Hacienda Pública Española/Review of Public Economics

Manuscript 1996

A multi-step method for estimating returns to tenure in the public sector in Spain.

Patricia Moreno-Mencía

Juan Rodriguez-Poo

David Cantarero-Prieto

A multi-step process approach for estimating public sector wages.

The Spanish experience.

Patricia Moreno-Mencia[1]

ORCID ID: 0000-0002-8210-4932

Group of Health Economics and Health Services Management.

Department of Economy, University of Cantabria- IDIVAL

E39005 Santander, Spain

Juan M. Rodríguez-Póo

ORCID ID: 0000-0001-8751-3025

Department of Economy, University of Cantabria

E39005 Santander, Spain

David Cantarero-Prieto

ORCID ID: 0000-0001-8082-0639

Group of Health Economics and Health Services Management.

Department of Economy, University of Cantabria- IDIVAL

E39005 Santander, Spain

[1] Corresponding author

Abstract

We propose a semiparametric model including problems of sample selection and endogeneity. Concretely, we focus on studying the effect of tenure in public sector wages. Here we face two problems: Public wages are only observed for public employees and the tenure is endogenous. The chosen method allows the introduction of a nonparametric control function and relaxing usual strong assumptions. This approach is applied to a wage model for the Spanish Public Sector, using the wage structure survey of 2016.

Keywords: Control function estimators; sample selection; semilinear model; Spanish wage structure survey.

JEL Classification: C31, C41; C50; J30; H00

1 Introduction

In this article we are concerned with the extension of standard models about human capital theory by allowing simultaneously for sample selection and endogeneity. The human capital theory (Mincer (1958), Becker (1964)) established that education, training and tenure at job produce an increment in the productivity of individuals improving their skills, knowledge and consequently derive in higher earnings. Hence, under human capital theory, experience, education and tenure are the most important factors explaining the economic situation of workers. In this context, the so-called Mincerian wage model has been used in several studies to analyse the differences in wages by sectors or between males and females (Quinn (1979) and Shapiro and Stelcner (1989) among many others). The objective of researchers is to find a relationship between the variable of interest and other factors for a given population. Only if the samples used in the research are obtained randomly the results are able to be extrapolated. In the Mincerian literature about wages, Gronau (1974) was the pioneer in introducing the problem of sample selection bias. In this way, the problem of censored variables have been studied for several decades in Applied Econometrics and Labor Economics. We face sample selection problem because wages in Public Sector are only observed for people working in this sector. Moreover, the introduction of tenure at job as regressor in the structural model creates problems in its specification and estimation. Endogeneity is the main problem in most studies concerned with studying the returns to tenure (see Woodcock (2015)). Thus, in this study is necessary to deal with two different problems and it is necessary to specify a three equation labor supply model (wage equation, participation process and endogeneity process). In a general framework of sample selection models, the most popular approach is the so-called sample selection model (see Heckman (1974)) in a parametric setting. Therefore, he proposed a

two step method in which the binary selection is estimated firstly by using a probit model and then obtaining the so-called Inverse of Mills Ratio. In a second step, this term is inserted in the structural equation as an additional regressor. The main advantage of this kind of parametric estimators is that they are easily and quickly computed. Other advantages are that they are quite easy to interpret and generally more efficient than the semiparametric ones. On the other side, they are inconsistent if the assumed joint error distribution is not correct or if the functional form assumptions are not adequate. Taking this approach as starting point other variations have been proposed to cope with this kind of problems (see Vella (1993) for a survey).

In reference to the econometric methods, nonparametric estimation in sample selection models or with endogenous regressors have been used to provide more flexibility by reducing the assumptions and avoiding misspecification by using smoothness and regular conditions on the densities and functions. The disadvantages of these techniques are, as we have mentioned above, that nonparametric estimation are usually more imprecise and its interpretation is also more difficult. The so-called control function approach is one of the preferable methods to solve these problems (Wooldridge (2015)). Using it as a possible solution, the final model becomes a partially linear regression model. Therefore, partially linear additive models are becoming so popular in semiparametric regression analysis, as is the case of Fernandez et al. (2001) who showed some semiparametric extensions of the Tobit models. For sample selection models, under nonparametric specification of the selection equation and unknown form of the distribution of the errors, some interesting proposals can be found in Ahn and Powell (1993), who proposed a root -N consistent estimator of the parameters of interest in the labor supply function. Furthermore, if we are not willing to impose any parametric functional restriction in the structural labor supply function in Das et al. (2003) is proposed a nonparametric estimator.

In this paper, we are concerned with getting of a root -N consistent estimator for the structural parameters of a sample selection model with an endogenous regressor. To this end, the control function used is treated as a nonparametric component, which is left unrestricted and it is also built with nonparametric first step estimates. In order to do so, we carry out a first-stage estimation with kernel regression methods. Those first step estimates can be built as conditional expectations (such as the so-called propensity score or first-stage residuals or sometimes a mixture of both). Then, in a second step we extend the "pairwise difference" approach of Honore and Powell (1994) to estimate the effect of the regressors of interest by using as correction an unknown function of the first step estimates.

The main contribution of this paper is the econometric modelling, which is adequate to manage simultaneously with endogeneity and selection in a flexible framework. Moreover, we are not willing to impose usual rather strong assumptions and then, the selection equation and the endogenous mechanism are assumed to be unknown non parametric functionals. Differently to Das et al. (2003), our function is not fully nonparametric, we have considered a semiparametric alternative. With this objective, we obtain a new estimator through a two step regression procedure; Firstly, correction terms for endogeneity and sample selection are obtained and secondly, using pairwise techniques, a root-N consistent estimator is proposed. To our knowledge, there is not other study focused on analysing the effect of tenure in Spanish Public Sector wages as the one proposed here, accounting for endogeneity and selection in a flexible framework. We have extended the proposal of Honore and Powell (1994) with a control function approach for correcting more than one problem simultaneously. Our results confirm that sample selection bias is present in this model. We have also detected the presence of endogeneity associated with the variable "Tenure" and then, it is necessary controlling for both problems simultaneously in order to estimate consistently.

The remainder of this paper is organized as follows. In next section we present the econometric model associated with our empirical exercise and we discuss the main problems we have to deal with, and the possible solutions. Then, we describe the estimation procedure and the estimator is proposed. In section 3 the data is described and the method is applied to the study of variations in public sector wages. After that, the effect that tenure has on individuals it is discussed, and also the importance of selection bias and endogeneity in our results. Finally, conclusions and policy implications are presented.

2 Model and estimation procedure

2.1 Econometric model

Let us to start with the consideration of a censored model which has an endogenous regressor. To this end, we are going to specify the relationship among public wages, Public Sector participation and other covariates. In this context, we define the so-called public sector wage equation as:

$$y_i = \begin{cases} y_i^* & for \quad p_i = 1, \\ 0 & otherwise. \end{cases} \tag{1}$$

$$p_i = 1(p_i^* > 0), \tag{2}$$

where

$$y_i^* = l(x_{1i}, s_i) - \eta_{1i}, \tag{3}$$

$$p_i^* = m(x_{1i}, x_{3i}) - \eta_{3i} \tag{4}$$

We assume that the decision to work in Public Sector is a prior precess determined endogenously with the interest outcome. Moreover, the participation decision is going to be positive if public sector wages are expected to be larger than in the private one. So that, the participation equation is defined in the equation above. The variable y_i is the outcome variable of interest, which is the logarithm of the wage for the $i-th$ individual in Spanish Public Sector. The functions $l(.)$ and $m(.)$ are unknown functions which need to be estimated. The starred variables,(y_i^*, p_i^*), stand for unobservable variables and s_i is the endogenous regressor, defined as:

$$s_i = f(x_{1i}, x_{2i}) - \eta_{2i} \tag{5}$$

for $i = 1, \cdots, N$; where (x_{1i}, x_{2i}, x_{3i}) is a $(k_1 + k_2 + k_3)$-vector of observed explanatory variables and the vector $(\eta_{1i}, \eta_{2i}, \eta_{3i})$ is an unobserved set of idiosyncratic error. $1(A)$ denotes the indicator function, i.e. $1(A) = 1$ if A is true, and 0 otherwise. Finally $f(.)$ is a non parametric function which is left unrestricted. This specification is often used in many theoretical problems being really popular in labor economics since the seminal work of Gronau (1974). Let us starting considering that wages equation is linear. $y_i^* = x_{1i}^{\top}\beta_1 + s_i\lambda_1 + \eta_{1i}$. Then, taking conditional expectations we have that:

$$E(y_i|x_{1i}, s_i, p_i^* > 0) = x_{1i}^{\top}\beta_1 + s_i\lambda_1 + E(\eta_{1i}|x_{1i}, s_i, p_i^* > 0), \;\; i = 1, \cdots, N. \tag{6}$$

Under some assumptions, i.e.

