ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Fernández-Sainz, Ana I.; Rodríguez-Póo, Juan M.

Article

An empirical investigation of parametric and semiparametric estimation methods in sample selection models

Revista de Métodos Cuantitativos para la Economía y la Empresa

Provided in Cooperation with:
Universidad Pablo de Olavide, Sevilla

Suggested Citation: Fernández-Sainz, Ana I.; Rodríguez-Póo, Juan M. (2010) : An empirical investigation of parametric and semiparametric estimation methods in sample selection models, Revista de Métodos Cuantitativos para la Economía y la Empresa, ISSN 1886-516X, Universidad Pablo de Olavide, Sevilla, Vol. 10, pp. 99-120

This Version is available at:
https://hdl.handle.net/10419/59086

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

http://creativecommons.org/licenses/by-sa/3.0/es/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

REVISTA DE MÉTODOS CUANTITATIVOS PARA LA ECONOMÍA Y LA EMPRESA (10). Páginas 99–120.
Diciembre de 2010. ISSN: 1886-516X. D.L: SE-2927-06.
URL: http://www.upo.es/RevMetCuant/art44.pdf

An Empirical Investigation of Parametric and Semiparametric Estimation Methods in Sample Selection Models

Fernández-Sainz, Ana I.
Departamento de Econometría y Estadística
Universidad del País Vasco (Spain)
Correo electrónico: ana.fernandez@ehu.es

Rodríguez-Póo, Juan M.
Departamento de Economía
Universidad de Cantabria, Santander (Spain)
Correo electrónico: rodrigjm@unican.es

ABSTRACT

In this paper we analyze empirically different specifications of a sample selection model. We are interested in how the estimates vary across alternative assumptions concerning the joint conditional distribution of the sample selection equation errors, such us the specification of error distribution, the functional relationship of the index function and heteroskedasticity. To do this, we estimate a wage equation for the Spanish labor market using two different approaches: Maximum Likelihood and Two-Step Methods. For the latter, three alternative semiparametric procedures are used to compute the sample selection mechanism, and thus three alternative two-step estimators of the parameters of the wage equation are obtained. We compare theses estimates with Heckman's approach.

Keywords: sample selection models; distributional assumptions; semiparametric two-step estimation methods.
JEL classification: C14; C25; J64.
MSC2010: 62P20; 91B40.

Artículo recibido el 20 de julio de 2010 y aceptado el 10 de diciembre de 2010.

Investigación empírica de métodos de estimación paramétricos y semiparamétricos de modelos de selección muestral

RESUMEN

En este trabajo se analizan empíricamente distintas especificaciones de un modelo de selección muestral. Estamos interesados en conocer cómo las estimaciones de los parámetros varían en función de supuestos alternativos sobre la distribución condicional conjunta de los errores de la ecuación de selección, de la forma funcional de la función índice y la heteroscedasticidad. Para el análisis, estimamos una ecuación de salarios para el mercado de trabajo español usado dos enfoques distintos: máxima-verosimilitud y métodos en dos etapas. Para el caso de la estimación en etapas, consideramos tres procedimientos semiparamétricos alternativos para el cómputo del mecanismo de selección. Así, se obtienen tres estimadores en dos etapas de los parámetros de la ecuación de salarios. Comparamos las estimaciones con la obtenidas siguiendo el método de Heckman.

Palabras clave: modelos de selección muestral; hipótesis distribucionales; métodos de estimación en dos etapas semiparamétricos.
Clasificación JEL: C14; C25; J64.
MSC2010: 62P20; 91B40.

1 Introduction

In any microeconometric study of the labor market, two facts are readily apparent: many individuals do not work, and wages are not available to non-working people. This introduces a serious bias in the estimation of many behavioral equations, since only a non-randomly chosen subsample is available to estimate the parameters of interest. This is pointed out in Gronau (1974) and Heckman (1974). In their papers, a sample selection model is introduced, consisting of two equations: a wage equation, explaining the potential log-wage rate of every individual, including non-workers, and a selection equation indicating whether or not someone is employed and therefore the wage is observed. Since then, many estimation techniques have been developed in econometrics literature to account for this issue. As pointed out in Vella (1998), in empirical literature, the main extension of Heckman's seminal paper has been the use of semiparametric and nonparametric methods in the estimation of relationships of interest. In fact, these techniques have allowed empirical researchers to relax some rather strong assumptions that were introduced in the early papers of this literature: Mainly those involving the form of the distribution of the selection mechanism and those involving the statistical relationship between errors and explanatory variables.

Although these new estimation procedures for sample selection models have received a lot of attention in theoretical econometric literature (see among others Ahn and Powel, 1993; Andrews and Schafgans, 1998; Chen and Lee, 1998; Das, Newey and Vella, 2000 and Lewbel, 2007) very few applications are available. Furthermore, with the exceptions of Melenberg and Van Soest (1993), Vella (1998), Martins (2001) and Coelho, Veiga and Veszteg (2005), no comparison of the different estimation techniques is available.

In this paper we seek to empirically study different specifications of a sample selection model. More precisely, we are interested in analyzing the behavior of the different estimates under alternative assumptions regarding the sample selection equation, such as the specification of error distribution, the functional relationship between selection and explanatory variables (index function) and the statistical relationship between error and explanatory variables (heteroskedasticity). To do this, we estimate a wage equation for the Spanish labor market using two different approaches: Maximum Likelihood and Two-Step Methods. The first technique is used as a benchmark. For the second, three alternative procedures are used to compute the sample selection mechanism, in order to estimate the parameters of the wage equation. Depending on whether the selection equation is fully parametric (where both conditional distribution and index function are known), semiparametric (where only the form of the index is known) or nonparametric, three alternative two-step estimators of the parameters of the wage equation are obtained: Those proposed by Heckman (1979), Powell (1987) and Ahn and Powell (1993). The impact of omitted heteroskedasticity in the selection equation is analyzed in a fully parametric approach through standard Lagrange multiplier tests. Finally, a consistent specification test of Heckman's model is also implemented. This test is based on a general specification test developed by Horowitz and Härdle (1994).

The paper is organized as follows. In Section 2 we introduce the model and the data. In Section 3, we develop the estimation methods and we present the main results. Finally, in Section 4 we conclude.

2 Data and Model

In order to estimate a wage equation for the Spanish labor market we have available data obtained from the *Encuesta de Población Activa (EPA)*, the Spanish quarterly Labor Force Survey. This survey has taken place every quarter since 1975 and is collected by the National Bureau of Statistics (INE). It covers approximately 60,000 households and contains information about 150,000 individuals aged over 16. It provides information at different levels of disaggregation at both national and regional level. From these surveys, in the second quarter of 1990 the National Bureau of Statistics randomly selected a cross-section of 4,989 individuals (1,010 are unemployed looking for work) and provided additional information about some variables that were considered relevant for labor market participation analysis.

The variables included in this data set are defined in Table 1, where we also include some descriptive statistics.

Variable	Description	Whole Sample	Worker Sample
AGE16-19	dummy, 1 if age 16 to 19	0.1317 (0.3383)	0.1111 (0.3145)
AGE20-25	dummy, age 20 to 25	0.2653 (0.4417)	0.2565 (0.4371)
AGE26-35	dummy, age 26 to 35	0.2782 (0.4483)	0.2614 (0.4398)
AGE>45	dummy, older than 45	0.1386 (0.3457)	0.1437 (0.3511)
ELEMENTARY	dummy, elementary school	0.3550 (0.4773)	0.3399 (0.4740)
H.SCHOOL	dummy, high school	0.1158 (0.3202)	0.1062 (0.3083)
UNIVERSITY	dummy, university	0.0643 (0.2455)	0.0392 (0.1943)
U-RATE	unemployment rate	0.1718 (0.0693)	0.1714 (0.0710)
NOT HEAD OF HOUSE	dummy, 1 if person is not head of household	0.7039 (0.4567)	0.6160 (0.4867)
SEXF	dummy, 1 if female	0.6802 (0.4666)	0.6258 (0.4843)
SINGLE	dummy, 1 if single	0.6891 (0.4631)	0.7255 (0.4466)
PARTICIPATING	dummy, 1 if participating	0.6059 (0.4888)	... (...)
SIZE		1010	612

Table 1: *Comparative statistics of explanatory variables, mean and standard deviation (in brackets).*

Before specifying the wage equation, we need to stress some issues related to both the characteristics of the Spanish labor market and the data. In 1990 most contracts in the Spanish labor market were signed for forty hours per week. Almost no part time contracts were allowed at this time. In a standard wage equation model (Heckman, 1974), this would imply that offered wages are not affected by hours worked. In this sense, a most accurate specification of a wage equation for our purposes

would be the one proposed in Gronau (1973), where in fact offered wages are not assumed to depend on hours worked. A second issue that also affects the standard specification of a wage equation is the fact that in our data set we only observe the wage of the individual that has been randomly selected. No information about other incomes in the household is available. Nor do we know the number of components of the family or other related issues, for example children's ages. In our analysis we will not therefore specify different wage equations for males and females.

