UNIVERSIDAD DE CANTABRIA, UNIVERSIDAD DE OVIEDO Y UNIVERSIDAD DEL PAÍS VASCO

Programa de Doctorado de Economía: Instrumentos del Análisis Económico

TESIS DOCTORAL

Ensayos sobre modelos econométricos sujetos a selección muestral y endogeneidad en el campo de la Economía de la Salud y del Trabajo

DOCTORAL PHD

Essays about econometric models subject to sample selection and endogeneity in the field of Health Economics and Labour

Patricia Moreno Mencía

Directores: Prof. Dr. Juan Rodriguez Poo y Prof. Dr. David Cantarero Prieto

Escuela de Doctorado de la Universidad de Cantabria

Santander 2018

UNIVERSIDAD DE CANTABRIA, UNIVERSIDAD DE OVIEDO Y UNIVERSIDAD DEL PAÍS VASCO

Programa de Doctorado de Economía: Instrumentos del Análisis Económico

TESIS DOCTORAL

Ensayos sobre modelos econométricos sujetos a selección muestral y endogeneidad en el campo de la Economía de la Salud y del Trabajo

DOCTORAL PHD

Essays about econometric models subject to sample selection and endogeneity in the field of Health Economics and Labour

Patricia Moreno Mencía

Directores: Prof. Dr. Juan Rodriguez Poo y Prof. Dr. David Cantarero Prieto

Escuela de Doctorado de la Universidad de Cantabria

Santander 2018

A mi familia

Agradecimientos

El desarrollo de esta tesis doctoral no ha sido un camino fácil, sin embargo nunca olvidaré cada apoyo que me encontré en esta trayectoria. Es por ello que no puedo más que agradecer cada ayuda, palabra de aliento y ánimo incondicional que me han sido otorgados.

En primer lugar, a mis directores de tesis, los profesores Juan Rodriguez-Poo y David Cantarero por su labor de guías en este largo camino. Ellos me animaron a continuar a pesar de la fatiga y me iluminaron cuando aparecía la oscuridad. No puedo olvidarme también de mi querida Ana Fernández, que ha sido un gran pilar para mí en el desarrollo de esta tesis. De ella he aprendido muchas cosas, tanto de ciencia como de la vida misma.

Asimismo, a mis compañeros del área de Econometría por su valioso apoyo a lo largo este tiempo, así como al resto de miembros del Departamento de Economía por la ayuda prestada. Quisiera agradecer especialmente a mis compañeras y amigas Sandra, Carla y Ana los consejos prestados y las situaciones de evasión que me proporcionaban en los momentos más necesarios.

No puedo tampoco olvidar a mis amigas: Raquel, Alba, Conchi, Maria e Isa que han soportado muchos de mis altibajos y siempre me han animado tratando de ofrecerme momentos de diversión para coger más impulso cuando más falta me hacía.

Reservo un lugar especial para toda mi familia. En especial a mis padres, Justo y Maria Antonia que me enseñaron a ser constante, responsable y me inculcaron los grandes valores de la vida. A mi marido, Froy, por su comprensión y apoyo incondicional todos estos años porque me dió fortaleza para seguir adelante. Finalmente, a mi luz, mi hijo Mario, porque junto a él las cosas malas se convierten en buenas y la tristeza se convierte en alegría. Mi dedicación en la tesis le quitó tiempo de juegos junto a su mami, pero sin proponérselo ha sido mi mayor incentivo.

Contenidos

Índice general

Índice de cuadros

Índice de figuras

Introducción

Introducción

El mercado laboral y su evolución en nuestro país en los últimos años constituye un gran reto para la Economía Española. Su análisis constituye pues un tema prioritario para los investigadores de nuestra sociedad. Desde la seminal propuesta de Heckman (1974) el problema de especificación y estimación de modelos de oferta laboral, los cuales habitualmente están sujetos a selección muestral han sido ampliamente estudiados en el campo de la Econometría. Tradicionalmente, la especificación econométrica de este tipo de modelos se ha basado en la imposición de supuestos bastante restrictivos sobre las formas funcionales y las distribuciones de los errores, generalmente normal. La idea basicamente consistía en estimar a través de métodos de máxima verosimilitud y el popular método de estimación en dos etapas. En las últimas décadas muchos han sido los autores que han trado de proponer nuevos enfoques para relajar algunos de los rígidos supuestos necesarios para la estimación descrita anteriormente.

La introducción de algunas variables de interés en los modelos de oferta laboral, como puede ser la cronicidad, la depresión o la antigüedad del trabajador introducen adicionalmente en el modelo estructural endogeneidad. Si dicha endogeneidad no es tratada de manera adecuada las estimaciones propuestas estarán sesgadas. En este caso, muchos investigados han propuesto el uso de variables intrumentales. (véase Heckman (1978), Angrist y Krueger (2001) y Vytlacil y Yildiz (2007)).

Las investigaciones empíricas habitualmente tienen como objetivo estudiar como cambia una variable de interés ante variaciones de ciertos factores explicativos. En este sentido, el modelo de regresión lineal es uno de los más empleados en el anásis empírico. El método de minimo cuadrados ordinarios permite obtener estimaciones de los efectos y permite una intuitiva interpretación así como la consecuente inferencia estadística. Sin embargo, se basa en supuestos muy rigidos que no siempre se sostienen. Los modelos de elección discreta y con variable dependiente censurada han sido ampliamente estudiados en los últimos años. Campos como el de Economía Laboral y Economía de la Salud se han visto frecuentemente afectados por los problemas econométricos que convellan estos tipos de modelos. Así pues, en esta tesis se considera la especificación y estimación de este tipo de modelos que si no son correctamente tratados conllevan a estimaciones inconsistentes. En concreto, es habitual que este tipo de modelos adolezcan de sesgo de selección si la población a estudiar no es elegida de manera aleatoria. En esta dirección, muchos trabajos se han centrado en ampliar el popular trabajo de estimación en dos etapas de Heckman (1979) como se ha mencionado anteriormente. Dicho enfoque consiste en obtener estimadores consistentes de la ecuación de selección en una primera etapa y en una segunda fase se estiman los parametros de la ecuación estructural incluyendo una funcíon de corrección con los estimadores de la primera etapa. A partir de este popular enfoque, basado en un contexto totalmente paramétrico, diversos autores han tratado de flexibilizar los supuestos necesarios en ese contexto, que resultan demasiado restrictivos. Literatura más reciente propone la utilización de modelos semi parametricos con la finalidad de dejar sin restringir la función de corrección y evitando así los problemas de mala especificación funcional de la misma. (Veáse Powell (1987)and Ahn y Powell (1993)). Este enfoque de modelización se basa principalmente en que sean los propios datos los que determinen la forma funcional adecuada.

Por otro lado, si alguno se los regresores es endógeno se incumple uno de los supuestos básicos para utilizar minimos cuadrados ordinarios siendo necesaria algún tipo de correción previamente. Modelos que adolezcan de ambos problemas simultaneamente han sido menos estudiados y requieren pues un análisis más profundo. A lo largo de esta tesis se irán proponiendo modelos de oferta laboral en distintos contextos que requerirán una adecuada especificación y unos métodos económetricos específicos para esta problematica, nada standard.

Ambos problemas, la selección y la endogeneidad resultan de especial interés en esta tesis, en la que se propone abordar dicha problemática de especial relevancia en problemas empíricos de oferta laboral de ciertos colectivos especiales como personas con discapacidad, mujeres casadas o empleados públicos. El enfoque de la función de control ha sido analizada en profundidad para lidiar con este tipo de cuestiones, es el caso de Wooldridge (2015) o Vella (1993). Las bases de este enfoque son nuestras premisas de partida, la construcción de términos de correción a partir de esperanzas condicionadas desconocidas, ya sea a través de probabilidades estimadas de la primera etapa, de residuos, o de una mezcla de ambos. En este contexto, las técnicas de regresión no paramétricas han ido ganando peso para solventar este tipo de problemas relajando cualquier supuesto restrictivo. Más aún, se trata de no asumir una forma funcional predeterminada a esta función de control pudiendo dejarla entrar en el modelo de manera no paramétrica, ampliando la propuesta de Ahn y Powell (1993) para resolver cada uno de los modelos propuestos a lo largo de los capítulos de esta tesis. El uso de este tipo de técnicas también tienen sus desventajas y es que aunque provean estimadores robustos a la incorrecta especificación de la función de regresión, estarán sujetos a la denominada maldición de la dimensionalidad además de que la interpretación de las estimaciones es menos intuitiva. Para manejar este tipo de desventajas se utilizan también los modelos semiparamétricos, que son como una mezcla de ambos enfoques.

Los modelos considerados en este tipo de estudios se componen de una ecuación estructural, que especifica el objetivo principal de partida a investigar y en el cual se ponen de manifiesto los problemas de selección muestral y/o endogeneidad. Dado este caso, se añaden una o varias ecuaciones reducidas para especificar cada uno de los procesos, obteniendo así un sistema de ecuaciones simultaneas. Si se da el caso de presencia de regresor endógeno de tipo continuo es algo habitual el uso del enfoque de la función de control, Blundell y Powell (2003). En este marco, se usa un modelo lineal para especificar la relación existe entre el regresor endógeno y el resto de covariables exógenas del modelo. La estimación de la primera etapa consiste pues en la obtención de los residuos de dicho proceso y su posterior inclusión como un regresor adicional en la ecuación estructural con el objetivo de controlar la mencionada endogenidad. Si bien es cierto que este procedimiento es el más habitual no funciona si el regresor endógeno no es continuo. En este segundo caso, Vytlacil y Yildiz (2007) propusieron una estrategia de estiamción para la situación de tener un regresor endógeno binario.

En este contexto expuesto, la motivación de esta tesis tiene una doble vertiente. Por un lado, se trata de desarrollar nuevas técnicas de estimación consistentes con la presencia de selección muestral y endogeneidad. Por otro lado, se basa en la aplicación de dichas propuestas capaces de manejar los incovenientes subrayados anteriormente a campos de interés en la Economía de nuestro país. La cronicidad, la discapacidad y la dependencia son problemas de actualidad en la sociedad española. Los cambios demográficos actuales conllevan implicaciones muy relevantes para el conjunto de la sociedad. Algunos ejemplos son el incremento de población jubilada, lo que amenaza el sistema de pensiones público. Una población cada vez más envejecida y con más disapacidades que demanda cada vez más recursos tanto de cuidados, como de gasto sanitario. Además, el aumento de las enfermedades crónicas y discapacidad demanda a su vez una mayor flexibilidad laboral para poder integrarse en el mercado de trabajo. Dada la relevancia del tema expuesto, resulta pues, de obligada necesidad investigar y analizar cómo afectan dichas situaciones a los individuos de nuestro país y como prevenirlas. En España más de 19 millones de personas están afectadas por alguna enfermedad crónica, la mayoría son mayores de 55 años de edad y requieren más del 75 % del gasto sanitario. En nuestro país, el presupuesto destinado a cronicidad ascendió a más de $50,000$ millones y además se espera que incremente un 45 % para 2020 dados los cambios demográficos y el aumento de la cronicidad. Así pues, el significativo crecimiento de las enfermedades crónicas en los últimos años hace necesario el estudio de todas sus consecuencias. Los posibles tratamientos y la prevención, la mejora de la calidad de vida y la inserción social de las personas con enfermedades crónicas son asuntos importantes para tratar. Este tipo de cuestiones relacionadas con la discapacidad, la dependencia y la cronicidad de los españoles tienen un importante impacto en nuestra sociedad tanto por razones sanitarias como económicas.

Son muchos los efectos de la cronicidad, la discapacidad y la dependencia en la economía española. Basicamente el deterioro de la salud de un individuo afecta a su educación, a su situación en el mercado laboral y a un aumento del gasto. También podría analizarse como afecta a la sociedad en general, en cuanto al incremento del gasto público y de las necesidades de atención. En esta tesis vamos a centrar nuestro objetivo en el ámbito laboral y cómo afectan a dichas decisiones laborales ciertos factores. Por otro lado, hemos de tener en cuenta la aplicación de una correcta metodología dadas las particularidades expuestas. Así pues, en este trabajo proponemos la estimación de los parámetros de interés a través de novedosos métodos econométricos capaces de solventar los problemas de sesgo e inconsistencia de los enfoques más tradicionales. Con este objetivo, la tesis doctoral se divide en cuatro capítulos y la estructura de la misma es la siguiente:

En el capítulo 1 se propone un modelo de maximización de utilidad que trata de explicar la baja participación laboral de las personas con discpacidad. En este caso, se tiene en cuenta la heterogeneidad en las preferencias y se considera que el tiempo destinado a cuidados para las personas con discapacidad está incluído en su función de utilidad. Las horas de trabajo deseables no son pues homogenas (dependen de características inobservables) ni tampoco continuas (hay ciertos paquetes elegibles unicamente).

En el capítulo 2 nos centramos en la estimación del impacto de la cronicidad en la oferta laboral de la población española. Con este objetivo, se propone un modelo de ecuaciones simultaneas que incluye una variable endógena binaria que está presente tanto en el modelo estructural como en el mecanismo de selección. En este contexto, no estamos dispuestos a la imposición de supuestos demasiado rigidos que no son requeridos por la teoria economica, por tanto los parámetros de la ecuación estructural se estiman semiparamétricamente. Para ello, tanto la selección muestral como la endogeneidad se tienen en cuenta a través de la llamada función de control. Bajo las condiciones impuestas en este articulo, el modelo econométrico resultante es un modelo parcialmente lineal con un componente no parametrico. Adicionalmente comparamos los resultados obtenidos bajo nuestra propuesta metodológica frente a algunas de las más usuales.

En el capítulo 3 se propone analizar si existe una diferencia significativa entre los salarios del sector público y privado. En este sentido, se considera que la decisión de entrar en un proceso de acceso al empleo público es un proceso previo que se determina endogenamente en el modelo y por ello deben propenserse métodos econométricos adecuados a esta situación. Además dado que el género es una variable con mucha influencia en las diferencias salariales consideramos también la brecha existente por causa de género en ambos sectores. Se propone usar técnicas de regresión por cuantiles para analizar la brecha salarial a lo largo de toda la distribución. El objetivo es contrastar la hipótesis de que las diferencias salariales por género son mayores de la parte alta de la distribución salarial (techos de cristal) y en la parte más baja de la misma. Más aún, el objetivo del capitulo se centra en corroborar si dicha brecha salarial es significativamente menor en el Sector Público que en el Privado, como cabría esperar a priori.

Por último, en el capítulo 4, se presenta la especificación y estimación de un modelo capaz de lidiar con selección muestral y la endogeneidad al mismo tiempo. En este caso, se propone un modelo semiparamétrico para analizar el efecto que tiene la antiguedad en el salario de los empleado públicos. Resulta claro que en este análisis el problema de selección surge porque los salarios públicos son sólo observables para los trabajadores que efectivamente trabajen en el Sector Público y la decisión de participación en el mismo se determina de manera endógena. Además adicionalmente, en el modelo la variable explicativa de interés es endógena porque hay variables inobservables que influyen en la variación de los salarios que están claramente correlacionados con la antiguedad del trabajador. Con estos objetivos, se plantea una extensión del modelo de dos etapas de Heckman relajando los supestos de errores normalmente distribuídos y forma funcional preseleccionada. Se concluye que existe una prima positiva en el Sector Público a favor de los hombres, de los trabajadores con estudios de mayor nivel y también para los que cuentan con más tiempo de antiguedad.

Finalmente, se concluye esta tesis doctoral presentando las principales conclusiones que se obtienen a lo largo de estos cuatro capítulos y se destacan las líneas de investigación a seguir en el fututo. En último lugar se encuentran las pruebas de los principales resultados obtenidos en la tesis y algunos resultados adicionales que han sido relegados al apéndice de la misma.

Versiones preliminares de estos capítulos han sido presentados en diversos congresos científicos tanto nacionales como internacionales. Asimismo el capítulo 1 de la misma ha sido publicado en la revista Applied Economics y el resto han sido enviados para evaluación y publicación en revistas economicas de relevancia.

Introduction

Introduction

The labor market and its evolution in our country in the last few years constitutes a great challenge for the Spanish Economy. Its analysis constitutes a priority issue for researchers in our society. From the seminal proposal of Heckman (1974) the problem of specification and estimation of models of labor supply, which usually are subjects to sample selection have been widely studied in the field of Econometrics. Traditionally, the econometric specification of this type of models has been based on the imposition of rather restrictive assumptions about functional forms and distribution of the errors, usually normal. The idea basically consisted on estimating through methods of maximum likelihood and the popular method of estimation in two stages. In recent years, many authors have been proposing new approaches to relax some of the strong assumptions necessary for those estimation proposals.

The introduction of some variables of interest in the labor supply models, such as the chronicity, the depression or the tenure of the worker additionally introduce in the structural model endogeneity. If such endogeneity is not adequately treated, the proposed estimates will be biased. In this case, many researchers have proposed the use of instrumental variables. (see Heckman (1978), Angrist y Krueger (2001) and Vytlacil y Yildiz (2007)).

Empirical research usually aims to study how the variable of interest changes in relation to variations of certain explanatory factors. In this sense, the linear regression model is one of the most used in the empirical analysis. The method of ordinary least squares allows us to obtain estimations of these effects and give us an intuitive interpretation and the consequent statistical inference. However, it is based on very rigid assumptions that is not always realistic. Discrete choice models with censored dependent variable have been widely studied in recent years. Fields such as the Labor Economics and Health Economics have often been affected by the econometric problems associated with these types of models. In this thesis is considered the specification and estimation of these type of models that if are not correctly treated lead to inconsistent estimates. In particular, it is something common that these type of models suffer from selection bias if the population of interest is not chosen in a random way. In this direction, many studies have focused on expanding the popular work of estimation in two stages of Heckman (1979) as we have mentioned above. This approach consists of obtaining consistent estimators of the selection equation in a first stage and in a second step the parameters of the structural equation are estimated including a correction function with first step estimates. From this popular approach, based on a totally parametric context, several authors have tried to make the necessary assumptions more flexible in that context, which actually are too restrictive. Most recent literature proposes the use of semi parametric models with the purpose of leaving the correction function unrestricted and thus, avoiding problems of functional mispecification. (see Powell (1987) and Ahn y Powell (1993)). This modeling approach is mainly based on the idea that is just the data which determine the functional form.

On the other hand, if any of the regressors is endogenous, then one of the basic assumptions to use least ordinary squares is violated. Models suffering of both problems simultaneously have been less studied and thus it require a deeper analysis. Throughout this thesis, we will be proposing labor supply models in different contexts that require a proper specification and adequate econometric methods for this non-standard problems.

Both problems, the selection and the endogeneity are of special interest in this thesis, which proposes to address this problem of special relevance in empirical labor supply problems of certain groups, such as people with disabilities, women married or public employees. The Control Function Approach has been analyzed in depth to deal with this type of issues, Wooldridge (2015), Vella (1993). The basis of this approach is our starting premise, the construction of correction terms from unknown conditional expectations, whether through estimated first stage probabilities, waste or a mixture of both. In this context, non-parametric regression techniques have been gaining weight to solve these types of problems by relaxing any restrictive assumptions. Moreover, the objective is to left the control function unrestricted, in a non parametric context extending the idea of Ahn y Powell (1993) and adapting the approach to any of the situations studies in this thesis. The use of these types of techniques also have their disadvantages and is that although they provide robust estimators to the incorrect specification of the function of regression, are subject to the so-called Çurse of the dimensionality", moreover the interpretation the obtained estimates it is more complicated. To handle this type of disadvantages, it is also used a semi-parametric specification, which are like a mixture of both approaches.

The models considered in this type of studies are composed of a structural equation, which specifies the main starting target to be investigated and in which the problems of sample selection and/or endogeneity are manifested. Given this case, one or several reduced equations are added to specify each one of the processes and thus, obtaining a system of simultaneous equations. In the case of presence of a continuous endogenous regressor, the use of the control function approach is something usual, Blundell y Powell (2003). In this framework, a linear model is used to specify the relationship between the endogenous regressor and the rest of the exogenous covariates of the model. The estimation of the first stage consists in obtaining the residuals of this process and its subsequent inclusion as an additional regressor in the structural equation with the objective of controlling the aforementioned endogeneity. Although it is true that this procedure is the most usual, it does not work if the endogenous regressor is not continuous. In this second case, Vytlacil y Yildiz (2007) proposed a strategy of estimating the situation of having a binary endogenous regressor.

In this described context, the motivation of this thesis is twofold. On the one hand, it is about developing new estimation techniques consistent with the presence of selection and endogeneity. On the other hand, it is based on the application of some proposals able of handling with problems previously highlighted and applying it to fields of interest in the Economy of our country. Chronicity, disability and dependency are current problems in Spanish society. The current demographic changes have really relevant implications for our society as a whole. Some examples are the increase in the retired population, which threatens the public pension system. A population that is becoming more and more aged and with more disabilities demanding more and more resources for both, care and health expenditure. Additionally, the increase in chronic diseases and disability demands also greater labor flexibility to be able to integrate into the labor market. Given the relevance of the subject, it is therefore necessary to investigate and analyze how these situations affect the individuals of our country and how to prevent them. In Spain, more than 19 million people are affected by a chronic disease, most of them are older than 55 years and require more than 75 % of health expenditure. In our country, the budget destined to chronicity amounted to 50, 000 million and in addition, it is expected to increase a 45 % more to 2020 due to demographic changes and the increase of chronicity. The significant growth of the chronic diseases in the last few years makes it necessary to study all its consequences. The possible treatments and prevention, the improvement of the quality of life and the social insertion of people with chronic diseases are important issues to deal with. This type of issues related to disability, dependence and chronicity of Spaniards have an important impact on our society for both health and economic reasons.

There are many effects of chronicity, disability and dependence on the Spanish economy. Basically the health deterioration of an individual affects his education, his situation in the labor market and also produces an increase in spending. It could also be analyzed as an effect for the society in general, producing an increase in public spending and more attention needs. In this thesis we will focus our objective on the working environment and how certain factors affect these labor decisions. On the other hand, we must take into account the application of a correct estimation method given the particularities exposed. Thus, we propose the estimation of the parameters of interest through novel economical methods able of solving the problems of bias and inconsistency of the most traditional approaches. With this objective, the doctoral thesis is divided into four chapters and the structure of the same is as follows:

In chapter 1 an utility maximization model is proposed for explaining the low labor participation ratio of people with disabilities. In this case, the heterogeneity in preferences is taken into account and also it is considered that the time allocated to health care of people with disabilities is included in its utility function. The desirable working hours are therefore not homogenous (they depend on unobservable characteristics) nor continuous (there are certain eligible packages only).

In the chapter 2 we focused on the estimation of the impact of chronicity on the labor supply of the Spanish population. To this end, we propose a model of simultaneous equations that includes a binary variable that is present both in the structural model and in the selection mechanism. In this context, we are not willing to impose too rigid assumptions that are not required by economic theory, so the parameters of the structural equation are estimated semiparamically. For this end, both the sampling selection and the endogeneity are taken into account through the so-called control function. Under the conditions imposed in this article, the resulting econometric model is a partially linear model with a non-parametric component. Additionally, we compare the obtained results under our methodological proposal with some of the most usual methods.

In chapter 3, it is proposed a model to analyze if there is a significant difference between the salaries of the public and private sector. In this sense, it is considered that the decision to enter Public Sector is a prior process that is endogenously determined in the model and for this reason, the adequate econometric methods should be chosen. In addition, given that the gender is a variable with a great influence on wage differences, we also consider the existing gender-gap in both sectors. We proposed to use regression quantile techniques to calculate the wage gap throughout the whole distribution. The objective is to test the hypothesis that gender wage differences are greater in the upper part of the salary distribution (glass ceilings) or in the lower part. Moreover, the interest of this chapter focuses on corroborating whether the gender wage gap is significantly lower in the Public Sector than in the Private Sector.

Finally, in the chapter 4, the specification and estimation of a model able of dealing with sample selection and endogeneity at the same time is presented. In this case, a semi-parametric model is proposed to analyze the effect that the tenure has on the salary of the public employees. It is clear that in this analysis the selection problem arises because public wages are only observable for workers who actually work in the Public Sector and this participation decision is determined in an endogenous manner. In addition, in the model, the explanatory variable of interest is endogenous because there are unobservable variables that influence the variation of wages that are clearly correlated with the worker's tenure. With these objectives, an extension of the Heckman two-stage model is proposed but relaxing the normally distributed errors assumption and no imposing a preselected functional form. It is concluded that there is a positive premium in the Public Sector in favor of men, of workers with higher level studies and also for those with more tenure.

Finally, we conclude this doctoral thesis by presenting the main conclusions that are obtained throughout these chapters and highlight the research lines to follow in the future. In the last place, we summarize the main results obtained in the thesis and some additional results and mathematical proofs have been relegated to the appendix.

Preliminary versions of these chapters have been presented in several national and international scientific congresses. Likewise, Chapter 1 has been published in the Applied Economics journal and the rest have been sent for evaluation and publication in some relevant economic journals.

Capítulo 1

Capítulo 1

A new approach for understanding labour supply of disabled people.

The main interest of this chapter is to propose an individual utility maximization model to explain the low participation of disabled people. We account for heterogeneity of preferences and furthermore time of self caring for disabled individuals is considered as an argument in the utility function. The hours of work decided by disabled individuals are neither homogeneous (they depend on unknown characteristics) nor continuous (discrete choice sets). We use data of 4790 households from the Spanish Survey of Disability, Personal Autonomy and Dependency and find association between time of informal care and labour participation and consequently, the choice between jobs.

1.1. Introduction

The main interest of this article is to explain both the low participation rate and the small percentage of disabled people that joins the labour force. To drive it, we propose to develop an individual utility maximization framework. This model accounts for heterogeneity of preferences and it considers self-care time for disabled individuals as an argument included in the individual utility function. The basic idea is that people with some health conditions and taking prescription medications may not be able to work full-time or in shifts jobs. It is necessary to take into account that disabled people require more time for self-care than non-disabled people and they have different preferences about time assignations.

According to the Observatory of Disability and Labor Market in Spain (ODISMET), during the period of 2008-2009, only 28 % of disabled people had a job. This ratio is below the OECD average, which was 32 %. On the contrary, if we look at the participation rate, according to the Spanish National Bureau of Statistics (INE), only 36.6 % of disabled people participated in the labour market in that period, about 40 points lower than population without disability. These facts are a clear challenge to theorists and applied economists to explain these discrepancies both in terms of participation and employment rates. There is not only Active Labor Market Policies (ALMP) that have an impact on labour integration but also other variables such as workplace accessibility, education level for disabled people, etc. Thus, the results obtained in the paper should be of great interest for organizations of disabled people that are working both to ensure their labour integration and to promote changes in the employment opportunities for disabled people.

Activity and occupation rates among disabled people are really worrying, so the first of our goals is to analyse the causes of this low participation. Traditionally, low participation rates associated with disabled people have been explained from supply side (e.g., Livermore et~al. (2000)). Not only standard costs in the employment search are important. People with disabilities often incur in additional costs for transport, rehabilitation, assistance, etc. All those circumstances may reduce their labour participation rate.

Greve (2009), in the report for the Academic Network of European Disability Experts, pointed out that most countries pursue active strategies to integrate disabled people in the labour market. However, the success of these policies is not always evaluated. As an example of legislation, these report highlights the flex-jobs in Denmark that help to make it possible for people without full-work ability to enter the labour market. Additionally, Parodi y Sciulli (2008) applying a logit model carried out a labour participation model in households with disabled persons. In order to increase the income of the households with disabled members they explained policy recommendations such as the provision of care services and structural policies to improve employment.