1. The quantity $E(\eta_{1i}|x_{1i}, s_i, p_i^* > 0)$ is constrained to be linear and depending on a finite vector of parameters.

2. The vector of error terms $(\eta_{1i}, \eta_{2i}, \eta_{3i})$ is multivariate normally distributed with zero mean and homoskedastic variance-covariance matrix.

Adding some convenient restrictions in the variance-covariance matrix, the quantity $E(\eta_{1i}|x_{1i}, s_i, p_i^* > 0)$ is known up to certain constants and therefore it can be shown that the ordinary least squares estimators of β_1 and λ_1 are consistent and asymptotically normal (see Kim (2006) for details). If we are not willing to assume none of the previous assumptions the quantity $E(\eta_{1i}|x_{1i}, s_i, p_i^* > 0)$ remains unknown and the least squares estimation of the parameters of interest is inconsistent.

Moreover, like misspecification of the parametric form of the index functions results in inconsistency it is useful to relax strong assumptions. There exist at least two well known reasons. Probit maximum likelihood estimators of the first stage, under non-normal disturbances, are inconsistent and second, the Mill's ratio is misspecified and therefore second stage estimators are asymptotically biased. For identification purposes we assume that $Pr(x_{1i}^{\top}\beta_1 + s_i\lambda_1 + \mu_1(\eta_{2i}, p_{1i}) = 0|p_i = 1) = 1$, that is there exits a constant, c such that: $Pr(x_{1i}^{\top}\beta_1 + s_i\lambda_1 = c|p_i = 1) = 1$. This is an identifying assumption needed to identify the parameters of interest, β_1 and λ_1. Moreover, there must exist an exact functional relationship between the linear component of $E(y_i|x_{1i}, s_i, p_i^* > 0) = x_{1i}^{\top}\beta_1 + s_i\lambda_1$ and the unknown function ($E(\eta_{1i}|x_{1i}, s_i, p_i^* > 0)$ (see also Vytlacil (2002) and Depalo and Pereda (2018)). To sum up, the main idea to identify and to estimate the model is to include a function of the estimated correction terms in the main equation. Then, the parameters β_1 and λ_1 of the structural model may be estimated through the corrected regression equation. So that, we need to compute $E(\eta_{1i}|x_{1i}, s_i, p_i^* > 0)$. As is well known, the necessary identification constraint in this framework is that at least one of the explanatory variables included in the participation equation be excluded from the wage equation. This is because the reduced equation could be linear and the second stage could then involve collinear regressors. In the case which first step estimates are obtained by a non-linear model, the correction term will not

be perfectly correlated with X, even in the absence of exclusion restrictions. In our case, the first step is non parametric and the control function is a non-linear function of them, thus, the second stage equation (wage function) is identified because of this non-linearity.

We assume that $E[\eta_{1i}|s_i] = E[\eta_{1i}|x_{1i}, x_{2i}, \eta_{2i}] = E[\eta_{1i}|\eta_{2i}] = f(\eta_{2i})$. Thus, we can estimate a kernel regression for $s_i = f(x_{1i}, x_{2i}) + \eta_{2i}$ and obtain the residuals, $\hat{\eta}_{2i}$. After this, following Das et al. (2003), if p_j is a binary variable, the selection correction can be expressed as a propensity score. Being j an observation different from i, let $p_j = E[p_j|x_{1j}, x_{3j}, x_j] = Pr(p_j = 1|x_{1j}, x_{3j}, x_j)$ denote the propensity score, p_1.

$$
\begin{aligned}
&E(\eta_1|x_1, s, p = 1) \\
&= E(\eta_1|x_1, s, \eta_3 > -m(x_1, x_3)) \\
&= E(\eta_1|\eta_2, u < p_1) \qquad (7)
\end{aligned}
$$

$$= \frac{\int_{-\infty}^{p_1} (\int \eta_1 f(\eta_1, u|\eta_2) d\eta_1) du}{\int_{-\infty}^{p_1} (f(\eta_1, u|\eta_2) d\eta_1) du} = \mu_1(p_1, \eta_2).$$

Where:

$$\hat{\eta}_{2i} = f(\eta_{2i}|x_{1i}, x_{2i})$$

$$\hat{p}_{1i} = E(p_i|x_{1i}, x_{3i}) = m(x_{1i}, x_{3i})$$

We denote the propensity score as p_1, which is a conditional expectation that can be estimated with nonparametric methods. In a general form, the propensity can be estimated non-parametrically with a multivariate kernel. The following Nadaraya-Watson nonparametric estimator for the link function $m(.)$ is used. Nadaraya (1965) and Watson (1964) therefore proposed that those link

function be estimated by replacing $m(.)$ by $\widehat{m}(.)$ where the density estimator is then the kernel estimator. So that, we define the estimated probability as:

$$\widehat{m}(x_{1i}, x_{3i}) = \hat{p}_{1i} = [\sum_{j=1}^{n} K\left(\frac{z_i - z_j}{h_1}\right) p_j][\sum_{j=1}^{n} K\left(\frac{z_i - z_j}{h_1}\right)]^{-1} \quad (8)$$

where we have denoted $z_i = (x_{1i}, w_{3i})$, $K(.)$ is a Kernel function which tends to zero as the magnitude of its argument increases and $h > 0$ is the smoothing parameter, which converges to 0 as the sample size increases to infinity. The propensity is sufficient for identification of structural relationships of selection mechanism (see Ahn and Powell (1993)). Furthermore, under standard conditions on the bandwidth and kernel function (integrates to one, has mean zero and that $h_1 \to 0$ as $n \to \infty$ and $nh_1 \to \infty$), this is a consistent estimator of $E(y|x_1, s, p = 1))$.