Taking into account the above restrictions, we propose the following sample selection model, also referred in Amemiya (1985) as the Type II Tobit model:

$$z_i^* = f_1(x_{1i}) + u_{1i}, \quad i = 1, \ldots, n, \tag{1}$$

$$y_i^* = f_2(x_{2i}) + u_{2i}, \quad i = 1, \ldots, n, \tag{2}$$

$$y_i = y_i^*, \quad \text{if} \quad z_i^* > 0, \tag{3}$$

$$y_i = 0, \quad \text{if} \quad z_i^* \leq 0. \tag{4}$$

Here, $f_1(x_1)$ and $f_2(x_2)$ are real functions and (u_1, u_2) are random variables whose realizations are unobserved by the researcher. The observed variables are y_i, z_i, x_{1i} and x_{2i}. x_1 and x_2 might contain common variables. z_i denotes a dummy variable indicating whether the i-th individual has a paid job ($z_i = 1 \quad \text{if} \quad z_i^* > 0$) or not ($z_i = 0 \quad \text{if} \quad z_i^* \leq 0$), and y_i is the wage someone receives if he/she is employed. It is only observed iff $z_i = 1$. Equation (2) is the so called market wage equation. The explanatory variables in this equation, x_1, are the standard ones in this type of models (see Vella, 1998), i.e. one dummy variable for the gender differential effect, and three dummy variables referring to education level. In a first attempt at specification we also included age as a proxy of experience, but it turned out that this variable was more relevant in explaining participation, and therefore since we needed a exclusion restriction in order to identify the parameters of the market wage equation we decided to remove this variable from the wage equation.

Equation (1) reflects the difference between the market and the reservation wage. It is a reduced form participation equation. Therefore, among the explanatory variables in this equation, x_2, we can find variables related to both market and individual characteristics: One dummy variable for the gender differential effect, four dummy variables associated with age and three dummy variables referring to education level. Education level is used as an indicator of potential earnings of individuals. We also used the unemployment rate in the area of residence since participation may depend on cyclical conditions of the economy. We decided also to include a dummy variable that indicates marital status. This last variable approximates the reservation wage.

The selection problem comes from the fact that we are interested in understanding the relationship represented in equation (2), but we observe only a subsample of observations due to the observability rule that is represented in equations (3) and (4). Finally, it is important to note that equations (1) to (4) alone do not restrict the distribution of (y, z) conditional on (x_1, x_2). An econometric model takes on content when restrictions are imposed on $f_1(\cdot)$, $f_2(\cdot)$ and the distribution of (u_1, u_2) conditional on x_1, x_2.

3 Estimation Methods and Results

Any first attempt to introduce an estimation method of the sample selection model must go through the analysis of identification conditions. Following Manski (1993), three different types of identifying restrictions can be imposed on the above specification.

First, we can assume that u_1 and u_2 are statistically independent conditionally on (x_1, x_2). In this case, $f_2(\cdot)$ can be consistently estimated without taking into account the information contained in the other equation. A further restriction not necessary to identify the conditional probability model is that $f_2(\cdot)$ should fall within a specified class of linear parametric models. In this case, standard least squares techniques provide consistent estimates for the parameters of the wage equation.

A second group of identifying restrictions is to assume that the joint distribution of (u_1, u_2) conditionally on (x_1, x_2) belongs to a pre-specified family of parametric density functions. Moreover, $f_1(\cdot)$ and $f_2(\cdot)$ are assumed to be linear parametric functions. These restrictions identify the parameters of the wage equation that can be estimated through maximum likelihood methods. Under the conditions detailed above it is also possible to consistently estimate the parameters of interest by a two-step method proposed in Heckman (1979). This method estimates in a first step the parameters of the selection equation and then, in a second step, the parameters of the wage equation, incorporating a correction term, are estimated by standard weighted least squares. Note that a two-step sample selection estimator with a linear correction term can be consistent for the regression coefficients despite misspecification of distribution (see Olsen, 1981 and Newey, 1999).

Maximum likelihood estimators of the wage equation are extremely sensitive to misspecification in the joint conditional distribution of (u_1, u_2) on (x_1, x_2). Hurd (1979) shows the consequences of omitted heteroskedasticity and Goldberger (1983) describes the effects of non-normality. Newey (1999) analyzes the impact of misspecification of distribution in two-step estimators, Fernández, Rodríguez-Póo and Villanua (2002) show the impact of ignoring heteroskedasticity and Nawata and Nagase (1996) compare the performance of the two estimators by a simulation study.

In order to weaken certain distributional assumptions, a third group of identifying restrictions has been introduced in the literature of sample selection models. The identifying restriction consists of assuming that the conditional distribution of the selection equation error, u_1, depends on a certain function of x_1 (single index function), through an unknown relationship, $h(x_1)$.

Based on this single index restriction, several semiparametric estimation methods have been proposed. The main advantage of these estimation procedures over the ones above is that knowledge of the conditional distribution of u_1 given x_1 is not required, and therefore they are robust to misspecification in error distribution. All are based on two-stage procedures. Powell (1987) and Newey (1991) additionally assume that $h(\cdot)$ is a linear parametric function, and they propose estimating the parameters of this index function through a semiparametric estimation method (Klein and Spady, 1993 or Horowitz and Härdle, 1996). Ahn and Powell (1993) do not impose linearity restrictions on index $h(\cdot)$, but they assume some conditions that guarantee the possibility to estimate nonparametrically this index function.

Note that by estimating a wage equation model with different estimation procedures, maximum likelihood, parametric and semiparametric two-step methods, we can obtain very important information

about possible specification errors in the econometric model. Thus, if we estimate the same wage equation model alternatively with maximum likelihood and with Heckman's two-step procedure, and the estimation results are similar, we can conjecture that the conditional distribution of u_2 given x_2 is approximately normal. Furthermore, by estimating the same econometric model by parametric and semiparametric two-step methods we could guess that in fact the conditional distribution of the sample selection mechanism, u_1 given x_1, is Gaussian if the estimation results are close. Finally, the impact of omitted heteroskedasticity in the parameter estimates of the wage equation can also be considered within the framework of conditionally Gaussian distributions.

In the estimation under these different identifying restrictions we can compare the parameter estimates, and decide, if possible, what sort of specification best fits the data structure. In what follows we will estimate the model described in equations (1) to (4) by using the different estimation techniques described above. To do this we will add the identification conditions already discussed for each of these estimation procedures.