However, standard approaches in individual maximization utility approach fail to consider several problems that are rather relevant when modeling participation decisions of disabled people. We assume that people with some declared disability (physical or psychical) need more time to improve their health; that is they have to go to medical consultations, rehabilitation or do specific exercises with more frequency than non-disabled people. This requires time, so we think that the preferences in time uses are different having disabilities or not. Indeed, previous studies consider too simple discrete choice models where, as an outcome, disabled people plan either not to work or to work a number of hours that is a continuous function of the offered wages. Unfortunately, other unobservable issues different from the wage and other individual characteristics in a labour participation model exists, such as location, accessibility or flexibility of the job. This set of unknown characteristics can create discontinuities in the labour supply and desired working hours will not depend only on offered wages, but also in those non-observed factors of each job-type that affect the utility function for disabled people.

In this paper, we consider a more general model where the heterogeneity in preferences between households is included. We suppose that individuals are not free to choose the working hours because labour markets are rigid. They only can choose between jobs which are seen as sets of several characteristics, one of which is effectively the timetable. Then, in order to allow for discreteness in labour supply functions we propose to use the framework developed in Dagsvik (1994)[1]. During the last few years, this approach and its extensions have been increasingly popular as in Van~Soest (1995), who offers an alternative approach in which the number of working hours were discretised and the error term was supposed to have an extreme value distribution. Assuming random utility maximization, he obtained the multinomial logit model (see (McFadden1984)). The same idea was proposed in a different framework, for nurses, (e.g.,Saether (2004) and Di~Tommaso et~al. (2009)) or for married women (Aaberge et~al. (1999) and Dagsvik y Strom (2006)). Also, for married women, Aaberge (1995) used simulation techniques to build the choice set.

An additional problem that the previous models present when dealing with disabled people's behaviour is that one may end up using empirical models where the decision of participation or not in the labour market is considered at the same level as the individual's decision on their labour supply. In statistical and empirical studies these models are specified as multinomial logit ones and in order to implement inference, it is necessary to assume the Independence of Irrelevant Alternatives property (IIA). Unfortunately, we believe that in these models, this assumption is not correct because some of the alternatives are more closely related than others. Hence, to solve it we propose using the so-called nested logit model where the individual decision is divided in two steps. In the first one, individuals maximize their utility by choosing to participate or not in the labour market while in the second one, they choose their ideal job type conditioned on participation.

We use data from the Spanish Survey on Disability, Personal Autonomy and Dependency Situations (2008). This survey is the most recent source of information about disability in Spain having detailed information about disabled people and their families.

The rest of the paper is organized as follows. We start with a presentation of our approach applying a random utility model, detailing the specification (based on supply side), the method and variables used. Then, we present the empirical results and we conclude summarizing the main findings, policy implications and possible directions for

[1] He proposed a discrete choice framework in which the agent's choice problem is based on the choice among feasible jobs and the distribution of desired working hours is discrete due to the choice opportunities distribution.

further research.

1.2. Methods

Let $U_n(G_i, tc_i, b_i)$ be the utility function for a household n, choosing the job-alternative i. Here $i = 1$ if there is no participation, $i = 2$ if working full-time, $i = 3$ if working part-time and $i = 4$ if working shift job (includes both long-term night shifts and work schedules in which employees change or rotate shifts). Besides, G is the disposable income corresponding to job i, tc is the time per day devoted to informal care (such as rehabilitation, doing exercise, tidiness...) for the disabled person in this alternative, and b_i are other characteristics conforming the job type. For given working hours, the budget constraint for a household is defined as:

$$G_i = f(wh_i, I) \tag{1.1}$$

where h_i is the working time associated with job i, w is the wage per hour and I is the non-labour income for disabled people, which also includes the wages of the rest of the household members and other income. $f(.)$ is a general function which describes the transformation of gross income into after-tax household income and s are some characteristics of the disabled person. We also know that time is limited so, $h_i + tc_i = T$

The omission of leisure time in our model is partly due to the difficulty in separating care time and leisure for disabled people, taking into account the large number of borderline cases. A similar point of view was assumed in [15], where it is pointed out the difficulty of separating leisure and child-care using as borderline case the time for playing with a child in the case of married women. Because we do not observe all variables affecting the preferences, we assume the utility function to be random, thus;

$$U_n(G_i, tc_i, b_i; s) = v_n(G_i, tc_i; s) + \epsilon(b_i). \tag{1.2}$$

Where $i \in J$ [J=$1, 2, 3, 4$] and $b_i \in B$ [B=b_1, b_2, b_3, b_4].

$v(.)$ is a positive deterministic function and $\epsilon(b_i)$ is a random component for unobserved characteristics that affect the utility assumed to be independent and identically extreme value distributed with a probability distribution function such as:

$$Pr(\epsilon(b_i) \leq x) = e^{-e^{-x}}. \tag{1.3}$$

We will assume that a household n makes an election because the job-alternative chosen, i , maximizes its utility given other characteristics, that is:

$$U_n(G_i, tc_i, b_i; s) > U_n(G_j, tc_j, b_j; s) \quad i, j \in J. \tag{1.4}$$

Due to preferences not completely known to us, the best we can do in simulating the behaviour of a household with a disabled person is to calculate the probability of choosing a job type category, given the characteristics and the available alternatives[2]. The probability that household n chooses job-type i, would be:

$$P(U_n|G_i, tc_i, b_i; s) = max_{j\in J}U_n(G_j, tc_j, b_j; s). \tag{1.5}$$

Given the assumed probability distribution for the error term, we will get the multinomial logit model.

$$P(U_n|G_i, tc_i, b_i; s)) = \frac{exp(v_n((G_i, tc_i, b_i; s))}{\sum_{j=i}^{J} exp(v_n((G_j, tc_j, b_j; s))}. \tag{1.6}$$

This model, which is the most commonly used assumes that the choice from a given set of alternatives satisfies the IIA. This property implies that random elements in utility are independent across alternatives and identically distributed. IIA property avoids including heterogeneity in preferences and correlation between alternatives. Those impositions may result too restrictive in our study, in which three of the alternatives are referred to labour participation and the other one not, so we are going to relax it. Thus, we construct a two level model with a degenerate branch; the upper level is a participation and non-participation nest, while the lower level is composed of the three kind of jobs corresponding to the participation nest. Then,

$$U_n(J = i|p, G_i, tc_i, b_i; s) = U_n(p|G_j, tc_j, b_j; s) + U_n(J = i|p, G_i, tc_i, b_i; s). \tag{1.7}$$

In a first step, households choose to belong to a nest m from two available (to participate in labour market, p or not, p_0). For evaluating this first choice, they compare $w*$, which is the shadow price of the time out of the market with market wage. This shadow time is non-observable to us but we observe that when $w^* > w$ there is no participation ($m' = p_0$). Consequentially, the household would choose alternative $i = 1$, but if $w^* \leq w$ then they participate ($m' = p$), then choose the alternative which maximizes the utility from $i = 2, 3, 4$. Then they have, as a starting point, two possible solutions: a corner ($i = 1 \in p_0$) and a three interior ones ($i = 2, 3, 4 \in p$).

In the nested model we need to add an additional parameter to the joint distribution of the error terms in each nest, λ_m. This parameter represents a measure of the correlation of the random components of the alternatives belonging to that nest. So that, the utilities are re-scaled by the inverse of the mentioned dissimilarity parameter, λ_m attached to an index variable called Inclusive Value (IV) which is defined by a set of utility expressions associated with a partitioned set of alternatives. That idea lies in the grouping of similar alternatives into nests and thus creating a hierarchical structure of those alternatives (Ben-Akiva y Lerman (1985) and Train (2003)). In our case, the

[2] For more details of this type of modeling see Dagsvik y Strom (2004)

IV for the nest implying participation, p (IV_p) is defined as:

$$IV_p = \lambda_p = Ln \sum_{j \in p} e^{v_{nj}/\lambda_p} = Ln[e^{(v_{n2}/\lambda_p)+exp(v_{n3}/\lambda_p)+exp(v_{n4}/\lambda_p)}]. \tag{1.8}$$

Then, the probability that a household chooses an alternative can be written as:

$$\begin{aligned} & P(J = i | i \in p, G_i, tc_i, b_i; s) = \\ = \quad & \frac{exp(v_{np}(f(wh_j, I), tc_j, b_j; s) + \frac{\lambda_p}{\lambda_j} \sum_{j=2}^{4} exp(v_{nj|p}(f(wh_j, I), tc_j, b_j; s)/\lambda_p}{\sum_{m' \in m} exp((v_{nm'}(f(wh_k, I), tc_k, b_k; s) + \frac{\lambda'_m}{\lambda_k} \sum_{j=2}^{4} exp(v_{nk|p}(f(wh_k, I), tc_k, b_k; s)/\lambda_{m'}} \\ \times \quad & \frac{exp(v_{ni|p}(f(wh_i, I), tc_i, b_i; s)/\lambda_p}{\sum_{j=2}^{4} exp(v_{nj|p}(f(wh_j, I), tc_j, b_j; s)/\lambda_p}. \end{aligned} \tag{1.9}$$

We must take into account that we have a degenerate branch. It means that the IV for the non participation is 1. If the property of IIA is satisfied, our model could be simplified to a multinomial logit.

1.2.1. Model Specification and Estimation Procedure.

The deterministic part of the utility function requires choosing a functional form, but this choice is not imposed directly by economic theory so it is usually determined by the data. The common approach in this context is to find a flexible family of parametric or semi-parametric specifications. Dagsvik y Strom (2006) demonstrate that under general regularity conditions, the systematic part of the household utility function can have a form, similar to;

$$v_n(G_i, tc_i; s) = \beta_G \frac{G_i^{\alpha_G} - 1}{\alpha_g} + \beta(s) \frac{tc_i^{\alpha_{tc}} - 1}{\alpha_{tc}}. \tag{1.10}$$

Where $\beta(s_n) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

The explicative variables included in our study are: x_1 = Age, $x_2 = 1$ if the person with disability is a man and x_3 = Limitation grade of the person with disability. We also assume that $v_n(G, t_c; s)$ is additive separable in income and care time and each utility component is supposed to have a Box-Cox functional form. In that sense, we take the advantage of having an specification which is globally concave. For the utility function to be quasi-concave, we require $\alpha_G < 1$ and $\alpha_{tc} < 1$. Note that if $\alpha_G \to 0$ and $\alpha_{tc} \to 0$, the utility function converges to a log-linear function. This model could be estimated sequentially but there is a considerable loss of information resulting in inefficient estimation, (see [18]). So, we use the full-information maximum likelihood which is efficient and its expression for a nested model is formalized as:

$$logL = \sum_{n=1}^{N} \sum_{i \in J} \delta_n(J = i) log[P(J = i/G_i, tc_i, s, \beta, \lambda]. \tag{1.11}$$

Where $n = 1, ..., N$ are observations, $i \in J$ are the available alternatives, λ is the parameter estimated for the inclusive value, s are the exogenous attributes and β their utility parameters. $\delta_n(J = i) = 1$ if the alternative i is chosen and zero otherwise.

1.2.2. Data

The available surveys on disability in Spain cover most of the needs for information of disability, dependency, ageing and the population health[3]. People with disability is defined as those who reported to have at least a deficiency; that does not always imply having limitations on the development of daily activities. According to the Survey on Disability, Personal Autonomy and Dependency Situations 2008, in Spain, of the 7.4 million old people (more than 65 years old), 2,227,500 declared a disability in 2008, representing the 30.3 %. The overall disability rate stands at 8.5 %, with an absolute value of 3,847,900 people with disabilities, of which 1,547,300 are men (40 %) and 2,300,500 women (60 %). It should be noted that the age range in which most people with disabilities fall are between 55 and 64 years (14.18 % of the total).

From the 14706 individuals with disability that answered the questionnaire (about disabilities) only 1632 were workers. Usually, the idea is that having a disability makes the individual not capable for any kind of work. This is totally unfounded and a true effort is necessary to match their capabilities with the labour market opportunities. [4]

[3] Three surveys have been carried out: the Survey on Disabilities, Deficiencies and Disabilities (EDDM1986), the Survey on Disabilities, Deficiencies and State of Health (EDDS1999) and the Survey on Disability, Personal Autonomy and Dependency Situations 2008.

[4] In Spain there is a law based on the integration of Disabled People. This law of Social Integration of disabled people 13/1982 of the 7th of April (LISMI) established that for private and public firms with more than 50 workers, it was compulsory to hire at least a 2 % of disabled people. Despite the existence of this regulations, there are very few companies that fulfill the obligation. According to the Academic Network of European Disability Experts, in Spain only 14 % of business larger than 50 workers meet the requirements in 2008

Table 1.1: **Main results of the Survey on Disability, Personal Autonomy and Dependency Situations 2008.**

Total Survey	
Total Sample. Households	96075
People who answered the questionnaire	22795
People with limitations who answered the questionnaire	14706
People with limitations who have a job	1632
People with limitations who work : How did you find your current job?	41.4 % Friends/Family; 22.9 % Firm recruitment; 2.62 % public service of employment; 1.15 % disability association
People with limitations. Are you looking for a job? (<65 years old)	10.16 % (610)
Why do you think you cant find a job?	43.65 % (of the 10.16 %) for the disability
Why don't you look for a job?	23.85 % think it is difficult because of the disability; 50 % can't work
Do you feel discrimination in your job due to your disability?	9.42 %
Do you feel discrimination when you are looking for a job?	20.8 %

Signif. codes: 0.01 '***'0.05 '**'0.1 '*'

Source: Own elaboration from the Survey on Disability, Personal Autonomy and Dependency Situations 2008.

1.3. Results

The maximum likelihood method is used to estimate the parameters of the nested logit model [5]. Tables 2 and 3 report the estimated parameters of the utility function according to the final specification of the model. For the nested logit model, we also report the estimate of λ_p and its standard deviation.

[5] We also estimate a multinomial logit in order to compare these two models

Table 1.2: **Estimated parameters of the utility function. Nested Logit.**

Variable	**Coefficient**	**Standard error**	$b/St.Er.$
Disposable Income	0.344***	0.080	4.29
α_G	-1.19		
Care Time	-2.409***	0.210	-11.421
α_{tc}	0.40		
Non-participation			
Age	0.089***	0.005	17.576
Gender	-0.268*	0.157	-1.702
Limitation	-0.827***	0.088	-9.408
Full-time job			
Age	0.035***	0.006	5.314
Gender	0.239	0.226	1.059
Limitation	0.417***	0.119	3.481
Part-time job			
Age	0.011	0.007	1.578
Gender	0.336	0.249	1.350
Limitation	0.241*	0.131	1.836
λ_p	0.669***	0.059	11.290
Number of Observations	19160		
R^2	0.287		

Signif. codes: 0.01 '***'0.05 '**'0.1 '*'
Source: Own elaboration from the Survey on Disability, Personal Autonomy and Dependency Situations 2008.

The parameters α_G and α_{tc} are estimated to yield a quasi-concave utility function, both of them being less than 1 (-1.19 and 0.40 respectively). The results show a positive sign of the disposable income as we expected, because more income in the household is associated with higher utility. Then, they would prefer jobs reporting more earnings if all other factors are similar.

We have tested other taste-modifying variables but we have selected those having more significant effects on preference for care time, and those were related with theoretical findings.

Care time has a negative sign, that is, as more time is devoted to care, less is the utility in the alternative chosen. This is related with the results obtained by Van den Berg et al.[20], who found an overall negative correlation between care hours and well-being (utility).[6] The positive sign for age shows a preference for care time when the age increases. Thus, people with disabilities prefer jobs which allow more care time when they get older[7].

Non-participation is preferred to shift-work (base category) for older people, women and more limited people. As the individuals are less limited, they would prefer jobs allowing shorter care time. People with more level of limitation would have more preference for care time and prefer jobs that could allow that dedication. We also provide an estimate of McFadden's goodness of fit measure, which indicated that the model fits quite well. Most of variables are individually significant and the model is globally significant too.

Table 1.3: **Discrete choice. Multinomial Logit.**

Variable	**Coefficient**	**Standard error**	*b/St.Er.*
Disposable Income	2.596***	0.403	6.442
Care Time	-2.521***	0.200	-12.557
Non-participation			
Age	0.095***	0.006	15.011
Gender	-0.192	0.208	-0.922
Limitation	-0.770***	0.112	-6.829
Full-time job			
Age	0.036***	0.006	5.914
Gender	0.258	0.209	1.238
Limitation	0.391***	0.110	3.546
Part-time job			
Age	0.006	0.007	0.925
Gender	0.381*	0.238	1.603
Limitation	0.246**	0.125	1.967
Number of Observations	4790		
R^2	0.270		

Signif. codes: 0.01 '***'0.05 '**'0.1 '*'
Source: Own elaboration from the Survey on Disability, Personal Autonomy and Dependency Situations 2008.

[6] That can be explained by the direct relation between that variable and the poor health status that reduce welfare. It allows for less time for social relations and maybe more stress, causing less utility for the household in every alternative chosen.

[7] This may be caused by the fact that disabilities are more frequent in advanced ages and care time is more valued.

Table 3 presents the results of multinomial logistic regression on job choices. As we observe, the results are similar to before. Working part or full time is preferred to working shifts for men, for those who have moderate or no limitation and also for those who are older. On the other hand, non-participation is preferred to shift-work if the individual is old, but is less preferred if the person has less limitations. Men also prefer to participate, even if it meant shift-work, rather than not participate.

We carried out the test developed by Hausman y McFadden (1984) to test the validity of the IIA assumption and we have to reject the null hypothesis of the IIA property. Then, the relevant model is not the multinomial logit model but the nested model. The degree of independence in unobserved utility among the upper nest is estimated to 0.67. Thus, our model satisfies the necessary condition in that it must be between 0 and 1. The parameter is not 1 and corroborates our hypotheses that the nested logit model is better than the multinomial logit one.

Table 1.4: **Utilities for each kind of job.**

	Non-partipation	**Full-time**	**Part-time**	**Shifts**
Disposable Income	5.77 (1.26)	5.25 (1.30)	5.22 (1.16)	5.27 (1.16)
Care Time	1.01 (0.56)	0.96 (0.59)	1.00 (0.60)	0.88 (0.60)
Age	54.85 (32.41)	45.76 (29.65)	46.82 (29.10)	45.70 (31.42)
Gender	0.58 (0.66)	0.70 (0.66)	0.75 (0.66)	0.68 (0.46)
Limitation	2.02 (1.42)	2.58 (1.69)	2.52 (1.73)	2.71 (1.82)

Signif. codes: 0.01 '***'0.05 '**'0.1 '*'
Source: Own elaboration from the Survey on Disability, Personal Autonomy and Dependency Situations 2008.

Table 4 show a pondered average of the parameters of the utility function. Each attribute is weighted by the percentage of choices of that alternative and then all values from each alternative are summed up to obtain the average. People with disability who choose shift-work are on average, the youngest. That result is logical with our intuition because this kind of job is less flexible and requires more effort than the others. We consider that the results referred to care time are reasonable. People who devote more time to care chose non-participation more frequently. This alternative is followed by part-time jobs which also allow more time out of the labour market. We argue that conciliation between personal and professional life is more difficult for people undertaking shift-work or full-time jobs, and that is reflected in the data (people who choose to work shifts have the lowest figure of care time, 0.88). On average, people with more limitations (2.02) choose the non-participation alternative and contrarily, people working shifts are those having the minimum level (2.71) (remember the level 3 implies the less limitative).

For each household we predict the job status for the person with disability in order to be the one having the highest estimated probability, and we calculate the numbers of households for which the predicted status is equal to the actual status. In that sense, the results obtained about the prediction power of the model are encouraging.

Table 1.5: **Job-specific elasticities.**

Type of job	**Total effect (increase in income)**	**Total effect (increase in care time)**
Non- participation	0.699	-0.774
Full-time	-0.752	0.968
Part-time	-0.823	0.657
Shifts	-0.900	0.428
Non- participation	-0.318	0,418
Full-time	0.780	-0.953
Part-time	-0.827	0.379
Shifts	-0.699	0.482
Non- participation	-0.091	0.072
Full-time	-0.217	0.253
Part-time	1.549	-1.480
Shifts	-0.056	0.149
Non- participation	-0.045	0.021
Full-time	-0.077	0.055
Part-time	-0.154	0.065
Shifts	1.569	-0.860

Signif. codes: 0.01 '***'0.05 '**'0.1 '*'
Source: Own elaboration from the Survey on Disability, Personal Autonomy and Dependency Situations 2008.

Our estimates suggest that policy makers have several tools to affect these choice variables, such as social policies and firms grants, to improve labour supply of people with disabilities changing their decisions. We note that when the income level related with non-participation increases by 1 % the decision probability of this alternative increases about 0.7 % . The higher increase in the probability to choose an alternative associated with an income increment is the referent to shift-work. When the income associated with this alternative is raised by 1 %, then the probability of choice for that type of job increases by 1.56 %.

1.4. Discussion and Conclusion

We have estimated a job-type model on Spanish data for a sample of households where a person with disability lives. Our approach differs from previous research because as far as we know, it is the first time that a model of labour supply for disabled people is estimated taking into account household's choices, regarding job-types as a set enclosing several factors. Besides, the performance gains of our approach over previous studies is based on disabled people need more time for health care than people without disability and can affect their decisions. We suppose that a person with disability can require support or assistance for basic living activities and it varies depending on the age, individual needs, lifestyle and other circumstances which sometimes detract them from participating in the labour market. Consistent with this objective, we have carried out a discrete choice labour supply model in which conditioning on their participation in the labour market, they face a set of choices of their ”job-alternatives”. Therefore, a choice between participating or not is presented and afterwards, another between working in a full-time, part-time job or to work shifts.

By learning from the past and looking into the future, our approach has allowed us to learn about one important feature: an overall income increase seems to enhance the probability of working shifts and shows that non participation is the least flexible choice. Care time is more valuable with age and limitation-level, affecting the choice decisions about labour activities. This should not be very surprising in the model given that individuals are rational and have to choose between a set of job-packages that may not be as flexible as they want. Thus, policy makers should aim to try and connect older people and those with limitations with more flexible jobs which respect their care needs.

This paper clearly shows that a generalist policy against some types of jobs which could be undesirable in the overall population. People with disability have special needs and in most cases require more flexible jobs. Not only but offering telecommuting may be a good option for disabled people to develop their capabilities and competencies. Hence, it is necessary to promote policies for support in addressing the accommodation needs for disabled people, and consequently promoting their wellbeing. : In this sense, specific firms management and job placement agencies could help to overcome the existent barriers, being more flexible to establish a reasonable schedule for disabled people.

To sum up, further work is needed to evaluate health-care for people with disability and promote their hiring in more flexible jobs. All these issues requires more research in the future but this paper is a good start to undertake it.

Capítulo 2

Capítulo 2

Semiparametric estimation of a sample selection model with a binary endogenous regressor: The effect of chronicity in labor supply

The aim of this paper is to investigate the effect of chronic illness in labor supply and participation rates. One possible approach to analyze this impact is to specify a simultaneous equation model where a dummy endogenous variable that accounts for chronicity is introduced both in the endogeneity and in the sample selection equation. Traditionally, identification and estimation of these type of models requires the introduction of strong restrictions that are not demanded by Economic Theory. If some of these restrictions are wrong the resulting estimators might exhibit rather poor theoretical properties. In this paper we propose a root-N consistent estimator of the structural parameters of the so-called wage equation that is robust to several sources of misspecification. In fact the endogeneity and the sample selection mechanism are assumed to be nonparametric functions and furthermore, the conditional distribution of the errors is also left unspecified and only the existence of certain high order moments of this distribution is needed. Under this fairly weak assumptions, and using a control function approach, the resulting estimator of the parameters of interest is obtained through a simple pairwise differencing transformation. The asymptotic properties of the estimator are established and its empirical performance is studied using the Spanish Living Conditions Survey.

2.1. Introduction

In Spain more than 19 million of people are affected by some chronic illness, most of them older than 55 years old and requiring more than the 75 % of the health expenditure. The public budget for chronicity purposes was in 2011 of around $50,000$ millions of euros and moreover, it is expected to be a 45 % higher in 2020 due to demographical changes and also increment of people with chronic illnesses. The raise of chronic illnesses in last years makes necessary the multidisciplinary study of its consequences. The possible treatments and prevention systems, the quality of life of chronic ills, their social insertion are some of the more important questions to deal with. So that, chronicity is an important issue which is going to be more relevant next decades and collecting information about the impact of chronicity in our society will be necessary to manage this situation as well as possible in both the health and economical context.

In academics, the effect of chronic illness in labor supply and participation rates has attracted the attention of some researchers (see among others Zhang et˜al. (2009), Langley et˜al. (2011) and Boot et˜al. (2014)). Usually, the theoretical framework of analysis of this type of problems has been a standard labor supply model as it was originally proposed in (Heckman, 1974). However, the introduction of chronicity creates problems in the specification and the estimation of these models. The reason is twofold. First, chronicity affects simultaneously labor supply and participation rates, but at the same time, the chronic illness can be affected by a long period of non-participation in the labor market.

If we want to consider the previous issues in our analysis, it is necessary to specify a three equation labor supply model (wage equation, participation and endogeneity) where a common dummy variable (the presence of a chronic disease) is present in both wage and sample selection equation and furthermore, it is determined endogenously in the system. Dummy endogenous regressors in limited dependent variable models have been of large interest (see among others (Heckman, 1978), (Angrist y Krueger, 2001) and (Vytlacil y Yildiz, 2007)). In a general framework of simultaneous equations models with censored endogenous regressors, where the censored equation contains dummy endogenous variables, Vella (1993) introduces a simple estimator using a control function approach and generalized residuals. This approach is based on the estimation of nonlinear models with selectivity or endogenous regressors and then treat them as partially linear regression models with the correction terms appearing in the non linear component of the regression. Those correction terms can be built as conditional expectations (such as the so-called propensity score) or first-stage residuals or sometimes a mixture of both.

Furthermore, sample selection models have been studied intensively with many variations (see Vella (1998) for a survey), however, the situation where the selection equation and the censored equation contains a common endogenous dummy variable has been considerably less studied. In Kim (2006) this model is interpreted as an endogenous switching type II- Tobit model and a simple three step estimation procedure is developed. The model analyzed in the above paper is fully parametric and joint normality is also assumed. Relying on these assumptions it is proposed to estimate the parameters of interest either using a standard maximum likelihood approach or a simple three stage estimation procedure based in the well known Heckman's approach (see Heckman (1979)). The main problem of these estimation techniques is that rely on a fully parametric specification and therefore, the risk of misspecification is high. For example, in the two stage estimation procedure, under possibly non-normal disturbances, probit maximum likelihood estimators are inconsistent estimators and therefore, second stage estimators are also asymptotically biased. For the sample selection model, under nonparametric specification of the selection equation and unknown form of the distribution of the errors, in (Ahn y Powell, 1993) it is 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) it is proposed a nonparametric estimator of the structural model of labor supply. None of these contributions unfortunately allow for binary endogenous regressors and therefore they are of limited interest for our problem.

In this paper, we consider a root -N consistent estimator of the structural parameters of a sample selection model with a common dummy endogenous regressor. The so-called wage equation is specified as a linear function of the parameters of interest and both the selection equation and the endogenous mechanism are assumed to be unknown nonparametric functionals. Furthermore, estimate the parameters of the wage equation we do not need to pre-specify a joint distribution function for the vector of heterogeneity terms. We propose estimating our model in two stages. The first stage consists in the nonparametric estimation estimation of both the endogenous mechanism and the sample selection equation. The obtained estimates can then be used to evaluate the conditional probability of an individual to suffer a chronic illness and the conditional probability of an individual to participate in the labour market given he suffers, or not, a chronic illness. In the last stage, the parameters of interest in the wage equation are estimated through a pairwise differencing transformation (see Aradillas-Lopez et˜al. (2007)) that removes the (unknown) function of the conditional probabilities. We obtain the asymptotic distribution of the estimator through the use of standard statistical tools (see Ahn y Powell (1993) and Powell (2001)). Finally, we analyze empirically the performance of the estimator using data from the Spanish Living Conditions Survey. The results are compared against those obtained by using standard methods and it turns out that they are rather different.