Then, we can write;

$$E[y_i|x_{1i}, s_i, p_i^* > 0] = x_{1i}^{\top}\beta_1 + s_i\lambda_1 + \mu_1(\eta_{2i}, p_{1i}) \quad (9)$$

And the final model may be written as;

$$y_i = x_{1i}^{\top}\beta_1 + s_i\lambda_1 + \mu_1(\eta_{2i}, p_{1i}) + \epsilon_i \quad (10)$$

Let we call $Z = (x_2, x_3)$ and we are going to establish some restrictions such as: *Full Independence*, that is, (η_1, η_2, η_3) and Z are independent. *Common support*: For all $x_1 \in X$, the support of (η_2, p_1) conditional on X and selection equals the support of (η_2, p_1) conditional on selection, and finally, *Strict monotonicity in* η_2: If $\Pi(z, \eta_2) > \Pi(z, \eta_2')$ for some (z, η_2, η_2') then $(z', \eta_2) > (z', \eta_2')$ for all z'. In this model, if these assumptions are satisfied, then x_1 and η_1 are independent conditional on p_1 and η_2 given selection. Following Das et al. (2003) we assume that: $E(\eta_{1i}|x_{1i}, s_i, p_i = 1) = \mu_1(m(x_{1i}, x_{3i}), f(x_{1i}, x_{2i})) = \mu_1(p_{1i}, \eta_{2i})$. This assumption means that the conditional expectation of the random error given the selection process, depends only on

the propensity score and the residuals. Moreover, x_1 is included in the endogenous equation together with x_3 and also in the participation mechanism in addition to x_2, thus at least one variable is included there and omitted from the outcome equation (see assumption $A.4$ in the Appendix). Here, we assume exchangeability, that is we make pairs (i, j) that fulfill:

$$\mu_1(\eta_{2i}, p_{1i}) = \mu_1(\eta_{2j}, p_{1j}), \text{ for } i \neq j.$$

where μ_1 is the control function, which depends on p_{1i} and η_{2i} and j is an observation different from i. This correction is an alternative to popular inverse Mills ratio usage but in this case with multi-components specification and without assuming any error distribution. Moreover, we allow $\mu_1(.)$ to have an unknown functional form like proposed Das et al. (2003). Then, taking differences of distinct observations with similar propensity score and first-step residuals, the selection and endogeneity biases vanishes while the structural model remains identifiable up to the constant term which disappears with the differencing process. The idea behind this approach is to assign weights to each pair of observations with declining weights to those having larger values of $|p_{1i} - p_{1j}|$, $|\eta_{2i} - \eta_{2j}|$. As p_1 and η_2 are non observable quantities, we substitute its values for a consistent estimation. Thus, using a pairwise transformation, for $i \neq j$

$$y_i - y_j = (x_{1i} - x_{1j})^\top \beta_1 + (s_i - s_j)\lambda_1 + (\epsilon_i - \epsilon_j).$$

where

$$\epsilon_i = y_i - E(y_i | x_i, s_i, p_i^* > 0), \quad \epsilon_j = y_j - E\big(y_j | x_j, s_j, p_j^* > 0\big),$$

If we denote $\beta = (\beta_1, \lambda_1)$ and $w_1 = (x_1, s)$. Then, the parameters β in the structural wage equation can be estimated as follows,:

$$\hat{\beta} = \hat{S}_{ww}^{-1} \hat{S}_{wy} \tag{11}$$

where ;

$$\hat{S}_{ww} = \binom{N}{2}^{-1} \sum_i \sum_{i<j} \widehat{\omega}_{ij} p_i p_j (w_{1i} - w_{1j})(w_{1i} - w_{1j})^{\top} \quad (12)$$

$$\hat{S}_{wy} = \binom{N}{2}^{-1} \sum_i \sum_{i<j} \widehat{\omega}_{ij} p_i p_j (w_{1i} - w_{1j})(y_i - y_j) \quad (13)$$

Then, define the weights for $i, j = 1, \ldots, n$ as:

$$\widehat{\omega}_{ij} = K\left(\frac{\hat{p}_{1i} - \hat{p}_{1j}}{h}, \frac{\hat{\eta}_{2i} - \hat{\eta}_{2j}}{h}\right) p_i p_j \quad (14)$$

In order to compute these weights we propose a symmetric, twice differentiable Kernel of fourth order and bounded support. For practical purpose the choice of the bandwidth parameter is based on the generalized cross-validation criterion.

Theorem 1: Under assumptions (A.1) to (A.9) detailed in the Appendix, as N tends to infinity,

$$\hat{\beta}_1 = \beta_1 + \Sigma_1^{-1} \frac{2}{\sqrt{N}} \sum_i [p_i \rho_{1i} \gamma_i (w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))] \epsilon_i + o_p(1)$$

where ϵ_i is the error term and ρ_i,γ_i and $\lambda_1(\cdot)$ are defined in (A.6).

Under previous assumptions and following the arguments in Ahn and Powell (1993), it is possible to show that $\hat{S}_{ww} = S_{ww} + o_p(1)$, and $\sqrt{n}\hat{S}_{w\epsilon} = \sqrt{n}S_{w\epsilon} + \sqrt{n}(\hat{S}_{w\epsilon} - S_{w\epsilon})$ and under regularity conditions $\hat{\beta}$ is root-n consistent and asymptotically normal (see Appendix).

2.2 Steps to obtain the estimates for the structural parameters

1. First step: We estimate the equation defining the endogenous regressor process (Eq. 5), which is a continuous variable, and obtain the residuals, $\hat{\eta}_{2i}$.

2. Second Step: We estimate the public sector participation (Eq.4), non-parametrically, and obtain the predicted values, $\hat{p}_1$.

3. Third step: With predicted values from previous steps we can built the required control function for correcting endogeneity and selectivity. An unrestricted function is used as an additional regressor in the final model (Eq.10). Thus, the structural equation can be estimated consistently.

2.3 A Montecarlo Simulation

In the first table of the Appendix section, we report the simulations results through the estimated bias, standard deviation (Std), and RMSE for the estimators of some usual specifications and the proposed in this study. We report the Monte Carlo simulation results to get an idea about the small sample properties of the estimator. As usually, the design is chosen to illustrate the method so is not related with any particular data set. The model is:

$$y_i^* = \beta_1 x_i + \lambda_1 s_i + \eta_{1i} \tag{15}$$

$$p_i^* = m(x_{1i}, x_{3i}) - \eta_{3i} \tag{16}$$

$$s_i = f(x_{1i}, x_{2i}) - \eta_{2i} \tag{17}$$

Where due to endogeneity:

$$\eta_{1i} = \theta \eta_{3i} + u_i$$

Here, $p_i = 1(p_i^* > 0)$ and $y_i = y_i^*$ if$p_i = 1$. We have assumed that the true values of β_1 and λ_1 are $(0.5,0.5)$ (see [24] for detailed simulations process). $u_i \sim N(0,1)$ and $\theta = \rho \frac{\sigma_1}{\sigma_2}$. We can set $\rho = 0.5$ and $\eta_{1i} \sim N(0, \sigma_1^2)$. For estimating $\hat{p}_i$ and $\hat{\eta}_{2i}$ we have simulated fully

nonparametric functions. We have used in these simulations the biweight kernel and $\hat{h} = \hat{\sigma}N^{-1/5}$. The results from 1000 replications with sample sized $50{,}100 \quad and \quad 500$ are given in Table $B.1$ in Appendix B. It can be seen that the values of the standard deviation, bias and RMSE decrease when the sample size increases, for all the compared methods. The Heckman 2 steps model for correcting only selection bias has generally the highest biases and RMSE for both parameters compared to other methods, although it is not so far from OLS results. We can conclude that our estimator shows a good performance, as N increases the RMSE and the bias are lower, as is expected according to the asymptotic properties discussed above.

3 Data and Empirical Application for estimating Spanish Public Sector employees's wages

In this section we present an application to illustrate the usefulness of the proposed approach. It is something known that salary is one of the most important components of labor market decisions, Mortensen (1986). Economic theory pointed out that the most general model for the determination of the wages structure will depend on workers personal characteristics (demographic or productive factors such as age, tenure and education) and a combination of political and economic factors. Most of studies conclude that there exists a positive wage premium for public sector employees, as Giordano et al. (2015). We are interested in studying the effect that the tenure at job has on public sector wages. The fundamental idea is to consider the job tenure as the main regressor, it can be seen as the public sector specific human capital accumulation. Public sector wages can be seen as an instrument of economic policy, and it represents the 16% of the total workforce. Public employees in Spain may be subject to administrative regulation or labor legislation. In this sense, public employees conditions are different to the private ones and generally

are associated with more security at job. The access to public employment is often conditioned to pass open exams. Furthermore, standard methods for estimating wages equation are not correct if the sector selection is not random. It is considered that there is a selection process for entering Public Sector, being the first decision to participate in a Public selection process or not. Then, it is clear that if some unobservable characteristics affecting wages are correlated with non-observable factors determining the Sector choice, the Ordinary Least Squares estimation in the separate regressions will not be consistent. Firstly, the tenure and the participation equations are estimated non-parametrically in order to obtain the correction terms that will be included in the unrestricted control function. It has been considered the possibility that the tenure had a non linear effect on wages (quadratic or cubic) but the effect seems to be mostly linear.