3.1 Maximum Likelihood Estimators

As already remarked in the previous section, in order to implement the maximum likelihood estimators of the wage equation, in the setting described by equations (1), (2), (3) and (4) we add the following restrictions:

$$\left(\begin{array}{c} u_1 \\ u_2 \end{array} \middle| \, X = x \right) \sim N\left(0, \Sigma(x, \alpha)\right) \tag{5}$$

and

$$\Sigma(x, \alpha) = \left(\begin{array}{cc} \sigma_1^2(x_1; \alpha_1) & \rho\sigma_1(x_1; \alpha_1)\sigma_2(x_2; \alpha_2) \\ \\ \rho\sigma_1(x_1; \alpha_1)\sigma_2(x_2; \alpha_2) & \sigma_2^2(x_2; \alpha_2) \end{array} \right). \tag{6}$$

Note that $x = (x_1 \; x_2)$. Furthermore, we also assume that $f_1(x_1) = x_1^T\beta_1$, $f_2(x_2) = x_2^T\beta_2$ and the functions $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ are known by the researcher and belong to some family of parametric functions. Under these conditions, β_1 is identified up to a scale factor, β_2 and ρ are identified, and the nuisance parameters α_1, α_2 can also be identified only under some specific functional forms for heteroskedasticity. For example, if $\sigma_1^2(x_1; \alpha_1) = \kappa_1 \exp\left(x_1^T\alpha_1\right)$ and $\sigma_2^2(x_2; \alpha_2) = \kappa_2 \exp\left(x_2^T\alpha_2\right)$, $\kappa_1 > 0$ and $\kappa_2 > 0$, then the nuisance parameters, α_1 and α_2, and the second scale factor, κ_2, are identified. The vector of coefficients, β_1, will be identified up to the scale, κ_1.

The statistical model represented in equations (1)–(6) nests a great variety of specifications. For example, if we make $\rho = 0$, we are imposing statistical independence between the selection and the wage equation. This restriction, as indicated in Section 2, identifies the parameters of the wage equation regardless of the distribution of errors. We can also consider the case where errors are independent of explanatory variables. That is, $\sigma_1^2(x_1; \alpha_1) = \kappa_1$ and $\sigma_2^2(x_2; \alpha_2) = \kappa_2$.

The unrestricted likelihood function takes the following form,

$$\begin{aligned} \ln L \;=\; & \sum_{i=1}^{n} \left[(1 - y_i) \ln \left(1 - \Phi\left(\frac{x_{1i}^T\beta_1}{\sigma_1\left(x_{1i}; \alpha_1\right)} \right) \right) + y_i \ln \left(\Phi\left(\frac{x_{1i}^T\beta_1}{\sigma_1\left(x_{1i}; \alpha_1\right)} \right. \right. \right. \\ & \left. \left. \left. + \rho \left(\frac{y_i - x_{2i}^T\beta_2}{\sigma_2(x_{2i}; \alpha_2)} \right) \left(1 - \rho^2\right)^{-\frac{1}{2}} \right) \frac{1}{\sigma_1\left(x_{1i}; \alpha_1\right)} \phi\left(\frac{y_i - x_{2i}^T\beta_2}{\sigma_2(x_{2i}; \alpha_2)} \right) \right) \right], \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ stand respectively for the Gaussian density and distribution function. We compute maximum likelihood estimates of four nested wage equation models for Spanish labor market data. The models are the following:

Model I: $\rho = 0$, $\sigma_1^2(x_1;\alpha_1) = \kappa_1$ and $\sigma_2^2(x_2;\alpha_2) = \kappa_2$.

Model II: $\sigma_1^2(x_1;\alpha_1) = \kappa_1$ and $\sigma_2^2(x_2;\alpha_2) = \kappa_2$.

Model III: $\sigma_1^2(x_1;\alpha_1) = \kappa_1 \exp\left(x_1^T\alpha_1\right)$ and $\sigma_2^2(x_2;\alpha_2) = \kappa_2$.

Model IV: $\sigma_1^2(x_1;\alpha_1) = \kappa_1 \exp\left(x_1^T\alpha_1\right)$ and $\sigma_2^2(x_2;\alpha_2) = \kappa_2 \exp\left(x_2^T\alpha_2\right)$.

In Table 2 we show the estimates for the wage equation, and the correlation coefficient between this equation and the sample selection one. In all models, standard deviations of the maximum likelihood estimators have been computed using the variance-covariance matrix that is robust to misspecifications of the conditional distribution of the errors. It is obtained by considering the Maximum Likelihood estimator as a special case of M-estimators (see Gourieroux and Monfort, 1995; Vol. I, p. 213). Note that under Gaussian errors, this variance-covariance matrix is the inverse of the Fisher information matrix.

Variable	**Model I**	**Model II**	**Model III**	**Model IV**
Constant	6.155 (0.03)	6.261 (0.03)	6.245 (0.03)	6.227 (0.03)
Sexf	−0.076 (0.04)	−0.022 (0.04)	−0.051 (0.04)	−0.004 (0.03)
Elementary	−0.168 (0.04)	−0.112 (0.04)	−0.111 (0.04)	−0.029 (0.04)
H. School	0.079 (0.06)	0.149 (0.06)	0.112 (0.06)	0.224 (0.05)
University	0.419 (0.09)	0.565 (0.09)	0.501 (0.10)	0.707 (0.07)
ρ	... (...)	−0.624 (0.07)	−0.543 (0.09)	−0.677 (0.05)
κ_2	0.438 (0.01)	0.484 (0.02)	0.217 (0.02)	0.144 (0.01)

Table 2: *Maximum likelihood estimates of the wage equation, standard deviation in brackets.*

Across the different models, the estimation results do not change significantly in size or sign. In all cases the sex dummy variable is insignificant[1] and small. However, the educational dummy variables are all significant and they keep their sign unchanged in all models. In fact, the dummy elementary school variable has a significant negative impact and the dummy university variable has a significant, strong, positive effect. It is important to remark that in the last three cases, ρ is significantly different from zero. On these grounds, Model I does not appear to be a reasonable specification.

In Table 3 we show maximum likelihood estimates of the parameters of the selection equation. In doing this, our aim is to analyze the impact of a possible misspecification on the selection equation in the wage equation. For example, if Model III is the right specification, then maximum likelihood

[1]Significance level: 5%.

estimators of the parameters in models I and II are inconsistent. A very interesting issue is studying how the parameter estimates of the wage equation change across different specifications of the selection equation. Note that we have chosen the absolute value of the parameter associated with the unemployment rate in the participation equation as a normalization scale instead of using more usual normalization scales such as $\sigma_1 = 1$. This is to allow for comparisons against semiparametric estimation methods.

Variable	**Model I**	**Model II**	**Model III**	**Model IV**
Constant	2.170 (2.57)	2.294 (2.59)	1.129 (0.99)	3.06 (0.53)
Sexf	−0.584 (0.76)	−0.595 (0.74)	−0.310 (0.33)	−1.066 (0.24)
Age16-19	−1.054 (1.39)	−1.292 (1.61)	−3.726 (4.75)	−3.202 (0.57)
Age20-25	−0.546 (0.77)	−0.359 (0.53)	1.088 (0.69)	0.652 (0.19)
Age26-35	−0.384 (0.56)	−0.313 (0.46)	−0.633 (0.72)	−0.411 (0.17)
Age>45	−0.886 (1.17)	−0.825 (1.03)	−0.916 (0.98)	−1.504 (0.30)
Elementary	0.065 (0.24)	0.233 (0.36)	3.531 (6.56)	0.915 (0.23)
H. School	−0.243 (0.42)	−0.127 (0.32)	6.022 (7.88)	−0.089 (0.16)
University	−1.082 (1.43)	−1.051 (1.32)	−0.131 (1.75)	−2.174 (0.41)
U-rate	1.0 (..)	1.0 (..)	1.0 (..)	1.0 (..)
Single	1.237 (1.59)	1.038 (1.27)	1.828 (2.37)	1.816 (0.34)
Not Head of House	−1.756 (2.26)	−1.932 (2.36)	−1.237 (1.38)	−3.149 (0.56)
κ_1	0.877 (0.12)	0.942 (0.12)	1.456 (0.74)	1.819 (0.087)

Table 3: *Maximum likelihood estimates of the selection mechanism.*

As expected, the estimates of the parameters in the sample selection equation present more significant changes across models, in both sign and size than those of the wage equation. This is particularly true in the dummies Age 20-25 and High School. Moreover, the dummies Single and Not head of house are only significant in the last column. These changes have also been remarked in other studies such as Gerfin (1996), Fernández and Rodríguez-Póo (1997) and Martins (2001). One possible explanation may be the presence of heteroskedasticity in the selection equation. If this is the case, the estimators in models I and II are inconsistent and their results are meaningless. To analyze this issue more precisely, we show the maximum likelihood estimates of the variance parameters both in the selection and wage equation in Table 4.