Our paper is organized as follows. In Section 2, we describe the structural model and we introduce the estimation procedure. In Section 3 we present the main asymptotic properties of the estimators proposed in the previous section. Section 4 analyzes the impact of chronicity in labor supply using data from the Spanish Labor Force Survey. In Section 5, we present the conclusion. The proof of the main results is relegated to the Appendix.

2.2. Model and estimation procedure

2.2.1. Econometric model

We start by considering a standard model of labor supply see ((Fernandez et˜al., 2001)) where we define the so-called wage equation as:

$$r_i = \begin{cases} r_i^* & for \quad z_i = 1, \\ 0 & otherwise. \end{cases} \tag{2.1}$$

$$z_i = 1\left(z_i^* > 0\right), \tag{2.2}$$

where

$$\begin{aligned} r_i^* &= w_{1i}^\top \beta + \lambda d_i + \eta_{1i}, \qquad (2.3) \\ z_i^* &= g_z(w_{1i}, w_{2i}, d_i) - \eta_{2i} \qquad (2.4) \end{aligned}$$

for $i = 1, \cdots, N$. Equation (2.1) stands for the so-called wage equation and equation (2.2) is the sample selection mechanism. The random variable d is a dummy variable that stands for the chronic illness. That is, $d_i = 1$ whether the i-th individual suffers from chronic illness and 0 otherwise. The k_1-vector of parameters β and the scalar λ are unknown objects that need to be estimated and $g_z\left(\cdot\right)$ is an unknown nonparametric function. Chronicity is explained through the following mechanism,

$$d_i = 1\left(g_d(w_{1i}, w_{2i}, w_{3i}) - \eta_{3i} > 0\right), \tag{2.5}$$

for $i = 1, \cdots, N$. $g_d(\cdot)$ is an unknown nonparametric function. Let $w_i = (w_{1i}, w_{2i}, w_{3i})$ be a $(k_1 + k_2 + k_3)$-vector of explanatory variables such as the age, the gender or the education. $\eta_i = (\eta_{1i}, \eta_{2i}, \eta_{3i})$ is the vector of idiosyncratic error terms. This model can be generalized using a switching model. For this, we extend the models (2.1) and (2.2) into two regimes respectively determined by the state of d_i, i.e.

$$\begin{aligned} r_{1i}^* &= w_{1i}^\top \beta + \lambda + \eta_{11i}, \qquad (2.6) \\ r_{0i}^* &= w_{1i}^\top \beta + \eta_{10i}, \qquad (2.7) \\ z_{1i}^* &= g_z(w_{1i}, w_{2i}, 1) - \eta_{21i}, \qquad (2.8) \\ z_{0i}^* &= g_z(w_{1i}, w_{2i}, 0) - \eta_{20i}. \qquad (2.9) \end{aligned}$$

and define accordingly

$$\begin{aligned} r_i^* &= d_i r_{1i}^* + (1 - d_i) r_{0i}^*, \qquad (2.10) \\ z_i^* &= d_i z_{1i}^* + (1 - d_i) z_{0i}^* \qquad (2.11) \end{aligned}$$

and $r_{1i} = 1\left(r^*_{1i} > 0\right)$, $r_{0i} = 1\left(r^*_i > 0i\right)$, $z_{1i} = 1\left(z^*_{1i} > 0\right)$, $z_{0i} = 1\left(z^*_i > 0i\right)$. Note that our aim is to estimate the parameter set $(\beta,\ \lambda)$ in the presence of some unknown objects $g_z(\cdot)$ and $g_d(\cdot)$, and without assuming any prespecified distribution for $(\eta_{10i}, \eta_{11i}, \eta_{20i}, \eta_{21i}, \eta_{3i})$.

2.2.2. Estimation techniques

Following the same ideas as in Heckman (1979) and Vella (1993) the parameters of interest, β and λ can be estimated through a three-stage procedure. In order to do so, here we derive the correction term which comprises two parts. One corrects the bias due to the endogenous switching and the other part corrects the sample selection bias for each state depending on d. To define these correction terms note that substituting (2.6)-(2.11) into (2.1)-(2.5) we obtain

$$\begin{aligned} r^*_i &= w^\top_{1i}\beta + \lambda d_i + d_i\eta_{11i} + (1-d_i)\eta_{10i}, \\ z^*_i &= g_z\left(w_{1i}, w_{2i}, 0\right) + \Delta_i d_i - d_i\eta_{21i} - (1-d_i)\eta_{20i}, \end{aligned} \tag{2.12}$$

and

$$\Delta_i = g_z\left(w_{1i}, w_{2i}, 1\right) - g_z\left(w_{1i}, w_{2i}, 0\right).$$

Now, taking conditional expectations on (2.1) and using (2.12) we obtain

$$\begin{aligned} E\left(r_i | w_i, d_i, z^*_i > 0\right) &= w^\top_{1i}\beta + \lambda d_i + d_i E\left(\eta_{11i} | w_i, z^*_i > 0, d^*_i > 0\right) \\ &\quad + (1-d_i)\, E\left(\eta_{10i} | w_i, z^*_i > 0, d^*_i \leq 0\right). \end{aligned} \tag{2.13}$$

The previous expression suggests to obtain a consistent estimator of the parameters of interest by calculating a (extended) least squares regression of r_i on d_i, w_{1i} and the functions $E\left(\eta_{11i} | w_i, z^*_i > 0, d^*_i > 0\right)$ and $E\left(\eta_{10i} | w_i, z^*_i > 0, d^*_i \leq 0\right)$. Under some assumptions, i.e.

1. The functions $g_z\left(\cdot\right)$ and $g_d\left(\cdot\right)$ are constrained to be linear and depending on a finite vector of parameters.

2. The vector of error terms $(\eta_{10i}, \eta_{11i}, \eta_{21i}, \eta_{20i}, \eta_{3i})$ is multivariate normally distributed with zero mean and homoskedastic variance-covariance matrix.

Adding some convenient restrictions in the variance-covariance matrix, the quantities $E\left(\eta_{11i} | w_i, z^*_i > 0, d^*_i > 0\right)$ and $E\left(\eta_{10i} | w_i, z^*_i > 0, d^*_i \leq 0\right)$ are known up to certain constants and therefore it can be shown that the ordinary least squares estimators of β and λ are consistent and asymptotically normal (see Kim (2006) for details).

If we are not willing to assume none of the previous assumptions the quantities $E\left(\eta_{11i} | w_i, z^*_i > 0, d^*_i > 0\right)$ and $E\left(\eta_{10i} | w_i, z^*_i > 0, d^*_i \leq 0\right)$ remain unknown and the least squares estimation of the parameters of interest is unfortunately unfeasible. We will try to solve this problem by following the same ideas as in (Das et˜al., 2003). To this end, let $w = (w_1, w_2, w_3)$ and $w_{12} = (w_1, w_2)$. Define the following propensity scores,

$$p_3 = E\left(d\,|w\right), \quad p_{21} = E\left(z\,|w_{12}, d = 1\right), \quad p_{20} = E\left(z\,|w_{12}, d = 0\right), \tag{2.14}$$

and assume the random variables $(\eta_{10}, \eta_{11}, \eta_{20}, \eta_{21}, \eta_3)$ are independent of the vector w. Then, for the CDF's of η_{20}, η_{21} and η_3 respectively $G_{\eta_2|d=1}(\cdot)$, $G_{\eta_2|d=0}(\cdot)$ and $G_{\eta_3}(\cdot)$, we have $p_3 = G_{\eta_3}(g_d(w))$ and

$$p_{21} = G_{\eta_2|d=1}\left(g_z\left(w_1, w_2, 1\right)\right), \quad p_{20} = G_{\eta_2|d=0}\left(g_z\left(w_1, w_2, 0\right)\right),$$

for $i = 1, \cdots, N$. Also, assuming that the CDF's are one-to-one, for $u_3 = G_{\eta_3}(\eta_3)$, $u_{21} = G_{\eta_2|d=1}(\eta_{21})$ and $u_{20} = G_{\eta_2|d=0}(\eta_{20})$ we can write

$$\begin{aligned} & E\left(\eta_{11}|w, z^* > 0, d^* > 0\right) \\ & = E\left(\eta_{11}|w, \eta_{21} < g_z(w_1, w_2, 1), \eta_3 < g_d(w)\right) \\ & = E\left(\eta_{11}|w, u_{21} < p_{21}, u_3 < p_3\right) \end{aligned} \tag{2.15}$$

$$= \int \int^{p_{21}} \int^{p_3} \frac{\eta_{11} f\left(\eta_{11}, u_{21}, u_3|w\right) d\eta_{11} du_{21} du_3}{\int^{p_{21}} \int^{p_3} f\left(u_{21}, u_3|w\right) du_{21} du_3} = \mu_1\left(p_{21}, p_3\right).$$

Proceeding in the same way

$$\begin{aligned} & E\left(\eta_{10}|w, z^* > 0, d^* < 0\right) \\ & = E\left(\eta_{10}|w, \eta_{20} < g_2(w_1, w_2, 0), \eta_3 > g_3(w)\right) \\ & = E\left(\eta_{10}|w, u_{20} < p_{20}, u_3 > p_3\right) \end{aligned} \tag{2.16}$$

$$= \int \int^{p_{20}} \int_{p_3} \frac{\eta_{10} f\left(\eta_{10}, u_{20}, u_3|w\right) d\eta_{10} du_{20} du_3}{\int^{p_{20}} \int_{p_3} f\left(u_{20}, u_3|w\right) du_{20} du_3} = \mu_0\left(p_{20}, p_3\right).$$

Hence we can rewrite (2.13) as

$$\begin{aligned} E\left(r_i|w_i, d_i, z_i^* > 0\right) &= w_{1i}^\top \beta + \lambda d_i + \mu_1\left(p_{21i}, p_{3i}\right) d_i \\ & +\mu_0\left(p_{20i}, p_{3i}\right)\left(1 - d_i\right). \end{aligned} \tag{2.17}$$

Note that $\mu_0(p_{20}, p_3)$ and $\mu_1(p_{21}, p_3)$ are unknown quantities that depend only on the respective propensity scores. However, it is interesting to realize that, following Powell (2001), if for some $i \neq j$, $p_{21i} = p_{21j}$, $p_{20i} = p_{20j}$ and $p_{3i} = p_{3j}$ then $\mu_1(p_{21i}, p_{3i}) = \mu_1(p_{21j}, p_{3j})$ and $\mu_0(p_{20i}, p_{3i}) = \mu_0(p_{20j}, p_{3j})$.

Thus, using a pairwise transformation, for $i \neq j$

$$\begin{aligned} r_i - r_j &= (w_{1i} - w_{1j})^\top \beta + \lambda(d_i - d_j) + \mu_1(p_{21i}, p_{3i})d_i - \mu_1(p_{21j}, p_{3j})d_j \\ & +\mu_0(p_{20i}, p_{3i})(1 - d_i) - \mu_0(p_{20j}, p_{3j})(1 - d_j) + (v_i - v_j), \end{aligned}$$

where

$$v_i = r_i - E\left(r_i | w_i, d_i, z_i^* > 0\right), \quad v_j = r_j - E\left(r_j | w_j, d_j, z_j^* > 0\right),$$

and for the pairs (i, j) that fulfill

(a) $\mu_1(p_{21i}, p_{3i}) = \mu_1(p_{21j}, p_{3j})$, for $i \neq j$ and $d_i = d_j$,

(b) $\mu_0(p_{20i}, p_{3i}) = \mu_0(p_{20j}, p_{3j})$ for $i \neq j$ and $d_i = d_j$,

we obtain

$$r_i - r_j \quad = \quad (w_{1i} - w_{1j})^\top \beta + v_{1i} - v_{1j}, \quad i \neq j. \tag{2.18}$$

Clearly, the least squares estimator of β in (2.18) is consistent using only those observations that fulfill (a) and (b). A feasible version of this idea uses all pairs of observations and a weighted least squares estimator of β that gives larger weights to pairs for which $p_{21i} \approx p_{21j}$, $p_{20i} \approx p_{20j}$ and $p_{3i} \approx p_{3j}$, i.e.

$$\widetilde{\beta} = S_{ww}^{-1} S_{wr} \tag{2.19}$$

where

$$\begin{aligned} S_{ww} &= \binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} (w_{1i} - w_{1j})(w_{1i} - w_{1j})^\top \\ S_{wr} &= \binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} (w_{1i} - w_{1j})(r_i - r_j). \end{aligned} \tag{2.20}$$

We propose the following weights

$$\begin{aligned} \omega_{ij} &= \frac{1}{h_2^2} K\left(\frac{p_{1i} - p_{1j}}{h}\right) z_i z_j d_i d_j \\ &\quad + \frac{1}{h_2^2} K\left(\frac{p_{0i} - p_{0j}}{h}\right) z_i z_j (1 - d_i)(1 - d_j), \end{aligned} \tag{2.21}$$

$$\begin{aligned} p_{1i} - p_{1j} &= \begin{pmatrix} p_{21i} - p_{21j} & p_{3i} - p_{3j} \end{pmatrix}^\top \\ p_{0i} - p_{0j} &= \begin{pmatrix} p_{20i} - p_{20j} & p_{3i} - p_{3j} \end{pmatrix}^\top, \end{aligned}$$

$K(\cdot)$ is a bivariate kernel function, h is the bandwidth. Note that $\widetilde{\beta}$ is unfeasible because it is a function of the propensity scores that are unknown functions that need to be estimated. As estimators for these quantities, given (2.5), (2.12) and (2.14), we propose

$$\widehat{p}_{21}(w_{12}) \equiv \frac{\widehat{t}_1(w_{12})}{\widehat{b}_1(w_{12})} = \frac{\frac{1}{N h_1^{k_1+k_2}} \sum_{j=1}^N K_1\left(\frac{w_{12} - w_{12j}}{h_1}\right) d_j z_j}{\frac{1}{N h_1^{k_1+k_2}} \sum_{j=1}^N K_1\left(\frac{w_{12} - w_{12j}}{h_1}\right) d_j}, \tag{2.22}$$

$$\widehat{p}_{20}\left(w_{12}\right) \equiv \frac{\widehat{t}_0(w_{12})}{\widehat{b}_0(w_{12})} = \frac{\frac{1}{Nh_1^{k_1+k_2}}\sum_{j=1}^{N} K_1\left(\frac{w_{12}-w_{12j}}{h_1}\right)\left(1-d_j\right)z_j}{\frac{1}{Nh_1^{k_1+k_2}}\sum_{j=1}^{N} K_1\left(\frac{w_{12}-w_{12j}}{h_1}\right)\left(1-d_j\right)}, \tag{2.23}$$

$$\widehat{p}_{3}\left(w\right) \equiv \frac{\widehat{t}_3(w)}{\widehat{b}_3(w)} = \frac{\frac{1}{Nh_1^{k_1+k_2+k_3}}\sum_{j=1}^{N} K_1\left(\frac{w-w_j}{h_1}\right)d_j}{\frac{1}{Nh_1^{k_1+k_2+k_3}}\sum_{j=1}^{N} K_1\left(\frac{w-w_j}{h_1}\right)}. \tag{2.24}$$

Replacing the unknown quantities by their corresponding estimators we propose the following two stage estimator

$$\widehat{\beta} = \widehat{S}_{ww}^{-1}\widehat{S}_{wr}, \tag{2.25}$$

where

$$\widehat{S}_{ww} = \binom{N}{2}^{-1}\sum_i\sum_{i<j}\widehat{\omega}_{ij}(w_{1i}-w_{1j})(w_{1i}-w_{1j})^{\top} \tag{2.26}$$

$$\widehat{S}_{wr} = \binom{N}{2}^{-1}\sum_i\sum_{i<j}\widehat{\omega}_{ij}(w_{1i}-w_{1j})(r_i-r_j),$$

and

$$\begin{aligned}\widehat{\omega}_{ij} &= \frac{1}{h^2}K\left(\frac{\widehat{p}_{1i}-\widehat{p}_{1j}}{h}\right)z_iz_jd_id_j\\ &\quad+\frac{1}{h^2}K\left(\frac{\widehat{p}_{0i}-\widehat{p}_{0j}}{h}\right)z_iz_j(1-d_i)(1-d_j),\end{aligned} \tag{2.27}$$

Note that the difference between $\widetilde{\beta}$ and $\widehat{\beta}$ are the propensity scores. In (2.25) they are assumed to be unknown and replaced by consistent estimators of these quantities given in (2.22)-(2.24). To our knowledge, our approach represents the first application of pairwise-differencing methods to solve simultaneously sample selection and the presence of the dummy endogenous regressor in both the main equation and the selection mechanism with nonparametric control functions. The advantage is that even if the control functions are completely unknown, root-n consistent estimation is possible.

There is also an important disadvantage to discuss here, is the fact that the pairwise difference procedure wipe out one of the parameters of interest, λ. In some cases the knowledge of this quantity might be of interest. For example, in this paper we might be interested in analyzing the effect of having a chronic illness in expected wages. This is not exceptional, in the semiparametric literature is something usual that the nonparametric correction function absorbs the intercept. In some cases, it may be possible to use a subgroup to identify the intercept ((Andrews y Schafgans, 1998)) or estimating counterfactuals and then using matching techniques ((Rosenbaum y Rubin, 1983)). Following this last approach we propose here an estimator for λ. Using (2.1)-(2.2), (2.6)-(2.11) note that λ can be obtained as

$$\lambda = E\left[r_{1i}|w_{1i}, d_i^* > 0, z_i^* > 0\right] - E\left[r_{0i}|w_{1i}, d_i^* < 0, z_i^* > 0\right]. \tag{2.28}$$

The main problem arises because we would like to know the difference between the outcome with and without chronic illness. Obviously, we cannot observe both results for the same individual simultaneously. The unobserved outcome is usually called counterfactual. For solving this kind of problem one may use the matching approach for trying to approximate the research to an experiment. So that, the idea is to look for an individual without chronic illness, j, which is similar to an individual who is ill, i, in all the other relevant characteristics w. By this way, finally, differences in outcomes between both groups can be attributed to the fact of having a chronic illness. This approach assumes that, conditionally on a set of confounders, $w_i, z_i = 1$,

$$d_i \perp (r_{1i}, r_{0i}) | w_i, z_i = 1. \tag{2.29}$$

or equivalently

$$d_i \perp (r_{1i}, r_{0i}) \, |p_{3i}, p_{21i}, p_{20i} \, . \tag{2.30}$$

Applying this assumption to (2.28), for all matched pairs, i,

$$\lambda = E\left(r_{1i} \, |p_{3i}, p_{21i}\right) - E\left(r_{0i} \, |p_{3i}, p_{20i}\right). \tag{2.31}$$

Both outcomes are non observable for the same individual. Then, we are only going to have the wage in one of the cases, i.e. r_{1i}, if $d_i = 1$, or r_{0i}, if $d_i = 0$. The case r_{0i}, if $d_i = 1$, is not observed. The way to solve this problem is to estimate the unobserved outcome by using the outcome of individuals from the control group with nearest propensities in order to estimate the counterfactuals.

Proceeding as for the estimation of β, we look for individuals (i, j) with $p_{21i} \approx p_{21j}$, $p_{20i} \approx p_{20j}$ and $p_{3i} \approx p_{3j}$, but contrary to before, we match them with $d_i \neq d_j$, that means that they are similar in all characteristics except that one receives the treatment $(d_i = 1)$ and the other not $(d_j = 0)$. Expectations are replaced by sample means, and we condition on p_{21i},p_{20i}, p_{3i} by matching each unit i to a set of comparable units, j on the basis of their predicted probabilities. So that, the purpose is to associate to the outcome r_{1i} of a treated unit i a matched outcome given by a kernel-weighted average of the outcome of all non-treated units, where the weight given to non-treated unit j is in proportion to the closeness between i and j. Once matches are made, we can calculate the impact of the binary regressor by comparing the means of outcomes across participants, i and controls, j.

That is,

$$\begin{aligned} \widehat{\lambda} &= (N_1/N)\widehat{\lambda}_1 + (N_0/N)\widehat{\lambda}_0, && (2.32) \\ \widehat{\lambda}_1 &= \frac{1}{N_1} \sum_{i=1}^{N} \left(r_i - \sum_{j=1}^{N} \widehat{\omega}_{0ij} r_j \left(1 - d_j\right) \right) d_i, && (2.33) \\ \widehat{\lambda}_0 &= \frac{1}{N_0} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} \widehat{\omega}_{1ij} r_j d_j - r_i \right) \left(1 - d_i\right), && (2.34) \end{aligned}$$

where

$$\widehat{\omega}_{0ij} = \frac{1}{Ng^2} K_3 \left(\frac{p_{0i} - p_{0j}}{g} \right), \tag{2.35}$$

$$\widehat{\omega}_{1ij} = \frac{1}{Ng^2} K_3 \left(\frac{p_{1i} - p_{1j}}{g} \right), \tag{2.36}$$

and $N_1 = \sum_{i=1}^{N} d_i$ and $N_0 = N - N_1$.

2.3. Large sample properties of the estimator

In this section we establish some asymptotic properties for $\widehat{\beta}$. This estimator was already defined in (2.25). The proof of these results basically follow the same lines as in Ahn y Powell (1993), Powell (2001) and Aradillas-Lopez et˜al. (2007). The derivation of the large sample properties of $\widehat{\beta}$ proceeds in two logical steps. First the asymptotic behavior of the related estimator $\widetilde{\beta}$ defined in (2.19) is analyzed. Then, in a second step, the difference between $\widetilde{\beta}$ and $\widehat{\beta}$ is analyzed. The proofs of all results are relegated to the Appendix.

The expressions for $\widetilde{\beta}$ and $\widehat{\beta}$ can be rewritten as,

$$\sqrt{N} \left(\widetilde{\beta} - \beta \right) = N^{1/2} S_{ww}^{-1} S_{wu}, \tag{2.37}$$

where S_{ww} is defined in (2.20) and

$$\begin{aligned} S_{wu} &= \binom{n}{2}^{-1} \sum_{i} \sum_{i<j} K_h \left(\frac{p_{1i} - p_{1j}}{h} \right) d_i d_j z_i z_j (w_{1i} - w_{1j}) \\ &\quad \times (u_{1i} - u_{1j}) && (2.38) \\ &+ \binom{n}{2}^{-1} \sum_{i} \sum_{i<j} K_h \left(\frac{p_{0i} - p_{0j}}{h} \right) (1 - d_i)(1 - d_j) z_i z_j (w_{1i} - w_{1j}) \\ &\quad \times (u_{0i} - u_{0j}), && (2.39) \end{aligned}$$

where now $K_h(\cdot) = \frac{1}{h} K(\cdot)$ and

$$\begin{aligned} u_{1i} &\equiv \mu_1(p_{21i}, p_{3i}) + v_i, \\ u_{0i} &\equiv \mu_0(p_{20i}, p_{3i}) + v_i. \end{aligned}$$

It is also clear that the estimator $\widehat{\beta}$ defined in (2.25) satisfies

$$\sqrt{N} \left(\widehat{\beta} - \beta \right) = N^{1/2} \widehat{S}_{ww}^{-1} \widehat{S}_{wu}, \tag{2.40}$$

where $\widehat{S}_{ww}$ is defined in (2.26) and

$$\widehat{S}_{wu} = \binom{n}{2}^{-1} \sum_{i} \sum_{i<j} K_h \left(\frac{\widehat{p}_{1i} - \widehat{p}_{1j}}{h} \right) d_i d_j z_i z_j (w_{1i} - w_{1j})$$

$$
\begin{aligned}
& \times (u_{1i} - u_{1j}) && (2.41) \\
+ \quad & \binom{n}{2}^{-1} \sum_i \sum_{i<j} K_h \left(\frac{\widehat{p}_{0i} - \widehat{p}_{0j}}{h} \right) (1-d_i)(1-d_j) z_i z_j (w_{1i} - w_{1j}) \\
& \times (u_{0i} - u_{0j}) \,. && (2.42)
\end{aligned}
$$

In order to show the results for $\widetilde{\beta}$ we need the following regularity conditions:

(A.1) The vectors $\left(r_i, z_i, d_i, w_i^\top\right)^\top$ satisfying (2.1), (2.2) and (2.5) are independent and identically distributed across i having bounded support and the remaining components having finite sixth-order moments.

(A.2) The propensity scores p_3, p_{21} and p_{20} have joint and marginal densities, $f(\cdot)$, that are absolutely continuous and bounded from above.

(A.3) Let $\Sigma_{ww} = \Sigma_{ww1} + \Sigma_{ww0}$, where

$$
\begin{aligned}
\Sigma_{ww1} \;=\; & E\Big[\rho_{1i}^2 \gamma_i^2 f(p_{21i}, p_{3i})(w_{1i} - \phi_1(p_{21i}, p_{3i})) \\
& \times (w_{1i} - \phi_1(p_{21i}, p_{3i}))^\top\Big] && (2.43)
\end{aligned}
$$

and

$$
\begin{aligned}
\Sigma_{ww0} \;=\; & E\Big[\rho_{0i}^2 \gamma_i^2 f(p_{20i}, p_{3i})(w_{1i} - \phi_0(p_{20i}, p_{3i})) \\
& \times (w_{1i} - \phi_0(p_{20i}, p_{3i}))^\top\Big] \,. && (2.44)
\end{aligned}
$$

Moreover,

$$
\rho_{1i} = E\left[z_i | p_{21i}, d_i = 1\right], \quad \rho_{0i} = E\left[z_i | p_{20i}, d_i = 0\right],
$$

$$
\gamma_i = E\left[d_i | p_{3i}\right],
$$

and

$$
\begin{aligned}
\phi_1(p_{11i}, p_{3i}) &= \frac{E\left[d_i z_i w_{1i} | p_{21i}, p_{3i}\right]}{\rho_{1i}\gamma_i}, && (2.45) \\
\phi_0(p_{20i}, p_{3i}) &= \frac{E\left[(1-d_i) z_i w_{1i} | p_{20i}, p_{3i}\right]}{\rho_{0i}\gamma_i}. && (2.46)
\end{aligned}
$$

We assume that Σ_{ww0} and Σ_{ww1} are nonsingular.

(A.4) The bivariate kernel function in (2.21) is compactly supported, bounded kernel such that $\int uu^\top K(u)du < \infty$. In addition, all odd-order moments of K vanish, that is $\int u_1^{\iota_1} u_2^{\iota_2} K(u)du = 0$ for all non-negative integers ι_1, ι_2 such that their sum is odd. Furthermore, $K(u)$ is twice continuously differentiable with

$$
\partial^2 K\left(u_1, u_2\right) / \partial u_l^2 < \kappa_0, \quad \text{for} \quad l = 1, 2.
$$

(A.5) The bandwidth sequence h used to define the weights $\widehat{\omega}_{ij}$ and ω_{ij} is of the form

$$h \sim N^{-\delta},$$

where $\delta \in \left(\frac{1}{8}, \frac{1}{6}\right)$.

(A.6) The density functions, $f(\cdot)$, defined in (A.2), the functions ρ_1, ρ_0, ϕ_1 and ϕ_0 (defined in Assumption (A.3) above), and the functions μ_1 and μ_0 and its derivatives are all fourth-order continuously differentiable, with derivatives that are bounded.

Assumptions (A.1) to (A.6) are rather similar to Assumptions (A.1) to (A.7) in Ahn y Powell (1993), pp. 11-12, or Assumptions 5.1 to 5.7 in Powell (2001), pp. 178-179. (A.3) is an identification assumption. It is needed to identify $\widetilde{\beta}$. (A.4) is a standard set of assumptions on the kernel function. The bandwidth rate imposed in Assumption (A.5) is needed to achieve the root-N consistency of $\widehat{\beta}$. Under these assumptions we can prove the following result:

Lemma 2.3.1 *Under assumptions (A.1) to (A.6), as N tends to infinity,*

(i)

$$S_{ww} = 2\Sigma_{ww} + o_p(1), \tag{2.47}$$

where Σ_{ww} is defined in assumption (A.3).