Figure 1: **Logarithm of wages by sector.**

Source: Own elaboration with the Annual Wage Structure Survey, 2016.

As can be seen in figure 1, the difference in the annual wage average, between both sectors, is over 5500 euros more in public sector. The (log) wage in Public sector have higher mean and lower standard deviation than in the private one. Moreover, the tails of the distribution are larger in the Private Sector. That is, there are more workers in the extreme of the wage distributions (much lower salaries and higher ones), on average, in the Private Sector.

3.1 The Survey

Our analysis sample comes from the Annual Wage Structure Survey. Concretely, we use the microdata provided by the Spanish Statistical Office. The main purpose of this Survey is to identify the average annual gross income per worker classified by working journey and other

sociodemographic and related to the occupation variables. The Survey also provides information on the average earnings and the distribution of wages. The Annual Wage Structure Survey is performed annually by the Spanish Statistical Office and it is the result of the combination of different statistical and administrative sources. This survey has a sample for Spain which is composed by all regular employees included in Social Security. The information may be also disaggregated by Autonomous communities. The data set available consist in $209{,}436$ workers from which $32{,}892$ were Public Sector employees. As we can observe in Table 1, the annual wage is on average of $29{,}458$ Euros in the Public Sector (it was $23{,}911$ Euros in the Private Sector).

Table 1: **Statistics of the variables: mean, and standard deviation in brackets.**

Source: Own Elaboration from Annual Wage Structure Survey, 2016.

The variables included in this data set are defined in Table 1 including some basic statistics. These data show that, on average, in Public Sector employees are older than in the Private one, the 31.5% of public sector workers are between 50 and 59 years old (compared to 18.7% of the private one). In Public Sector there is also more education level, tenure and women incidence. Annual earning are on average higher in Public Sector than in the Private one. The control variables, x_1, are $Age\ from\ 50\ to\ 59$, Sup and $Male$ and the endogenous regressor s, is $Tenure$. Although we think it would be better to use the variable age as continuous, it was not available. Then, and taking into account the problem of the curse of dimensionality we can only choose one of the age interval. This interval of age is the most representative in this context to make the comparisons. Older workers usually have achieved more wage complements, they have

usually completed their studies and, finally they have collected more years of tenure, making these variables to control better for wage variations. The variable $comsal$ is the so-called x_3 in our specification. From a theoretical point of view, it is reasonable to think that monetary complements are associated with tenure, and with high probability these complements are going to increase with tenure. However, once the control variables and first step residuals are considered, it is probable that this variable is uncorrelated with other omitted factors affecting wages. On the other hand, to have elementary studies is the variable chosen to be the so-called variable, x_2. It is reasonable to think that having elementary studies influence the decision to participate in one sector or another, basically because the selection process is usually related with exams or merits. However, once the decision to participate in the public sector has been made, it no longer affect wages considering that to have high level education acts as control. On the other hand, if we may think that $Cov(z,u)$ is not 0, is < 0, for example, thinking that the ability of workers who hold primary education is lower than the ability of the other workers. This means that β will be biased downward and therefore that one under-estimates the impact of a change in the explanatory variables. The exogeneity condition is difficult to achieve and it is well known that in practice, valid instruments are often difficult to find if not impossible (see Escanciano et al. (2016)).

4 Results

In this section we present the results obtained for the wages model. In order to implement our estimator, it is necessary to obtain “first-step" estimates of p_1 and η_2. After having obtained these estimates non parametrically, we can proceed with the next step. To give further insights we have estimated the wage equations under some of the most usual methods. In Table 2 we can observe the estimates for the uncorrected Ordinary Least Squares estimation in the first column. In the second column, the results for Heckman two step standard approach are calculated with the

Inverse Mills ratio as additional regressor to correct for the sample selection. Finally, in the third column we have calculated the estimates in a two-steps least squares by including the residual from first step as regressor. After controlling for selection bias, the coefficient of tenure increases compared to the ordinary least squares estimates while its effect is smaller if only correct for the endogeneity. The problem with this usual methods is that when the normality assumption is wrong, the estimates may be even worse than using ordinary least squares. The results obtained reflects something similar to the presented in the previous literature, usually referred to the whole labor market and not only for Public Sector. The majority of studies find large and significant tenure effects while a few papers estimate the effect small or insignifficant (see Abraham and Farber (1987),Altonji and Shakotko (1987), Buchinsky et al. (2010) for the whole labor market). We have included some controls in the analysis such as the education, the gender or the age. We would like to include more covariates but practical applications of nonparametric estimators with more than three covariates suffer a great deal from the well-known curse of dimensionality: convergence slows down as dimension increases so that, we have included few explanatory variables (see Pagan and Ullah (1999) for more detailed explanation).

We can observe in Table 2 that in all specifications, there exists a favorable premium for men in relation to women, people with higher level of studies have a higher wage and tenure is positive related with wages, as its consistent with the beliefs of the economic theory. Furthermore, it has been shown that the inverse of Mills ratio is statistically significant in column 2 and also the residual component in column 3, so that it is suspected that a correction is needed because we have evidence of sample selection and endogeneity. This gives an important relevance to the method proposed in this article, in which a correction term built under flexible assumptions is included, which is quite different from standard proceeds.

Table 2: **Estimation Results for the Logarithm of wages in Public Sector with some standard parametric approaches.**

Source: Own Elaboration from Annual Wage Structure Survey, 2016.

To sum up, Table 2 show two important features: First, the presence of sample selection bias is supported by the statistical significance of the Inverse Mills Ratio, and second, the statistical significance of the term involving the tenure residuals supports our idea about this variable is endogenous in wage equation. Additionally Hausman test suggested that the variable tenure at job was endogenous. Hence, we propose to introduce both corrections terms for selection and endogeneity and we proved some different specifications in order to corroborate the possibility that this correction function could be nonlinear.

As we have described above we can relax parametric assumptions on first step estimates while the sample selection and endogeneity problem will be treated in the second stage. So that, we propose to estimate the first step equations in a more flexible way.

Table 3: **Public Sector wages estimates under some standard specifications for the control functions.**

Source: Own Elaboration from Annual Wage Structure Survey, 2016.

Table 3 supports our idea of maintaining unrestricted the control function, under a linear

specification we have not enough evidence against selection bias because the propensity score is not statistically significant. However, this also may reflect the necessity of including high order terms to capture the selection bias. In the last column, we have introduced the quadratic terms and it seems to provide evidence about selection bias with the significance of the propensity score squared. To avoid misspecification, as we have explained in this article it is preferable to left the control function unrestricted, and using the pairwise difference technique to obtain consistent estimates in wages equation. To estimate the model we use a nonparametric estimation for the reduced equations, $f(.)$ and $m(.)$ which allowed us to obtain $\hat{p}_{1i} = E[p_i|f(.)]$ and $\hat{\eta}_{2i} = s_i - E[s_i|m(.)]$. In addition to our initial bandwidth choice guided by cross validation criterion we have selected other bandwidth parameters to check the robustness, the estimates were very close over the bandwidths. In a second stage, we focus on the estimation of wages for public sector employees which were 32.892 workers.

Table 4: **Estimation Results for the Logarithm of wages in Public Sector under pairwise difference method. Unrestricted Control Function.**

Source: Own Elaboration from Annual Wage Structure Survey, 2016.

