$w_1^i < \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)$$

and given that high is the final level of education considered the option has value equal to zero in this case:

$$O_1^i = 0 \quad \forall i$$

Therefore, the probability of choosing each level of education is:

$$\begin{aligned}
\Pr(D_0^i = 1) &= 1 - \Pr(w_0^i < \frac{1}{(1+r)^{T_1}} E_0 \max\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)\}) \quad (10) \\
&\Pr(D_1^i = 1) = \\
&= (1 - \Pr(w_1^i < \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i) | w_0^i < \frac{1}{(1+r)^{T_1}} E_0 \max\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)\}) \\
&\Pr(w_0^i < \frac{1}{(1+r)^{T_1}} E_0 \max\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)\}) \\
&\Pr(D_2^i = 1) = \\
&= \Pr(w_1^i < \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i) | w_0^i < \frac{1}{(1+r)^{T_1}} E_0 \max\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)\}) \\
&\Pr(w_0^i < \frac{1}{(1+r)^{T_1}} E_0 \max\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)\})
\end{aligned}$$

Assume, for instance, that individuals are myopic, i.e., in deciding about continuing studying from the low level of education they fall prey to the making the easier choice of entering the labor market after completing the medium level. Moreover, assume that the following two conditions are satisfied:

$$\begin{aligned}
w_0^i &> \frac{1}{(1+r)^{T_1}} E_0(w_1^i) \\
w_0^i &> E_0 \max\left\{\frac{1}{(1+r)^{T_1}} w_1^i, \frac{1}{(1+r)^{T_2}} E_1(w_2^i)\right\}
\end{aligned}$$

If this is true, then, decisions made by myopic individuals and those that decide sequentially would be the same. It is important to note that, in this case, there is an option value to continue studying from the low level of education which is different from zero but individuals are not changing their decisions because of it. On the other hand, if both of the following conditions are satisfied:

$$\begin{aligned}
w_0^i &> \frac{1}{(1+r)^{T_1}} E_0(w_1^i) \\
w_0^i &< E_0 \max\left\{\frac{1}{(1+r)^{T_1}} w_1^i, \frac{1}{(1+r)^{T_2}} E_1(w_2^i)\right\}
\end{aligned}$$

A binary static model predicts that individuals that fall under the above category will stop studying at the low level, whereas, a sequential model says that these individuals will continue studying solely for the purpose of achieving higher earnings from the high level of education.

Assume for an instance that there is no uncertainty. Then, in a multinomial (static) model the probability of each level of education can be written as:

$$\begin{aligned}
\Pr(D_0^i = 1) &= \Pr\left(\frac{w_1^i}{(1+r)^{T_1}} < w_0^i, \frac{w_2^i}{(1+r)^{T_2}} < w_0^i\right) = \\
&= \Pr\left(w_0^i > \frac{1}{(1+r)^{T_1}} \max\left\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} w_2^i\right\}\right) \\
\Pr(D_1^i = 1) &= \Pr\left(\frac{w_1^i}{(1+r)^{T_1}} > \frac{w_2^i}{(1+r)^{T_2}}, \frac{w_1^i}{(1+r)^{T_1}} > w_0^i\right) = \\
&= \Pr\left(w_1^i > \frac{w_2^i}{(1+r)^{T_2}} \mid \frac{w_1^i}{(1+r)^{T_1}} > w_0^i\right) P\left(\frac{w_1^i}{(1+r)^{T_1}} > w_0^i\right) \\
\Pr(D_2^i = 1) &= \Pr\left(\frac{w_1^i}{(1+r)^{T_1}} < \frac{w_2^i}{(1+r)^{T_2}}, \frac{w_2^i}{(1+r)^{T_2}} > w_0^i\right)
\end{aligned}$$

Although people don't make decisions this way if they know everything, the sequence is not important; the sequential model gives the same decisions as in a multinomial static model with people deciding at $t = 0$.

On the other hand, if there is something that changes from one stage to the other then the probability of each decision will be modelled by (10). Let π_0^i and π_1^i denote the probability of continuing studying from the low and medium level of education. Then, the probabilities of each level of education can be rewritten in the following way:

$$\begin{aligned}\Pr(D_0^i = 1 | z, z') &= (1 - \pi_0^i(z)) \\ \Pr(D_1^i = 1 | z, z') &= \pi_0^i(z)(1 - \pi_1^i(z')) \\ \Pr(D_2^i = 1 | z, z') &= \pi_1^i(z')\pi_0^i(z)\end{aligned}$$

where:

$$\begin{aligned}\pi_0^i(z) &= \Pr(w_0^i < \frac{1}{(1+r)^{T_1}} E_0 \max\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)\}) \\ \pi_1^i(z') &= \Pr(w_1^i < \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i) | w_0^i < \frac{1}{(1+r)^{T_1}} E_0 \max\{w_1^i, \frac{1}{(1+r)^{T_2-T_1}} E_1(w_2^i)\})\end{aligned}$$

The value of the variables in z can be different of the value of the variables in z' as some uncertainty is resolved as additional years of education are completed.

The likelihood of a sample of size N has the following form:

$$L(\dots|x) = \prod_{i=1|D_0^i=1}^N (1 - \pi_0^i(z)) \prod_{i=1|D_1^i=1}^N (1 - \pi_1^i(z')) \prod_{i=1|D_2^i=1}^N \pi_1^i(z')\pi_0^i(z)$$

Which, taking logarithms, results in the following expression:

$$\begin{aligned}
\log L(\dots|x) &= \sum_{i=1|D_0^i=1}^N \log(1 - \pi_0^i(z)) + \sum_{i=1|D_1^i=1}^N \log(1 - \pi_1^i(z')) + \sum_{i=1|D_2^i=1}^N \log(\pi_0^i(z)) + \\
&\quad + \sum_{i=1|D_2^i=1}^N \log(\pi_1^i(z')) + \sum_{i=1|D_2^i=1}^N \log(\pi_0^i(z)) \\
&\quad \Downarrow \\
\log L(\dots|x) &= \sum_{i=1|D_0^i=1}^N \log(1 - \pi_0^i(z)) + \sum_{i=1|D_1^i=1 \cup D_2^i=1}^N \log(\pi_0^i(z)) + \\
&\quad + \sum_{i=1|D_1^i=1}^N \log(1 - \pi_1^i(z')) + \sum_{i=1|D_2^i=1}^N \log(\pi_1^i(z'))
\end{aligned}$$

Then , we can estimate the following reduced form models:⁴³

1. Considering all the sample estimate a model for the probability of continuing studying from the low level.
2. With those who have achieved at least the medium level estimate a model for the probability of achieving the high level.