(ii)

$$\begin{aligned} S_{wu} &= \frac{2}{N}\sum_i [d_i z_i \rho_{1i}\gamma_i f(p_{21i}, p_{3i})\,(w_{1i} - \phi_1(p_{21i}, p_{3i})) \\ &+ (1-d_i) z_i \rho_{0i}(1-\gamma_i) f(p_{20i}, p_{3i})\,(w_{1i} - \phi_0(p_{20i}, p_{3i}))] v_i \\ &+ o_p(\frac{1}{\sqrt{N}}). \end{aligned} \tag{2.48}$$

Now, we need to introduce some additional assumptions to show the asymptotic representation for $\widehat{\beta}$. As it was previously mentioned we obtain the asymptotic representation for this estimator in two steps. The first by obtaining the asymptotic representation of $\widetilde{\beta}$ (see Lemma 4.2.1 above) and next, analyzing the divergence between $\widehat{\beta}$ and β. In order to do so, we need to introduce some additional assumptions that enables us to achieve some uniform bounds for $\widehat{p}_{21}(w_{12})$, $\widehat{p}_{20}(w_{12})$ and $\widehat{p}_3(w)$.

(A.7) The bivariate Mth-order kernel function, $K_1(u)$, satisfies:

- $K_1(u)$ is bounded and integrable. $|K_1(u)| \leq \bar{K} < \infty$ and

$$\int_{R^{k_1+k_2+k_3}} |K_1(u)|\, du \leq \tau < \infty.$$

- For some $\Lambda_1 < \infty$ and $L < \infty$, either $K_1(u) = 0$ for $\|u\| > L$ and for all $u, u' \in R^{k_1+k_2+k_3}$

$$\|K_1(u) - K_1(u')\| \leq \Lambda_1 \|u - u'\|.$$

- $\int |u|^{M+1} |K_1(u)| \, du < \infty$.

(A.8) Let $f(w)$ the joint density of the vector w. We assume that the M-th derivative of $f(w)$ is uniformly continuous and that $\sup_w \|w\|^q f(w) < \infty$, for some $q > 0$.

(A.9) The second derivatives of p_{21}, p_{20} and p_3 are uniformly continuous and bounded.

(A.10) For some $s > 2$,

$$\begin{aligned}
&E\left(\left| d_i z_i \right|^s \middle| w_{12} = w_{120}\right) f(w_{120}) < B_1 < \infty, \\
&E\left(\left| z_i \right|^s \middle| w_{12} = w_{120}\right) f(w_{120}) < B_2 < \infty, \\
&E\left(\left| d_i \right|^s \middle| w_{12} = w_{120}\right) f(w_{120}) < B_3 < \infty, \\
&E\left(\left| d_i \right|^s \middle| w = w_0\right) f(w_0) < B_4 < \infty.
\end{aligned}$$

(A.11) The bandwidth sequence h_1 used to define the estimators (2.22)-(2.24) is of the form,

$$h_1 \sim N^{-M/2},$$

for M given in assumption (A.7).

Assumptions (A.7) to (A.11) are sufficient conditions to obtain uniform bounds for the estimators (2.22)-(2.24). See for example Hansen (2008).

Teorema 2.3.1 *Under assumptions (A.1)-(A.11)*

$$\widehat{S}_{ww} = S_{ww} + o_p(1), \tag{2.49}$$

$$\begin{aligned}
\sqrt{N}\left(\widehat{S}_{wu} - S_{wu}\right) &= \frac{2}{\sqrt{N}} \sum_{i=1}^{N} \rho_{1i}^2 \gamma_i f(p_{21i}, p_{3i}) D\mu_1(p_{21i}, p_{3i}) \\
&\quad \times (w_{1i} - \phi_1(p_{21i}, p_{3i})) \left[z_i - p_{21i}\right] \left[d_i - p_{3i}\right] \qquad (2.50) \\
&+ \frac{2}{\sqrt{N}} \sum_{i=1}^{N} \rho_{0i}^2 (1 - \gamma_i) f(p_{20i}, p_{3i}) D\mu_0(p_{20i}, p_{3i}) \\
&\quad \times (w_{1i} - \phi_0(p_{20i}, p_{3i})) \left[z_i - p_{20i}\right] \left[d_i - p_{3i}\right] \qquad (2.51) \\
&\quad + o_p\left(N^{-1/2}\right),
\end{aligned}$$

where

$$\begin{aligned}
D\mu_0(p_{20i}, p_{3i}) &= \partial\mu_0\left(p_{20i}, p_{3i}\right)/\partial p_{21i} + \partial\mu_0\left(p_{20i}, p_{3i}\right)/\partial p_{3i}, \\
D\mu_1(p_{21i}, p_{3i}) &= \partial\mu_1\left(p_{21i}, p_{3i}\right)/\partial p_{21i} + \partial\mu_1\left(p_{21i}, p_{3i}\right)/\partial p_{3i}.
\end{aligned}$$

Corolario 2.3.1 *Under assumptions (A.1)-(A.11) the estimator $\widehat{\beta}$ in (2.25) has the following asymptotic linear representation,*

$$\begin{aligned}
\sqrt{N}\left(\widehat{\beta}-\beta\right) &= \Sigma_{ww}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\rho_{1i}\gamma_i f(p_{21i},p_{3i})\left(w_{1i}-\phi_1(p_{21i},p_{3i})\right)\{d_i z_i v_i \\
&+ \rho_{1i}D\mu_1(p_{21i},p_{3i})\left(z_i-p_{21i}\right)\left(d_i-p_{3i}\right)\} \qquad (2.52)\\
&+ \Sigma_{ww}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\rho_{0i}(1-\gamma_i)f(p_{20i},p_{3i})\left(w_{1i}-\phi_1(p_{20i},p_{3i})\right)\{(1-d_i)z_i v_i \\
&+ \rho_{0i}D\mu_0(p_{20i},p_{3i})\left(z_i-p_{20i}\right)\left(d_i-p_{3i}\right)\}+o_p(1). \qquad (2.53)
\end{aligned}$$

and is asymptotically normal,

$$\sqrt{N}\left(\widehat{\beta}-\beta\right)\to_d N\left(0,\Sigma_{ww}^{-1}\left(\Omega_1+\Omega_0\right)\Sigma_{ww}^{-1}\right),$$

where

$$\begin{aligned}
\Omega_1 &= E\left\{\rho_{1i}^2\gamma_i^2 f^2(p_{21i},p_{3i})\left[d_i z_i v_i+\rho_{1i}D\mu_1(p_{21i},p_{3i})\right.\right.\\
&\quad \left.\left.\times\left(z_i-p_{21i}\right)\left(d_i-p_{3i}\right)\right]^2\left(w_{1i}-\phi_1(p_{21i},p_{3i})\right)\left(w_{1i}-\phi_1(p_{21i},p_{3i})\right)^{\top}\right\},\\
\Omega_0 &= E\left\{\rho_{0i}^2\left(1-\gamma_i\right)^2 f^2(p_{20i},p_{3i})\left[(1-d_i)z_i v_i+\rho_{0i}D\mu_0(p_{20i},p_{3i})\right.\right.\\
&\quad \left.\left.\times\left(z_i-p_{20i}\right)\left(d_i-p_{3i}\right)\right]^2\left(w_{1i}-\phi_0(p_{20i},p_{3i})\right)\left(w_{1i}-\phi_0(p_{20i},p_{3i})\right)^{\top}\right\}.
\end{aligned}$$

2.4. Empirical results

2.4.1. Data description and summary statistics

In this section, we use the estimators proposed above to analyze empirically the effects that some relevant variables have on expected wages while accounting for sample selection and chronicity. Of course we are also interested in the effect of chronicity in labor supply or participation rates. The application is based on the Survey on Living Conditions from the Spanish Institute of Statistics. This survey allows us to use microdata in two possible units of analysis: the private households and the individuals that are in those households. In our application we use individual microdata and we have then the possibility of having information about socioeconomic characteristics and chronic illnesses affecting Spanish people. This survey collects information in a harmonized way for the EU countries since 2004. We use the microdata available for the more recent year, which is 2016. The sample is composed by $30,688$ observations referent to individuals interviewed in the spring of 2016. Due to our interest in labor market, we restrict the sample for people between 16 and 65 years old which is the population in age of working. The dependent variable r_i is the wage which is specified to depend on the level of studies, the age, the stability at job and the chronicity. The last one is the binary endogenous regressor of interest, which is defined as the onset of some chronic illness.

Table 2.1: **Definition of variables.**

	Definition
r = Wages	In Euros
w1= Age	Age in years
w2= Sup	1 if the person has high level studies
w3 = Single	1 If the person is single
w4= Indefinite-Term contract	1 if the person has a indefinite-Term c.
w5= Elementary	1 if the person only has elementary studies
d = Chronic illness	1 if the person has a chronic illness
z = Employed	1 if the person has a job

Source: Living Conditions Survey (INE), 2016

The descriptive statistics are also based on the Living Conditions Survey data of 2016. Statistics about chronicity indicate that the 24 % of the respondents reported having chronic illness. In this study, we understand as chronic the long-term illness which is not due to isolate string process. The survey consider as chronic, only the illnesses diagnoses by a health professional. Before estimating the parameters of the structural model of labor supply, we need to highlight some important characteristics of the Spanish population and their relationship with labor market. To have a more general perspective, we show in Table 2.2 the main descriptive statistics of the selected variables. About the half of the respondents (45,7 %) reported having a job and a 37,5 % reported to have high level of education. Slightly over the 45 % of respondents reported to have an indefinite-term contract.

Table 2.2: **Descriptive Statistics**

Variable	**Mean**	**Standard Deviation**
Annual wages	9339.3	12277.08
Age	43.21	13.46
Single	0.375	0.48
Indefinite-Term contract	0.457	0.49
Sup	0.317	0.33
Elementary	0.129	0.11
Chronic Illness	0.244	0.42
Employed	0.457	0.49

Source: Living Conditions Survey (INE), 2016

Traditionally, it was considered that chronic illnesses were a problem for elderly population. Nowadays it is a general belief that young people as the middle-aged ones are affected by chronicity (see Table 2.3). The great increase of chronic conditions might have important economic implications as the reduction of wages, labor participation and productivity.

Table 2.3: **Chronicity Prevalence.** %

Gender/Age	**16-29**	**30-49**	**50-65**
Men	10.80	17.64	36.56
Women	11.34	20.62	38.06

Source: Living Conditions Survey (INE), 2016

We observe that women are more affected by chronicity than men in all ranges of age. Women reported a higher percentage of chronic illness (26 %) compared to (23 %) for men but this difference is not so large. Biological differences are the possible reason for this gap. This may be also explained because women usually live more years, so the possibility of developing chronic illness is higher.

Table 2.4: **Average Income for Spanish workers.**

Chronic Illness/Age	**16-29**	**30-49**	**50-65**
With chronic illness	5030.4	13814	16320
Without Chronic Illness	6427.7	15467	19439

Source: Living Conditions Survey (INE), 2016

Income for people with chronic illness is, in average, lower than for the rest of population. The difference between both groups of population is $1,397$,3 euros for the first range of age, $1,653$ for the second and $3,119$ for the last range, so the difference is more significant with the increment of the age, as we mentioned before, this group has also more chronic illnesses.

2.4.2. Estimation results

We are going to concentrate now on the estimation of the econometric model already detailed in Section 2. Firstly, to estimate the propensity scores (2.14) we use (2.22)-(2.24). In this first step we use quartic (biweight) kernels and we choose the bandwidth using a cross validation criteria. Therefore we obtain $\hat{p}_{20i}, \hat{p}_{21i}, \hat{p}_{3i}$. For the selection equation case, eq. (2.1)-(2.2), to control for generic factors more than having a chronic illness, we have included the age, high level of studies, the stability at job and the civil status of individuals. Finally, to calculate the probability of having a chronic illness we have chosen these variables and also if the individual only has elementary studies.

In a second place, we have calculated the parameter vector β using (2.25). The kernel function chosen here is the "biweight kernel". It is also necessary to choose a bandwidth. There is a large literature on choosing the bandwidth in nonparametric estimation techniques. In this case a simple way to solve the choice is to set the bandwidth equal to the absolute value of the distance to the x-th nearest neighbour (i.e $h_n = |\hat{p}_{20i} - \hat{p}_{20j}|; |\hat{p}_{21i} - \hat{p}_{21j}|; |\hat{p}_{3i} - \hat{p}_{3j}|$), Additionally it is possible, and something quite usual, to choose a fixed bandwidth since the $p's$ lies between 0 and 1 an appropriate bandwidth could be 0,2 or 0,5. We had tried different bandwidths in order to check the robustness of our results, we present the more relevant ones, for $h = 0{,}2, 0{,}5, 0{,}75$. Finally, we show the estimates of the wage equations, as was our main interest. The results are given in Table 2.5 along with their estimated standard errors.

Table 2.5: **Wage equation estimates. (Standard deviation in brackets)**

	Coefficient (h=0.2)	**Coef. (h=0.5)**	**Coef. (h=0.75)**
Indefinite c.	8259,63 (360.8)	8572,66 (261.28)	8733,22 (255.26)
Sup	10002,45 (348.011)	8950,33 (486.6)	8985,47 (480.76)
Age	−35,05 (18.17)	−34,57 (18.17)	186,02 (94.44)

Source: Own Elaboration from Living Conditions Survey (INE), 2016.

As presented in Table 2.5, there is a significant positive relationship between an additional year of education and expected wage. The coefficient of superior studies is too large, compared to the coefficient of the age variable, which shows a rather small effect which is only significant at a 5 % in the third case, (h=0.75). According to previous empirical and theoretical economic evidence, education has a positive and direct relationship with expected wages; this means that, ceteris paribus, the higher the educational level is, the higher will be the expected earnings. Secondly, our results show as we expected, a positive correlation between the expected wage and the fact of having a permanent contract, and this coefficient is also quite large. This indicate that there is a substantial wage differential between job-contracts that cannot be explained by other observable characteristics of individuals. Moreover, temporary jobs are usually known to pay less, offer less training, and be less satisfying than regular jobs ((Kahn, 2007)). In the previous study it was also found that young people and women are more concentrated in temporary jobs. One of the possible hypothesis in explaining this fact is that having a permanent job is seen by employers as a promotion, that is workers without experience and less studies level must receive training in the temporary job with starting wages below the permanent workers.

Table 2.6: **Impact of chronicity on expected wages**

	Controls	**Treated**	**Difference (ATT)**	**T-stat**	**ATE**
$\hat{\lambda}(h = 0{,}2)$	7371.03	10126.8	-2755.8	-14.33	-2418.86
$\hat{\lambda}(h = 0{,}05)$	7370.24	10124.25	-2754.02	-14.32	-2417.56
$\hat{\lambda}(h = 0{,}75)$	7370.78	10126.55	-2755.76	-14.33	-2419.16

Source: Own Elaboration from Living Conditions Survey (INE), 2016.

Now we estimate the impact of chronicity on expected wages. In order to do this we use (2.32)-(2.34). The estimates of the average treatment effect, λ, are provided in Table 2.6. Also the standard deviation estimates are given in the table. They have been computed using bootstrap techniques. For comparison, we present the results under different choices of the bandwidth. Our results confirm the idea that having a chronic illness has a negative effect in wages (the same as under the joint normality assumption), concretely of around 2,500 Euros less than other individual without chronic illness. For the whole population, chronicity has a slight negative effect, but has a stronger negative effect for those with a chronic illness (Treated). This suggests that having a chronic illness is perverse on earnings. These findings are related with the ones of (Meenan et˜al., 1982) and (Duguet y Le-Clainche, 2014), who showed that chronic illnesses had a negative impact in the income, due mainly to wages losses more than health expenditures.

For the sake of comparison in Table 2.7 we provide the parameter estimates using the proposal in (Kim, 2006).

Table 2.7: **Wages equation estimates under standard assumptions.**

	Coefficient	**Std. Err.**	**t**	$P > \|t\|$
Indefinite-Term contract	10163.4	140.7	72.24	0.000
Sup	2419.3	230.6	10.49	0.000
Age	484.9	18.8	25.7	0.000
Chronic Illness	-2313.9	155.9	-14.84	0.000
Cons	3627.2	273.7	13.25	0.000

Source: Own Elaboration from Living Conditions Survey (INE), 2016.

As we expected, having a chronic illness affects negatively the expected wage and it is significant. People with any chronic illness earn about $2,314$ Euros less than people without chronic illness ceteris paribus. The rest of estimated effects have the same sign that in our proposal while the magnitudes are somehow different. Summarizing, our results confirm that having more level of studies, indefinite contract and being older show positive influences on the expected wages. The standard errors under our nonparametric first step estimates approach are larger than using the joint normality assumption. This may seems discouraging but we think is a more real estimation due to the uncertain in the form of the distributions of the reduced equations.

Table 2.8: **Comparison of different estimates of λ**

	Coefficient	**Std. Err.**	**t**	$P > \|t\|$
Effect of Chronic Illness				
OLS	-3813.537	283.44	-13.45	0.000
Two stage sample selection without endogeneity	-3821.711	285.06	-13.41	0.000
Propensity Score Matching (without considering the selection)	-2716.14	677.930	-4.01	0.000
Kim's Approach	-2313.9	155.9	-14.84	0.000
Our Approach with parametric first step estimates	-2094.82	189.064	-11.08	0.000
Our Approach with nonparametric first step estimates	-2755.7	192.3	-14.33	0.000

Source: Own Elaboration from Living Conditions Survey (INE), 2016.

Just for comparison purposes we had included in Table 2.8 the estimates obtained for

λ using other approaches such as simple OLS, Heckman's fully parametric two stage sample selection without endogeneity, the usual propensity score matching method and two different versions of proposal. Using probit maximum likelihood estimators of the propensity scores and fully nonparametric ones. We can observe that our results are quite close from the obtained with propensity score matching and under Kim's estimation technique. As we have mentioned before, standard errors in our approach are somewhat larger than in the case of parametric first step estimates, this is in part for the imprecision of the estimates from the first stage which are nonparametric equations instead of probit models. In our opinion this is not discouraging because we think they give a more realistic view which is not sensitive to particular specification assumptions. So that, the advantage of our proposal is the flexibility and robustness to misspecification.

2.5. Conclusions

In this paper, we have presented a new method for estimating a structural model of labor supply in which a binary endogenous variable is included. The estimator is obtained in a two stage regression procedure. In the first stage, corrections from endogeneity and sample selection are derived. In the second stage, using pairwise differencing techniques root-N consistent estimators of the parameters of interest are obtained. The main novelty of this technique is that, to identify and estimate the parameters of interest, there is no need to pre-specify the sample selection mechanism and the endogeneity. Finally we have shown an empirical application to explain variations on wages. Given the results, it is necessary to highlight the importance of managing chronic disease in a right way. Thus, policies are needed to reduce risks and minimize the negative effects of chronicity as well as prevention and early detection. In order to do so, governmental evaluations of chronicity effects in most of the developed countries are gaining popularity due to the changes in the demography and the reasons explained above.

Capítulo 3

Capítulo 3

Is there a gender pay gap by employment sectors in Spain?

Using microdata from the Wage Structure Survey, we analyse the gender pay gap per sector considering the whole wage distribution. The main contribution of this paper is to assume that the decision to work in the private/public sector is a prior process determined endogenously in the model. In this case, the usual Ordinary Least Square estimation is inconsistent and it is necessary to use alternative techniques. We use quantile regression techniques to calculate how much of the gap is due to differences in returns between men and women and sectors, taking into account the sample selection bias. We find that the size of the gap attributed to different returns varies substantially across the wage distribution. Public sector employees are paid a wage premium, on average, related to similar counterparts in the private sector and the gap is wider for women. Moreover, the proportion of the gender pay gap explained (due to different characteristics) tends to be greater for workers who are at the bottom of the wage distribution in both the private and public sectors. It can also be observed that the conditional distribution of wages in the public sector is more compressed than the wage distribution in the private sector, i.e. the wage distribution has a lower standard deviation in the public sector.

In short, a look at the whole wage distribution reveals that the discrimination in the gender pay gap is typically higher at its top than at its bottom, suggesting that glass ceilings are more prevalent than sticky floors for both men and women.

3.1. Introduction

Most studies on earnings gaps between men and women conducted by economists divide gender differences in payment into two components: the part that is *explained* by worker characteristics and the rest *unexplained* ((Blau y Kahn, 2007)). The latter part is often used as a proxy for discrimination. According to (Suh, 2017), modernization theory unfolds a map that there is a positive relationship between female labor force participation and development increasing demand for labor and social acceptance of women's employment.

Despite the gradual incorporation of women into the labour market since the World War II, the fact is that, on average, women earn less than men. Moreover, women have lower participation rates and higher levels of unemployment. As a result, the entry of women into the labour force has attracted scholarly interest in the wage discrimination field. As is well known, wages depend on several characteristics of individual workers and at the same time on other factors associated directly with the market itself, such as trade union power and minimum wage legislation, so there could be heterogeneity in wages among workers. The popular neoclassical interpretation of wage gaps is associated with differences in productivity. According to that hypothesis, with the same productivity level there should be equal pay. But in fact it is known that this is not happening and there is a large body of literature analysing wage gaps ((Mandel y Semyonov, 2014)).

A variety of techniques have been used to estimate these gender earnings gaps and to estimate how much of them is due to wage-determining factors and how much to unexplained reasons (discrimination). There are currently a great many studies of unequal wages for different labour force groups in the labour market. Most authors interested in studying variation in wages have adopted the human capital model as the theoretical basis for the earnings function ((Becker, 1964)). Some of them use variants of (Oaxaca, 1973) decomposition to identify the causes of gender wage gaps.

It is widely believed that although it has narrowed in the last few decades, there is still a substantial earnings gap between male and female employees. Recently the size of both components has declined in the labour market due to several changes such as a reduction in discrepancies in relevant labour force attributes (education and work experience) and changes in the earnings returns on such attributes ((Arulampalam et~al., 2007)).

Nevertheless, sociologists argue that occupational segregation is one of the main factors explaining earnings disparities between men and women. This means that although the differences in work-related characteristics between men and women have decreased over time, and with them the wage pay gap, occupational segregation is still one of the biggest factors explaining actual earnings disparities. According to this point of view, womens earnings are lower than men's because women are selected (either denied access or self-selected) into female-typed low-paying jobs and occupations ((Bielby y Baron, 1986); (Petersen, 1995); (Treiman y Hartmann, 1981)). However, in recent decades the labour market has seen a steady decline in rates of occupational segregation, which is especially evident among highly educated workers. The major cause of this decline is the growing integration of women into new occupations, particularly managerial positions, from which they were traditionally absent ((Burris y Wharton, 1982); (Cotter et˜al., 2004); (Jacobs, 1992); (Mandel, 2012); (Mandel, 2013); (Weeden, 2004)).

Sectoral segregation is another important factor to take into account in the gender pay gap. (Alaez y Ullibarri, 2001) show that sectoral and occupational segregation of women in the Spanish labour market is the main source of gender wage gaps. In this paper we seek to analyse the differences in wages between the public and private sectors in Spain and calculate what proportion of those differences is not due to explainable factors taking into account that the decision to work in the private/public sector is a process determined endogenously in the model. As a second objective, taking as our starting point the fact that a gender wage gap exists, we seek to test whether that gap is smaller in the public sector than in the private one.

There are several reasons for earning differences between public and private sector workers. The main one is that in the private sector payments are related to corporate profits while the public sector is subject, to some extent, to political constraints. In that sense, in recent years the Spanish government has been pressured to reduce a large budget deficit due to the economic crisis and one of the measures taken was to cut back expenditure, including the wages of government employees. Governments can thus use public sector wages as an instrument of economic policy. Public sector wages are of great concern to policy makers because public employees account for nearly 16 % of the total Spanish work force. In this sense, it is necessary to analyse carefully the payment gaps between the two sectors.

But if the selection of a sector is not a random decision by workers the conventional approach to analysing discrimination is not correct because there is an endogeneity problem. If it is considered that there is a prior selection process for entering the public sector, the first choice for an individual is whether to take part in a public selection process to opt for a public job or not. Thus, if non-observable characteristics affecting wages are correlated with non-observable factors determining the choice of sector, the usual Ordinary Least Squares estimation will not be consistent. In this case, the so-called endogenous switching model may be a suitable procedure. Moreover, using the entire wage distribution and applying quantile regression, it is possible to supplement previous results for the mean analysing the whole wage distribution. To our knowledge, no prior studies have addressed the issue of differences in the gender wage gap between the public and private sectors in Spain, also considering the whole wage distribution . The decline in gender pay gaps can be attributed to three major trends: a reduction in men's and women's measured and unmeasured wage-related characteristics, a decrease in the rate of occupational/sectoral segregation and a decline in pay discrimination against women ((Alaez y Ullibarri, 2001)).

Thus, to sum up, the two main objectives of the paper are to examine wage gaps in Spain between the public and private sectors and to test the hypothesis that gender discrimination is lower in the public sector considering that the decision to work in the private or public sector is a prior process determined endogenously in the model. We report an analysis on some aspects of wage gaps; specifically we consider payment differences by gender and sector. We expect to bring new insights to explain wage structure in both ways and investigate why workers with apparently identical characteristics receive different wages across firms but also in the public sector. In this paper, we first develop theoretical expectations regarding the sources of the decline in earnings gaps between men and women in the private and public sectors and then empirically estimate our expectations in the labour market in both cases, as a whole and separately for different sectors.

The structure of the paper is as follows: Section 2 sets out the data and the econometric methodology used. Section 3 presents the principal results and, finally, Section 4 addresses the main conclusions and implications.

3.2. Data and Methods

3.2.1. Data and variables

The source of the data used is the Annual Wage Structure Survey (2016), specifically the micro-data 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 day and other social and demographic variables, related to occupation variables. The Survey also provides information on average earnings and the distribution of wages. The Annual Wage Structure Survey is conducted annually by the Spanish Statistical Office and it results from the combination of data from different statistical and administrative sources. It uses a sample for Spain which is made up of all regular employees included in the Social Security system. The information may be also broken down by regions (Autonomous Communities). Average annual earnings in Spain in 2016 were $20,131{,}41$ Euros for female and $25,924{,}43$ Euros for men, according to the Annual Wage Structure Survey (INE), so on average womens earnings were 77,7 % of those of men. In Europe as a whole, the gap in wages by gender is also quite significant, but it decreases when similar situations such as having the same occupation, working day and type of contract, among others, are considered. Nevertheless, in 2016, most European Countries had a bigger gender pay gap in the private sector than in the public one, as expected (Table A1 in the Annex). On average, European women earn 11,8 % less than men in the public sector and 16,9 % less in the private sector. In Spain the figures are 13 % and 19 %. That is, Spanish women who work in the private sector have wages almost 20 % lower than their male colleagues.

The variables selected for estimating earnings equations are those conventionally used in labour productivity models for predicting earnings: gender, age, level of education, work experience, tenure and full time job . Earnings, the dependent variable, is measured as usual by annual wage (in logs). The variables and their means are listed in Table 3.1.

Table 3.1: **Statistics of the variables: mean and standard deviation (in brackets).**

	Variable definition	**Public Sector**	**Private Sector**
Wage	Annual Wage, in Euros	29458,72 (17480.54)	23911,49 (22521.98)
Age1	Age from 20 to 29	0,049 (0.216)	0,120 (0.325)
Age2	Age from 30 to 39	0,235 (0.424)	0,336 (0.472)
Age3	Age from 40 to 49	0,320 (0.466)	0,309 (0.462)
Age4	Age from 50 to 59	0,314 (0.462)	0,187 (0.390)
Age5	Age more than 60	0,079 (0.270)	0,04 (0.204)
Primary	1, if primary studies	0,049 (0.216)	0,154 (0.361)
High School	1, if High School	0,467 (0.498)	0,568 (0.495)
College	1, if College	0,478 (0.499)	0,262 (0.440)
Tenure	Years of experience	14,25 (10.76)	9,18 (9.27)
Full-time	1, if full-time job	0,90 (0.298)	0,80 (0.392)
Male	1, if male	0,468 (0.499)	0,592 (0.491)

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

As can be seen, public sector workers in Spain are older: 31,5 % of them are aged between 50 and 59 years old, compared to 18,7 % in the private sector. Moreover, in the public sector the percentage of workers with higher education qualifications is 47,8 %, compared to 26,2 % in the private sector. The number of women is much higher in the public sector (53,2 %) than in the private sector (40,8 %), the proportion of full-time contracts is very high in both sectors and, finally, the average job tenure is greater than in the private sector by just over 5 years.