The final results are presented in Table 4, it is shown that older people earn more than younger employees in Public Sector. This is something expected due to older workers are also the ones with more experience and acquired skills. In the same direction, tenure is positively related with public wages usually due to increments in payment for periods of time being Public employee. An additional year of tenure increases the expected wage in a $2{,}2\%$, holding the rest of factors

equal. The education level is one of the most important determinants of Public wages, because having superior studies is a necessary condition to access to some official scales and to promote. So that, people with higher education are expected to increase their wage in a 32% more than employees without superior studies. Finally, men are expected to earn in public sector about a 25% more than women, being the rest of factors the same. This shows that despite the progressive incorporation of women to labor market there exists an important gap in payment and on average, women earn less than men, also in Public Sector.

5 Conclusions

The objective of this article is to analyse the returns of tenure at job for public employees in Spain. For this purpose we have estimated an extension of Mincerian wage model for analysing factors associated with Public Sector wages. The wage premium received by public employees in relation to the private ones is positive, as in most OECD countries in the last four decades. Alternatively, during the last decades estimation techniques have been progressively adapting to deal with the implicit problems of economic models. Using the wage structure wage survey for 2016, we have to deal with two sources of bias in the proposed model, the sample selection and the existence of an endogenous regressor.

There are two main differences between this study and previous ones. Here, we have specified a model which allows for sample selection and endogeneity in estimating the returns to seniority within a job, concretely in public sector. Our results clearly demonstrate the importance of this joint estimation of the wage equation along with the public sector participation and the reduced equation of the endogenous regressor. These two decisions have signifficant effects on the observed outcomes of wages. Thus, while the empirical findings are interesting, the key contribution of this article is the modeling approach and implementation. We have reached several conclusions in this article. The first one is that a new approaches to estimate a standard labor supply models which are subject to sample selection and endogeneity are needed to obtain consistent results. In this sense, the most remarkable finding is that we have specified our model as a multi-equation system which involves all processes and using non-parametric estimates for the correction terms. Within this set up we propose an estimator based on pairwise differences of observations with close values of its non-parametric first step estimates. So that, our approach do not require

strong exogeneity for the regressors in the main equation. The advantages of our proposal are the model flexibility and the provision of a root-n consistent estimator for the structural parameters, even if the involved control function is unknown. We consider that is of outstanding interest for several economic models and it is possible to compare our estimates with the ones obtained from fully parametric settings.

The empirical application proposed in this paper shows that our estimator performs well and produces the most plausible estimates, which was our main interest. Despite the several strengths of this study, it is necessary to mention some of the limitations. Due to the cross-sectional nature of the results presented, it was not possible to establish temporal relations. Moreover, lacking family composition controls is a huge limitation in order to study gender wage gaps. Additionally, due to the nonparametric estimation is not possible to include more control variables in order to avoid the curse of dimensionality problem. Despite these limitations, this study provides important information, and the empirical results give us valid conclusions. Moreover, our results support the idea that returns to tenure at job are positive and significant. Additionally, we have shown that men, high-skilled workers and older ones earn more money. Then, this corroborates that disadvantaged labor force groups are expected to earn less. Finally, the evidence suggests that it is necessary to deal with selection and endogeneity of tenure to obtain consistent estimates. To show that, we have compared our approach with others in order to present the main differences.

Bibliographical References:

Abraham, K. and Farber, H. (1987). "Job duration, seniority, and earnings". *American Economic Review 77* (3), 278–297.

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

Altonji, J. and Shakotko, R. (1987). "Do wages rise with job seniority?" *Review of Economic Studies*. 54 (3), 437–459.

Becker, G. (1964). "Human capital". New York: Columbia university press.

Buchinsky, M., Fougere, D., Kramarz, F. and Tchernis, R. (2010). "Interfirm mobility, wages and the returns to seniority and experience in the united states". *Review of Economic Studies 77*, 972–1001.

Das, M., Newey, W. and Vella, F. (2003). "Non parametric estimation of sample selection models". *Review of Economic Studies. 70*(1), 33–58.

Depalo, D. and Pereda, S. (2018). "Consistent estimates of public/private wage gap". *Empirical Economics*, 1–11.

Escanciano, J., Jacho-Chavez, D. and Lewel, A. (2016). "Identification and estimation of semi-parametric two-step models". *Quantitative Economics 7*, 561–589.

Fernández, A., Rodríguez-Poo, J. and Sperlich, S. (2001). "A note on the parametric three step estimator in structural labor supply models". *Economic Letters 74*, 31–41.

Giordano, R., Depalo, D., Coutinho, M., Eugene, B., Papapetrou, E., Perez, J., Reiss, L. and Roter, M. (2015). "The public sector pay gap in a selection of euro area countries in the pre-crisis period". *Hacienda Pública Española - Review of Public Economics*,11–34.

Gronau, R. (1974). "Wage comparisons-a selectivity bias". *Journal of Political Economy 82*(6), 1119–43.

Heckman, J. (1974). "Shadow prices, market wages and labor supply". *Econometrica 42*, 679–693.

Honore, B. and Powell, J. (1994). "Pairwise difference estimators of censored and truncated regression models". *Journal of Econometrics 64*, 241–278.

Kim, K. (2006). "Sample selection models with a common endogenous regressor in simultaneous equations; a simple two-step estimation". *Economic Letters. 91*, 280–286.

Mincer, J. (1958). "Investment in human capital and personal income distribution". *Journal of Political Economy 66*(4), 281–302.

Mortensen, D. (1986). "Models of search in the labor market". *Handbook of Labor Economics Amsterdam: North-Holland.*

Nadaraya, E. (1965). "On nonparametric estimates of density functions and regression curves". *Theory of Applied Probability 10*, 186–190.

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Themes in Modern Econometrics. Cambridge University Press.

Powell, J. (1987). "Semiparametric estimation of bivariate latent variable models". *Working paper. Social System Research Institute, University of Wisconsin-Madison.*

Powell, J. Stock, J. and Stoker, T. (1989). "Semiparametric estimation of weighed average derivatives". *Econometrica. 57*, 1403–1430.

Quinn, J. (1979). "Wage differentials among older workers in the public and private sectors". *Journal of Human Resources. 14*, 41–62.

Shapiro, D. and Stelcner, M. (1989). “Canadian public-private sector earnings differentials, 19701980”. *Industrial Relations 28*, 72–81.

Vella, F. (1993). “A simple estimator for models with censored endogenous regressors”. *International Economic Review. 34,* 441–457.

Verbeek, M. (2004). “*A guide to modern econometrics”* (2nd. Ed.). Chichester: John Wiley and Sons.

Vytlacil, E. (2002). “Independence, monotonicity, and latent index models: An equivalence result”. *Econometrica 70*, 331–341.

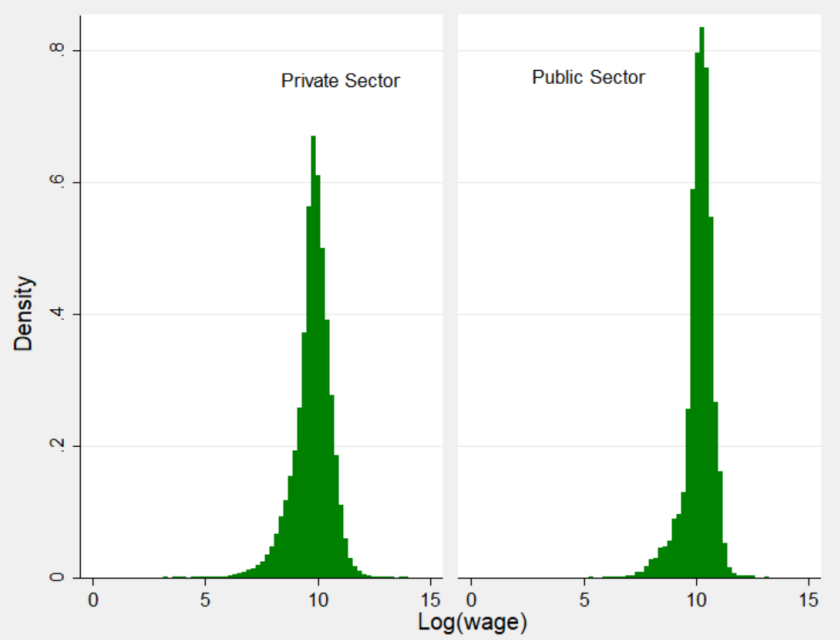
Watson, G. (1964). “Smooth regression analysis”. *Sankhya. 26* (15), 359–372.