⁴³ Assuming that there are neither restrictions between the parameters nor the unobservables in each choice.

References

- [1] Alcaide, I (2003), *Evolución Económica de las Regiones y Provincias Españolas en el Siglo XX*, in Fundación BBVA, eds.
- [2] Andrews, D and M. Schafgans (1998), “Semiparametric Estimation of the Intercept of a Sample Selection Model”, *The Review of Economic Studies* 65, no. 3, 497-517.
- [3] Blundell, R., Dearden, L. and B. Sianesi (2001), “Estimating the Returns to Education: Models, Methods and Results”, unpublished.
- [4] Card, D. (2001), “Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems”, *Econometrica* 69, no. 5, 1127-1160.
- [5] Card, D. (1999), “The Causal Effect of Education on Earnings”, in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, vol. 3A, cap. 30, Amsterdam, North-Holland.
- [6] Carneiro, P, Hansen, K. and J. Heckman (2003), “Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice”, *International Economic Review* 44, no. 2, 361-422.
- [7] Carneiro, P., Heckman, J. and E. Vytlacil (2003), “Understanding What Instrumental Variables Estimate-Estimating Marginal and Average Returns to Education”, unpublished.
- [8] Comay, Y, Melnik, A. and M.A. Pollatschek (1973), “The Option Value of Education and the Optimal Path for Investment in Human Capital”, *International Economic Review* 14, no. 2, 421-435.
- [9] Cosslett, S. (1993), “Estimation from Endogenously Stratified Samples”, in G.S. Maddala, C. R. Rao and H.D. Vinod, eds., *Handbook of Statistics* vol. 11, Cap. 1, Amsterdam, Elsevier.

- [10] Deaton, A.(2000), *The Analysis of Household Surveys. A Microeconometric Approach to Development Policy*, in The World Bank, eds.
- [11] Heckman, J. (2001), “Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture”, *Journal of Political Economy* 109, no. 4, 673-737.
- [12] Heckman, J. (1997), “Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations”, *The Journal of Human Resources* 32, no. 3, 441-462.
- [13] Heckman, J. (1990), “Varieties of Selection Bias”, *The American Economic Review* 80, no. 2, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, 313-318.
- [14] Heckman, J.(1979), “Sample Selection Bias as a Specification Error”, *Econometrica* 47, no. 1, 153-161.
- [15] Heckman, J., Lochner, L. and P. Todd (2003), “Fifty Years of Mincer Earnings Regressions”, *National Bureau of Economic Research Working Paper* 9732.
- [16] Heckman, J., Smith, J., and N. Clements (1997), “Making the Most Out of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts”, *Review of Economic Studies* 64, 487-535.
- [17] Heckman, J., Tobias, J. and E. Vytlacil (2001), “For Parameters of Interest in the Evaluation of Social Programs”, *Southern Economic Journal* 68, no.2, 210-223.
- [18] Heckman, J. and E. Vytlacil (2000), “Local Instrumental Variables”, in C. Hsiao, K. Morimune, and J. Powell, eds., *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, Cambridge, Cambridge University Press.
- [19] Heckman, J. and E. Vytlacil (1999), “Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects”, *Economic Sciences* 96, 4730-4734.
- [20] Imbens, G. and J. Angrist (1994), “Identification and Estimation of Local Average Treatment Effects”, *Econometrica* 62, no. 2, 467-475.

- [21] Newey, W. (1988), “Two Step Estimation of Sample Selection Models”, unpublished.
- [22] Newey, W., Powell, J. and J. Walker (1990), “Semiparametric Estimation of Selection Models: Some Empirical Results.”, *The American Economic Review* 80, no. 2, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, 324-328.
- [23] Mansky, C. and S. Lerman (1977), “The Estimation of Choice Probabilities from Choice Based Samples”, *Econometrica* 45, no. 8, 1977-1988.
- [24] Pons, E. and M. Gonzálo (2001), “Returns to Schooling in Spain. How reliable Are IV Estimates?”, Queen Mary University of London Working Paper 446.
- [25] Powell, J. (1994), “Estimation of Semiparametric Models”, in R. F. Engle and D.L. McFadden, eds., *Handbook of Econometrics*, Vol. 4, cap. 41, Amsterdam, North-Holland.
- [26] Stute, W. (1997), “Nonparametric Model Checks for Regression”, *The Annals of Statistics* 25, no. 2, 613-641.
- [27] Stute, W., González Manteiga, W. and M. Presedo Quindimil. (1998), “Bootstrap Approximations in Model Checks for Regression”, *Journal of the American Statistical Association* 93, no. 441, 141-149.
- [28] Taber, C. (2001), “The Rising College Premium in the Eighties: Return to College or Return to Unobserved Ability”, *The Review of Economic Studies* 68, no. 3, 665-691.
- [29] Taber, C. (2000), “Semiparametric Identification and Heterogeneity in Discrete Choice Dynamic Programming Models”, *Journal of Econometrics* 96, 201-229.
- [30] Vytlačil, E. (2002), “Independence, Monotonicity, and Latent Index Models: An Equivalent Result”, *Econometrica* 70, no. 1, 331-341.
- [31] Willis, R. and S. Rosen (1979), “Education and Self-Selection”, *Journal of Political Economy* 87, no. 5, 7-36.

- [32] Wooldridge, J. (2001), “Asymptotic Properties of Weighted M-Estimators for Standard Stratified Samples”, *Econometric Theory* 17, 451-470.
- [33] Wooldridge, J. (1999), “Asymptotic Properties of Weighted M-Estimators for Variable Probability Samples”, *Econometrica* 67, no. 6, 1385-1406.