3.2.2. Pay gap

The Oaxaca decomposition is the most popular technique for analysing gender wage discrimination. It breaks down the average wage gap between two demographic groups, usually, but not only, men and women. The idea is to estimate the earnings functions for each group separately and the results are used to calculate the percentages of the (log) earnings gaps attributable to the explained part (due to different characteristics) and the unexplained part (discrimination). It provides a simple method for learning whether public sector workers are paid as well as their private sector counterparts, but it says little about the underlying wage distribution.

In order to calculate the Oaxaca index, consider the standard model of earnings:

$$w_{ij} = x_i^\top \beta_j + u_{ij} \tag{3.1}$$

where $j = 1$ for public sector and $j = 2$ for the private one. (Blinder, 1973) and (Oaxaca, 1973) propose calculating the wage gap and its causes by subtracting the group 1 income equation from the group 2 income equation, assuming that the difference between the parameters of the equation corresponds to discrimination; so it can be computed

$$\overline{x}_1^\top \hat{\beta}_1 - \overline{x}_2^\top \hat{\beta}_2 = \underbrace{(\overline{x}_1 - \overline{x}_2)^\top}_{\textit{explained}} \hat{\beta}_2 + \overline{x}_1^\top \underbrace{(\hat{\beta}_1 - \hat{\beta}_2)}_{\textit{discrimination}}$$

where $\overline{x}_1; \overline{x}_2$ are the values of covariates, on average, in each sector. If there is discrimination, i.e. $\beta_1 - \beta_2 \neq 0$, the discriminant coefficient ((Oaxaca, 1973)) can be calculated as:

$$D = \frac{[\overline{w}_1/\overline{w}_2] - [\overline{w}_1/\overline{w}_2]^0}{[\overline{w}_1/\overline{w}_2]^0} \tag{3.2}$$

where $[\overline{w}_1/\overline{w}_2]^0$ is the proportion that would exist if both groups were paid with the same criteria and $\overline{w_1}, \overline{w_2}$ are the average wages.

Expression 2 has conventionally been used in the standard literature, but it has some important drawbacks; (i) if the selection process is not random and it is thus endogenously determined, the Ordinary Least Squares estimates will be biased. However, this problem can be solved by developing some correction terms (control function approach) as the next section proposes to overcome this bias; (ii) The Oaxaca decomposition only takes into account divergences in the mean, but in general, more information is provided by analysing the whole wage distribution and letting the effects differ at any point of the wage distribution. Considering this problem, we study the gender wage gap by sectors, including correction terms to solve selection bias and extending the analysis over the whole distribution and not just the mean. Considering the whole distribution is important in our context because there are major discrepancies in the tails of the wage distribution, so it is relevant to examine whether the wage gap is larger in the upper tail of the distribution or among those with lower salaries.

Empirical quantile regression literature in economics has become popular recently. The theoretical framework for this estimation is based on the quantilic regression methodology developed by (Koenker y Bassett, 1978) and applied, in the context of wage equations by (Chamberlain, 1994) and (Machado y Mata, 2005) among others. (Amemiya, 1985) was the first to consider quantile regression methods in the presence of endogenous covariables. He shows the consistency and asymptotic normality of two-step median regression estimators. Assume this model:

$$w_{ij} = x_i^\top \beta_j^\theta + u_{\theta i} \tag{3.3}$$

We assume that the quantile $\theta - th$ of the conditional distribution of the endogenous variable (wages) is a linear function of covariates (worker's characteristics, x_i). Quantile regression is based on minimising weighted absolute deviations to estimate conditional quantile functions, so it can be employed to explain the determinants of the dependent variable at any point of the distribution. Thus, the difference in the (log) wages between the public and private sectors can be written as:

$$Q_\theta(w_{i1}) - Q_\theta(w_{i2}) = \sum \hat{\beta}_2^\theta(\overline{x}_1 - \overline{x}_2) + \sum(\hat{\beta}_1^\theta - \hat{\beta}_2^\theta)\overline{x}_1 \tag{3.4}$$

(Koenker y Bassett, 1978) have shown that quantiles can be estimated by minimising $(\beta_1^\theta; \beta_2^\theta)$. But this approach does not consider sample selection bias. We can, however, combine the decomposition technique with quantile regressions to determine the effects of interest at various points in the wage distribution. We employed the method proposed by (Melly, 2006) to decompose the gap in way similar to an Oaxaca-Blinder decomposition but using the complete marginal density distribution.

3.2.3. Econometric Methodology

To account for the sample selection problem, we start by considering a structural econometric model of labour supply. The earnings function is a useful framework for summarising the relationship between wages and observed productivity characteristics. The simplest form of this model is the (Mincer y Polacheck, 1974) human capital earnings equation, which states that the individual (log) wage depends on education, labour market work experience and a random unobservable component. We adapt this model for considering the public and private sectors (see also (Arulampalam et˜al., 2007)). Following this approach, the model takes the form:

$$\begin{aligned} w_{ij} &= x_i^\top \beta_j + u_{ij}, \quad iff \quad s_{ij} = 1, && (3.5)\\ where \quad s_{ij} &= 1\left(s_{ij}^* > 0\right) = 1\left(w_{ij}\beta_w + z_i^\top \gamma + \eta_{ij} > 0\right) \\ &= 1\left((x_i^\top \beta_j + u_{ij})\beta_w + z_i^\top \gamma + \eta_{ij} > 0\right) = 1\left(x_i^\top \beta_j \beta_w + z_i^\top \gamma + \beta_w u_{ij} + \eta_{ij} > 0\right) \\ &= 1\left(x_i^\top \alpha_j + z_i^\top \gamma + v_{ij} > 0\right), && (3.6) \end{aligned}$$

The variable w_{ij} is the outcome variable of interest: it is the (log) annual wage for the $i-th$ individual in Sector j (public or private sector). x is a vector of individual characteristics, the variable s_{ij} is a binary indicator variable set to one if person i is in the j sector and 0 otherwise, $1\,(A)$ denotes the indicator function, i.e. $1\,(A) = 1$ if A is true, and 0 otherwise. As usual, the β_j, β_w and γ are parameters to be estimated (we define $\alpha_j = (\beta_j \beta_w)$ and $v_{ij} = (\beta_w u_{ij} + \eta_{ij}))$ and the vector (u_{ij}, η_{ij}) is an unobserved idiosyncratic error.

This econometric model is used in many theoretical problems and is very well known for example in labour economics. In this context, equation (3.5) stands for the so-called wage equation, and equation (3.6) specifies the participation model in each sector j, which depends simultaneously on the wage that an individual expects in public sector and other personal characteristics. One of the most important characteristics included in x affecting wages is gender, which is a binary variable that we denote by d.

The first problem is that if in equation 3.5 the $E\left[u_{ij}|x_i, s_{ij}^* > 0\right]$ is different from 0, then Ordinary Least Squares Estimates are asymptotically biased. The estimation procedure proposed is therefore based on the idea of two-step selection models, with the following switching introduced to the importance of gender in all types of discrepancies concerning wages:

$$\begin{aligned} If \quad d_i = 0, \quad s_{i1} = 1 \quad &: w_{ij} = x_i^\top \beta_1^0 + u_{i1}^0, && (3.7)\\ If \quad d_i = 0, \quad s_{i2} = 1 \quad &: w_{ij} = x_i^\top \beta_2^0 + u_{i2}^0, && (3.8)\\ If \quad d_i = 1, \quad s_{i1} = 1 \quad &: w_{ij} = x_i^\top \beta_1^1 + u_{i1}^1, && (3.9)\\ If \quad d_i = 1, \quad s_{i2} = 1 \quad &: w_{ij} = x_i^\top \beta_2^1 + u_{i2}^1, \quad i = 1, \cdots, N. && (3.10) \end{aligned}$$

And then:

$$\begin{aligned}
E\left[w_{ij}|x_i, d_i=0, s_{i1}=1\right] &= x_i^\top \beta_1^0 + E\left[u_{i1}^0|d_i=0, s_{i1}=1\right], && (3.11)\\
E\left[w_{ij}|x_i, d_i=0, s_{i2}=1\right] &= x_i^\top \beta_2^0 + E\left[u_{i2}^0|d_i=0, s_{i2}=1\right], && (3.12)\\
E\left[w_{ij}|x_i, d_i=1, s_{i1}=1\right] &= x_i^\top \beta_1^1 + E\left[u_{i1}^1|d_i=1, s_{i1}=1\right], && (3.13)\\
E\left[w_{ij}|x_i, d_i=1, s_{i2}=1\right] &= x_i^\top \beta_2^1 + E\left[u_{i2}^1|d_i=1, s_{i2}=1\right], \quad i=1,\cdots,N && (3.14)
\end{aligned}$$

If there is evidence of non-normal disturbances in the reduced equation then it is not correct to estimate the participation decision by Probit Maximum Likelihood and in that case the so-called Inverse of Mills Ratio (IMR) is misspecified. In such a case, a partially linear model could be estimated using a non parametric function instead of IMR ((Arellano y Bonhomme, 2017)).

On the other hand, if the participation error (v_{ij}) is distributed as a standardised normal, probit maximum likelihood estimators of the first stage are consistent, and then the Mills ratio is suitable to correct for the bias. In that situation, we introduce those correction terms as additional regressors in the structural equations, thus making it possible to estimate the parameters by OLS consistently. The procedure proposed for estimating the structural parameters of this model is as follows:

- **Step 1:** The participation reduced equation, eq. (3.6) is estimated by Probit Maximum Likelihood.

- **Step 2:** With the parameters estimated from the first step the inverse Mills ratio terms required to approximate the unknown correction terms and estimate the wage equation to obtain consistent estimates can be estimated (equations from (3,11) to (3,14)).

- **Step 3:** Finally, to estimate the effects in the structural equation, a decomposition of the wage gaps applied to both standard regression model and Quantile Regression estimates is presented.

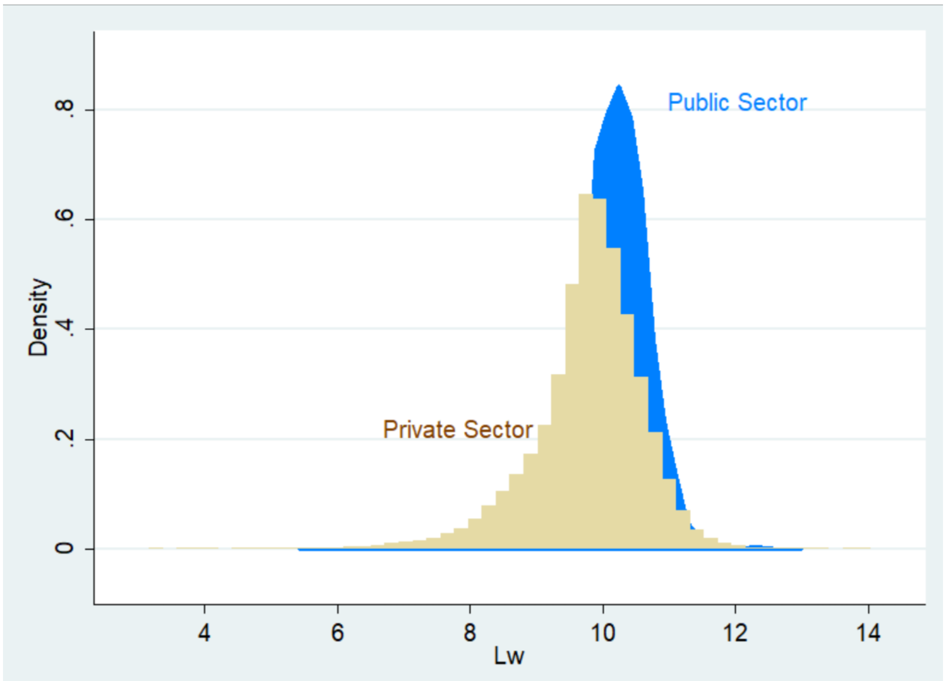
Following the seminal papers in the field, we use the standard methodology for analysing sector wage gaps, i.e. we decompose the observed wage gap into two components: a difference in average characteristics between sectors or gender and a difference in the returns of these individual characteristics between sectors or gender which is treated as a residual effect ((Oaxaca, 1973)). Moreover, considering the evidence observed, the public-private wage gap may be higher at the lower end of the wage distribution and one main difference between Ordinary Least Squares and Quantile Regression is that the former are focused on obtaining the estimated wage at the sample average vector of characteristics and the latter generates counterfactual densities at each quantile of the distribution. Melly (Melly, 2006), proposes an intuitive procedure for decomposing differences at different quantiles of the unconditional distribution and shows that this estimator is consistent and asymptotically normally distributed. Consistent estimators of variances are also proposed. Accordingly. the cdeco STATA command is used to estimate the conditional quantiles at each percentile of the test score distribution ((Chernozhukov et˜al., 2013)). Bootstrapped standard errors are bootstrapped using 100 replications, using the asymptotic properties by (Chernozhukov et˜al., 2013).

3.3. Results

3.3.1. Summary statistics

As can be seen in Table 3.1, the difference in average annual wages between the two sectors is just over 5, 500 Euros in favour of the public sector. Moreover, almost half of the workers in the public sector have higher education qualifications (25 % in the private sector), average job tenure is almost 5 years greater in the public sector and approximately 47 % of public sector workers are men, compared to 59,2 % of private sector workers. These data seem to indicate that in the public sector women, older workers and holders of higher education qualifications predominate.

Figura 3.1: **Density function by Sector.**

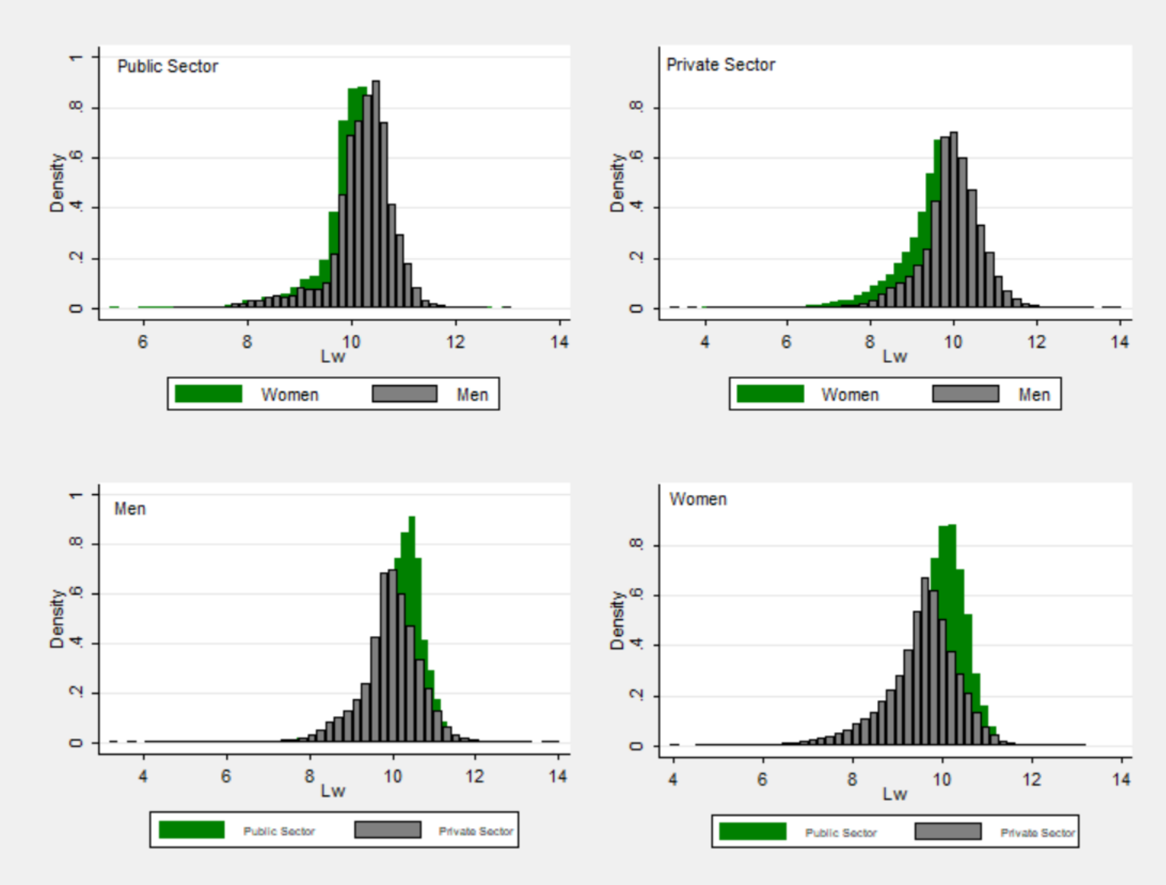


Source: Own elaboration with the Annual Wage Structure Survey, 2016.

Figure 1 shows the distribution of annual (log) wages. The (log) wages in the public sector have a higher mean and a lower standard deviation than those in the private sector. Moreover, the tails of this distribution are substantially larger in the private sector, i.e. there are workers with much lower and higher wages, on average, in the private sector. Therefore, it seems reasonable to think that the sector where an individual works is a determining factor for the wages received.

Figure 2 show the density of annual (log) wages by gender and sector. The first graph of Figure 2 reveals that in the public sector women have lower wages than male since the left tail is longer for women; the second graph in figure 2 shows that the average wage is higher in the public sector than in the private one for both men and women.

Figura 3.2: **Wage Density by Sector.**



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

Taking into account that we use the dependent variable in logarithm form, there is no reason to suspect non-normality in the structural equation based on the above graphs. Then, we can use the so-called Inverse of Mills Ratio to approximate the correction terms. Additionally, as can been observed in both graphs, there is quite heterogeneous behaviour along the wage distribution, so it seems more appropriate to analyse all the points of the distribution and not only the mean.

3.3.2. Estimation results

To estimate the gender pay gap, we first distinguish between the explained part (differences in endowments or characteristics) and the unexplained part (discrimination) in the two sectors. Subsequently, we compare the components of the pay gap over the quantiles and across sectors. Estimates are calculated with STATA statistical software.

Firstly, in Table A2 in the appendix, we show the results of equation (3.5) using standard Ordinary Least Squares estimation for each sector with no correction. As argued throughout the paper, this is used merely to give an initial intuition of the effects, because they will be inconsistent.

In the first step of the estimation procedure proposed, we regress the selection binary variable (Public/Private job) on some explanatory variables as proposed in equation (3.6) by calculating a Probit model, thus obtaining the so-called Inverse Mills Ratio. Thus, Table A3 shows the estimated probability of working in the public sector conditionally on personal characteristics and considering the expected wage. These results help us to describe the process of selecting between the two sectors and provide an initial estimator for the sample selection correction procedure. As can be seen in Table A3, the probability of working in the public sector increases with education, as expected due to the entrance process, decreases with age and is lower for men. Once again, all the variables are statistically significant.

As pointed out in the previous section, it is possible to re-estimate equation (3.5) by including the necessary sample selection correction term in each case. These results are presented in Table A4, desegregating the four possible cases $(d_i = 1, s_{i1} = 1), (d_i = 1, s_{i2} = 1), (d_i = 0, s_{i1} = 1), (d_i = 0, s_{i2} = 1)$ developed in equations (3.7)-(3.10). In most cases age, education, experience and full-time work are factors expected to increase wages. It interesting to highlight that further education has similar effects in the public sector for men and women, but that effect is considerably higher for men in the private sector. Additionally, working full-time has a higher effect on wages for men than for women in both sectors. Furthermore, the private sector pays much more for further education qualifications than the public sector for men, but not for women. Finally, there is a statistically significant self-selection term in the 4 groups considered (Table A4).

Table 3.2: **Estimated Wage Gender Gap by Sector. Euros**

	Mean	**Std. Dev.**
In Public Sector	-3,929.06	2993.3
In Private Sector	-5,159.66	4210.80

Source: Annual Wage Structure Survey, 2016.

Table 3.2 shows the gender pay gap estimated in the two sectors once the model proposed is estimated and it can be seen that in the private sector men earn 5, 159 Euros more than women but in the public sector the figure is only 3, 929. That is, gender pay discrimination is estimated to average around 14 % in the public sector and around 19,9 % in the private sector. Moreover, there is more dispersion in the private sector.

Next we analyse the causes of the differences by decomposing the different effects (Table 3.3).

Table 3.3: **Estimated Wage Differences by Sector (Oaxaca Decomposition).**

	Male	Female	Differences				
			Total	Explained		Not Explained	
				With correction	Without	With correction	Without
Public Sector	10,213 (0.005)	10,050 (0.004)	0,162 (0.006)	0,475 (0.012)	0,031 (0.004)	−0,313 (0.013)	0,131 (0.005)
Private Sector	9,952 (0.002)	9,546 (0.003)	0,405 (0.003)	0,492 (0.004)	0,219 (0.003)	−0,087 (0.004)	0,186 (0.003)

Standard deviation in brackets.
Source: Own Elaboration from Annual Wage Structure Survey, 2016.

The Oaxaca decomposition (Table 3.3) shows the mean predictions by sectors and gender and their differences. Both the differences explained by endowments and the unexplained differences (discrimination) are also corrected by endogeneity. In our sample, the geometric mean of wages in the public sector is around 27, 269 Euros for men and 23, 177 for women, giving a gap of about 16, 2 %. This means that mens wages in the public sector are on average about 16 % higher than those of women in the same sector. Adjusting womens endowments levels in regard to those of men increases women wages by 3,1 % but a gap of 13,1 % remains unexplained.

If a correction term is included due to the endogenous choice of sector, the effect due to differences in characteristics would increase womens wages by 47,5 % but a gap of −31,3 % remains unexplained [1].

The mean wage for women is round 13, 999 Euros in the private sector and about 21, 001 for men, so wages in the private sector are about 40,5 % higher for men than for women. Adjusting mens endowments levels in regard to those of women would increase womens wages by 21, 8 %. A gap of 18, 6 % remains unexplained. If the correction term is introduced the results change, and adjusting mens endowment levels to those of women would increase womens wages by 49, 2 %. A gap of −8,7 % remains unexplained. Next, the complete distribution of wages is analysed using Quantile Regression and the correction method proposed in this paper.

Table 3.4: **Estimated Wage Differences by Sector (Quantile Decomposition).**

Quantile	Sector	Differences					
		Total		Explained		Unexplained	
		With correction	Without	With	Without	With	Without
0.1	Public	−0,165 (0.197)	−0,218 (0.024)	−0,185 (0.04)	−0,166 (0.02)	0,020 (0.04)	−0,051 (0.02)
	Private	−0,575 (0.011)	−0,556 (0.010)	−0,515 (0.010)	−0,473 (0.006)	−0,060 (0.008)	−0,08 (0.008)
0.3	Public	−0,165 (0.006)	−0,148 (0.005)	−0,005 (0.005)	−0,010 (0.005)	−0,160 (0.006)	−0,138 (0.006)
	Private	−0,445 (0.004)	−0,417 (0.005)	−0,267 (0.004)	−0,231 (0.003)	−0,178 (0.003)	−0,186 (0.003)
0.5	Public	−0,167 (0.006)	−0,180 (0.005)	0,062 (0.06)	0,012 (0.04)	−0,229 (0.05)	−0,193 (0.05)
	Private	−0,351 (0.004)	−0,349 (0.003)	−0,139 (0.008)	−0,122 (0.008)	−0,212 (0.007)	−0,227 (0.006)
0.7	Public	−0,173 (0.006)	−0,171 (0.005)	0,117 (0.006)	0,026 (0.004)	−0,291 (0.006)	−0,198 (0.007)
	Private	−0,311 (0.004)	−0,339 (0.004)	−0,052 (0.003)	−0,082 (0.002)	−0,258 (0.005)	−0,256 (0.005)
0.9	Public	−0,159 (0.010)	−0,174 (0.009)	0,178 (0.03)	0,025 (0.025)	−0,337 (0.011)	−0,200 (0.010)
	Private	−0,311 (0.004)	−0,309 (0.005)	0,049 (0.005)	−0,042 (0.001)	−0,360 (0.005)	−0,266 (0.005)

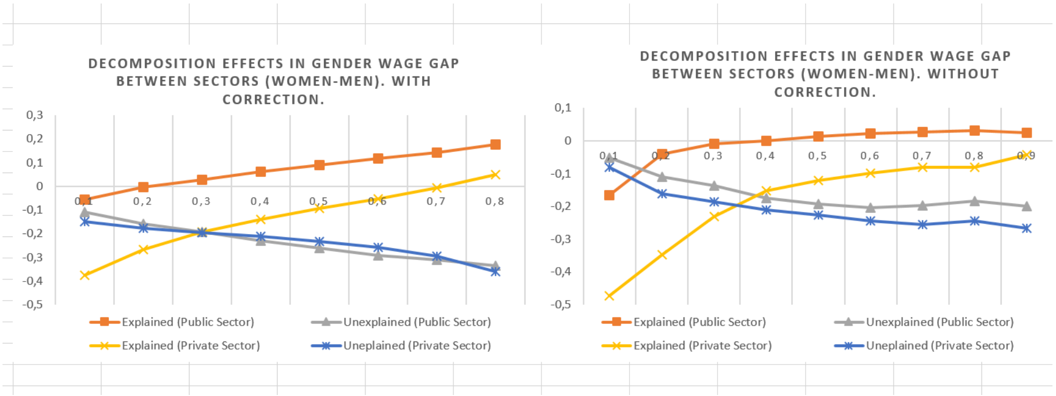
Pointwise Standard error in brackets.
Source: Own Elaboration from Annual Wage Structure Survey, 2016.

[1] We use bootstrap standard errors with 100 replications for all estimates in Blinder-Oaxaca decompositions, because (Jann, 2005) shows that bootstrap standard errors of the decomposition results are unbiased when sampling distributions are not known and regressors are stochastic.

Table 3.4 shows the gender gaps (female-male) by sectors. As can be seen, in both sectors there is a negative wage gap for women throughout the wage distributions. The biggest difference between the observable distributions is a 17,3 % in the 70th quantile , while the smallest is at the top end of the distribution, in the 90th quantile. It is also possible to observe that this gap is greater in the private sector for all points of the distribution, at 57,5 % in the 10th and decreasing as one moves up the wage distribution.

As can be seen, in the public sector the effect of characteristics (explained differences) against women is declining at the top of the wage distribution. At this end of the distribution it is possible to observe that gaps are attributable more to the coefficients effect (unexplained differences). In the private sector the pattern is similar but stronger, with larger differences.

Figura 3.3: **Estimated Gender Wage Gap between Sectors.**



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

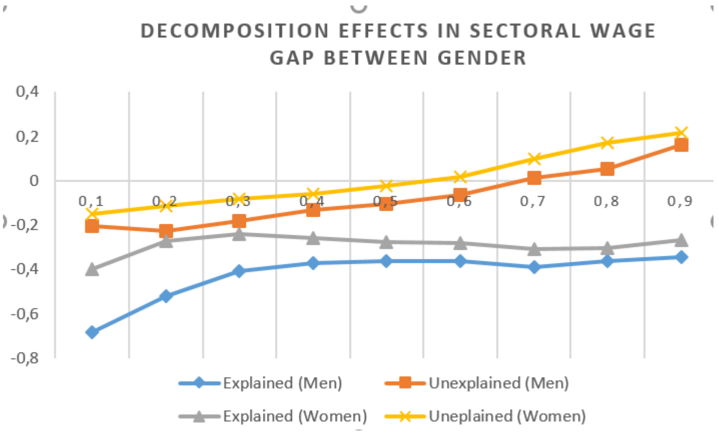
Table 3.5: **Unexplained differences by sector (Private-Public)**

Quantile	Men		Women	
	Effect	Pointwise St. Error	Effect	Pointwise St. Error
0.1	-0.551	0.020	-0.890	0.016
0.2	-0.385	0.006	-0.744	0.008
0.3	-0.321	0.006	-0.590	0.005
0.4	-0.314	0.005	-0.504	0.004
0.5	-0.300	0.005	-0.469	0.004
0.6	-0.265	0.005	-0.428	0.004
0.7	-0.209	0.004	-0.376	0.006
0.8	-0.133	0.005	-0.306	0.006
0.9	-0.05	0.009	-0.184	0.007

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

Table 3.5 shows that the first decile of the private sector wage distribution is around 55 % lower than the first decile of the public sector wage distribution. At the median, it is 30 % lower. Note that these gaps narrow at the top end of the wage distribution. For women, the pattern throughout the distribution is similar to that for men but with higher figures. At the median the private sector wage distribution is 46, 9 % lower than that of the public sector.