Woodcock, S. (2015). “Match effects”. *Research in Economics 69*, 100–121.

Wooldridge, J. (2015). “Control function methods in applied econometrics”. *Journal of Human resources. 50*, 420–445.

Tables and Figures:

Figure 1: **Logarithm of wages by sector.**



Source: Own elaboration with the Annual Wage Structure Survey, 2016.

Table 1: **Statistics of the variables: mean, and standard deviation in brackets.**

	Variables definition	**Public Sector**
Wage	Annual Wage, in Euros	29458.72 (17480.54)
Age1	less than 19	0.000 (0.025
Age2	Age from 20 to 29	0.049 (0.216
Age3	Age from 30 to 39	0.235 (0.424
Age4	Age from 40 to 49	0.320 (0.466
Age5	Age from 50 to 59	0.314 (0.462
Age6	Age more than 60	0.079 (0.270
Primary	1, if having elementary studies	0.049 (0.216
High School	1, if High School	0.467 (0.498

Sup	1, if having Higher Education	0.478 (0.499
Tenure	Time working with current employer	14.25 (10.76
Comsal	Complements to wage, in Euros	1, 000.71 (1,102.26)
Male	1, if the person is male	0.468 (0.499)

Source: Own elaboration with the Annual Wage Structure Survey, 2016.

Table 2: **Estimation Results for the Logarithm of wages in Public Sector with some standard parametric approaches.**

	OLS	**Heckman 2-steps for selection**	**Residual correction**
Intercept	9.53^{***} (0.0006)	9.62^{***} (0.013)	9.48^{***} (0.0006)
Age from 50 to 59	0.035^{***} (0.035)	0.114^{***} (0.114)	0.048^{***} (0.006)
Male	0.156^{***} (0.0006)	0.062^{***} (0.0126)	0.172^{***} (0.005)
Sup	0.382^{***} (0.0006)	0.568^{**} (0.023)	0.351^{***} (0.005)
Tenure	0.023^{***} (0.0003)	0.032^{***} (0.001)	0.021^{***} (0.000)
Mills		-0.782^{***} (0.092)	
Residual correction			0.0002^{***} (0.000)

Source: Own Elaboration from Annual Wage Structure Survey, 2016.

Table 3: **Public Sector wages estimates under some standard specifications for the control functions.**

	OLS	**Linear**	**Quadratic**
Intercept	9.53^{***} (0.006)	9.486^{***} (0.016)	9.432^{***} (0.019)
Age from 50 to 59	0.035^{***} (0.007)	0.047^{***} (0.0128)	0.052^{***} (0.0131)
Male	0.156^{***} (0.006)	0.173^{***} (0.008)	0.184^{***} (0.008)
Sup	0.382^{***} (0.006)	0.349^{**} (0.018)	0.350^{***} (0.017)
Tenure	0.023^{***} (0.003)	0.021^{***} (0.0007)	0.021^{***} (0.0007)

propensity score, $\hat{p}_1$		0.022 (0.177)	0.522 (0.223)
Residual correction, $\hat{\eta}_2$		0.0002^{***} (0.000)	0.0003^{***} (0.0000)
$\hat{p}_1^2$			-1.135^{***} (0.458)
$\hat{\eta}_2^2$			-0.009^{***} (0.0000)

Source: Own Elaboration from Annual Wage Structure Survey, 2016.

Table 4: **Estimation Results for the Logarithm of wages in Public Sector under pairwise difference method. Unrestricted Control Function.**

	Coef.	**Std.Err.**	**t-statistic**
Age from 50 to 59	0.014	0.010	1.35
Male	0.249	0.009	27.31
Sup	0.329	0.009	35.83
Tenure	0.022	0.0004	47.92

Source: Own Elaboration from Annual Wage Structure Survey, 2016.

Appendix

Appendix A: Proofs of the main Results

Some additional assumptions are needed to estimate the parameters of interest. Basically, we assume the following,

(A.1) The vectors $(y_i, s_i, p_i, x_i^\top)^\top$ satisfying (1) to (5) are realizations from measurable independent and identically distributed across i random variables.

(A.2) The random variables have bounded support and finite $2+\gamma$ order moments for some $\gamma > 0$.

(A.3) Define the parameter vector $\beta = (\beta_1, \lambda_1)$ and the parameter space Θ. Then $\beta \in \Theta$, Θ is a compact set, and β is an interior point of Θ.

(A.4) Following Das et al. (2003) we assume that: $E(\eta_{1i}|x_{1i}, s_i, p_i = 1) = \mu_1(m(x_{1i}, x_{3i}), f(x_{1i}, x_{2i})) = \mu_1(p_{1i}, \eta_{2i})$. This assumption means that the conditional expectation of the random error given the selection process, depends only on the propensity score and the residuals.

(A.5) For any random variables $l(x_1, s)$ and $\mu_1(p_1, \eta_2)$, $Pr(l(x_1, s) + \mu_1(p_1, \eta_2) = 0|p = 1)$. This implies that $l(x_1, s)$ is constant. (see Das et al. (2003) for more detail).

(A.6) Let us define, Σ_{ww}, where $w_{1i} = (x_{1i}, s_i)$,

$$\Sigma_{ww} = E[\rho_{1i}^2 \gamma_i (w_{1i} - \lambda_1(p_{1i}, \eta_{2i})) \times (w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))^{\top}] \quad (18)$$

Σ_{ww} is nonsingular.

where

$$\rho_{1i} = E[p_i|p_{1i}, \eta_{2i}]$$
$$\gamma_i = \mu_1(p_{1i}, \eta_{2i})$$

and

$$\lambda_1(p_{1i}, \eta_{2i}) = \frac{E[p_i w_{1i}|p_{1i}, \eta_{2i}]}{\rho_{1i}\gamma_i} \quad (19)$$

(A.7) Assumptions on the kernel function:

$K(u)$ is twice differentiable $K''(u) < K_0$, for some K_0; $K(u) = K(-u)$; $K(u) = 0$ if $|u| > l_0$ for some $l_0 > 0$ $\int K(u)du = 1$, $\int u^l K(u)du = 0$ for $l = 1,2,3$.

(A.8) The bandwidth sequence h_N verifies that $h_N = c_N N^{-\theta}$ where $c_0 <$

$c_N < c_0^{-1}$ for some $c_0 > 0$ and $\theta \in (1/8,1/6)$.

(A.9) We allow to use preliminary estimators of the quantities involved in the correction terms. The preliminary estimators of p_1, and η_2 converge uniformly at the rate $o_p\left(N^{-1/4}\right)$, as N tends to infinity.

The first three assumptions are rather standard in the literature. (A.5) is an identifying assumption necessary to identify the parameters β_1 and λ_1, more concretely, that there is not a perfect relation between the linear part of the equation and the unknown function. A sufficient condition is that the reduced equations have al least a significant regressor not included in the structural equation. The necessary conditions for developing the limiting distribution of the estimator are similar to the ones of Ahn and Powell (1993) and also the analysis of the large sample properties of S_{ww} and $S_{w\epsilon}$ are based on the results of Powell (1987).