Table 1. Levels of Education by cohorts. (%)

<i>Level of Education</i> <i>Year of Birth</i>	0	1	2	3	4	5
1918-1922	5,68	25,27	53,07	5,35	4,95	5,68
1923-1927	4,04	26,74	52,63	5,66	4,97	5,96
1928-1932	3,90	26,07	50,65	6,38	5,64	7,37
1933-1937	2,85	23,11	51,58	7,44	6,91	8,11
1938-1942	1,98	20,29	48,62	10,21	9,11	9,78
1943-1947	0,89	12,60	48,09	13,56	12,10	12,76
1948-1952	0,71	9,08	42,31	17,67	16,37	13,85
1953-1957	0,51	5,52	33,48	22,65	21,70	16,14
1958-1962	0,16	3,74	25,17	29,28	26,78	14,87
1963-1967	0,27	1,50	16,83	38,58	29,69	13,13

Note: (i) Source: EPF 80, EPF 90 and Encuesta Continua de Presupuestos Familiares 95.

(ii) Level 0 refers to illiterate people, those who did not complete any level of education but can read and write are in Level 1. Level 2 represents the education of those who left school when they were 10 years old, level 3 is for those who dropped out at 14, level 4 is for those who finished high school or the vocational studies, and level 5 is for people who have a college degree.

Table 2. Sample Statistics

	$D_0 = 1$	$D_1 = 1$	$D_2 = 1$
age	42.56	35.50	37.53
female	61.73	53.76	44.93
illiterate	48.60	22.70	9.18
primary	47.45	55.70	48.79
high school	2.67	11.31	15.46
college	1.28	10.30	26.57
white collar	2.13	11.73	22.46
blue collar	16.16	29.54	41.79
agriculture	32.81	14.18	8.45
mining	8.99	9.70	3.14
manufacture	33.41	29.37	17.87
unemployed	6.62	5.74	7.00
supervisory	4.56	15.19	25.36
distance	158.95	157.18	145.67
high distance	8.26	6.41	7.97
family income-14	0.66	0.86	0.85
employment-14	0.38	0.36	0.38
family income-18		0.94	0.93
employment-18		0.36	0.37
n.obs	1646	1185	414

Table 3. Mincer equation. OLS & GMM estimates

<i>Variable</i>	<i>OLS with weights</i>	<i>OLS without weights</i>	<i>GMM with weights</i>
age	0.049*** (0.0093)	0.039*** (0.0077)	0.050*** (0.0088)
age2	-0.00053*** (0.00011)	-0.00041*** (0.000090)	-0.00049*** (0.00011)
tenure	0.013** (0.0053)	0.014*** (0.0036)	0.014*** (0.0052)
tenure2	-0.00022 (0.00015)	-0.00021** (0.000093)	-0.00025* (0.00014)
female	-0.20*** (0.030)	-0.17*** (0.023)	-0.17*** (0.033)
schooling	0.044*** (0.0041)	0.046*** (0.0035)	0.086*** (0.022)
constant	5.22*** (0.20)	5.20*** (0.15)	4.35*** (0.39)

Note: (i) Number of observations: 1653, (ii) Source: ECBC, (iii) Standard errors in parenthesis. *** means significance at a 99% level, ** at a 95% level and * at a 90% level, (iv) 11 dummies for Spanish regions, as well as, 4 occupational dummies and 10 industry dummies were also included in these estimates, (v) These estimates are performed using as instruments: income of the family at 14 and 18 years old, distance to the nearest university when the individual was 18 years old and a dummy for the head of household doing a supervisory job. Sargan identification test: 2.90 (p-value=0.57)

Table 4. Levels of schooling OLS & GMM estimates

<i>Variable</i>	<i>OLS with weights</i>	<i>OLS without weights</i>	<i>GMM with weights</i>
age	0.049 *** (0.0097)	0.039 *** (0.0075)	0.054 *** (0.016)
age2	-0.00057 *** (0.00012)	-0.00043 *** (0.000091)	-0.00057 *** (0.00016)
tenure	0.014 *** (0.0055)	0.015 *** (0.0036)	0.020 *** (0.0056)
tenure2	-0.00023 *** (0.00016)	-0.00022 *** (0.000093)	-0.00036 ** (0.00016)
female	-0.20 *** (0.031)	-0.17 *** (0.023)	-0.15 *** (0.036)
medium	0.18 *** (0.030)	0.23 *** (0.027)	0.65 ** (0.31)
high	0.47 *** (0.041)	0.44 *** (0.037)	0.83 ** (0.38)
constant	5.39 *** (0.20)	5.58 *** (0.14)	4.78 *** (0.43)

Note: (i) Number of observations: 1653, (ii) Source: ECBC, (iii) Standard errors in parenthesis. *** means significance at a 99% level, ** at a 95% level and * at a 90% level, (iv) 11 dummies for Spanish regions, as well as, 4 occupational dummies and 10 industry dummies were also included in these estimates, (v) These estimates are performed using as instruments: income of the family at 14 and 18 years old, distance to a city that traditionally have a university (Madrid, Barcelona, Salamanca, Valladolid and Granada are the provinces considered), employment in the province of residence at the age of 18, a dummy for studying in the system after the 1970 reform and variables about the head of household when the individual was a child (three dummies for the level of education, three dummies for the kind of job and a dummy for doing a supervisory job). Sargan identification test: 6.93 (p-value=0.50).

Table 5A. Probability of continuing studying from the low level (π_0)

Variable	Without weights	With weights
age	-0.020*** (0.0043)	-0.012** (0.0052)
illiterate	-1.49*** (0.15)	-1.61*** (0.20)
primary	-0.96*** (0.15)	-1.14*** (0.19)
high school	-0.40** (0.17)	-0.40* (0.23)
blue collar	-0.14 (0.13)	-0.018 (0.17)
agriculture	-0.75*** (0.13)	-0.63*** (0.18)
mining	-0.54*** (0.15)	-0.26 (0.19)
manufacture	-0.57*** (0.13)	-0.41** (0.17)
unemployed	-0.24 (0.15)	-0.22 (0.19)
supervisory	0.34*** (0.095)	0.26** (0.13)
employment-14	0.40 (0.61)	-0.73 (0.71)
family income-14	0.23 (0.14)	0.75*** (0.17)
female	-0.32*** (0.050)	-0.31*** (0.061)
high distance	-0.10 (0.10)	-0.13 (0.10)
constant	2.06*** (0.29)	1.38*** (0.37)

Note: (i) Number of observations: 3245, (ii) Source: ECBC, (iii) Standard errors in parenthesis.
*** means significance at a 99% level, ** at a 95% level and * at a 90% level.