The results estimated from Models III and IV are not conclusive. In fact, the main interest of these estimates is that they allow us to construct general specification tests for nested models. More precisely, we are interested in testing Model I against Model II ($H_0 : \rho = 0$), for independence between the two equations; Model II against Model III ($H_0 : \sigma_1(x_1; \alpha_1) = \kappa_1$), for homoskedasticity

Variable	**Model III**	**Model IV**
Heter. Sample selection		
Sexf	0.687 (0.50)	0.705 (0.09)
Age16-19	−0.274 (2.51)	1.066 (0.10)
Age20-25	−0.196 (1.35)	0.675 (0.10)
Age26-35	−1.323 (0.71)	0.156 (0.10)
Age>45	−0.581 (0.74)	0.553 (0.09)
Elementary	3.996 (2.82)	1.596 (0.10)
H. School	4.710 (1.54)	1.650 (0.10)
University	2.845 (3.84)	0.559 (0.10)
U-rate	−2.740 (3.03)	0.570 (0.10)
Single	0.723 (1.03)	−0.881 (0.09)
Not Head of House	2.949 (0.80)	1.231 (0.09)
Heter. Wage		
sexf	... (...)	0.259 (0.08)
Elementary	... (...)	0.685 (0.08)
H. School	... (...)	0.529 (0.09)
University	... (...)	0.550 (0.09)

Table 4: *Variance parameters in selection and wage equations.*

	Model II	**Model III**	**Model IV**
Model I	31.31 (2.2e-08)	76.76 (1.7e-11)	90.29 (2.2e-12)
Model II	... (...)	45.45 (4.0e-06)	58.98 (3.8e-07)
Model III	... (...)	... (...)	13.53 (0.009)

Table 5: *Likelihood ratio statistic and p-values in brackets.*

in the selection equation; Model III against Model IV ($H_0 : \sigma_2(x_2; \alpha_2) = \kappa_2$), for homoskedasticity in the wage equation, and it is also interesting to test Model II against Model IV ($H_0 : \sigma_1(x_1; \alpha_1) = \kappa_1$ and $\sigma_2(x_2; \alpha_2) = \kappa_2$), for homoskedasticity in both equations. In Table 5 we present the values of the different likelihood ratio tests for the specifications.

Within the framework of conditionally Gaussian models (this assumption is maintained throughout the four models), the different likelihood ratio tests support the idea of a specification close to a Gaussian conditionally distributed model with heteroskedasticity in the participation equation and significant sample selection bias (Model III). As remarked in previous studies (see Fernández and Rodríguez-Póo, 1997), heteroskedasticity is present in the participation equation and therefore a specification based on heteroskedasticity appears to be a more reasonable structure. The coefficient estimates of the wage equation are rather sensitive to the specification of the participation equation. Mainly, these parameter estimates change significantly when the correlation between the two equations is different from zero and when we assume exponential heteroskedasticity for the participation equation.

3.2 Two Stage Estimation Procedures

This estimation procedure relies on the following expression for the wage equation that can be easily obtained from the structural model represented in equations (1) to (4),

$$E\left(y|z^{*}>0,x_{1},x_{2}\right)=x_{2}^{T}\beta_{2}+m\left(x_{1},x_{2}\right), \tag{7}$$

where

$$m\left(x_{1},x_{2}\right)=E\left(u_{2}|x_{2},f_{1}(x_{1})+u_{1}>0\right). \tag{8}$$

Then the parameter vector, β_2, of the wage equation can be estimated through the following corrected regression equation,

$$y_{i}=x_{2i}^{T}\beta_{2}+\hat{m}\left(x_{1i},x_{2i}\right)+\upsilon_{i}, \tag{9}$$

where

$$\upsilon_{i}=y_{i}-m\left(x_{1i},x_{2i}\right)-\left\{\hat{m}\left(x_{1i},x_{2i}\right)-m\left(x_{1i},x_{2i}\right)\right\},$$

and $\hat{m}\left(x_{1},x_{2}\right)$ can be any estimator of $m\left(x_{1},x_{2}\right)$.

In considering two-step methods, it is interesting to categorize them in different groups according to the restrictions that are imposed in order to estimate the function $m\left(x_{1},x_{2}\right)$. In the first group we will include those that fully explode parametric assumptions. This is the case of the estimator proposed in Heckman (1979). Since the selection equation error is assumed to be conditionally Gaussian then, considering equations (1), (2), (3) and (4) and $f_{1}\left(x_{1}\right)=x_{1}^{T}\beta_{1}$, the following expression can be obtained

$$m\left(x_{1},x_{2}\right)=\rho\sigma_{2}\lambda\left(\frac{x_{1}^{T}\beta_{1}}{\sigma_{1}}\right), \tag{10}$$

and

$$\lambda(u)=E\left[\frac{u_{1}}{\sigma_{1}}\middle|\frac{u_{1}}{\sigma_{1}}>-z\right]=\frac{\phi(z)}{\Phi(z)}. \tag{11}$$

Here $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and the distribution function. The function $m(x_1,x_2)$ can be estimated in a first stage by a probit maximum likelihood technique, and then

$$\hat{m}\left(x_{1},x_{2}\right)=\rho\sigma_{2}\lambda\left(x_{1}^{T}\hat{\beta}_{1}^{*}\right), \tag{12}$$

where $\beta_{1}^{*}=\frac{\beta_{1}}{\sigma_{1}}$.

From (9), (10) and (11) it is clear that $E\left(\upsilon|z^{*}>0,x_{1},x_{2}\right)=0$ and the variance $\mathrm{Var}\left(\upsilon|z^{*}>0,x_{1},x_{2}\right)$ is not constant. Therefore, O.L.S. estimators of β_2 are consistent, but unfortunately the standard

Variable	**Heckman**	**Powell**	**Ahn-Powell**	**Newey**
Constant	6.021 (0.08)	... (...)	... (...)	... (...)
Sexf	−0.444 (0.12)	−0.455 (0.18)	−0.386 (0.22)	−0.402 (0.16)
Elementary	−0.121 (0.08)	−0.083 (0.08)	0.012 (0.11)	−0.1 (0.02)
H. School	0.184 (0.12)	0.103 (0.13)	0.209 (0.15)	0.112 (0.14)
University	0.405 (0.19)	0.441 (0.22)	0.591 (0.25)	0.425 (0.21)
λ	0.156 (0.09)	... (...)	... (...)	... (...)

Table 6: *Two-step estimates of the sample selection model. Wage equation.*

Variable	**Probit M.L.**	**Horowitz-Härdle**
Constant	2.176 (2.47)	... (...)
Sexf	−0.586 (0.74)	−0.029 (0.08)
Age $16-19$	−1.050 (1.34)	−0.403 (0.25)
Age $20-25$	−0.544 (0.75)	−0.220 (0.22)
Age $26-35$	−0.384 (0.55)	−0.223 (0.19)
Age >45	−0.889 (1.12)	−0.169 (0.09)
Elementary	0.065 (0.23)	−0.104 (0.12)
H. School	−0.240 (0.40)	−0.270 (0.27)
University	−1.078 (1.36)	−0.722 (0.25)
U-rate	1 (...)	1 (...)
Single	1.236 (1.52)	0.483 (0.12)
Not Head of House	−1.762 (2.17)	0.027 (0.19)
κ_1	0.879 (0.12)	... (...)

Table 7: *Two-step estimates of the sample selection model. Selection Equation.*

errors computed in the traditional way are inconsistent (Amemiya, 1985). Several methods have been proposed for estimating these standard errors consistently (Newey, 1987). In the first column of Table 6 we represent O.L.S. estimates of the corrected wage equation. The standard deviations of the parameter estimates have been computed using the method proposed in White (1980), which provides estimators that are robust to heteroskedasticity.

If we compare the two-step estimates of the wage equation with the maximum likelihood estimates already shown in Tables 2 and 6, we can observe strong disagreements between the two groups of results. The sex variable is insignificant and small in the M.L.E. case, whereas in the two-step

method it is large and significant. A similar discrepancy occurs for the impact of the highest degree, i.e. the dummy elementary school has a significantly negative impact for all the maximum likelihood estimates but one, and is not significant for the two-step estimator. These disagreements may be due to possible misspecification errors that come from the sample selection equation. In other words, even if the wage equation is correctly specified, excluding the case when sample selection and wage equations are uncorrelated, a specification error in the selection equation can cause misleading estimation results in the wage equation.