Thus being $\beta = (\beta_1, \lambda_1)$,

$$\hat{\beta} = \beta + S_{ww}^{-1} S_{w\epsilon},$$

where

$$S_{w\epsilon} = \frac{1}{h^2} N_2^{-1} \sum_i \sum_{i<j} K\left(\frac{p_{1i}-p_{1j}}{h}, \frac{\eta_{2i}-\eta_{2j}}{h}\right) p_i p_j \times (w_{1i} - w_{1j})\{(\mu_1(p_{1i}, \eta_{2i}) - \mu_1(p_{1j}, \eta_{2j})) + \epsilon_i - \epsilon_j\}$$

Lemma 1: Under assumptions (A.1) to (A.9), as N tends to infinity,

$$S_{ww} = 2\Sigma_{ww} + o_p(1) \tag{20}$$

where Σ_{ww} had been defined above. Furthermore,

$$S_{w\epsilon} = \frac{2}{N}\sum_i \left[p_i\rho_{1i}\gamma_i(w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))\right]\epsilon_i + o_p(\frac{1}{\sqrt{N}})$$

where ϵ_i is the error term and $\rho_i, \gamma_i, \lambda_1(\cdot)$ are defined in (A.6). We can deduce the asymptotic normal distribution of $\tilde{\beta}_1$ from $Lemma1$ ($\tilde{\beta}_1$ would be the estimator if we would know p_1 and η_2, as these quantities are unknown and we need to estimate it, then the estimator is $\hat{\beta}_1$. In relation with all this, we can similarly derive the large-sample properties of the feasible estimator, $\hat{\beta}_1$.

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = \left[\hat{S}_{ww}\right]^{-1}\sqrt{n}\hat{S}_{w\epsilon} \tag{21}$$

Proofs of **Theorem 1**:

Being $w_1 = (x_1, s)$ and $\beta = (\beta_1, \lambda_1)$;

$$\hat{\beta} = \left[\binom{N}{2}^{-1}\sum_i \sum_{i<j} \omega_{ij}p_ip_j\left[w_{1i} - w_{1j}\right]\left[w_{1i} - w_{1j}\right]^T\right]^{-1} \tag{22}$$

$$\left[\binom{N}{2}^{-1}\sum_i \sum_{i<j} \omega_{ij}p_ip_j\left[w_{1i} - w_{1j}\right]\left[w_{1i} - w_{1j}\right]^T\right]\beta_1 +$$

$$+\left[\binom{N}{2}^{-1}\sum_i \sum_{i<j} \omega_{ij}p_ip_j\left[w_{1i} - w_{1j}\right]\left[w_{1i} - w_{1j}\right]^T\right]^{-1}\left[\binom{N}{2}^{-1}\sum_i \sum_{i<j} \omega_{ij}p_ip_j\left[w_{1i} - w_{1j}\right][\epsilon_i - \epsilon_j]\right]$$

Thus;

$$\hat{\beta} - \beta = \left[\binom{N}{2}^{-1}\sum_i \sum_{i<j} \omega_{ij}p_ip_j\left[w_{1i} - w_{1j}\right]\left[w_{1i} - w_{1j}\right]^T\right]^{-1} \tag{23}$$

$$\left[\binom{N}{2}^{-1}\sum_i \sum_{i<j} \omega_{ij}p_ip_j\left[w_{1i} - w_{1j}\right][\epsilon_i - \epsilon_j]\right] = S_{ww}^{-1}S_{w\epsilon}$$

To proof theorem 1 it is necessary to establish some lemmas which are included and detailled in the Appendix Online. Finally, appliying these lemmas we can show that,

$$||\hat{S}_{ww} - S_{ww}|| \leq \left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} [w_{1i} - w_{1j}][w_{1i} - w_{1j}]^T\right] sup_{ij}(\hat{\omega}_{ij} - \omega_{ij}) =$$

$O_p(1)o_p(1) = o_p(1)$.

And the asymptotic distribution for $S_{w\epsilon}$, which is developped in the Appendix online is:

$$S_{w\epsilon} = \frac{2}{N}\sum_i [p_i\rho_{1i}\gamma_i(w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))]\eta_{11i} + o_p(\frac{1}{\sqrt{N}}) \quad (24)$$

Then,

$$\hat{\beta} = \beta + \Sigma^{-1}\frac{2}{\sqrt{N}}\sum_i [p_i\rho_{1i}\gamma_i(w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))]\eta_{11i} + o_p(1) \quad (25)$$

where η_{11i} is the error term and ρ_i,γ_i and $\lambda_1(\cdot)$ are defined in (A.6).

Appendix B:

Table B1: **Montecarlo Simulations**

		OLS			Heckman 2-Steps			Our Estimator		
		Mean Bias	SD	RMSE	Mean	SD	RMSE	Mean	SD	RMSE
N=50	$\beta 1$	0.046	0.138	0.145	0.104	0.132	0.168	-0.082	0.032	0.089
	$\lambda 1$	-0.065	0.183	0.194	-0.082	0.171	0.189	-0.042	0.039	0.059
N=100	$\beta 1$	-0.036	0.094	0.100	0.083	0.115	0.141	0.025	0.029	0.037
	$\lambda 1$	-0.013	0.116	0.116	-0.066	0.128	0.144	-0.034	0.016	0.0381
N=500	$\beta 1$	-0.019	0.050	0.053	-0.011	0.054	0.055	0.024	0.014	0.027
	$\lambda 1$	0.012	0.050	0.051	0.060	0.055	0.081	0.020	0.011	0.024

Source: Own Elaboration

For the **Appendix Online:**

Proofs of **Theorem 1**:

Being $w_1 = (x_1, s)$ and $\beta = (\beta_1, \lambda_1)$;

$$\hat{\beta} = \left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j [w_{1i} - w_{1j}][w_{1i} - w_{1j}]^T\right]^{-1} \tag{22}$$

$$\left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j [w_{1i} - w_{1j}][w_{1i} - w_{1j}]^T\right] \beta_1 +$$

$$+ \left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j [w_{1i} - w_{1j}][w_{1i} -$$

$$w_{1j}]^T\Big]^{-1} \left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j [w_{1i} - w_{1j}][\epsilon_i - \epsilon_j]\right]$$

$$= \beta + \left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j [w_{1i} - w_{1j}][w_{1i} -$$

$$w_{1j}]^T\Big]^{-1} \left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j [w_{1i} - w_{1j}][\epsilon_i - \epsilon_j]\right]$$

Thus;

$$\hat{\beta} - \beta = \left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j [w_{1i} - w_{1j}][w_{1i} - w_{1j}]^T\right]^{-1} \tag{23}$$

$$\left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j [w_{1i} - w_{1j}][\epsilon_i - \epsilon_j]\right] = S_{ww}^{-1} S_{w\epsilon}$$

To proof theorem 1 it is necessary to establish the following lemmas;

Lemma 1

The proof of this theorem is based in a generalization of lemma 5.1 and theorem 5.1 in [20]. Consider a second order U-Statistics of the form:

$$U_n = \binom{N}{2}^{-1} \sum_{i=1}^{n} p_n(\xi_i, \xi_j) \tag{24}$$

where the sum is taken over the n_2 combinations of 2 distinct elements $[i, j]$ from the sample set. Without loss of generality the kernel function $p_n(.)$ can be taken to be symmetric in its arguments. Additionally we define:

$$\theta_n = E[r_n(\xi_i)] = E[p_n(\xi_i, \xi_j)] r_n = E[p_n(\xi_i, \xi_j)]$$

Then;

$$\widehat{U}_n = \theta_n + \frac{2}{n} \sum_{i=1}^{n} [r_n(\xi_i) - \theta_n]$$