Table 5B. Probability of continuing studying from the low level (π_0)

System before 1970 Variable	Without weights	With weights
age	-0.024*** (0.0046)	-0.011* (0.0058)
illiterate	-1.58*** (0.16)	-1.55*** (0.22)
primary	-1.03*** (0.15)	-1.05*** (0.22)
high school	-0.49*** (0.18)	-0.26 (0.27)
agriculture	-0.63*** (0.081)	-0.63*** (0.097)
mining	-0.43*** (0.12)	-0.28* (0.15)
manufacture	-0.54*** (0.079)	-0.44*** (0.10)
supervisory	0.31*** (0.12)	0.26* (0.16)
family income-14	0.13 (0.17)	0.73*** (0.21)
high distance	-0.10 (0.10)	-0.22* (0.12)
female	-0.50*** (0.061)	-0.49*** (0.076)
constant	2.53*** (0.34)	1.03** (0.45)

Note: (i) Number of observations: 3245, (ii) Source: ECBC, (iii) Standard errors in parenthesis.
*** means significance at a 99% level, ** at a 95% level and * at a 90% level.

Table 5B. Probability of continuing studying from the low level (π_0)
(continuation)

System after 1970 Variable	Without weights	With weights
age	0.065*** (0.012)	0.024* (0.014)
illiterate	-1.72*** (0.33)	-2.15*** (0.33)
primary	-1.22*** (0.32)	-1.71*** (0.32)
high school	-0.52 (0.36)	-0.99*** (0.37)
agriculture	-0.56*** (0.14)	-0.47*** (0.16)
mining	-0.32** (0.15)	-0.13 (0.18)
manufacture	-0.26** (0.10)	-0.22* (0.13)
supervisory	0.41** (0.17)	0.35 (0.22)
family income-14	0.12 (0.22)	0.18 (0.26)
high distance	-0.08 (0.11)	-0.06 (0.23)
female	-0.032 (0.086)	-0.073 (0.10)
constant	0.031 (0.55)	1.18* (0.61)

Note: (i) Number of observations: 3245, (ii) Source: ECBC, (iii) Standard errors in parenthesis.
*** means significance at a 99% level, ** at a 95% level and * at a 90% level.

Table 6A. Probability of continuing studying from the medium level (π_1)

Variable	Without weights	With weights
age	0.0078* (0.0039)	0.010** (0.0039)
illiterate	-0.97*** (0.16)	-0.96*** (0.19)
primary	-0.52*** (0.13)	-0.52*** (0.16)
high school	-0.40*** (0.14)	-0.49*** (0.17)
blue collar	0.17 (0.12)	0.092 (0.15)
agriculture	-0.16 (0.16)	-0.19 (0.19)
mining	-0.58*** (0.20)	-0.62*** (0.22)
manufacture	-0.15 (0.14)	-0.23 (0.17)
unemployed	0.20 (0.18)	0.17 (0.21)
supervisory	0.17* (0.10)	0.20* (0.12)
islands	-0.20 (0.19)	-0.29 (0.20)
distance	-0.063** (0.027)	-0.079** (0.030)
employment-18	-0.16 (0.87)	-2.06* (1.098)
female	-0.27*** (0.071)	-0.22*** (0.080)
constant	-0.15 (0.30)	-0.29 (0.33)

Note: (i) Number of observations: 1599, (ii) Source: ECBC, (iii) Standard errors in parenthesis.
*** means significance at a 99% level, ** at a 95% level and * at a 90% level.

Table 6B. Probability of continuing studying from the medium level (π_1)

System before 1970 Variable	Without weights	With weights
age	−0.0051 (0.0052)	−0.012* (0.0060)
illiterate	−0.72*** (0.18)	−0.73*** (0.21)
primary	−0.31** (0.14)	−0.37** (0.18)
high school	−0.26 (0.17)	−0.40* (0.21)
agriculture	−0.34** (0.15)	−0.43*** (0.16)
mining	−0.66*** (0.23)	−0.50** (0.24)
manufacture	−0.40*** (0.13)	−0.44*** (0.15)
islands	−0.29 (0.20)	−0.44** (0.21)
distance	−0.092*** (0.028)	−0.11*** (0.031)
supervisory	0.12 (0.10)	0.17 (0.12)
employment-18	1.39 (1.11)	0.40 (1.28)
female	−0.44*** (0.11)	−0.44*** (0.11)
constant	0.33 (0.49)	0.33 (0.49)

Note: (i) Number of observations: 1599, (ii) Source: ECBC, (iii) Standard errors in parenthesis.
*** means significance at a 99% level, ** at a 95% level and * at a 90% level.

Table 6B. Probability of continuing studying from the medium level (π_1)
(continuation)

System after 1970 Variable	Without weights	With weights
age	0.13*** (0.018)	0.11*** (0.023)
illiterate	-1.20*** (0.23)	-1.11*** (0.25)
primary	-0.31** (0.14)	-0.55*** (0.20)
high school	-0.27 (0.17)	-0.49** (0.23)
agriculture	-0.36 (0.23)	-0.14 (0.24)
mining	-0.66*** (0.23)	-0.90*** (0.26)
manufacture	-0.40*** (0.13)	-0.29** (0.15)
islands	-0.30 (0.18)	-0.44** (0.21)
distance	-0.092*** (0.028)	-0.11*** (0.031)
supervisory	0.12 (0.10)	0.17 (0.12)
employment-18	-4.83*** (1.53)	-3.59** (1.61)
female	-0.11 (0.11)	-0.069 (0.12)
constant	-1.67*** (0.61)	-2.13*** (0.61)

Note: (i) Number of observations: 1599, (ii) Source: ECBC, (iii) Standard errors in parenthesis.
*** means significance at a 99% level, ** at a 95% level and * at a 90% level.