Figura 3.4: **Estimated differentials by type of effect between Sectors.**



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

The estimated unexplained wage differential (coefficient effect) varies depending on which quantile is chosen. For men it varies from −20 % in $\theta = 0{,}1$ to 16 % in $\theta = 0{,}9$. For women it varies from −15 % in $\theta = 0{,}1$ 21, 6 % in $\theta = 0{,}9$. On the other hand, the characteristics differential seems to be more stable over the distribution although for men there is more variation with θ than for women. The unexplained part of the sector wage gap is decreasing along the wage distribution up to the median for women, and up to quantile 0,6 for men. Finally, accounting also for the characteristics effect, our results show that, as expected, the wage gap for public employees is especially large for low-income earners. At the upper end of the wage distribution the gap is narrower.

The results show that the gender pay gap as a whole and the unexplained gap are larger in the private sector than in the public one. Moreover, the reduction in the gross gender pay gap is larger in the private sector than in the public sector (one half compared with one-quarter).

3.4. Conclusions

In this paper we investigate the public-private sector gender pay gap using Spanish micro data, seeking to provide new evidence on the gender pay gap between the public and private sectors in Spain. To that end, we propose a model that corrects for the selectivity bias of the sample considering the prior process of choosing the sector and additionally switching it by gender.

Our findings suggest that there is discrimination against women in both the public and private sectors, but less so in the former. Women are paid less than men, even when the distribution of characteristics is kept constant in the analysis. This result is consistent with the findings of recent reviews. The size of the gaps attributed to different returns varies substantially across the distribution of wages.

The first result suggests that public sector employees are paid a wage premium, on average, in relation to their counterparts in the private sector. The gap is greater for women, as detailed by the Oaxaca decomposition. Then, using quantile regression, we show that the gender gap tends to be greater for workers at the bottom end of the wage distribution. That is, the public /private wage gap is sensitive to the choice of quantile, so the hypothesis of a constant wage differential (implied in OLS methods) is rejected and the pattern of premia varies with both gender and skill.

It can also be observed that the conditional distribution of wages in the public sector is more compressed than the wage distribution in the private sector, i.e. the wage distribution has a lower standard deviation in the public sector. We suggest that the endogenous selection problem is likely to contribute to these differences. The inclusion of the correction terms in the estimation suggests that taking into account the selection mechanism increases the size of the gap due to characteristics and reduces the unexplained part. Examining the whole wage distribution, it emerges that the gaps with the corrections are in general somewhat smaller than without them, but very similar. On the other hand, the explained part of the gaps is larger when the correction is used.

According to sector pay gap, the unexplained wage gap (coefficient effect) estimated varies with the quantile chosen, and the characteristics gap seems to be more stable. Moreover, the unexplained part is decreasing along the wage distribution up to the median for women (and up to quantile 0,6 for men). Finally, also taking into account the characteristics effect, our results show that, as expected, the wage gap for public employees is especially large for low-income earners. At the upper end of the wage distribution the gap is narrower. In Spain, low-skilled public sector workers are paid higher wages than their private sector counterparts, but the reverse is true for high-skilled workers. These effects are more pronounced for women.

In short, the main findings are that both men and women earn more in the public sector, but this difference is much greater for women. However, the gender pay gap between the public and private sectors decreases as one moves up the wage distribution to higher levels. In general, wage gap estimates suggest that individuals are better off working in the public sector, especially in the lowest deciles. The opposite is true for men in the highest deciles. The evidence seems to indicate that there is a 'glass ceiling effectín private sector pay for women (in the top deciles) and a "low floor effect"for the private sector pay of low-skilled women.

These results have important policy implications: first, empirical evidence confirms that the public sector is a *fair employer*, as it has both a lower gender pay gap and a more compressed pay dispersion than the private sector, but governments should consider how to continue to reduce gender gaps in the public sector. Moreover, the existence of a positive public-private pay gap across most of the wage distribution also means that the public sector pays more than the opportunity wage for low-skilled labour, but less than what is needed to attract, retain and motivate high-skilled workers. Secondly, the differences in the private sector are highly significant, so policy-makers need to propose measures to achieve equal pay in private companies.

Capítulo 4

Capítulo 4

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

The objective of this paper is to analyse the effect of the tenure at job in Public Sector wages. To this end, a semiparametric model is proposed to solve a problem subject to sample selection and endogeneity. Here we face two problems: firstly, the public sector wages are only observed for public employees and in this sense decisions about participation are endogenously determined. Moreover, the second problem is that the regressor of interest is endogenous because there exists unobserved components related with wages which are also correlated with tenure. The chosen method is based on the usage of predicted covariates and also allowing the introduction of a nonparametric control function. To this end, we provide an estimation procedure based on an extension of the popular two-step model but relaxing usual strong assumptions about the functional form of the reduced equations in the first step. This approach is applied to a wage determination model for the Spanish Public Sector. Using the recent wage structure survey of 2016, we find a positive wage premium in Public Sector for men, those with more education and also with more tenure.

4.1. 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. Human capital theory ((Mincer, 1958), (Becker, 1964)) establish that education and training or also tenure at job produce an increment in the productivity of individuals improving their skills and knowledge and consequently deriving in higher earnings. That means that under human capital theory, experience and education are the more important factors for the economic situation of workers. In this context, the so-called Mincerian wage model has been used in several studies to analyse the wages differences due to working in public or private sector and between males and females ((Quinn, 1979) and (Shapiro y Stelcner, 1989) among many others). Usually, the purpose of researchers is to find a relationship between the variable of interest and some explanatory variables for a concrete population. Only if the samples are obtained randomly the results are able to be extrapolated to the entire population. 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 the field of Applied Econometrics and Labor Economics. The difficulty here occurs because wages in Public Sector are only observed for people actually working in this sector. Moreover, the introduction of tenure at job in the structural model creates problems in its specification and estimation. It is necessary to deal with two different problems, so that we have to specify a three equation labor supply model (wage equation, participation process and endogeneity). 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 the problem (see Vella (1993) for a survey).

In reference to the econometric methods, nonparametric estimation in sample selection models or with endogenous regressors are used to provide more flexibility by reducing the necessary assumptions and avoiding mispecification by using smoothness and regular conditions on the densities and functions. The disadvantages of these techniques are, as we have mentioned above, that nonparametric estimates are usually more imprecise having a convergence rate slower that $N^{1/2}$, and additionally its interpretation is also more difficult. The so-called control function approach is one of the preferable methods to solve the mentioned problems ((Wooldridge, 2015)). Using it as a possible solution, the final model becomes a partially linear regression model. Partially linear additive models are becoming so popular in semiparametric regression analysis, as is the case of (Fernandez et˜al., 2001) who showed 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 y 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) it is proposed a nonparametric estimator of the structural model of labor supply.

In this paper, we are concerned with getting of a root -N consistent estimation of the structural parameters of a sample selection model with an endogenous regressor. To this end, the nonparametric component we are going to obtain in the final model is the so-called control function, which is left unrestricted and is also built with nonparametric first step estimates. The estimation procedure is thus, an extension of the usual two-step approach but relaxing the assumed distribution for the error terms and no assuming a concrete functional form for the control function. In order to do so, we carry out a first-stage estimation of the correction terms with kernel regression methods. It can be built as conditional expectations (such as the so-called propensity score) or first-stage residuals or sometimes a mixture of both. Then, is a second step we extend the pairwise difference approach of (Honore y Powell, 1994) to estimate the effect of the regressors of interest by using as correction an unknown function of the first step estimations.

As empirical application, our proposal consists in extending human capital theory to analyse the influence of factor associated with Public Sector earnings. Thus, the selectivity problem arises because factors determining public sector participation are correlated with those determining wages. Besides, endogeneity is due to the presence of the variable "tenure"which usually depends on unobserved factors as the ability of workers which is also affecting wages and thus included in the disturbance term. We use microdata of Spanish population to estimate the effect of tenure on Public Sector wages accounting for sample selection of public employees. People working in public sector are differently paid, usually higher, than the ones in private sector. Government is directly associated with wages fixation in public sector, and contrary to firms which want to maximize profits, have other objectives as obtain votes and adjusting public budget. In order to limit the influence of Politicians about Public wages, its fixation is usually rigid and based on professional scales. Moreover the analysis of public employment is relevant because it represents about the 16 % of Spanish workforce. Additionally, the economic conditions that arose in the 2008 financial crisis have motivated the analysis of the earnings of public employees. It is something evident that Public employment is attractive for workers due to its benefits and stability, so that compensation here consists in something more than just wages. Furthermore, the entrance process and the tenure practices serve employees to obtain more security. According to this, the goal of our empirical application is to characterize how Public wages change in response to changes in the tenure of workers. It is expected that due to the working of the professional carrier incentives and public promotion the relationship between tenure and wages would be positive. In general, government focus on the recruitment of the more qualifies candidates and thus, the entrance to Public Sector in Spain is based on the realizations of some exam and the associated with the education level. In the other side, labor experience is expected to be negative related with more years of education due to opportunity costs. There exists studies which focus on the analysis of private-public wage gap differences as in (Smith, 1976) in the case of the United States, (Melly, 2006) in the case of Germany or (Garcia-Perez y Jimeno, 2007) for the Spanish regions.

The main contribution of this paper is related with the statistical modelling proposing a multi-step process to tackle simultaneously with endogeneity and selection. To our knowledge there is not an study focused on analysing the effect of tenure in Spanish Public Sector wages as is proposed here accounting for endogeneity and selection in a flexible framework. Our results show that sample selection bias is present in this model. We have also detected the presence of endogenity associated with the variable "Tenure.ªnd then controlling for both problems simultaneously is necessary to obtain consistent estimates.

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 faced an the possible solutions. In this section, we describe estimation procedure and then the estimator is proposed. In section 3 the data we use is described and we apply our approach to the study of public sector wages. Using our main findings we discuss the effect of tenure has on individuals and the importance of selection bias and endogeneity in our results. Finally, conclusions and policy implications are presented.

4.2. Model and estimation procedure

4.2.1. 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:

$$y_i = \begin{cases} y_i^* & for \quad p_i = 1, \\ 0 & otherwise. \end{cases} \tag{4.1}$$

$$p_i = 1\left(p_i^* > 0\right), \tag{4.2}$$

where

$$\begin{aligned} y_i^* &= l(w_{11i}, s_i) - \eta_{1i}, & (4.3) \\ p_i^* &= m_1(w_{11i}, w_{3i}) - \eta_{3i} & (4.4) \end{aligned}$$

We assume that individuals decide to work in Public Sector or not before the realization of 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 equation 4.2. 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 starred variables, (y_i^*, p_i^*), stand for unobservable variables and s_i is the endogenous regressor, defined as:

$$s_i^* = m_2(w_{11i}, w_{2i}) - \eta_{2i} \tag{4.5}$$

for $i = 1, \cdots, N$. Here, $(w_{11i}, w_{2i}, w_{3i})$ is a $(k_1 + k_2 + k_3)$-vector of observed explanatory variables and the vector $\eta_i = (\eta_{1i}, \eta_{2i}, \eta_{3i})$ is an unobserved set of idiosyncratic error. This kind of specification is often used in many theoretical problems and is really popular in labor economics and it is an extension of the seminal work of (Gronau, 1974).

Let us to consider that wage equation is linear, $w_i^* = w_{11i}^\top \beta_1 + s_i\lambda_1 + \eta_{1i}$. Then, taking conditional expectations we have that:

$$E\left(y_i | w_{11i}, s_i, p_i^* > 0\right) \quad = \quad w_{11i}^\top \beta_1 + s_i\lambda_1 + E\left(\eta_{1i} | w_{11i}, s_i, p_i^* > 0\right), \quad i = 1, \cdots, N \tag{4.6}$$

Under some assumptions, i.e.

1. The functions $m_1(\cdot)$ and $m_2(\cdot)$ are 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.

Then, adding some usual restrictions in the variance-covariance matrix, the quantity $E\left(\eta_{1i} | w_{11i}, s_i, p_i^* > 0\right)$ 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 more details). On the other side, if we are not willing to assume previous assumptions, then the quantity $E\left(\eta_{1i} | w_{11i}, s_i, p_i^* > 0\right)$ remains unknown and the popular least squares estimation of the parameters of interest is unfortunately unfeasible.

That means that like mispecification of the parametric form 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 then misspecified and therefore second stage estimators are asymptotically biased. Based on single index methods, several approaches have been proposed in recent literature to solve those problems (see (Ahn y Powell, 1993)). The main advantage is that in these methods we do not have to know the conditional distribution of the errors given the explanatory variables, so that they are robust to mispecification in the error distribution.

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 model. Then, the parameters β_1 and λ_1 of the main equation may be estimated through the corrected regression equation. So that, we need to compute $E\left(\eta_{1i}|w_{11i}, s_i, p_i^* > 0\right)$. The objective for estimation and identification of this model is to include the propensity score jointly with the estimated η_{2i} in the correction mechanism.

To this end, we assume that $E[\eta_{1i}|s_i] = E[\eta_{1i}|w_{11i}, w_{2i}, \eta_{2i}] = E[\eta_{1i}|\eta_{2i}] = f_1(\eta_{2i})$. So that, we can estimate a kernel regression for $s_i = m_2(w_{11i}, w_{2i}) + \eta_{2i}$ and obtain the residuals, $\hat{\eta}_{2i}$. After this, following the ideas of Das et al. (Das et~al., 2003), if p_j is a binary variable, the selection correction can be expressed as a propensity score. Let $p_j = E[p_j|w_{11j}, w_{3j}, w_j] = Pr(p_j = 1|w_{11j}, w_{3j}, w_j)$ denote the propensity score, p_1.

$$\begin{aligned} & E\left(\eta_1|w_{11}, s, p^* > 0\right) \\ & = E\left(\eta_1|w_{11}, s, \eta_3 > -m_1(w_{11}, w_3)\right) \\ & = E\left(\eta_1|\eta_2, u < p_1\right) \end{aligned} \tag{4.7}$$

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

Then, we can write 4.6;

$$E[y_i|w_{11i}, s_i, p_i^* > 0] = w_{11i}^\top\beta_1 + s_i\lambda_1 + \mu_1(\eta_{2i}, p_{1i}) \tag{4.8}$$

And the final model may be written as;

$$y_i = w_{11i}^\top\beta_1 + s_i\lambda_1 + \mu_1(\eta_{2i}, p_{1i}) + \epsilon_i \tag{4.9}$$

where μ_1 is the control function, which depends on p_{1i} and η_{2i}. 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). Here, we make pairs (i, j) that fulfill

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

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 differences process. It seems adequate to use the kernel weighting method covering the whole sample. Thus, 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 $i \neq j$

$$y_i - y_j = (w_{11i} - w_{11j})^\top \beta_1 + (s_i - s_j)\lambda_1 + (\epsilon_i - \epsilon_j). \tag{4.10}$$

where

$$\epsilon_i = y_i - E\left(y_i|w_i, s_i, p_i^* > 0\right), \quad \epsilon_j = y_j - E\left(y_j|w_j, s_j, p_j^* > 0\right),$$

If we denote $\beta = (\beta_1, \lambda_1)$ and $w_1 = (w_{11}, s)$. Then, an estimator the parameters β in the structural wage equation is:

$$\tilde{\beta} = \tilde{S}_{ww}^{-1}\tilde{S}_{wy} \tag{4.11}$$

where:

$$S_{ww} = \binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j (w_{1i} - w_{1j})(w_{1i} - w_{1j})^\top \tag{4.12}$$

$$S_{wy} = \binom{N}{2}^{-1} \sum_i \sum_{i<j} \omega_{ij} p_i p_j (w_{1i} - w_{1j})(y_i - y_j) \tag{4.13}$$

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

$$\omega_{ij} = \frac{1}{h^2} K\left(\frac{z_i - z_j}{h}\right) p_i p_j \tag{4.14}$$

where $z_i - z_j = (p_{1i} - p_{1j}\eta_{2i} - \eta_{2j})^\top$, $K(\cdot)$ is a Kernel function and h is the corresponding bandwidth. In this case, $\tilde{\beta}$ is unfeasible because it is a function of unknown quantities. So that we propose some estimators for those unknown quantities:

$$\hat{p}_{1i} = E\left(p_i|w_{11i}, w_{3i}\right) = m_1\left(w_{11i}, w_{3i}\right) \tag{4.15}$$

$$\hat{\eta}_{2i} = s_i - m_2\left(w_{11i}, w_{2i}\right) \tag{4.16}$$

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 nonparametrically 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 $\hat{m}(.)$ where the density estimator is then the kernel estimator. So that, we define the estimated probability as:

$$\hat{m_1}(w_{11}, w_3) = \hat{p}_1(x) = \frac{\hat{t}(x)}{\hat{b}(x)} = \frac{\frac{1}{Nh^{k_1+k_2}} \sum_{j=1}^n K\left(\frac{x-x_j}{h}\right) p_j}{\frac{1}{Nh^{k_1+k_2}} \sum_{j=1}^n K\left(\frac{x-x_j}{h}\right)} \tag{4.17}$$

where we have denoted $x = (w_{11}, w_3)$, $K(.)$ 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 y Powell, 1993)).

$$s_i - \hat{m}_2(w_{11}, w_{21}) = \hat{\eta}_2(w) = \frac{\hat{t}(w)}{\hat{b}(w)} = \frac{\frac{1}{Nh^{k_1+k_2}} \sum_{j=1}^{n} K\left(\frac{w-w_j}{h}\right) s_j}{\frac{1}{Nh^{k_1+k_2}} \sum_{j=1}^{n} K\left(\frac{w-w_j}{h}\right)} \tag{4.18}$$

Replacing the unknown quantities by its estimates we propose the following estimator:

$$\hat{\beta} = \hat{S}_{ww}^{-1} \hat{S}_{wy} \tag{4.19}$$

where ;

$$\hat{S}_{ww} = \binom{N}{2}^{-1} \sum_{i} \sum_{i<j} \hat{\omega}_{ij} p_i p_j (w_{1i} - w_{1j})(w_{1i} - w_{1j})^{\top} \tag{4.20}$$

$$\hat{S}_{wy} = \binom{N}{2}^{-1} \sum_{i} \sum_{i<j} \hat{\omega}_{ij} p_i p_j (w_{1i} - w_{1j})(y_i - y_j) \tag{4.21}$$

And:

$$\hat{\omega}_{ij} = \frac{1}{h^2} K\left(\frac{\hat{z}_i - \hat{z}_j}{h}\right) p_i p_j \tag{4.22}$$

4.2.2. Large sample properties

In this section we are going to establish the asymptotic properties of the proposed estimator, $\widetilde{\beta}$. In order to develop this proofs we follow the lines explained in Ahn y Powell (1993), Powell (2001) and Aradillas-Lopez et˜al. (2007). The derivation of the large sample properties of $\widehat{\beta}$ proceeds in two steps. In a first moment, the asymptotic behavior of the related estimator $\widetilde{\beta}$ defined in (4.11) is analyzed. Secondly, it is necessary to investigate the difference $\widetilde{\beta}$ and $\widehat{\beta}$. The proofs of all results are detailed in the Appendix. The expressions for $\widetilde{\beta}$ and $\widehat{\beta}$ can be rewritten as,

$$\sqrt{N}\left(\widetilde{\beta} - \beta\right) = \sqrt{N} S_{ww}^{-1} S_{wu}, \tag{4.23}$$

where

$$\begin{aligned} S_{wu} &= \binom{N}{2}^{-1} \sum_{i} \sum_{i<j} K_h\left(\frac{z_i - z_j}{h}\right) p_i p_j \\ &\times \ (w_{1i} - w_{1j}) \left\{ \left(\mu_1 \left(p_{1i}, \eta_{2i}\right) - \mu_1 \left(p_{1j}, \eta_{2j}\right)\right) + \epsilon_i - \epsilon_j \right\} \\ &= \binom{N}{2}^{-1} \sum_{i} \sum_{i<j} K_h\left(\frac{z_i - z_j}{h}\right) p_i p_j \\ &\times \ (w_{1i} - w_{1j}) \left(u_{1i} - u_{1j}\right) \end{aligned} \tag{4.24}$$

where now $K_h(\cdot) = \frac{1}{h}K(\cdot)$ and

$$u_{1i} \equiv \mu_1(p_{1i}, \eta_{2i}) + \epsilon_i \tag{4.25}$$

It is logical that the estimator $\hat{\beta}$ satisfies:

$$\sqrt{N}\left(\hat{\beta} - \beta\right) = \sqrt{N}\hat{S}_{ww}^{-1}\hat{S}_{wu}, \tag{4.26}$$

where

$$\begin{aligned} \hat{S}_{wu} &= \binom{N}{2}^{-1} \sum_i \sum_{i<j} K_h\left(\frac{\hat{z}_i - \hat{z}_j}{h}\right) p_i p_j \\ &\times \left(w_{1i} - w_{1j}\right)\left(u_{1i} - u_{1j}\right) \end{aligned} \tag{4.27}$$

In order to proof the properties for $\widetilde{\beta}$, it is necessary to establish some regularity conditions:

(A.1) The vectors $\left(y_i, p_i, s_i, w_i^\top\right)^\top$ satisfying (4.1), (4.2) and (4.3) are independent and identically distributed across i having bounded support and the remaining components having finite sixth-order moments.

(A.2) The propensity score p_1 and the residuals η_2 have joint and marginal densities: $f(\cdot)$, that are absolutely continuous and bounded from above.

(A.3) Let us define, Σ_{ww}, where:

$$\begin{aligned} \Sigma_{ww} &= E\left[\rho_{1i}^2 \gamma_i f(p_{1i}, \eta_{2i})\left(w_{1i} - \phi(p_{1i}, \eta_{2i})\right)\right. \\ &\left. \times (w_{1i} - \phi(p_{1i}, \eta_{2i}))^\top\right] \end{aligned} \tag{4.28}$$

Σ_{ww} is nonsingular. Moreover,

$$\begin{aligned} \rho_{1i} &= E\left[p_i | p_{1i}, \eta_{2i}\right], \\ \gamma_i &= E\left[s_i | p_{1i}, \eta_{2i}\right], \end{aligned} \tag{4.29}$$

and

$$\phi(p_{1i}, \eta_{2i}) = \frac{E\left[p_i w_{1i} | p_{1i}, \eta_{2i}\right]}{\rho_{1i}\gamma_i} \tag{4.30}$$

(A.4) The kernel function is compactly supported, bounded kernel such that $\int uu^\top K(u)du < \infty$. In addition, all odd-order moments of K vanish, that is $\int u_1^{\iota_1} u_2^{\iota_2} K(u)du = 0$ for all non-negative integers ι_1, ι_2 such that their sum is odd. Furthermore, $K(u)$ is twice continuously differentiable with

$$\partial^2 K\left(u_1, u_2\right) / \partial u_l^2 < \kappa_0, \quad \text{for} \quad l = 1, 2.$$

(A.5) The bandwidth sequence h used to define the weights $\widehat{\omega}_{ij}$ and ω_{ij} is of the form

$$h \sim N^{-\delta},$$

where $\delta \in \left(\frac{1}{8}, \frac{1}{6}\right)$.

(A.6) The density functions, $\phi(\cdot)$, ρ_1 and γ (defined in Assumption (A.3) above), and the function μ_1 and its derivatives are all fourth-order continuously differentiable, with derivatives that are bounded.

Assumptions from (A.1) to (A.6) are similar to the ones in Ahn y Powell (1993). The third assumption is for identification, needed to identify $\widetilde{\beta}$. The fourth is for the kernel function and it is also necessary to impose the bandwidth rate to achieve the root-N consistency of $\widehat{\beta}$. Under these assumptions:

Lemma 4.2.1 *Under assumptions (A.1) to (A.6), as N tends to infinity,*

(i)

$$S_{ww} = 2\Sigma_{ww} + o_p(1), \tag{4.31}$$

where Σ_{ww} is defined in assumption (A.3).

(ii)

$$\begin{aligned} S_{wu} &= \frac{2}{N}\sum_i \left[p_i \rho_{1i}\gamma_i f(p_{1i}, \eta_{2i})\left(w_{1i} - \phi(p_{21i}, \eta_{2i})\right)\right]\epsilon_i \\ &+ o_p(\frac{1}{\sqrt{N}}). \end{aligned} \tag{4.32}$$

Now, we need to introduce some additional assumptions to show the asymptotic representation for $\widehat{\beta}$. The first step is to obtain the asymptotic representation of $\widetilde{\beta}$ (see Lemma 4.2.1 above) and secondly, to analyze the difference between $\widehat{\beta}$ and β. In order to do so, we need to introduce some additional assumptions that enables us to achieve some uniform bounds for $\widehat{p}_1$ and $\widehat{\eta}_2$.

(A.7) The Mth-order kernel function, $K_1(u)$, satisfies:

- $K_1(u)$ is bounded and integrable. $|K_1(u)| \leq \bar{K} < \infty$ and

$$\int_{R^{k_1+k_2+k_3}} |K_1(u)|\, du \leq \tau < \infty.$$

- For some $\Lambda_1 < \infty$ and $L < \infty$, either $K_1(u) = 0$ for $\|u\| > L$ and for all $u, u' \in R^{k_1+k_2+k_3}$

$$\|K_1(u) - K_1(u')\| \leq \Lambda_1 \|u - u'\|.$$

- $\int |u|^{M+1} |K_1(u)|\, du < \infty$.

(A.8) Let $f(w)$ the joint density of the vector w. We assume that the M-th derivative of $f(w)$ is uniformly continuous and that $\sup_w \|w\|^q f(w) < \infty$, for some $q > 0$.

(A.9) The second derivatives of p_1 and η_2 are uniformly continuous and bounded.

(A.10) For some $s > 2$,

$$\begin{aligned} E\left(|\, p_i\,|^s\middle|\, x{=}x_1\right) f(x_1) < B_1 < \infty, \\ E\left(|\, s_i\,|^s\middle|\, w = w_1\right) f(w_1) < B_2 < \infty. \end{aligned}$$

(A.11) The bandwidth sequence h_1 used to define the first step estimators is of the form,

$$h_1 \sim N^{-M/2},$$

for M given in assumption (A.7).

Assumptions (A.7) to (A.11) are sufficient conditions to obtain uniform bounds for the estimators. See for example Hansen (2008).