If the kernel $p_n(.)$ satisfies; $E[||p_n(\xi_i, \xi_j)||^2] = o(n)$, then;

$$U_n = \widehat{U}_n + o_p(n^{-1/2}) \tag{25}$$

$$\widehat{U}_n = 2U_n + o_p(1) \tag{26}$$

According to this we can write:

1. $S_{ww} = 2\Sigma_{ww} + o_p(1)$
2. $\hat{S}_{w\epsilon} = S_{w\epsilon} + o_p(n^{-1/2})$

Lemma 2

From lemma 3.1 of [19]. For an iid sample of random variables, we have:

$$E\left[\omega_{ij}p_i p_j (w_{1i} - w_{1j})(w_{1i} - w_{1j})^T | s_i, w_{1i}, p_{1i}, \eta_{2i}, p_{1j}, \eta_{2j}\right] \quad (27)$$

Now, operating we have:

$$E\left[K_h\left(\frac{p_{1i}-p_{1j}}{h}, \frac{\eta_{2i}-\eta_{2j}}{h}\right) E\left[p_i p_j (w_{1i} - w_{1j})(w_{1i} - w_{1j})^T | p_{1j}, \eta_{2j}\right]\right]$$

We chose a function $\lambda_1(p_{1j}, \eta_{2j})$ such that:

$$E\left[p_j (w_{1j} - \lambda_1(p_{1j}, \eta_{2j})) | p_{1j}, \eta_{2j}\right] = 0$$

And has this expression:

$$\lambda_1(p_{1i}, \eta_{2i}) = \frac{E[w_{1i}p_i | p_{1i}, \eta_{2i}]}{E[p_i | p_{1i}, \eta_{2i}]}$$

We can also denote:

$$E[p_i | p_{1i}, \eta_{2i}] = \rho_{1i}$$

Then:

$$E\left[K_h\left(\frac{p_{1i}-p_{1j}}{h}, \frac{\eta_{2i}-\eta_{2j}}{h}\right) E\left[p_i p_j (w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))(w_{1j} -\right.\right.$$

$$\left.\left.\lambda_1(p_{1j}, \eta_{2j}))^T | p_{1i}, \eta_{2i}, p_{1j}, \eta_{2j}\right]\right] =$$

$$E\left[K_h\left(\frac{p_{1i}-p_{1j}}{h}, \frac{\eta_{2i}-\eta_{2j}}{h}\right) E\left[p_i p_j (w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))(w_{1j} -\right.\right.$$

$$\left.\left.\lambda_1(p_{1j}, \eta_{2j}))^T | p_{1i}, \eta_{2i}, p_{1j}, \eta_{2j}\right]\right] =$$

$$= E\left[K_h\left(\frac{p_{1i}-p_{1j}}{h}, \frac{\eta_{2i}-\eta_{2j}}{h}\right) E[p_i | p_{1i}]^2 (w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))(w_{1j} - \lambda_1(p_{1j}, \eta_{2j}))^T\right]$$

That is:

$$= E[\rho_i^2 \gamma_i (w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))(w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))^T]$$

Finally we will have:

$$\Sigma_{ww} = E[\rho_{1i}^2 \gamma_i (w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))(w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))^T]$$

Where:

$$\gamma_i = \mu_1(p_{1i}, \eta_{2i}) \tag{28}$$

So that;

$$S_{ww} = 2\left[E[\rho_i^2 \gamma_i (w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))(w_{1i} - \lambda_1(p_{1i}, \eta_{2i}))^T]\right] + o_p(1)$$

And note here that,

$$||\hat{S}_{ww} - S_{ww}|| \leq \left[\binom{N}{2}^{-1} \sum_i \sum_{i<j} [w_{1i} - w_{1j}][w_{1i} - w_{1j}]^T\right] sup_{ij}(\widehat{\omega}_{ij} - \omega_{ij}) =$$

$O_p(1)o_p(1) = o_p(1)$.

So this closes the proof.

We are going to obtain now the asymptotic distribution for $S_{w\epsilon}$

$$S_{w\epsilon} = \binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j (w_{1i} - w_{1j})(\epsilon_i - \epsilon_j)$$

given that:

$$(\epsilon_i - \epsilon_j) = (w_{11i} - w_{11j})'\beta_1 + (s_i - s_j)\lambda_1 + \mu_1(p_{1i}, \eta_{2i}) - \mu_1(p_{1j}, \eta_{2j}) + (\eta_{11i} - \eta_{11j})$$

Substituting:

$$S_{w\epsilon} = \binom{N}{2}^{-1} \sum_i \sum_{i<j} K_h\left(\frac{p_{1i} - p_{1j}}{h}, \frac{\eta_{2i} - \eta_{2j}}{h}\right) p_i p_j [w_{1i} - w_{1j}][w_{1i} - w_{1j}]^T \beta +$$

$$+\binom{N}{2}^{-1}\sum_i \sum_{i<j} K_h\left(\frac{p_{1i}-p_{1j}}{h},\frac{\eta_{2i}-\eta_{2j}}{h}\right) p_i p_j\left[\mu_1(p_{1i},\eta_{2i})-\mu_1(p_{1j},\eta_{2j})\right]+$$

$$+\binom{N}{2}^{-1}\sum_i \sum_{i<j} K_h\left(\frac{p_{1i}-p_{1j}}{h},\frac{\eta_{2i}-\eta_{2j}}{h}\right) p_i p_j(\eta_{11i}-\eta_{11j})$$

Reordering terms, we will have:

$$S_{w\epsilon}=\binom{N}{2}^{-1}\sum_i \sum_{i<j} \omega_{ij} p_i p_j (w_{1i}-w_{1j})(w_{1i}-w_{1j})^T\beta+$$

$$+\sum_i \sum_{i<j} K_h\left(\frac{p_{1i}-p_{1j}}{h},\frac{\eta_{2i}-\eta_{2j}}{h}\right) p_i p_j (w_{1i}-w_{1j})(w_{1i}-w_{1j})^T\left[\mu_1(p_{1i},\eta_{2i})-\right.$$

$$\left.\mu_1(p_{1j},\eta_{2j})\right]+\sum_i \sum_{i<j} \omega_{ij} p_i p_j (w_{1i}-w_{1j})(\eta_{11i}-\eta_{11j})$$

It is clear that:

$$E\left[K_h\left(\frac{p_{1i}-p_{1j}}{h},\frac{\eta_{2i}-\eta_{2j}}{h}\right) p_i p_j (w_{1i}-w_{1j})\left[\mu_1(p_{1i},\eta_{2i})-\mu_1(p_{1j},\eta_{2j})|p_{1j},\eta_{2j}\right]\right]=0$$

So finally, we will have:

$$E\left[K_h\left(\frac{p_{1i}-p_{1j}}{h},\frac{\eta_{2i}-\eta_{2j}}{h}\right) p_i p_j (w_{1i}-w_{1j})(\eta_{11i}-\eta_{11j})|p_{1j},\eta_{2j}\right]=p_i\rho_{1i}\gamma_i(w_{1i}-$$

$$\lambda_1(p_{1i},\eta_{2i}))\eta_{11i}$$

So now, we have the asymptotic expansion for $S_{w\epsilon}$;

$$S_{w\epsilon}=\frac{2}{N}\sum_i\left[p_i\rho_{1i}\gamma_i(w_{1i}-\lambda_1(p_{1i},\eta_{2i}))\right]\eta_{11i}+o_p\left(\frac{1}{\sqrt{N}}\right) \tag{29}$$

Finally,

$$\hat{\beta}=\beta+\Sigma^{-1}\frac{2}{\sqrt{N}}\sum_i\left[p_i\rho_{1i}\gamma_i(w_{1i}-\lambda_1(p_{1i},\eta_{2i}))\right]\eta_{11i}+o_p(1) \tag{30}$$

where η_{11i} is the error term and ρ_i,γ_i and $\lambda_1(\cdot)$ are defined in (A.6).