Table 7. Linearity Test on the Propensity Scores

Statistic	Crámer Von-Mises	Kolmogorov
Value	$1.30 \cdot 10^{-4}$	0.41
Critical Value ($\alpha = 0.05$)	$0.92 \cdot 10^{-5}$	0.43

Table 8. OLS & GMM estimates. Control Function

Variable	OLS ¹ without weights	Variable	GMM ² with weights
age	0.058*** (0.0095)	age	0.061*** (0.011)
age2	-0.00059*** (0.00011)	age2	-0.00068*** (0.00012)
tenure	0.010*** (0.0038)	tenure	0.018*** (0.0053)
tenure2	-0.00017*** (0.000096)	tenure2	-0.00032** (0.00015)
female	-0.17*** (0.025)	female	-0.19*** (0.031)
medium	0.32* (0.17)	medium	0.69*** (0.18)
high	0.68 (0.45)	high	1.08*** (0.40)
λ_1^*	-0.72*** (0.25)	λ_1	0.32*** (0.11)
λ_2^*	0.66 (0.65)	λ_2	-0.55** (0.26)
λ_3^*	-0.038 (0.72)	λ_3	-0.053 (0.15)
		λ_4	0.41 (0.33)
constant	5.09*** (0.22)	constant	5.02*** (0.22)

Note: (i) Number of observations: 1653, (ii) Source: ECBC, (iii) Standard errors in parenthesis. *** means significance at a 99% level, ** at a 95% level and * at a 90% level, (iv) 11 dummies for Spanish regions, as well as, 4 occupational dummies and 10 industry dummies were also included in these estimates, (v) Sargan Identification Test: 7.49 (p-value=0.48)

1) $\lambda_1^* = -\phi(\gamma_1^*(z))(1 - \pi_1(z))$; $\lambda_2^* = -\phi(\gamma_1^*(z))\pi_1(z)$; $\lambda_3^* = \phi(\gamma_2^*(z))\pi_0$.

2) These estimates are performed using as external instruments: income of the family when the individual was 14 years old, distance to a city that traditionally have a university, three dummies for the level of education, three dummies for the kind of job, a dummy for doing a supervision job of the head of household when the individual was a child, a dummy for studying after the 1970 reform and regional employment at the age of 18. $\lambda_1 = \left(\frac{-\phi(\gamma_1^*(z))}{\Phi(\gamma_1^*(z))} \right) D_1$; $\lambda_2 = \left(\frac{\phi(\gamma_2^*(z))}{1 - \Phi(\gamma_2^*(z))} \right) D_1$; $\lambda_3 = \left(\frac{-\phi(\gamma_1^*(z))}{\Phi(\gamma_1^*(z))} \right) D_2$; $\lambda_4 = \left(\frac{-\phi(\gamma_2^*(z))}{\Phi(\gamma_2^*(z))} \right) D_2$.

Table 9. Treatment Effects for the Extension of the 1970 Reform

$\Delta\pi_0$	Low π_0 (0.08)	Medium π_0 (0.30)	High π_0 (0.80)
1.2	−0.031 ($\pi_1^{\min}=0.03, \pi_1^{\max}=0.10$)	−0.021 ($\pi_1^{\min}=0.06, \pi_1^{\max}=0.27$)	0.083 ($\pi_1^{\min}=0.21, \pi_1^{\max}=0.54$)
1.5	−0.036 ($\pi_1^{\min}=0.03, \pi_1^{\max}=0.14$)	−0.027 ($\pi_1^{\min}=0.06, \pi_1^{\max}=0.36$)	0.089 ($\pi_1^{\min}=0.21, \pi_1^{\max}=0.62$)
2.0	−0.039 ($\pi_1^{\min}=0.03, \pi_1^{\max}=0.19$)	−0.032 ($\pi_1^{\min}=0.05, \pi_1^{\max}=0.38$)	
4	−0.032 ($\pi_1^{\min}=0.03, \pi_1^{\max}=0.38$)		