Teorema 4.2.1 *Under assumptions (A.1)-(A.11)*

$$\widehat{S}_{ww} = S_{ww} + o_p(1), \tag{4.33}$$

$$\begin{aligned} \sqrt{N}\left(\widehat{S}_{wu} - S_{wu}\right) &= \frac{2}{\sqrt{N}} \sum_{i=1}^{N} \rho_{1i}^2 \gamma_i f(p_{1i}, \eta_{2i}) D\mu_1(p_{1i}, \eta_{2i}) \\ &\quad \times (w_{1i} - \phi(p_{1i}, \eta_{2i}))\,[p_i - p_{1i}]\,[s_i - \eta_{2i}] \\ &\quad + o_p\left(N^{-1/2}\right), \end{aligned} \tag{4.34}$$

where

$$D\mu_1(p_{1i}, \eta_{2i}) = \partial\mu_1\left(p_{1i}, \eta_{2i}\right)/\partial p_{1i} + \partial\mu_1\left(p_{1i}, \eta_{2i}\right)/\partial\eta_{2i}. \tag{4.35}$$

Corolario 4.2.1 *Under assumptions (A.1)-(A.11) the estimator $\widehat{\beta}$ in eq.* 4,18 *has the following asymptotic linear representation,*

$$\begin{aligned} \sqrt{N}\left(\widehat{\beta} - \beta\right) &= \Sigma_{ww}^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \rho_{1i}\gamma_i f(p_{1i}, \eta_{2i})\,(w_{1i} - \phi(p_{1i}, \eta_{2i}))\,\{p_i\,\text{či} \\ &+ \rho_{1i} D\mu_1(p_{1i}, \eta_{2i})\,(p_i - p_{1i})\,(s_i - \eta_{2i})\} \\ &+ o_p(1). \end{aligned} \tag{4.37}$$

and is asymptotically normal,

$$\sqrt{N}\left(\widehat{\beta} - \beta\right) \to_d N\left(0, \Sigma_{ww}^{-1}\Omega_1\Sigma_{ww}^{-1}\right), \tag{4.38}$$

where

$$\begin{aligned}\Omega_1 &= E\left\{\rho_{1i}^2\gamma_i^2 f^2(p_{1i},\eta_{2i})\left[p_i\epsilon_i + \rho_{1i}D\mu_1(p_{1i},\eta_{2i})\right.\right.\\ &\quad \left.\left.\times\left(p_i - p_{1i}\right)\left(s_i - \eta_{2i}\right)\right]^2\left(w_{1i} - \phi(p_{1i},\eta_{2i})\right)\left(w_{1i} - \phi(p_{1i},\eta_{2i})\right)^\top\right\}.\end{aligned}$$

4.2.3. Steps to obtain the estimates for the structural parameters

1. First step: We estimate the reduced equation for the endogenous regressor, which is a continuous variable and obtain the residuals, $(\hat{\eta}_{2i})$. In this step, we solve the problem of endogeneity.

2. Second Step: We estimate the public sector participation non-parametrically and obtain the predicted values, $\hat{p}_1$. We consider that expected wages affect participation decision so that here it is also necessary to account for the endogeneity of s. In this step we deal with the problem of selection bias.

3. Third step: With predicted values from previous steps we can built the required control function for correcting from both endogeneity and selectivity and use the unrestricted function as additional regressor in the final model. Thus, the structural equation can be estimated.

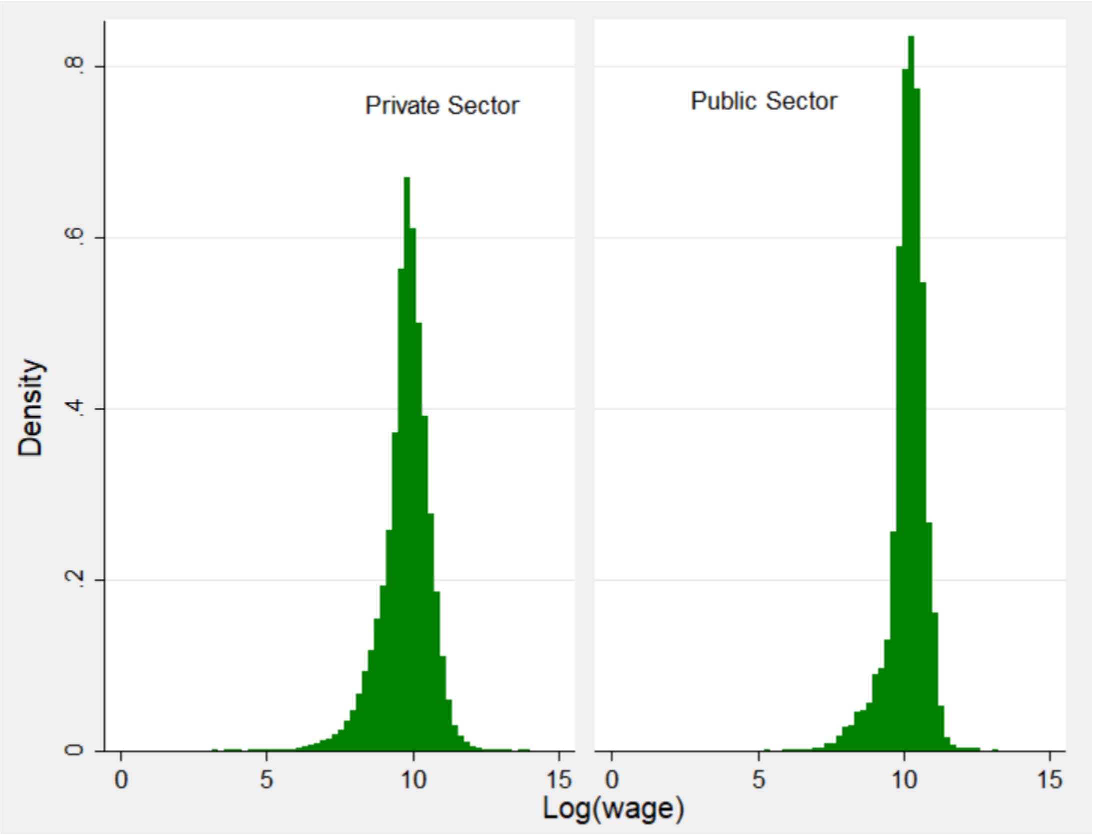
4.3. An Application for estimating Spanish Public Sector employees's wages

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 wage structure will depend on workers personal characteristics (demographic or productive such as age, tenure and education) but also on a combination of political and economic factors. Most of studies conclude that there exists a positive wage premium for public sector employees, (Giordano et˜al., 2011).

We are interested here in studying the effects that some variable, concretely the tenure at job, has on public sector wages. The interest of analysing the Spanish public sector is due to private sector payments are related with firm's profits while the public sector is subject to political constraints. So that, public sector wages can be seen as an instrument of economic policy, and it represents a great amount of Spanish population (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 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. We consider then, that there is a previous selection process for entering Public Sector, an individual first decision is 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. We are going to estimate firstly the tenure and the participation equations non parametrically in order to obtain the correction terms that will be included in the unrestricted control function. Public wages are determined by the employer which can be local corporations, regional authorities or the central government. There also exist two type of public employees, the civil servant and the so-called ”personal laboral”, which, as we mentioned above the later are subject to usual labor legislation.

Figura 4.1: **Logarithm of wages by sector.**



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

As can be seen in graph 4.1, 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), in average, in the Private Sector. Therefore, it seems reasonable to think that the sector: Public or Private is one of the most important factors determining the wages.

4.4. Data

4.4.1. 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 in average of $29,458$ Euros in the Public Sector and $23,911$ Euros in the Private Sector. The data have been initially segmented between the private and public sectors in order to analyse its main factors. Thus, the variables in the model are defined and summarized in table 4.1.

Table 4.1: **Statistics of the variables: mean, and standard deviation in brackets.**

	Variables definition	Public Sector	Private Sector
Wage	Annual Wage, in Euros	29458,72 (17480.54)	23911,49 (22521.98)
Age1	less than 19	0,000 (0.025)	0,001 (0.039)
Age2	Age from 20 to 29	0,049 (0.216)	0,120 (0.325)
Age3	Age from 30 to 39	0,235 (0.424)	0,336 (0.472)
Age4	Age from 40 to 49	0,320 (0.466)	0,309 (0.462)
Age5	Age from 50 to 59	0,314 (0.462)	0,187 (0.390)
Age6	Age more than 60	0,079 (0.270)	0,04 (0.204)
Primary	1, if having elementary studies	0,049 (0.216)	0,154 (0.361)
High School	1, if High School	0,467 (0.498)	0,568 (0.495)
Sup	1, if having Higher Education	0,478 (0.499)	0,262 (0.440)
Tenure	Years of experience	14,25 (10.76)	9,18 (9.27)
Comsal	Complements to wage, in Euros	1000,71 (1102.26)	534,41 (909.21)
Male	1, if the person is male	0,468 (0.499)	0,592 (0.491)

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

The variables included in this data set are defined in 4.1 including some basic statistics. Further, we certainly have the information whether a person work in PUBLIC Sector or not. The data presented in Table 1 in this initial diagnosis show that, in 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, more tenure and more incidence of women. Annual earning are on average higher in Public Sector than in the Private one.

4.5. Results

In this section we present the results we have obtained for the wages model. In order to implement our estimator, it is necessary to obtain ”first-step.$^{\text{es}}$timates of p_1 and η_2. After having obtained those estimates non parametrically, we can proceed with the next step. To give further insights to our approach we have estimated the wage equations under some of the most usual methods. In table 4.2 we can observed the estimates for the uncorrected Ordinary Least Squares estimation without any correction in the first column. In the second column we have the results for Heckman two step standard approach with the Inverse Mills ratio as additional regressor to correct for the sample selection separately. 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, these estimates may be even worse than using ordinary least squares. That is due to the fact that if bivariate normality distribution of errors fails and we have used the inverse mills ratio as control function, then the results are inconsistent.

The results we have obtained reflects something similar to the presented in the literature. We can see in table 4.2 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 also more tenure is positive related with wages, as its consistent with the beliefs of the economic theory. Furthermore, we find that the inverse of Mills ratio is statistically significant in column 2 and also the residual component in column 3, so that we suspect that a correction is needed because we have evidence of sample selection and endogenity. This gives an important relevance to the method proposed in this article, in which we include a correction term built under flexible assumptions quite different from standard proceeds.

Table 4.2: **Estimation Results for the Logarithm of wages in Public Sector with some standard parametric approaches.**

	OLS	**Heckman 2-steps. 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)
Men	$0,156^{***}$ (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.

To sum up, table 4.2 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. Due to those findings, 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 correcting sample selection and endogenity in the second stage. So that, we propose to estimate the reduced equations in a more flexible way.

Table 4.3: **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)
Men	$0,156^{***}$ (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 4.3 supports our idea of left 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 introduce the quadratic terms and it seems to proof some evidence about selection bias with the significance of the propensity score squared. We now consider the control function as nonparametric. To avoid mispecification, as we have explained in this article it is preferable to left the function unrestricted, and use the pairwise difference technique to obtain consistent estimates in wages equation. Then, we first estimate the tenure and participation equations. To estimate the model we use a nonparametric estimation for the reduced equations, $f(.)$ and $m(.)$ are estimated allowing us to obtain $\hat{p}_{1i} = E[p_i|m_1(.)]$ and $\hat{\eta}_{2i} = s_i - E[s_i|m_2(.)]$. In addition to our initial bandwidth choice guided by cross validation criterion we have selected other bandwidth parameters, 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 4.4: **Estimation Results for the Logarithm of wages in Public Sector under pairwise difference method. Unrestricted Control Function.**

	Coeff.	**Std.Err.**	**t-statistic**
Age from 50 to 59	0,014	0,010	1,35
Men	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.

The final results are presented in 4.4 and show 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 level of studies is one of the most important variables affecting Public wages, because having superior studies is a necessary condition to access to some official scales and to promote. So that, people with higher level of studies are expected to increase their wage in a 32 % 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 reflects 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. Moreover, women earn in average less than men also in Public Sector.

4.6. Conclusions

The objective of this article is to analyse returns to human capital 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 relative to private ones has been positive in most OECD countries in the last four decades. During the last decades estimation techniques have been progressively adapting to deal with the implicit problems that economic models contain. Using the wage structure wage survey for 2016, we have to deal with two sources of bias in our model, the sample selection and the existence of an endogenous regressor.

While the empirical findings are interesting, the key contribution of this article is the modeling approach and its implementation. Then, to demonstrate the usefulness or our approach we apply it to investigate the eects of some characteristics on public sector wages. 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 endogenity 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 of the main equation. The advantages of our proposal are the flexibility of the model 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 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. The results obtained support the idea that returns to tenure at job are positive and significant. Moreover, we have shown that men, high-skilled workers and older ones earn more money. The evidence suggests that it is necessary to deal with selection and endogeneity of tenure to obtain consistent estimates. To show that, we compare our approach with others in order to present the main differences. These results have also important policy implications, in this sense, policymakers have to ensure that public wages are determined for being competitive with private ones. In other way it would be difficult to hire the better qualified workers. This is an important implication because without productive workers, the quality of public services is going to decrease. On the other side, it is important for authorities to find a good equilibrium because a rise in public sector wages in expansion cycles may be difficult to maintain in periods of recession.

Conclusiones

Conclusiones

El avance en la investigación de nuevas propuestas metodológicas ha permitido considerablemnte mejorar la robustez de los resultados de estomación en la aplicaciones empíricas cientificas. Concretamente en modelos de oferta laboral de colectivos con características específicas como personas con dicapcidades, enfermedades crónicas requieren la aplicación de métodos apropiados que nos permitan explotar sus peculiaridades.

En este sentido, en esta tesis se ha abordado un doble objetivo. La propuesta de nuevos enfoques para especificar y estimar de manera consistente modelos sujetos a selección muestal y endogenidad en un marco teórico lo más flexible posible. En segundo lugar, demostrar la utilidad de dichos métodos econométricos a través de aplicaciones empíricas de interés en la actualidad. Concretamente nos centramos en cuatro aplicaciones que van desarrolandose en cada uno de los capítulos de esta tesis.

En esta sección, y a modelo de resumen general se recogen los principales objetivos marcados en cada uno de los capítulos y se presentan los principales resultados obtenidos en cada una de las partes que conforman la tesis doctoral.

En el primer capítulo se propone un modelo de oferta laborar para las personas con discapacidad. En este sentido, dada la baja participación de este colectivo se considera que su utilidad difiere de la usual. Dada la heterogenidad de las preferencias y el uso del tiempo dedicado a mejorar su salud resulta de interés analizar sus peculiaridades. Con este objetivo, se presenta un modelo de paquetes de trabajos en el cual no puede elegirse libremente cuánto o cómo se quiere trabajar sino unicamente elegir el empleo que mejor se adapte a nuestras preferencias de entre un conjunto de empleos disponibles. En este contexto, nuestro estudio a diferencia de los pre-existentes considera que las personas con discapacidad requieren más tiempo para dedicar a su salud y eso afecta claramente a sus deciones laborales. Nuestra conclusión principal es que la no participación es la elección menos flexible ante cambios de renta. El tiempo dedicado a cuidados tiene mucho peso para las personas con más edad y más limitaciones que prefieren no trabajar o hacerlo a tiempo parcial si no encuentran un "paquete"lo suficientemente flexible dadas sus preferencias. Fomentar opciones como el tele trabajo o la flexibildad horaria podrian ser soluciones ante la baja tasa de actividad y de empleo de este colectivo.

En el segundo capitulo se desarrolla un nuevo método de especificación y estimación de modelos sujetos a selección muestral y endogenidad, concretamente la existencia de un regresor endógeno binario. En este caso, se propone una extensión del modelo de Heckman con una función de control flexible que se deja sin restringir. Se desarrolla pues un modelo semiparamétrico en el cual no hay imposición de supuestos distribucionales entre los errores del sistema de ecuaciones involucradas. Se propone un estimador que acta diferenciando pares de observaciones parecidas en ciertas caracteristicas asociadas a la función desconocida. Para probar la utilidad de este estimador se estima una ecuación de salarios que contiene como regresor el hecho de padecer una enfermedad crónica, regresor binario endógeno. Se concluye que, las variaciones en los salarios ademas de las decisiones de participación en el mercado se ven afectadas por el nivel de educación y la estabilidad laboral. Adicionalmente se deduce que es necesario controlar la endogeneidad de la cronicidad en el modelo estructural. Dicha cronicidad muestra una relación negativa con el salario, con lo que acciones politicas han de considerarse dados los cambios demográficos y la incicipente evolución de la cronicidad.

En el capítulo 3, nuestro objetivo se centra en analizar a que se deben las diferencias salariales entre el sector Público y Privado. Más aún la finalidad del estudio es probar si la discrinación salarial por género es menor, como se espera a priori, en el Sector Público. Nuestros hallazgos muestran que efectivamente existe una discrimación en contra de las mujeres en ambos sectores, aunque de menor magnitud en el Público. Se propone un modelo sujeto a selección muestral y diferenciado por género en el cual se desagrgan los efectos de las discrepancias. Existe un diferencial positivo en los salarios de los empleados públicos frente a los del sector privado, este diferencial es mayor en el caso de las mujeres. Usando una regresión por cuantiles se muestra que esa diferencia es mayor para los que tienen salarios más altos (parte alta de la distribución). Además la distribución condicionada está más comprimida en el Sector Público que en el Privado. Se presentan pues algunas implicaciones políticas relevantes: Los gobiernos tienen que analizar la manera de reducir la becha salarial por género en el Sector Público, y en segundo lugar, deben tomarse medidas para llegar a la igual de pago y de oportunidades en las empresas privadas.

Una vez presentado en el capítulo 2 un modelo de especificación y estimación de un modelo de oferta laboral sujeto a selección y con un regresor binario endógeno, se presenta en el capitulo 4 el caso de que el regresor endógeno sea continuo. Concretamente nos centramos en el estudio del efecto que la antiguedad tiene en los salarios del Sector Público. El problema de la selección muestral viene dado porque solo observaremos salarios públicos para los empleados en el Secot público y la decisión de participar en dicho Sector no es aleatorio. La decisión de participar en el Secto Público conolleva un proceso de valorar por parte del individuo, si opta a entrar en él o no. Por otro lado, el regresor de interés que es la antiguedad del trabajador, es una variable endógena pues factores que influyen en el salario de lso empleados público no observables y por tanto incluídos en el componente de error, están relacioandos con la antiguedad de dicho trabajador (piensese por ejemplo en la habilidad que el trabajador he desarrolado en su puesto). El método de estiamción propuesto se basa de manera relacionada con el capítulo 2 de esta tesis en la introducció nde umna función de corrección no paramétrica en el modelo estructural. De esta manera se relajan los rigidos supuestos necesarios para el uso del popular modelo de Heckman en dos etapas. Los resultados del capítulo sugieren que existe un diferencial positivo en los salarios del Sector Público a favor de los hombres, de los trabajadores con más nivel de educación y también de los que cuentan con más tiempo de antiguedad.

Líneas de Investigación Futuras

A lo largo de esta tesis y una vez presentadas las ventajas que presenta la introducción de la función de control con forma desconocida en el modelo estructural han ido surgiendo futuras lineas de investigación. En este sentido, cuando se trata de analizar decisiones laborales de ciertos colectivos de interés en nuestra sociedad con ciertas peculiaridades como son las personas con discapacidad, con cronicidad o las mujeres han de tenerse en cuenta modelos econométricos capaces de lidiar con dichas particularidades. En los capítulos de esta tesis ha quedado demostrado la utilidad de emplear un estimador adecuado y flexible al marco que se pretende estudiar dadas las caracteristicas del conjunto de datos. En este sentido, se ha puesto de manifiesto que no es siempre necesario recurrir a supuestos sobre la especificacón del modelo tan restrictivos como en los modelos totalmente paramétricos. También se ha señalado que la flexibilidad mencionada tiene también su contrapunto, que es la obtención de estimadores con tasas de convergencia más lentas y más dificiles de interpretar.

En el primer capítulo sería interesante repetir el ejercicio propuesto para información más actualizada. Asimismo, tener en cuenta variables monetarias resultaría conveniente, sin embargo es complicado obtener bases que contemplen tiempo de cuidados y por ejemplo salarios para los mismos individuos. Tal vez, podra pensarse en imputar datos similares procedentes de otras fuentes.

En referencia al capítulo 2, un primer ejercicio puede consistir en la obtención de un método alternativo que permita la identificación directa del efecto del regresor binario endógeno, lo cuál no resulta trivial.

Una línea de investigación futura puede consistir en la extensión del modelo de especificación y estimación presentado en los capítulos 2 y 4 de esta tesis doctoral a un modelo de datos de panel. La disponibilidad de nuevos datos puede permitirnos tratar de ampliar nuestro enfoque, ahora centrado en la sección cruzada a métodos de panel.

Finalmente, en el capítulo 3 el análisis realizado pordía enriquecerse teniendo en cuenta las diferentes ocupaciones de los individuos. Desagregar si los diferenciales salariales son distintos entre diferentes países de la Unión Europea.

Conclusions

Conclusions

The progress in the investigation of new methodological proposals has considerably improved the robustness of the results of the estimation in the scientific empirical applications. Specifically in models of labor supply of groups with specific characteristics as people with disabilities or chronic diseases, it is required the application of appropriate methods that allow us to exploit their peculiarities.

In this sense, this thesis has addressed a twofold objective. The proposal of new approaches to specify and to estimate consistently models subject to sample selection and endogeneity in a theoretical framework as flexible as possible. Secondly, to demonstrate the usefulness of these econometric methods with the proposal of some empirical applications with interest nowadays. Specifically we focus our attention on four applications that are developed in each of the chapters of this thesis.

In this section, as a summarize we collect the main objectives pursued in each one of the chapters and we present a conclusion of the main results obtained in each one of the parts that composes this dissertation.

The first chapter proposes a model of labour supply for people with disabilities. In this sense, given the low participation of this group, it is considered that its utility differs from the usual. Given the heterogeneity of the preferences and the time spent to improve their health, it is of great interest to analyse their peculiarities. With this objective, there is a model of "work packagesïn which you can not choose freely how many time you want to work but only can choose the job-package that best suits your preferences among a set of jobs available. In this context, our study unlike the pre-existing ones, considers that people with disabilities require more time to devote to their health and that situation clearly affects their labor decisions. The main conclusion is that non-participation decision is the most rigid one in relation to income changes. That means, that time devoted to health care has an important weight in labor decisions for disabled and older people. In this context, if the job-packages are to rigid and they do not find any "package"so flexible given their preferences, they prefer not to work or only part-time. Promoting options such as tele-work or time-schedule flexibility could be solutions to the low rate of activity and employment of this group.

In the second chapter it is developed a new method of specification and estimation for models subject to sample selection and endogenity, specifically the existence of an endogenous regressor which is binary. In this case, an extension of the Heckman model is proposed with a flexible control function that is left unrestricted. It is developed a semiparametric model in which there is no imposition of distributional assumptions between the errors of the system of equations involved. Here, it is proposed an estimator that acts differentiating pairs of similar observations in certain characteristics associated with the unknown function. To prove the usefulness of this estimator is estimated a wage equation that contains as regressor the fact of suffering a chronic disease, which is a binary endogenous variable. It is concluded that, the variations in salaries in addition to the participation decisions in the labor market are affected by the level of education and the labor stability. Additionally it is deduced that it is necessary to control for the endogeneity of the chronicity in the structural model. This chronicity shows a negative relation with the salary, so that political actions are to be considered given the demographic changes and the increasing evolution of the chronicity.

In Chapter 3, our focus is on analyzing the wage differences between the public and private sector. Even more the purpose of the study is to test whether the salary discrimination by gender is less, as expected a priori, in the Public Sector. Our findings show that there is indeed a discrimination against women in both sectors, albeit of lesser magnitude in the Public one. It is proposed a model subject to sample selection and disaggregated by gender in which the causes of those discrepancies are analysed. There is a positive differential in salaries of public employees compared to those of the private sector, this differential is higher in the case of women. Using a quantile regression it is showed that this difference is greater for those who have higher wages (upper part of the distribution). Moreover, the conditioned distribution is more compressed in the Public Sector that in the Private. Some relevant policy implications are presented: governments have to analyse the way of reducing the wage gap between gender in Public Sector, and secondly it is necessary to implement some measures to obtain equal opportunities and payment in Private sector, in which women are clearly unfavourable treated.

Once it is presented in the chapter 2 a model of specification and estimation of a model of labor supply subject to sample selection and with a binary endogenous regressor, in chapter 4 it is presented the case that the endogenous regressor is continuous. Specifically, we focus on the study of the effect that tenure has on salaries in the Public Sector. The problem of the sample selection is given because we only observe public salaries for employees in the Public Sector and the participation decision is nor random. The decision to participate in the Public Sector it is guided by a process, the individual has to decide if it's worth to enter or not. On the other hand, the regressor of interest which is the tenure of the worker, is an endogenous variable because non observable factors that influence the salary of public employees and therefore included in the error component, are also correlated with the Employee's tenure (consider for example the skills that the worker has developed in his job with the pass of time). The proposed estimation method is based on the proposal of chapter 2 of this thesis and consist in the introduction of a non-parametric control function in the structural model. In this context, the idea is to relax some of usual strong assumptions in the popular two-step model of Heckman. The obtained results in this chapter, suggest that there exists a positive differential in Public Sector payment for men, workers with more education level and also for those with more years of tenure.

Future Research

Throughout this thesis and given the advantages of the introduction of the control function in an unknown form in the structural model, some future lines of research have arisen. In this sense, when we focus on analyzing the labor decisions of certain groups of interest in our society with certain peculiarities such as people with disabilities, with chronicity or women it is necessary to take into account econometric models able of dealing with such peculiarities. In the chapters of this thesis has been demonstrated the usefulness of using an estimator adequate and flexible given the features of the data set and the population of interest. In this sense, it has been shown that it is not always necessary to resort to assumptions so restrictive on model specification as in the fully parametric models. However, we have also shown that this flexibility has also its counterpoint, which is the obtention of estimators with slower convergence rates and more difficult to interpret.

In the first chapter it would be interesting to repeat the proposed exercise for more up-to-date information. Also, taking into account monetary variables would be convenient, however it is difficult to obtain databases that contemplate care time and for example salaries for the same individuals. Perhaps, one could think of imputing to similar observations, data from other sources.

With reference to chapter 2, a first exercise may consist in obtaining an alternative method that allows the direct identification of the effect of the endogenous binary regressor, which is not trivial. Additionally, a future research line may consist in the extension of the specification and estimation model presented in chapters 2 and 4 of this doctoral thesis to a panel data model. The availability of new data may allow us to try to broaden our focus, actually centered on cross-section, to panel methods. Finally, in chapter 3, the analysis carried out by the authors will be enriched taking into account the different occupations of the individuals. Disaggregate if wage differentials are different between European countries.

Anexos

Apéndice A

Apéndice Capitulo 2

For the proof of the stated results consider a second order U-statistic of the form

$$U_N = \binom{N}{2}^{-1} \sum_{i=1}^{N} p_N(\xi_i, \xi_j) \tag{A.1}$$

where $\{\xi_1, \cdots, \xi_N\}$ is an i.i.d. sample of random vectors and $p_n(\cdot, \cdot)$ is a symmetric kernel, i.e. $p_N(\xi_i, \xi_j) = p_N(\xi_j, \xi_i)$. Define

$$\begin{aligned} r_N(\xi_i) &= E[p_N(\xi_i, \xi_j)|\, \xi_i], \\ \theta_N &= E[r_N(\xi_i)] = E[p_N(\xi_i, \xi_j)], \end{aligned}$$

as well as the projection of the statistic U_N,

$$\hat{U}_N = \theta_N + \frac{2}{N} \sum_{i=1}^{N} [r_N(\xi_i) - \theta_N] \tag{A.2}$$

The following two lemmas will be used in the proofs of the main results:

Lemma A.0.1 *If the kernel $p_N(\cdot)$ satisfies; $E[||p_N(\xi_i, \xi_j)||^2] = o(N)$, then*

$$\begin{aligned} U_N &= \theta_N + o_p(1), && \text{(A.3)} \\ \widehat{U}_N &= U_N + o_p\left(N^{-1/2}\right), && \text{(A.4)} \end{aligned}$$

as N tends to infinity.

For a proof of Lemma C.0.1, see Powell et˜al. (1989), Lemma 3.1.