Before implementing further developments of two-step methods that require weaker assumptions on the sample selection specification equation, we perform an empirical comparative analysis of a standard probit maximum likelihood estimator with a semiparametric estimation method of the binary equation, in order to detect posible misspecifications in this equation. Much deeper research into this problem can be found in Fernández and Rodríguez-Póo (1997). Among other possible specification errors, omitted hetereoskedasticity and non-normality can cause inconsistency in probit maximum likelihood estimators. Since Manski (1975), much research has been devoted to the estimation of binary response models without assuming either homoskedasticity or knowledge of the conditional distribution of the error term. These semiparametric methods rely basically on three different groups of identifying restrictions (see Manski, 1988): quantile independence restrictions, single index restrictions and independence between errors and explanatory variables. Weighted average derivative estimators belong to the second group of identifying restrictions, and are of great interest for our purposes since, first, they allow for heteroskedasticity that depends on the index function and, second, there is no need to specify the form of the conditional distribution of the error term in order to estimate the parameters of index function. Horowitz and Härdle (1996) propose a weighted average derivative estimator that allows for discrete explanatory variables. For the sake of comparison, in Table 7 we show probit maximum likelihood estimates of the selection equation and the semiparametric estimates proposed by Horowitz and Härdle (1996). A detailed description of this last estimator is provided in the Appendix. In order to make the calculations we take $c_0 = 0.2$, $c_1 = 0.8$, a fourth order kernel with support $[-1, 1]$,

$$K(u) = \frac{105}{64}\left(1 - 5u^2 + 7u^4 - 3u^6\right) I(|u| < 1), \tag{13}$$

and the vector of parameters associated to continuous variables is computed as a weighted average of density weighted average derivative estimates with weights equal to the frequencies of the discrete variables. The bandwidth, h_n was chosen by least square cross-validation. All the above choices are justified by the assumptions introduced in the paper by Horowitz and Härdle (1996). The higher order kernel is needed to deal with the bias term in the semiparameteric estimator.

Comparing the two results, we observe significant changes in both the size and the sign of the coefficient estimates. More specifically, in the semiparametric estimates most variables, as expected, become significant, and the signs of both the *elementary* and *not head of household* dummy variables change their sign. A deeper analysis could be performed, but our interest here is to note that these significant changes can be due to misspecification errors in the parametric equation. Of course, as already remarked in a simulation study in Nawata and Nagase (1996) and Fernández, Rodríguez-Póo and Villanua (2002), these errors may seriously affect the properties of the estimators of the wage equation.

A second group of sample selection estimators relies on the so called single index assumption (see

Manski, 1993). Let $h(x_1)$ be a known index, and assume that $f_1(x_1)$ and (u_1, u_2) vary with x_1 only through $h(\cdot)$. Then, (8) can be written as

$$m(x_1, x_2) = E(u_2|x_2, f_1[h(x_1)] + u_1 > 0) = g[f_1[h(x_1)], x_2], \tag{14}$$

and hence, from (7),

$$E(y|z^* > 0, x_1, x_2) = x_2^T \beta_2 + g[f_1[h(x_1)], x_2]. \tag{15}$$

Note that if no assumption is made about the conditional distribution of the error terms, even if $h(\cdot)$ is known, the function $g(\cdot, \cdot)$ is unknown. One way to estimate consistently the parameters β_2 of the wage equation is to take advantage of the partial additive structure of (15) by noting that

$$y_i = x_{2i}^T \beta_2 + g[f_1[h(x_{1i})], x_{2i}] + \nu_i, \tag{16}$$

where

$$\nu_i = y_i - E(y_i|z^* > 0, x_{1i}, x_{2i}).$$

Then, if we take two different observations of indexes i and j such that

$$g[f_1[h(x_{1i})], x_{2i}] = g[f_1[h(x_{1j})], x_{2j}],$$

we obtain

$$y_i - y_j = \beta_2^T (x_{2i} - x_{2j}) + \nu_i - \nu_j. \tag{17}$$

Since

$$E[\nu_i - \nu_j | z^* > 0, x_1, x_2] = 0,$$

then a consistent estimator of β_2 is given by the following weighted least squares estimator:

$$\hat{\beta}_{2n} = \left[\binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \hat{w}_{ijn} (x_{2i} - x_{2j}) (x_{2i} - x_{2j})^T \right]^{-1} \times \binom{n}{2}^{-1} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \hat{w}_{ijn} (x_{2i} - x_{2j}) (y_i - y_j). \tag{18}$$

The sequence of weights $\hat{w}_{ijn}$ can take several forms. If the index function is assumed to be known and parametric, i.e. $f_1[h(x_1)] = x_1^T \beta_1$, then the weight function is

$$\hat{w}_{ijn} = \frac{1}{h_n} K\left(\frac{(x_{1i} - x_{1j})^T \hat{\beta}_{1n}}{h_n} \right) \qquad \text{for} \qquad i, j = 1, ..., n, \tag{19}$$

and $\hat{\beta}_{1n}$ is any root-n consistent semiparametric estimator of β_1 (for example, the one proposed in Horowitz and Härdle (1996)). $K(\cdot)$ is a kernel function and h_n is the window width.

The estimator defined in (18), jointly with the weights in (19), has already been proposed in Powell (1987), where it is shown that under some technical assumptions $\hat{\beta}_{2n}$ is consistent and asymptotically normal. One can relax the above assumption in the index function, $h(x_1)$, by assuming that this function is unknown and needs to be estimated by the researcher. Let us define the following composite function:

$$q(x_1) = f_1[h(x_1)].$$

Then, in this case, the weights in (18) take the following expression:

$$\hat{w}_{ijn} = \frac{1}{h_n} K\left(\frac{\hat{q}(x_{1i}) - \hat{q}(x_{1j})}{h_n} \right) \qquad \text{for} \qquad i, j = 1, ..., n, \tag{20}$$

and $\hat{q}(x_{1i})$ are multivariate nonparametric kernel regression estimators:

$$\hat{q}(x_{1i}) = \frac{\frac{1}{nl_n}\sum_{j=1}^{n}\prod_{k=1}^{K} L\left(\frac{x_{1ki}-x_{1kj}}{l_n}\right)\mathbf{1}\left(z_j^* > 0\right)}{\frac{1}{nl_n}\sum_{j=1}^{n}\prod_{k=1}^{K} L\left(\frac{x_{1ki}-x_{1kj}}{l_n}\right)}, \qquad i = 1, ..., n, \tag{21}$$

where $\mathbf{1}(\cdot)$ stands for the indicator function.

Note that, in order to implement this procedure, we need to use two different bandwidths, h and l. This can create several problems, and empirically it represents an important drawback of this method. It is also necessary to use two different kernels, $K(\cdot)$ and $L(\cdot)$. Ahn and Powell (1993) show that the parameters of the wage equation estimated under this technique are root-n consistent, and also calculate their asymptotic distribution.

In Table 6 we show parameter estimates of the wage equation obtained by the method proposed in Powell (1987) and Ahn and Powell (1993). For the first method, the semiparametric estimates of the sample selection equation (the β_1s) are computed following the method proposed in Horowitz and Härdle (1996), and the kernel function $K(\cdot)$ is

$$K(u) = \frac{\tau^3 k(u) - k\left(\frac{u}{\tau}\right)}{\tau\left(\tau^2 - 1\right)}, \tag{22}$$

for $\tau = \sqrt{2}$ and

$$k(v) = \frac{15}{16}\left(1 - v^2\right)^2 I\left(|v| < 1\right). \tag{23}$$

This is a so called higher order kernel, and it is necessary to eliminate a bias term in the asymptotic expression of the estimator for β_2 that would otherwise render it inconsistent. For the kernel function $L(\cdot)$ in the method proposed in Ahn and Powell (1993), we choose the Gaussian kernel, and finally, for the bandwidths the results obtained tend to be more sensitive to the choice of l than to the choice of h. On these grounds, we set the first bandwidth arbitrarily and calculate the second by least square cross-validation.