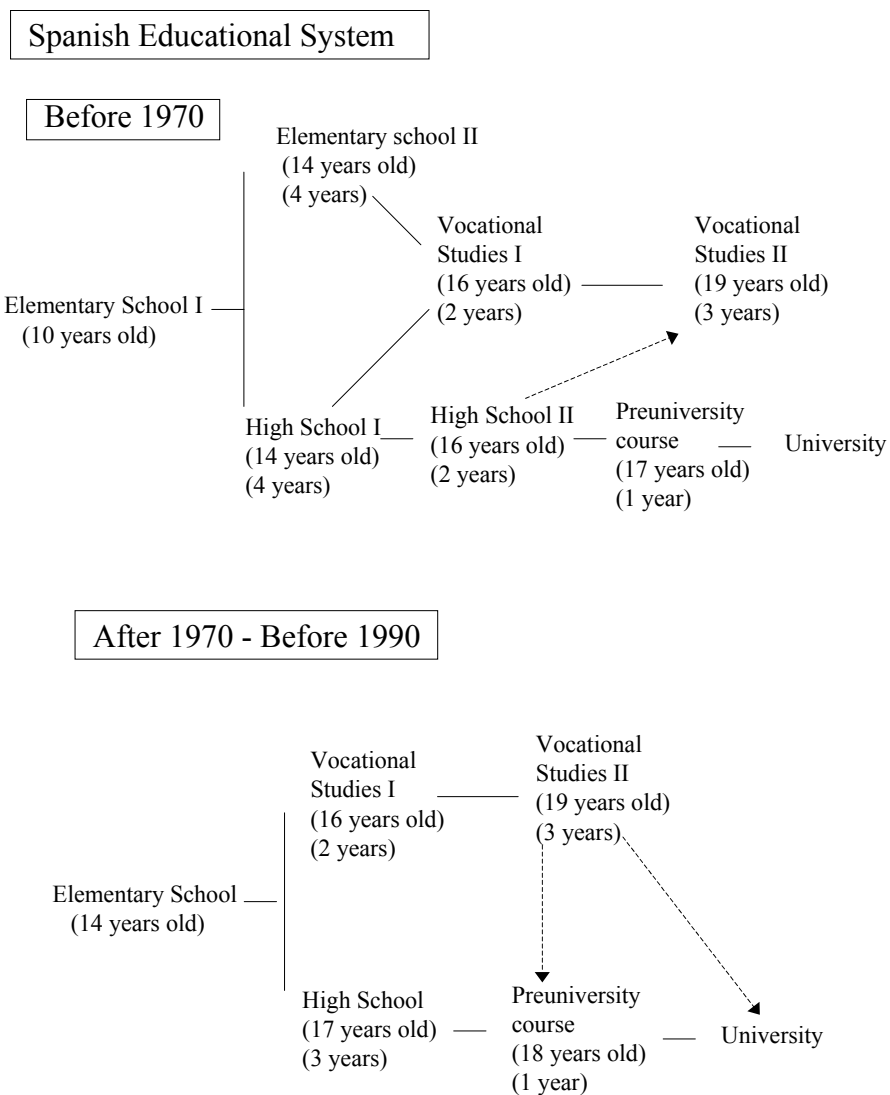


Figure 1. Spanish Educational System

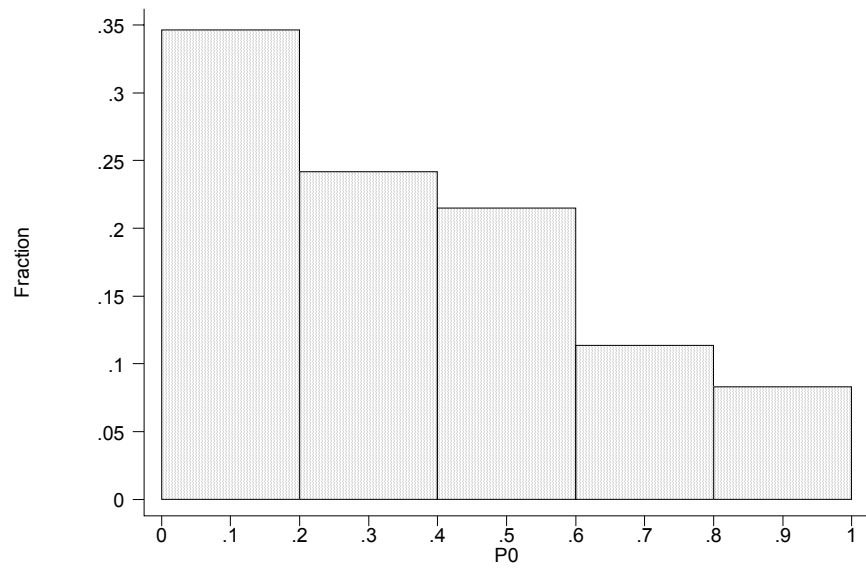


Figure 2. Estimated Probability of Continuing Studying After the Low Level of Education (π_0)

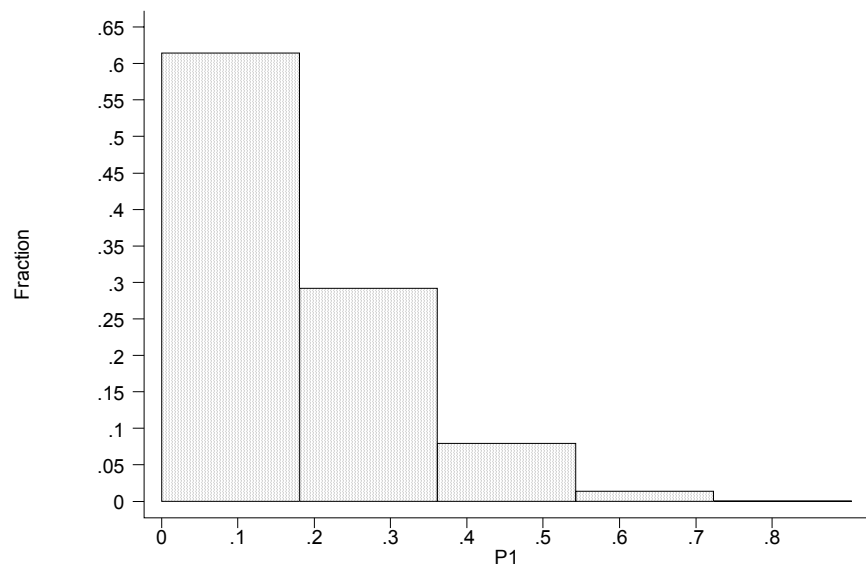


Figure 3. Estimated Probability of Achieving the High Level of Education for Those That Completed The Low Level (π_1)

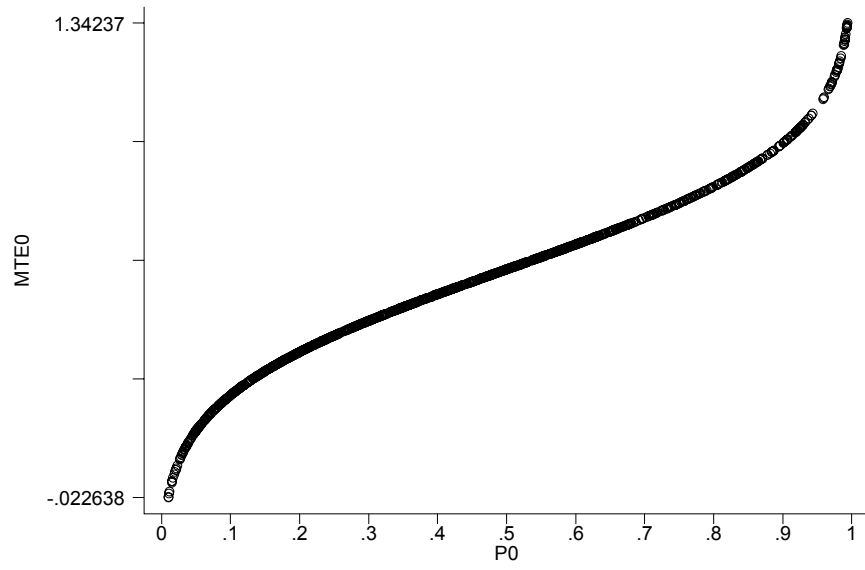


Figure 4. Control Function Approach. $MTE_0(\pi_0, \pi_1 = 0.10)$

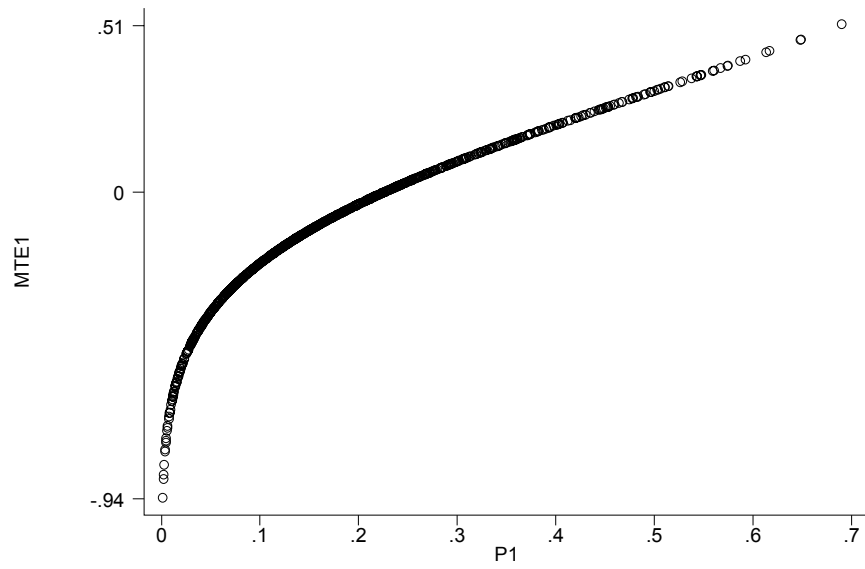


Figure 5. Control Function Approach. $MTE_1(\pi_0 = 0.43, \pi_1)$

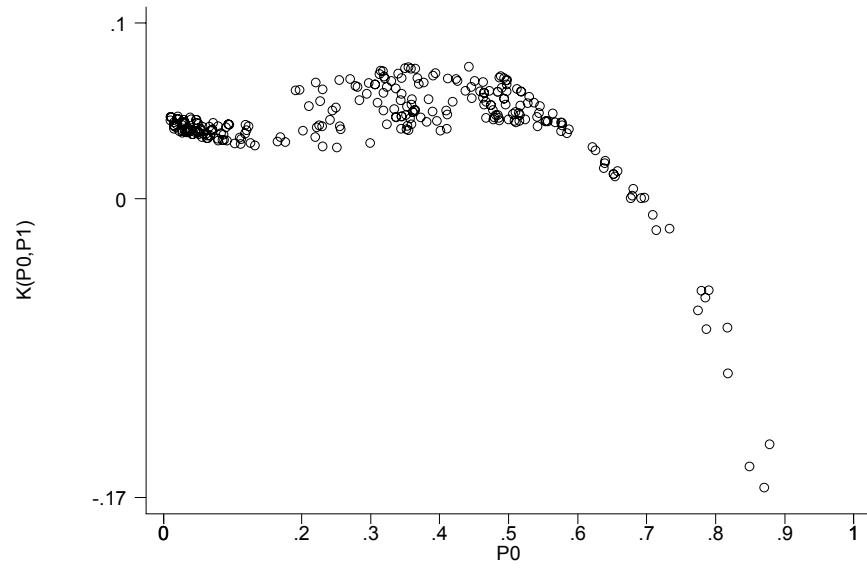


Figure 6. $K(\pi_0, \pi_1)$ for Values of the Probability of Continuing Studying after the Medium Level Smaller than 0.05 ($\pi_1 < 0.05$)

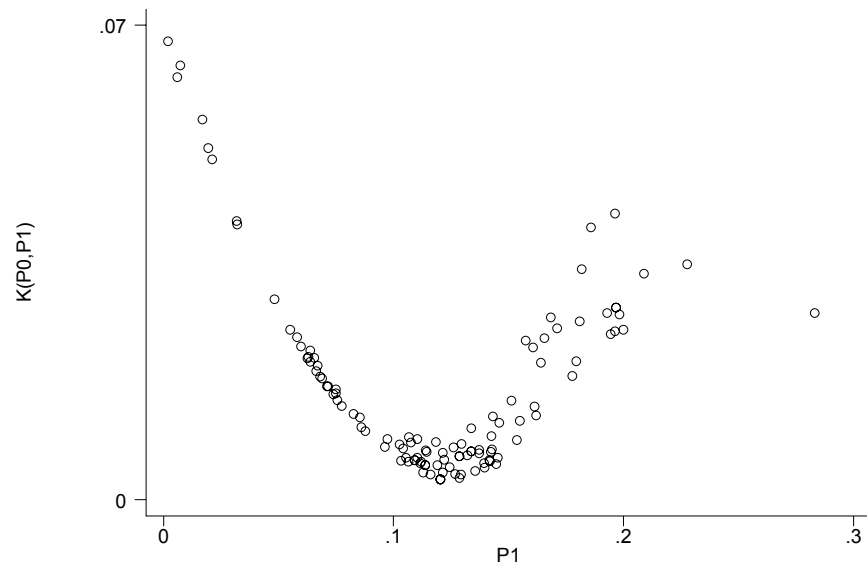


Figure 7. $K(\pi_0, \pi_1)$ for Values of the Probability of Continuing Studying after the Low Level between 0.20 and 0.30 ($0.20 < \pi_0 < 0.30$)