Standard deviations of the parameter estimators are computed according to the following expression:

$$\begin{aligned}\widehat{\mathrm{Var}}\left(\hat{\beta}_{2n}\right) = \quad & \left[\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\hat{w}_{ijn}\left(x_{2i}-x_{2j}\right)\left(x_{2i}-x_{2j}\right)^T\right]^{-1} \times \\ & \left[\binom{n}{2}^{-1}\sum_{i=1}^{n}\widehat{\mathrm{Var}}\left(y_i - E[y|z_i^* > 0, x_{1i}, x_{2i}]\right)\sum_{j=i+1}^{n}\hat{w}_{ijn}\left(x_{2i}-x_{2j}\right)\left(x_{2i}-x_{2j}\right)^T\right] \times \\ & \left[\binom{n}{2}^{-1}\sum_{i=1}^{n}\sum_{j=i+1}^{n}\hat{w}_{ijn}\left(x_{2i}-x_{2j}\right)\left(x_{2i}-x_{2j}\right)^T\right]^{-1},\end{aligned} \tag{24}$$

where the form of the weights depends on whether the form of the link function $h(\cdot)$ is known and parametric (Powell, 1987), or unknown (Ahn and Powell, 1993).

Finally, Newey (1991) proposes an efficient GMM-estimator that is different from the above approaches. Let $f_1\left[h(x_1)\right] = x_1^T\beta_1$, then, if we substitute this equality into (16) we get

$$y_i = x_2^T\beta_2 + g\left[x_{1i}^T\beta_1, x_{2i}\right] + \nu_i, \tag{25}$$

where

$$\nu_i = y_i - E\left(y|z_i^* > 0, x_{1i}, x_{2i}\right) = u_{2i} - g\left[x_{1i}^T\beta_1, x_{2i}\right],$$

and

$$E\left[\nu_i|z_I^* > 0, x_{1i}, x_{2i}\right] = 0. \tag{26}$$

Equation (26) implies that ν_i is uncorrelated with any vector of functions $\varphi(x_1, x_2)$ in the selected sample

$$E\left\{\mathbf{1}(z^* > 0)\left[\varphi(x_1, x_2) - E\left(\varphi(x_1, x_2)|z^* > 0, x_1, x_2\right)\right]\left[u_2 - g\left[x_1^T\beta_1, x_2\right]\right]\right\} = 0,$$

and furthermore, if we define a more general function $\psi(u_2, \nu)$, under the single index restriction, the following moment condition also holds:

$$E\left\{\mathbf{1}(z^* > 0)\left[\varphi(x_1, x_2) - E\left(\varphi(x_1, x_2)|z^* > 0, x_1, x_2\right)\right]\left[\psi(u_2, \nu) - E\left(\psi(u_2, \nu)|z^* > 0, x_1, x_2\right)\right]\right\} = 0.$$

Based on this set of moment conditions, Newey (1991) proposes taking the sample analog with $\varphi(x_1, x_2) = (x_1 \quad x_2)$ and

$$\psi_j(u_2, \nu; \gamma) = \left[\frac{u_2 - \mu_2}{\sigma_2}\right]^{\kappa_j}\left[\frac{\nu - \mu_\nu}{\sigma_2}\right]^{\lambda_j} \qquad j = 1, \cdots, J.$$

$\gamma = (\mu_2, \mu_\nu, \sigma_2, \sigma_\nu)$ are location and scale parameters of u_2 and ν, and conditional expectations are replaced by nonparametric regression estimators. Then the parameter vector β_2 is estimated by the Generalized Method of Moments where the optimal weighting matrix is taken as suggested in Hansen (1982). For the sake of comparison, we also present the results for the efficient GMM estimator proposed by Newey in Table 6 .

As can be observed in Table 6, there are no significant differences between the alternative two-step estimators. All variables are significant across the estimators, and the only change in sign is the dummy variable *Elementary School* which is positive in the case of the estimator proposed by Ahn and Powell. For the other estimators, the sign of the coefficient related to this variable is negative. These rather stable results can be justified by the robustness of these two-step methods to several misspecification errors. If we compare them with the results obtained for the maximum likelihood estimates (see Table 2), the differences are not really significant either. The coefficient associated with the *female sex* dummy variable is significantly smaller for the M.L.E. estimates than for the two-step ones, but for the others there are not significant divergences. As concluded in other studies, if there is correlation between the two equations, a specification of the wage equation that accounts for conditionally Gaussian errors and heteroskedasticity in the sample selection equation appears as a fairly reasonable model for an empirical problem such as the one proposed in this paper. Finally, as a guideline to enable the empirical researcher to discriminate among the different two-step estimation procedures we present a specification test where the null hypothesis is the model presented in Heckman (1979), and the alternative is semiparametric.

3.3 A Nonparametric Test for Sample Selection Models

This test relies on the idea that if the model under the null hypotheses is the true one, then a no nparametric estimate will deviate from the parametric specification only due to sampling error. In our case, the null hypothesis is the parametric model proposed in Heckman (1979). Recall that under some assumptions on the conditional distribution of the selection equation the correction term in the wage equation has an expression that is proportional to the inverse of Mill's ratio. The alternative is a correction term that is the negation of the null.

The model we aim to test is

$$E\left(y|z^* > 0, x_1, x_2\right) = x_2^T\beta_2 + \rho\sigma_2\lambda\left(\frac{x_1^T\beta_1}{\sigma_1}\right) \tag{27}$$

and

$$\lambda(u) = E\left[\frac{u_1}{\sigma_1}\middle|\frac{u_1}{\sigma_1} > -z\right] = \frac{\phi(z)}{\Phi(z)}. \tag{28}$$

Here $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and the distribution function.

The alternative is the following specification:

$$E\left(y|z^* > 0, x_1, x_2\right) = x_2^T\beta_2 + g\left[x_1^T\beta_1, x_2\right], \tag{29}$$

where $g(\cdot,\cdot)$ is an unknown smooth function. Then, based in the idea of Horowitz and Härdle (1994), we propose the following nonparametric test statistic

$$T_h = \sqrt{h}\sum_{i=1}^{n}\pi\left(x_{1i}^T\hat{\beta}_{1n}, x_{2i}\right)\left\{y_i - x_{2i}^T\hat{\beta}_{2n} - \hat{\rho}_n\hat{\sigma}_2\lambda\left(x_{1i}^T\hat{\beta}_{1n}^*\right)\right\}$$

$$\left\{\hat{\rho}_n\hat{\sigma}_2\lambda\left(x_{1i}^T\hat{\beta}_{1n}^*\right) - \hat{g}_h\left[x_{1i}^T\hat{\beta}_{1n}, x_{2i}\right]\right\}\mathbf{1}\left(z_i^* > 0\right).$$

$\pi(.)$ is a weight function that down weights the extreme observations and $\hat{g}_h(.)$ is a nonparametric estimator of the bias correction:

$$\hat{g}_h\left[u,v\right] = \frac{\sum_i^n K\left(\frac{u-x_{1i}^T\hat{\beta}_{1n}}{h}\right)L\left(\frac{v-x_{2i}}{h}\right)\mathbf{1}(z_i^* > 0)\left\{y_i - x_{2i}^T\hat{\beta}_{2n}\right\}}{\sum_i^n K\left(\frac{u-x_{1i}^T\hat{\beta}_{1n}}{h}\right)L\left(\frac{v-x_{2i}}{h}\right)\mathbf{1}(z_i^* > 0)}.$$

Finally, the vectors $\left(\hat{\beta}_{2n}, \hat{\rho}_n\hat{\sigma}_2, \hat{\beta}_{1n}\right)$ are the parameters estimated under the null (Heckman two-step estimates). The idea of the test is very simple. Under the null hypothesis the last term will be negligible since the nonparametric estimate will deviate from the parametric specification only due to sampling error.