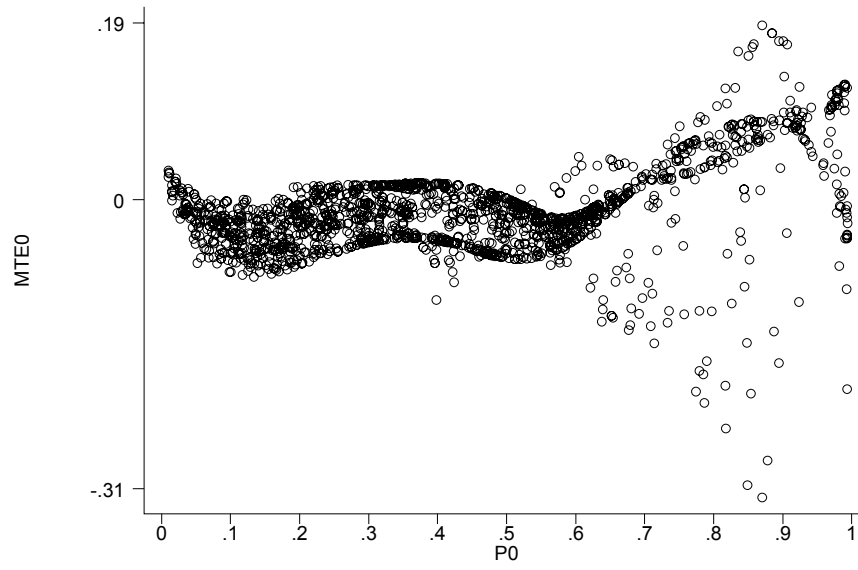


Figure 8. MTE_0 as a Function of the Probability of Continuing Studying after the Low Level (π_0)

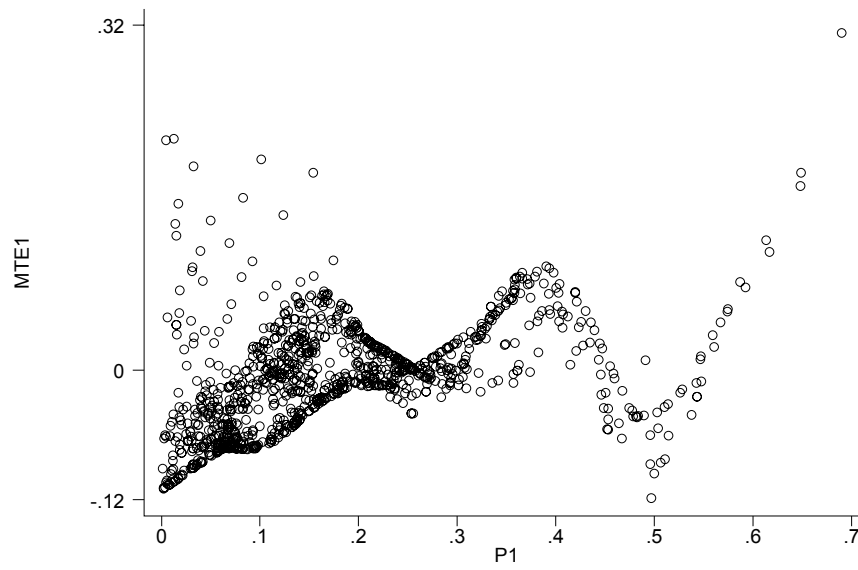


Figure 9. MTE_1 as a Function of the Probability of Continuing Studying after the Medium Level (π_1)

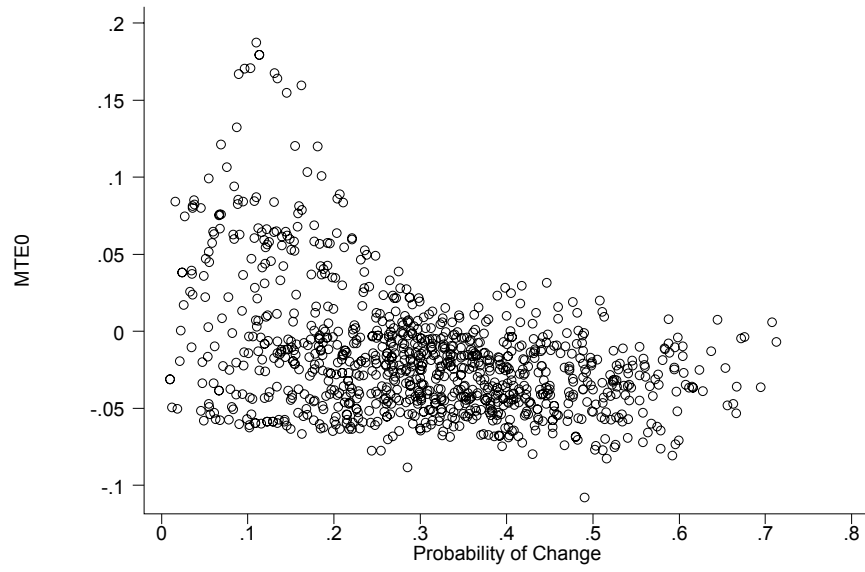


Figure 10. MTE_0 as a Function of the Probability of Being Affected by the 1970 Reform for those that Studied Before It

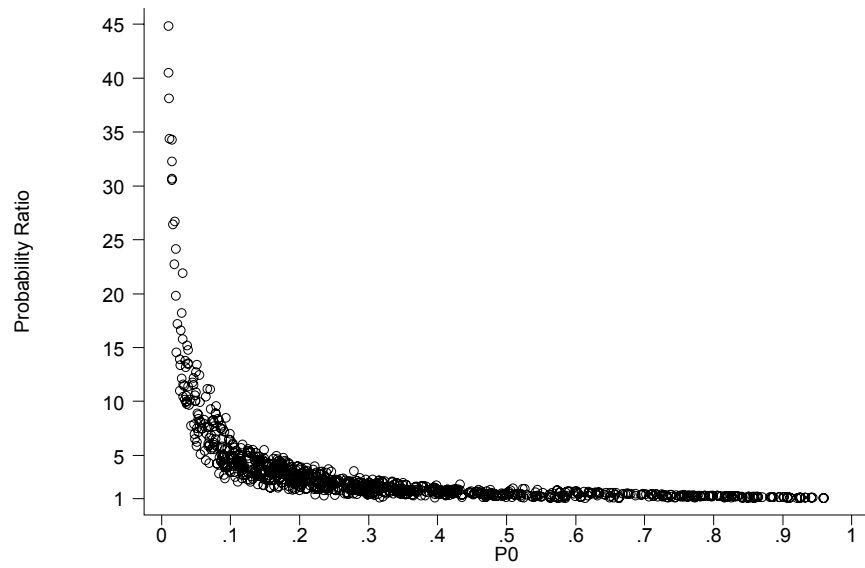


Figure 11. Change in the Probability of Continuing Studying from the Low Level (π_0) Because of the 1970 Reform for those that Studied Before It.