The central term under the null, using standard Central Limit Theorem arguments, will be bounded in distribution and therefore the whole statistic, once conveniently normalized, will tend to a normal density. Under the alternative, the last term is unbounded and this gives the consistency of the test. For a test for a binary logit model, Horowitz and Härdle (1994) show that under the null hypothesis their statistic converges in distribution to the normal density. They also provide a consistent estimator of asymptotic variance. Unfortunately, they do not provide a guide for computing the bandwidth h. Following Proença and Ritter (1994), we have computed the T_h-statistic for different bandwidth values. Since this type of statistic shows a bias that depends linearly on the bandwidth h, it is advisable to choose small bandwidth values, otherwise we might have a non-negligible bias. In Table 8 we present the results.

The p-values are computed using a normal distribution, and the results do not recommend the use of the statistical model proposed in Heckman (1979) for small bandwidth values.

4 Conclusions

In any microeconometric study of the labor market, the estimation of models with sample selection bias is very common from both empirical and theoretical points of view. Since the two-step sample selection estimation methods poposed in Heckman (1979), many estimation techniques have been developed to weaken some strong assumptions.

Bandwidth	T-Statistic	p-value
0.1	0.695	0.2514
0.3	0.763	0.2236
0.5	0.845	0.2035
0.7	0.412	0.3409
1.0	0.564	0.2877
1.3	3.291	0.0005
1.5	4.157	0.0000
1.8	3.412	0.0030
2.0	2.121	0.0170

Table 8: *Horowitz-Härdle test.*

In this paper we estimate a sample selection model taking into account different sets of identifying restrictions. First, the errors are statistically independent conditionally on regressors; second, if the equations are linear in the regressors, then maximum likelihood or Heckman two-step methods can be used. Third, if the conditional distribution of the selection equation error depends on a function h(.), then semiparametric estimation two-step methods are available . The main advantage is that no knowledge of conditional distribution is necessary, and so the estimators are robust to misspecification in error distribution.

Empirical results support the idea of a specification close to a Gaussian models with heteroskedasticity in the selection equation. If two-step methods are used, estimates of the parameters of the wage equation do not vary across different specifications (parametric and semiparametric). This is important since Heckman's estimator relies on normality whereas the semiparametric estimator does not requires this hypothesis.

Additionally, several specification tests have been performed which support the same conclusions that we achieve in the estimation part. However, the results obtained in the tests are not conclusive, and further research in tests for distributional assumptions in sample selection models is needed.

Appendix

In the sample selection equation model:

$$z_i^* = x_{1i}^T \beta_1 + u_{1i}, \quad i = 1, \cdots, n \tag{30}$$

$$\begin{aligned} z_i &= 1 \quad \text{if} \quad z_i^* > 0, \\ z_i &= 0 \quad \text{if} \quad z_i^* \leq 0, \end{aligned} \tag{31}$$

we distinguish between continuous and discrete variables by rewriting (30) as

$$z_i^* = x_{1di}^T \beta_{1d} + x_{1ci}^T \beta_{1c} + u_{1i}, \tag{32}$$

where $x_1^T = \begin{pmatrix} x_{1d}^T & x_{1c}^T \end{pmatrix}$ and $\beta_1^T = \begin{pmatrix} \beta_{1d}^T & \beta_{1c}^T \end{pmatrix}$. x_{1d} denotes a vector of discrete random variables and x_{1c} denotes a vector of continuous random variables.

If all variables in the index function are continuous, i.e. $\beta_{1d} = 0$, then root-n consistent semiparametric estimation of β_{1c} can be implemented by the so called Average Derivative Estimation method proposed in Härdle and Stoker (1989). This estimation method relies on the following ideas. Taking into account that

$$r(x_{1c}) = E\left(z|x_{1c}\right) = G\left(x_{1c}^T\beta_{1c}\right), \tag{33}$$

where by the single index restriction the function $G(\cdot)$ does not depend on x_{1c} and is not necessarily a distribution function, then, if $r(\cdot)$ is a.e. first differentiable in x_{1c} and this variable is continuously distributed with first differentiable density $f(x_{1c})$, the local effects of changing x_{1c} on z are given as the vector of derivatives $\nabla r(x_{1c}) = \partial r(x_{1c})/\partial x_{1c}$. The *average derivative* is the expectation of these effects over the population:

$$\delta = E\left(\nabla r(x_{1c})\right), \tag{34}$$

where the expectation is taken with respect to x_{1c}. If we substitute (33) into (34) and we make some straightforward computations, it is easy to show that

$$\delta = E\left[\partial G\left(x_{1c}^T\beta_{1c}\right)/\partial\left(x_{1c}^T\beta_{1c}\right)\right]\beta_{1c} = \theta\beta_{1c}. \tag{35}$$

Therefore, δ is proportional to β_{1c} and we can equivalently replace β_{1c} by δ (provided that $\theta \neq 0$) in (33) obtaining

$$r(x_{1c}) = G\left(x_{1c}^T\delta\right), \tag{36}$$

and $G(\cdot)$ is defined in such a way that $E\left[\partial G\left(x_{1c}^T\beta_{1c}\right)/\partial\left(x_{1c}^T\beta_{1c}\right)\right] = \mathbf{1}$.

Taking the sample counterpart of (34), Härdle and Stoker (1989) propose the simplest Average Derivative Estimator as

$$\hat{\delta} = \frac{1}{n}\sum_{i=1}^{n}\nabla\hat{r}_h(x_{1ci})\hat{I}(x_{1ci}), \tag{37}$$

where $\hat{I}(x_{1ci}) = 1\left[\hat{f}_h(x_{1ci}) \geq b\right]$ is an indicator that drops observations with small estimated density b, $\hat{f}_h(x_{1ci})$ is a standard Parzen-Rosenblatt nonparametric density estimator,

$$\hat{f}_h(x_{1c}) = \frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{x_{1c} - x_{1ci}}{h}\right), \tag{38}$$

$\hat{r}_h(x_{1c})$ is a Naradaya-Watson nonparametric regression estimator denoted as

$$\hat{r}_h(x_{1c}) = \frac{\hat{c}\left(x_{1c}\right)}{\hat{f}_h(x_{1c})}, \tag{39}$$

where

$$\hat{c}\left(x_{1c}\right) = \frac{1}{nh^d}\sum_{i=1}^{n}K\left(\frac{x_{1c} - x_{1ci}}{h}\right)z_i.$$

Finally,

$$\nabla\hat{r}(x_{1ci}) = \frac{\nabla\hat{c}\left(x_{1c}\right)}{\hat{f}_h(x_{1c})} - \hat{r}_h(x_{1c})\frac{\nabla\hat{f}_h(x_{1c})}{\hat{f}_h(x_{1c})}. \tag{40}$$

The main problem presented by this estimation technique, and other estimators derived from it (see Powell, Stock and Stoker, 1989), is the requirement that all the variables that appear in the index must have absolutely continuous density functions. This rules out many interesting cases such as qualitative variables. In order to overcome this problem, Horowitz and Härdle (1996) propose

a modified version of the A.D.E. method that accounts for discrete covariates. In order better to explain this estimator, let us now consider only one discrete random variable. Then, (33) is now

$$E\left(z|x_{1c},x_{1d}\right)=G\left(x_{1c}^{T}\beta_{1c}+x_{1d}\beta_{1d}\right). \tag{41}$$

The estimator proposed by Horowitz and Härdle (1996) estimates parameter β_{1d} according to the following steps. First, let us define $S_{1d}=\left\{x_{1c}^{(j)}:\ j=1,\cdots,M\right\}$ as the support of the discrete random variable x_{1d}. To estimate the parameter vector associated with the continuous variables β_{1c} (up to a normalization scale θ), we use standard average derivative methods for each x_{1d} in its support S_{1d}, and then average over all these estimators. The estimator for the parameter β_{1d}, associated with the discrete random variable works by deducing the horizontal distance between $G\left(\eta+x_{1d}^{(j)}\beta_{1d}\right)$ and $G\left(\eta+x_{1d}^{(1)}\beta_{1d}\right)$ for $j=1,\cdots,M$, on a set of η values in which $G\left(\eta+x_{1d}\beta_{1d}\right)$ is assumed to satisfy a weak monotonicity condition. That is, for their estimator to work they assume that there are finite numbers η_0, η_1, c_0 and c_1 such that $\eta_0<\eta_1$, $c_0<c_1$, and for each $x_{1d}\in S_{1d}$:

$$\begin{cases} G\left(\eta+x_{1d}\beta_{1d}\right)<c_0 & \text{if} \quad \eta<\eta_0; \\ \\ G\left(\eta+x_{1d}\beta_{1d}\right)>c_1 & \text{if} \quad \eta>\eta_1. \end{cases}$$

This assumption is crucial since then, in Horowitz and Härdle (1996), Lemma 1, it is shown that

$$J\left[x_{1d}^{(j)}\right]-J\left[x_{1d}^{(1)}\right]=\left(c_1-c_0\right)\left[x_{1d}^{(j)}-x_{1d}^{(1)}\right]\beta_{1d} \text{ for } j=1,\cdots,M, \tag{42}$$

where

$$\begin{aligned} J\left[x_{1d}\right]=\int_{\eta_0}^{\eta_1}\{&c_0 I\left[G\left(\eta+x_{1d}\beta_{1d}\right)<c_0\right]+c_1 I\left[G\left(\eta+x_{1d}\beta_{1d}\right)>c_1\right] \\ &+G\left(\eta+x_{1d}\beta_{1d}\right)I\left[c_0\leq G\left(\eta+x_{1d}\beta_{1d}\right)\leq c_1\right]\}\,d\eta. \end{aligned} \tag{43}$$

Equation (42) constitutes $M-1$ linear equations in the components of β_{1d}. These equations may be solved if a unique solution exists. To do this, define the $M-1$-vector

$$\Delta J=\begin{bmatrix} J\left[x_{1d}^{(2)}\right]-J\left[x_{1d}^{(1)}\right] \\ \vdots \\ J\left[x_{1d}^{(M)}\right]-J\left[x_{1d}^{(1)}\right] \end{bmatrix} \tag{44}$$

and the matrix

$$W=\begin{bmatrix} x_{1d}^{(2)}-x_{1d}^{(1)} \\ \vdots \\ x_{1d}^{(M)}-x_{1d}^{(1)} \end{bmatrix}. \tag{45}$$

Then an estimator for β_{1d} is that which solves the following system of equations:

$$\Delta J=\left(c_1-c_0\right)W\beta_{1d}, \tag{46}$$

with solution

$$\beta_{1d}=\left(c_1-c_0\right)^{-1}\left(W^{T}W\right)^{-1}W^{T}\Delta J. \tag{47}$$

Equation (47) is the basis for the estimation of β_{1d}. All that we have to do now is to replace population quantities by their sample analogs. More specifically, the function $G(\cdot)$ in (43) is unknown. We propose that it be replaced by the Naradaya-Watson regression estimator in (39). Under some conditions (see Horowitz and Härdle, 1996 for details), both consistency and asymptotic normality of the estimator can be shown.

Acknowledgements

This research was financially supported by the Dirección General de Investigación del Ministerio de Ciencia e Innovación under research grant SEJ2005-08269 and the Department of Education of the Basque Government through grant IT-334-07 (UPV/EHU Econometrics Research Group). We would like to thank the two anonymous referees for their extremely helpful comments and suggestions.

References

Ahn, H., Powell, J. (1993) Semiparametric estimations of censored selection models with a nonparametric selection mechanism. Journal of Econometrics 58: 3–29.

Andrews D.W., Schafgans, M. (1998) Semiparametric estimation of the intercept of a sample selection model. Review of Economic Studies 65: 497–515.

Amemiya, T. (1985) Advanced Econometrics. Cambridge MA: Harvard University Press.

Chen, S., Lee, L.F. (1998) Efficient semiparametric scoring of sample selection models. Econometric Theory 14: 423–462.

Coelho, D., Veiga, H., Veszteg, R. (2005) Parametric and semiparametric estimation of sample selection models: an empirical apllication to the female labour force in Portugal. Working paper 636.05. Unibersitat Autonoma de Barcelona.

Das, M., Newey, W.K., Vella, F. (2000) Nonparametric estimation of sample selection models. Unpublished manuscript.

Fernández, A., Rodríguez-Póo, J. (1997) Estimation and Specification Testing in Female Labor Participation Models: Parametric and Semiparametric Methods. Econometric Reviews 16: 229–248.

Fernández, A., Rodríguez-Póo, J., Villanua, I. (2002) Finite sample behaviour of two step estimators in sample selection models. Computational Statistics 76: 1–16.

Gerfin, M. (1996) Parametric and Semi-parametric Estimation of the Binary Response Model of Labor Market Participation. Journal of Applied Econometrics 11: 321–339.

Goldberger, A. (1983) Abnormal Selection Bias. In S. Karlin, T. Amemiya and L. Goodman (eds.). Studies in Econometrics, Time Series and Multivariate Statistics. Academic Press. New York.

Gourieroux, C., Monfort, A. (1995) Statistics and econometric models. Vol. I, Cambridge University Press.

Gronau, R. (1973) The effect of children on the housewifes value of time. Journal of Political Economy 81: 168–99.

Gronau, R. (1974) Wage comparisons –a selectivity bias. Journal of Political Economy 82: 1119–1143.

Hansen, L.P. (1982) Large sample properties of generalized method of moments estimators. Econometrica 50: 1029–1054.

Härdle, W., Stoker, T.M. (1989) Investigating Smooth Multiple Regression by the Method of Average Derivatives. J.A.S.A. 84: 986–995.

Heckman, J. (1974) Shadow prices, market wages and labor supply. Econometrica 42: 679–694.

Heckman, J. (1979) Sample selection bias as a specification error. Econometrica 47: 153–161.

Horowitz, J., Härdle, W. (1994) Testing a parametric model against a semiparametric alternative. Econometric Theory 10: 821–848.

Horowitz, J., Härdle, W. (1996) Direct semiparametric estimation of single index models with discrete covariates. J.A.S.A 91: 1632–1641.

Hurd, M. (1979) Estimation in truncated samples where there is heteroskedasticity. Journal of Econometrics 11: 247–258.

I.N.E. (1990) Encuesta de Población Activa y Encuesta Anexa sobre Ganancias y Subempleo.

Klein, R.L., Spady, R.H. (1993) An Efficient Semiparametric Estimator for the Binary Response Model. Econometrica 61: 387–421.

Lewbel, A. (2007) Endogenous selection or treatment model estimation. Journal of Econometrics 141: 777–806.

Manski, C.F. (1975) Maximum Score Estimator of the Stochastic Utility model of Choice. Journal of Econometrics 3: 205–228.

Manski, C.F. (1988) Identification of Binary Response Models. J.A.S.A. 8: 729–738.

Manski, C.F. (1993) The Selection Problem in Econometrics and Statistics. In G. S. Maddala, C.R. Rao and H.D. Vinod (eds.). Handbook of Statistics 11. Elsevier Science Publishers.

Martins, M.F.O. (2001) Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal. Journal of Applied Econometrics 16: 23–39.

Melenberg, B., van Soest, A. (1993) Semi-parametric estimation of the sample selection model. CentER Discussion Paper 9334, Tilburg University.

Nawata, K., Nagase, N. (1996) Estimation of sample selection bias models. Econometric Reviews 15: 387–400.

Newey, W.K. (1987) Specification Test for Distributional Assumptions in the Tobit Model. Journal of Econometrics 34: 125–145.

Newey, W.K. (1991) Two-step series estimation of sample selection models. Working paper, MIT.

Newey, W.K. (1999) Consistency of two step sample selection estimators despite misspecification of distributions. Economics Letters 63: 129–132.

Olsen, R.J. (1981) A least squares correction for selectivity bias. Econometrica 48: 1815–1820.

Powell, J.L. (1987) Semiparametric estimation of bivariate latent variable models. Working Paper 8704. University of Wisconsin-Madison.

Powell, J.L., Stock, J.H., Stoker, T.M. (1989) Semiparametric estimation of index coefficients. Econometrica 57: 1403–1430.

Proença, I., Ritter, Ch. (1994) Semiparametric Testing of the Link Function in Models for Binary Outcomes. Discussion Paper 17. Humboldt Uinversität zu Berlin. S.F.B.

Vella, F. (1998) Estimating models with sample selection bias: a survey. Journal of Human Resources, 33(1): 127–169.

White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica 48: 817–838.