Lemma A.0.2 *Let $r_{21N}(\xi_i)$, $r_{31N}(\xi_i)$, $r_{20N}(\xi_i)$ and $r_{30N}(\xi_i)$ be random vectors valued functions of the data vector $\xi_i = \left(r_i, z_i, d_i, w_i^\top\right)^\top$, which satisfy*

$$\begin{aligned} \|r_{21N}(\xi_i)\| &= 0 \quad \text{iff} \quad w_{12i} \notin \mathcal{W}_{12} \quad \text{and} \quad \frac{1}{N}\sum_i \|r_{21N}(\xi_i)\| = O_p(N^\alpha), \\ \|r_{31N}(\xi_i)\| &= 0 \quad \text{iff} \quad w_{12i} \notin \mathcal{W}_{12} \quad \text{and} \quad \frac{1}{N}\sum_i \|r_{31N}(\xi_i)\| = O_p(N^\alpha), \end{aligned}$$

$$\begin{aligned}
\left\|r_{20N}\left(\xi_i\right)\right\| &= 0 \quad \text{iff} \quad w_{12i} \notin \mathcal{W}_{12} \quad \text{and} \quad \frac{1}{N}\sum_i \left\|r_{20N}\left(\xi_i\right)\right\| = O_p\left(N^{\alpha}\right),\\
\left\|r_{30N}\left(\xi_i\right)\right\| &= 0 \quad \text{iff} \quad w_{12i} \notin \mathcal{W}_{12} \quad \text{and} \quad \frac{1}{N}\sum_i \left\|r_{30N}\left(\xi_i\right)\right\| = O_p\left(N^{\alpha}\right),
\end{aligned}$$

where α satisfies the inequality $\alpha < \frac{1}{6} + 2\delta$, for δ given in Assumption (A.5). Then

$$\begin{aligned}
\frac{1}{\sqrt{N}}\sum_{i=1}^{N} r_{21N}\left(\xi_i\right)\left(\widehat{p}_{21i} - p_{21i}\right) &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{21N}\left(\xi_i\right)}{b_1\left(w_{12i}\right)}\left(\widehat{t}_1\left(w_{12i}\right) - \widehat{b}_1\left(w_{12i}\right)p_{21}\left(w_{12i}\right)\right) + o_p(1)\\
\frac{1}{\sqrt{N}}\sum_{i=1}^{N} r_{31N}\left(\xi_i\right)\left(\widehat{p}_{31i} - p_{31i}\right) &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{31N}\left(\xi_i\right)}{b_3\left(w_{12i}\right)}\left(\widehat{t}_3\left(w_{12i}\right) - \widehat{b}_3\left(w_{12i}\right)p_{3}\left(w_{12i}\right)\right) + o_p(1)\\
\frac{1}{\sqrt{N}}\sum_{i=1}^{N} r_{20N}\left(\xi_i\right)\left(\widehat{p}_{20i} - p_{20i}\right) &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{20N}\left(\xi_i\right)}{b_0\left(w_{12i}\right)}\left(\widehat{t}_0\left(w_{12i}\right) - \widehat{b}_0\left(w_{12i}\right)p_{20}\left(w_{12i}\right)\right) + o_p(1)\\
\frac{1}{\sqrt{N}}\sum_{i=1}^{N} r_{30N}\left(\xi_i\right)\left(\widehat{p}_{30i} - p_{30i}\right) &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{30N}\left(\xi_i\right)}{b_3\left(w_{12i}\right)}\left(\widehat{t}_3\left(w_{12i}\right) - \widehat{b}_3\left(w_{12i}\right)p_{3}\left(w_{12i}\right)\right) + o_p(1)
\end{aligned}$$

for $b_0\left(w_{12i}\right)$, $b_1\left(w_{12i}\right)$ and $b_3\left(w_{12i}\right)$, $\widehat{b}_0\left(w_{12i}\right)$, $\widehat{b}_1\left(w_{12i}\right)$ and $\widehat{b}_3\left(w_{12i}\right)$, $\widehat{t}_0\left(w_{12i}\right)$, $\widehat{t}_1\left(w_{12i}\right)$ and $\widehat{t}_3\left(w_{12i}\right)$ defined in (2.22)-(2.24).

For a proof of this Lemma see Ahn y Manski (1989), Lemma 2.

Proof of Lemma 4.2.1

In order to show (i) we use Lemma C.0.1, equation (C.3). To do this, note that

$$p_n(\xi_i, \xi_j) = \omega_{ij}(w_{1i} - w_{1j})(w_{1i} - w_{1j})^{\top}, \tag{A.5}$$

with $\xi_i = \left(w_i^{\top}, z_i, d_i\right)^{\top}$. Following the same arguments of the proof of Lemma 5.1, pp. 188-189 in Powell (2001) it can be shown (i). In order to show (ii), we use equation (C.4) of Lemma C.0.1 noting that

$$\begin{aligned}
p_n(\xi_i, \xi_j) &= K_h\left(\frac{p_{1i} - p_{1j}}{h}\right) d_i d_j z_i z_j (w_{1i} - w_{1j})\\
&\quad \times (\eta_1(p_{21i}, p_{3i}) - \eta_1(p_{21j}, p_{3j})) \qquad \text{(A.6)}\\
&+ K_h\left(\frac{p_{0i} - p_{0j}}{h}\right)(1 - d_i)(1 - d_j) z_i z_j (w_{1i} - w_{1j})\\
&\quad \times (\eta_0(p_{20i}, p_{3i}) - \eta_0(p_{20j}, p_{3j})) \qquad \text{(A.7)}\\
&+ \omega_{ij} z_i z_j (w_{1i} - w_{1j})(v_i - v_j). \qquad \text{(A.8)}
\end{aligned}$$

Following again the proof of Lemma 5.1, pp. 191-192 in Powell (2001) it can be shown (ii).

■

Proof of Theorem 4.2.1

Define the matrix norm $\|A\|^2 = \operatorname{tr}\left(A^\top A\right)$. We will show first that under the conditions established in the theorem,

$$\left\|\widehat{S}_{ww} - S_{ww}\right\| = o_p(1). \tag{A.9}$$

This will show (4.33) in Theorem 4.2.1. In order to do this, note that

$$\begin{aligned} &\left\|\widehat{S}_{ww} - S_{ww}\right\| \\ &\leq \binom{N}{2}^{-1} \sum_i \sum_{i<j} \left|\widehat{\omega}_{ij} - \omega_{ij}\right| \left\|w_{1i} - w_{1j}\right\|^2 && \text{(A.10)} \\ &\leq \left(\binom{N}{2}^{-1} \sum_i \sum_{i<j} \left\|w_{1i} - w_{1j}\right\|^2\right) \max_{ij} \left|\widehat{\omega}_{ij} - \omega_{ij}\right|. && \text{(A.11)} \end{aligned}$$

The first term of the third line of the previous inequality is a U-statistic whose kernel has finite expectation, it converges to that expectation almost surely by the Strong Law of Large Numbers for U-statistics (see Serfling (1980), Theorem A, p. 190). We show now that under assumptions (A.1)-(A.11) $\max_{ij} |\widehat{\omega}_{ij} - \omega_{ij}| = o_p(1)$ and therefore (A.9) is proved. In order to show the last claim note that using a Taylor's series expansions,

$$\begin{aligned} \widehat{\omega}_{ij} - \omega_{ij} &= \left(\frac{1}{h_2^2}\right)^2 DK\left(\Delta_{1D}\right)^\top \left[\left(\widehat{p}_{1i} - p_{1i}\right) - \left(\widehat{p}_{1j} - p_{1j}\right)\right] z_i z_j d_i d_j \\ &+ \left(\frac{1}{h_2^2}\right)^2 DK\left(\Delta_{0D}\right)^\top \left[\left(\widehat{p}_{0i} - p_{0i}\right) - \left(\widehat{p}_{0j} - p_{0j}\right)\right] z_i z_j (1-d_i)(1-d_j) \\ &+ R_{ij}, \end{aligned} \tag{A.12}$$

where $DK\left(\Delta_{1D}\right)$ and $DK\left(\Delta_{0D}\right)$ are gradient functions, Δ_{1D} and Δ_{0D} are intermediate values between $\widehat{p}_{1i} - \widehat{p}_{1j}$ and $p_{1i} - p_{1j}$ and $\widehat{p}_{0i} - \widehat{p}_{0j}$ and $p_{0i} - p_{0j}$ respectively and

$$\begin{aligned} R_{ij} &= \left(\frac{1}{h_2^2}\right)^3 \left[(\widehat{p}_{1i} - p_{1i}) - (\widehat{p}_{1j} - p_{1j})\right]^\top H\left(\Delta_{1H}\right) \left[(\widehat{p}_{1i} - p_{1i}) - (\widehat{p}_{1j} - p_{1j})\right] \\ &+ \left(\frac{1}{h_2^2}\right)^3 \left[(\widehat{p}_{0i} - p_{0i}) - (\widehat{p}_{0j} - p_{0j})\right]^\top H\left(\Delta_{0H}\right) \left[(\widehat{p}_{0i} - p_{0i}) - (\widehat{p}_{0j} - p_{0j})\right]. \end{aligned}$$

$H\left(\Delta_{1H}\right)$ and $H\left(\Delta_{0H}\right)$ are hessian functions and again Δ_{1H} and Δ_{0H} are intermediate values between $\widehat{p}_{1i} - \widehat{p}_{1j}$ and $p_{1i} - p_{1j}$ and $\widehat{p}_{0i} - \widehat{p}_{0j}$ and $p_{0i} - p_{0j}$ respectively. Then, under the assumptions (A.1)-(A.11), we show that

$$\begin{aligned} &\max_{ij} \Big| DK\left(\Delta_{1D}\right)^\top \left[\left(\widehat{p}_{1i} - p_{1i}\right) - \left(\widehat{p}_{1j} - p_{1j}\right)\right] z_i z_j d_i d_j && \text{(A.13)} \\ &+ DK\left(\Delta_{0D}\right)^\top \left[\left(\widehat{p}_{0i} - p_{0i}\right) - \left(\widehat{p}_{0j} - p_{0j}\right)\right] z_i z_j (1-d_i)(1-d_j) \Big| = o_p(h_2^4), \end{aligned}$$

and

$$\max_{ij} |R_{ij}| \;\; = \;\; o_p\left(N^{-1/2}\right). \tag{A.14}$$

Following the same arguments as in Ahn y Powell (1993), p. 23, we have

$$\begin{aligned}
\text{l.h.s. of } (A{,}13) \;\; &\leq \;\; 2\kappa_1 h_2^{-4}\left\{\sup_{w_{12}}|\widehat{p}_1(w_{12})-p_1(w_{12})| + \sup_{w_{12}}|\widehat{p}_0(w_{12})-p_0(w_{12})|\right\} \\
&\leq \;\; O\left(N^{4\delta}\right)\times O_p\left(N^{-(1/3+\delta)}\right) = o_p(1). \\
\max_{ij}|R_{ij}| \;\; &\leq \;\; 2\kappa_0 h_2^{-6}\left\{\sup_{w_{12}}|\widehat{p}_1(w_{12})-p_1(w_{12})| + \sup_{w_{12}}|\widehat{p}_0(w_{12})-p_0(w_{12})|\right\}^2 \\
&\leq \;\; O\left(N^{6\delta}\right)\times O_p\left(N^{-(2/3+2\delta)}\right) = o_p(N^{-1/2}),
\end{aligned}$$

and (4.33) is shown. Now we show (4.34). By a usual Taylor's expansion,

$$\begin{aligned}
&\widehat{S}_{wu} - S_{wu} \\
&\quad = \binom{N}{2}^{-1}\sum_i\sum_{i<j}\left(\frac{1}{h_2}\right)^2 DK\left(\Delta_{1D}\right)^{\top}\left[(\widehat{p}_{1i}-p_{1i})-(\widehat{p}_{1j}-p_{1j})\right] \\
&\qquad \times z_i z_j d_i d_j\left(w_{1i}-w_{1j}\right)\left(u_{1i}-u_{1j}\right) \\
&\quad + \binom{N}{2}^{-1}\sum_i\sum_{i<j}\left(\frac{1}{h_2}\right)^2 DK\left(\Delta_{0D}\right)^{\top}\left[(\widehat{p}_{0i}-p_{0i})-(\widehat{p}_{0j}-p_{0j})\right] \\
&\qquad \times z_i z_j(1-d_i)(1-d_j)\left(w_{1i}-w_{1j}\right)\left(u_{0i}-u_{0j}\right).
\end{aligned}$$

Now, by (C.15) and a Strong Law of Large Numbers for U-statistics (see Serfling (1980), theorem A, p. 190),

$$\begin{aligned}
&\Big\|\widehat{S}_{wu} - S_{wu} \\
&\quad - \binom{N}{2}^{-1}\sum_i\sum_{i<j}\left(\frac{1}{h_2}\right)^2 DK\left(\Delta_{1D}\right)^{\top}\left[(\widehat{p}_{1i}-p_{1i})-(\widehat{p}_{1j}-p_{1j})\right] \\
&\qquad \times z_i z_j d_i d_j\left(w_{1i}-w_{1j}\right)\left(u_{1i}-u_{1j}\right) \\
&\quad - \binom{N}{2}^{-1}\sum_i\sum_{i<j}\left(\frac{1}{h_2}\right)^2 DK\left(\Delta_{0D}\right)^{\top}\left[(\widehat{p}_{0i}-p_{0i})-(\widehat{p}_{0j}-p_{0j})\right] \\
&\qquad \times z_i z_j(1-d_i)(1-d_j)\left(w_{1i}-w_{1j}\right)\left(u_{0i}-u_{0j}\right)\| \\
&\quad \leq \left\{\max_{ij} R_{ij}\right\}\binom{N}{2}^{-1}\sum_i\sum_{i<j}\|w_{1i}-w_{1j}\|\times\|u_{1i}-u_{1j}\| \\
&\quad = o_p\left(N^{-1/2}\right).
\end{aligned}$$

Thus, the normalized difference between $\widehat{S}_{wu}$ and S_{wu} is of the form

$$\begin{aligned}\sqrt{N}\left(\widehat{S}_{wu}-S_{wu}\right) &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N} r_{21N}\left(\xi_i\right)\left(\widehat{p}_{21i}-p_{21i}\right)+\frac{2}{\sqrt{N}}\sum_{i=1}^{N} r_{31N}\left(\xi_i\right)\left(\widehat{p}_{3i}-p_{3i}\right)\\ &+ \frac{2}{\sqrt{N}}\sum_{i=1}^{N} r_{20N}\left(\xi_i\right)\left(\widehat{p}_{20i}-p_{20i}\right)+\frac{2}{\sqrt{N}}\sum_{i=1}^{N} r_{30N}\left(\xi_i\right)\left(\widehat{p}_{3i}-p_{3i}\right)\\ &+ o_p(1), \qquad (A.15)\end{aligned}$$

where

$$\begin{aligned} r_{21N}\left(\xi_i\right) &= \frac{1}{N-1}\sum_{j=1}^{N}\left(\frac{1}{h_2}\right)^2 DK_{21}\left(\Delta_{1D}\right) z_i z_j d_i d_j \left(w_{1i}-w_{1j}\right)\left(u_{1i}-u_{1j}\right),\\ r_{31N}\left(\xi_i\right) &= \frac{1}{N-1}\sum_{j=1}^{N}\left(\frac{1}{h_2}\right)^2 DK_{31}\left(\Delta_{1D}\right) z_i z_j d_i d_j \left(w_{1i}-w_{1j}\right)\left(u_{1i}-u_{1j}\right),\\ r_{20N}\left(\xi_i\right) &= \frac{1}{N-1}\sum_{j=1}^{N}\left(\frac{1}{h_2}\right)^2 DK_{20}\left(\Delta_{0D}\right) z_i z_j (1-d_i)(1-d_j)\left(w_{1i}-w_{1j}\right),\\ &\times\left(u_{0i}-u_{0j}\right),\\ r_{30N}\left(\xi_i\right) &= \frac{1}{N-1}\sum_{j=1}^{N}\left(\frac{1}{h_2}\right)^2 DK_{30}\left(\Delta_{0D}\right) z_i z_j (1-d_i)(1-d_j)\left(w_{1i}-w_{1j}\right)\\ &\times\left(u_{0i}-u_{0j}\right).\end{aligned}$$

$DK_{20}\left(\Delta_{1D}\right)$, $DK_{30}\left(\Delta_{1D}\right)$, $DK_{21}\left(\Delta_{0D}\right)$, $DK_{31}\left(\Delta_{1D}\right)$ are the corresponding row vectors of $DK\left(\Delta_{0D}\right)$ and $DK\left(\Delta_{1D}\right)$ respectively. Thus substituting (2.22)-(2.24) into (C.15) and applying Lemma C.0.2 we have that

$$\begin{aligned}\sqrt{N}\left(\widehat{S}_{wu}-S_{wu}\right) &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{21N}\left(\xi_i\right)}{b_1\left(w_{12i}\right)}\left(\widehat{t}_1\left(w_{12i}\right)-\widehat{b}_1\left(w_{12i}\right)p_{21}\left(w_{12i}\right)\right)\\ &+ \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{31N}\left(\xi_i\right)}{b_3\left(w_{12i}\right)}\left(\widehat{t}_3\left(w_{12i}\right)-\widehat{b}_3\left(w_{12i}\right)p_{3}\left(w_{12i}\right)\right)\\ &+ \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{20N}\left(\xi_i\right)}{b_0\left(w_{12i}\right)}\left(\widehat{t}_0\left(w_{12i}\right)-\widehat{b}_0\left(w_{12i}\right)p_{20}\left(w_{12i}\right)\right)\\ &+ \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{30N}\left(\xi_i\right)}{b_3\left(w_{12i}\right)}\left(\widehat{t}_3\left(w_{12i}\right)-\widehat{b}_3\left(w_{12i}\right)p_{3}\left(w_{12i}\right)\right)\\ &+ o_p(1).\end{aligned}$$

Now, we treat each of the previous terms as in the proof of Theorem 3.1. (*ii*), pp. 25-28 in Ahn y Powell (1993) and we obtain,

$$\begin{aligned}\widehat{S}_{wu}-S_{wu} &= \frac{2}{N}\sum_{i=1}^{N}\rho_{1i}^2\gamma_i D\mu_1(p_{21i},p_{3i})\\ &\times\left(w_{1i}-\phi_1(p_{21i},p_{3i})\right)\left[z_i-p_{21i}\right]\left[d_i-p_{3i}\right]\end{aligned}$$

$$\begin{aligned} + & \frac{2}{N}\sum_{i=1}^{N}\rho_{0i}^{2}(1-\gamma_i)D\mu_0(p_{20i},p_{3i}) \\ & \times\left(w_{1i}-\phi_0(p_{20i},p_{3i})\right)\left[z_i-p_{20i}\right]\left[d_i-p_{3i}\right]+o_p\left(N^{-1/2}\right), \end{aligned}$$

where

$$\begin{aligned} D\mu_0(p_{20i},p_{3i}) &= \partial\mu_0\left(p_{20i},p_{3i}\right)/\partial p_{21i}+\partial\mu_0\left(p_{20i},p_{3i}\right)/\partial p_{3i}, \\ D\mu_1(p_{21i},p_{3i}) &= \partial\mu_1\left(p_{21i},p_{3i}\right)/\partial p_{21i}+\partial\mu_1\left(p_{21i},p_{3i}\right)/\partial p_{3i}. \end{aligned}$$

Apéndice B

Apéndice Capitulo 3

Table A1: **Wage Differences by Gender and Sector. 2016**

Country	Public	Private
Belgium	0.1	9.2
Bulgaria	20.6	12.3
Czech Republic	20.5	22.8
Denmark	11.6	15.6
Germany	13.0	24.0
Spain	13.0	19.0
Italy	4.4	17.9
Cyprus	-6.6	23.0
Latvia	16.9	14.1
Lithuania	13.7	17.6
Hungary	11.8	15.0
Netherlands	12.7	21.6
Poland	2.8	16.1
Portugal	13.4	22.6
Romania	9.9	6.8
Slovenia	11.3	7.9
Slovakia	12.7	20.4
Finland	17.7	17.2
Sweden	10.2	11.8
United Kingdom	24.4	22.2
Iceland	12.2	16.4
Norway	8.5	17.8
Switzerland	17.5	17.7
TOTAL	11.8	16.9

Source: Estimated by Eurostat.

Table A2: **Wages Equation by Sector**

	Public Sector	**Private Sector**
d (Male)	0.130*** (0.005)	0.178*** (0.003)
Age1	-0.201*** (0.016)	0.031*** (0.008)
Age2	0.088*** (0.011)	0.250*** (0.007)
Age3	0.134*** (0.010)	0.305*** (0.007)
Age4	0.106*** (0.010)	0.239*** (0.007)
Primary	-0.208*** (0.012)	-0.161*** (0.004)
Sup	0.368*** (0.005)	0.457*** (0.006)
Full-time	0.990*** (0.008)	0.985*** (0.004)
Tenure	0.018*** (0.000)	0.028*** (0.000)
Constant	8.662*** (0.013)	8.296*** (0.007)

Source: Own Elaboration from Annual Wage Structure Survey, 2016.
Standard deviation in brackets.
Signif. codes: 0.01 '***'0.05 '**'0.1 '*'

Table A3: **First Step Estimations: Probability of working in Public Sector**

	Public Sector (Std. Dev.)
Age1	-0.593*** (0.021)
Age2	-0.436*** (0.017)
Age3	-0.251*** (0.016)
Age4	-0.036*** (0.016)
Primary	-0.488*** (0.013)
Sup	0.528*** (0.008)
Tenure	0.020*** (0.000)
Full-time	0.405*** (0.011)
cotization	-0.0009*** (0.000)
Constant	-1.159*** (0.018)

Signif. codes: 0.01 '***'0.05 '**'0.1 '*'
Source: Own Elaboration from the Annual Wage Structure Survey, 2016.

Table A4: **Second Step Estimations: (log) wages**

	Men/Public	**Men/Private**	**Women/Public**	**Women/Private**
Age1	-0.572*** (0.020)	-0.324*** (0.010)	-0.835*** (0.027)	-0.407*** (0.015)
Age2	-0.207*** (0.020)	-0.090*** (0.009)	-0.518*** (0.022)	-0.144*** (0.014)
Age3	-0.109*** (0.018)	0.066*** (0.008)	-0.245*** (0.017)	0.066*** (0.013)
Age4	0.028** (0.013)	0.204*** (0.007)	-0.035** (0.013)	0.199*** (0.012)
Primary	-0.407*** (0.017)	-0.369*** (0.006)	-0.664*** (0.022)	-0.497*** (0.008)
Sup	0.733*** (0.016)	0.879*** (0.006)	0.991*** (0.019)	0.978*** (0.010)
Tenure	0.0337*** (0.000)	0.0422*** (0.000)	0.0449*** (0.000)	0.0565*** (0.000)
Full-Time	1.355*** (0.016)	1.243*** (0.005)	1.195*** (0.013)	1.200*** (0.006)
Constant	10.602** (0.089)	11.062*** (0.140)	11.683*** (0.086)	11.745*** (0.053)
IMR	-6.404*** (0.280)	-8.287*** (0.140)	-8.738*** (0.253)	-10.238*** (0.168)

Signif. codes: 0.01 '***'0.05 '**'0.1 '*'

Source: Annual Wage Structure Survey, 2016.

Apéndice C

Apéndice Capitulo 4

For the proof of the stated results consider a second order U-statistic of the form

$$U_N = \binom{N}{2}^{-1} \sum_{i=1}^{N} p_N(\xi_i, \xi_j) \tag{C.1}$$

where $\{\xi_1, \cdots, \xi_N\}$ is an i.i.d. sample of random vectors and $p_n(\cdot, \cdot)$ is a symmetric kernel, i.e. $p_N(\xi_i, \xi_j) = p_N(\xi_j, \xi_i)$. Define

$$\begin{aligned} r_N(\xi_i) &= E[p_N(\xi_i, \xi_j)|\, \xi_i], \\ \theta_N &= E[r_N(\xi_i)] = E[p_N(\xi_i, \xi_j)], \end{aligned}$$

as well as the projection of the statistic U_N,

$$\hat{U}_N = \theta_N + \frac{2}{N} \sum_{i=1}^{N} [r_N(\xi_i) - \theta_N] \tag{C.2}$$

The following two lemmas will be used in the proofs of the main results:

Lemma C.0.1 *If the kernel* $p_N(\cdot)$ *satisfies;* $E[||p_N(\xi_i, \xi_j)||^2] = o(N)$*, then*

$$U_N = \theta_N + o_p(1), \tag{C.3}$$

$$\widehat{U}_N = U_N + o_p\left(N^{-1/2}\right), \tag{C.4}$$

as N *tends to infinity.*

For a proof of Lemma C.0.1, see Powell et˜al. (1989), Lemma 3.1.

Lemma C.0.2 *Let* $r_{1N}(\xi_i)$*,* $r_{2N}(\xi_i)$ *be random vectors valued functions of the data vector* $\xi_i = \left(y_i, p_i, s_i, w_i^\top\right)^\top$*, which satisfy*

$$\begin{aligned} \|r_{1N}(\xi_i)\| &= 0 \quad \text{iff} \quad x_i \notin \S \quad \text{and} \quad \frac{1}{N}\sum_i \|r_{1N}(\xi_i)\| = O_p(N^\alpha), \\ \|r_{2N}(\xi_i)\| &= 0 \quad \text{iff} \quad w_i \notin \mathcal{W} \quad \text{and} \quad \frac{1}{N}\sum_i \|r_{2N}(\xi_i)\| = O_p(N^\alpha), \end{aligned}$$

where α satisfies the inequality $\alpha < \frac{1}{6} + 2\delta$, for δ given in Assumption (A.5). Then

$$\begin{aligned}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} r_{1N}(\xi_i)(\widehat{p}_{1i} - p_{1i}) &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{1N}(\xi_i)}{b_1(x_i)}\left(\widehat{t}_1(x_i) - \widehat{b}_1(x_i)p_1(x_i)\right) + o_p(1)\\ \frac{1}{\sqrt{N}}\sum_{i=1}^{N} r_{2N}(\xi_i)(\widehat{\eta}_{2i} - \eta_{2i}) &= \frac{2}{\sqrt{N}}\sum_{i=1}^{N}\frac{r_{2N}(\xi_i)}{b_2(w_i)}\left(\widehat{t}_2(w_i) - \widehat{b}_2(w_i)\eta_2(w_i)\right) + o_p(1)\end{aligned}$$

for $b_1(x_i)$ and $b_2(w_i)$, $\widehat{b}_1(x_i)$ and $\widehat{b}_2(w_i)$, $\widehat{t}_1(x_i)$ and $\widehat{t}_2(w_i)$ defined in chapter 4.

For a proof of this Lemma see Ahn y Manski (1989), Lemma 2.

Proof of Lemma 4.2.1

In order to show (i) we use Lemma C.0.1, equation (C.3). To do this,being $w_1 = (w_{11}, s)$ note that

$$p_n(\xi_i, \xi_j) = \omega_{ij}(w_{1i} - w_{1j})(w_{1i} - w_{1j})^\top, \tag{C.5}$$

with $\xi_i = \left(w_i^\top, p_i, s_i\right)^\top$. Following the same arguments of the proof of Lemma 5.1, pp. 188-189 in Powell (2001) it can be shown (i). In order to show (ii), we use equation (C.4) of Lemma C.0.1 noting that

$$\begin{aligned} p_n(\xi_i, \xi_j) &= K_h\left(\frac{z_i - z_j}{h}\right)p_i p_j(w_{1i} - w_{1j})\\ &\quad \times(\eta_1(p_{1i}, \eta_{2i}) - \eta_1(p_{1j}, \eta_{2j})) && \text{(C.6)}\\ &+ \omega_{ij}p_i p_j(w_{1i} - w_{1j})(\epsilon_i - \epsilon_j). && \text{(C.7)}\end{aligned}$$

Following again the proof of Lemma 5.1, pp. 191-192 in Powell (2001) it can be shown (ii). ∎

Proof of Theorem 4.2.1

Define the matrix norm $\|A\|^2 = \operatorname{tr}\left(A^\top A\right)$. We will show first that under the conditions established in the theorem,

$$\left\|\widehat{S}_{ww} - S_{ww}\right\| = o_p(1). \tag{C.8}$$

This will show (4.33) in Theorem 4.2.1. In order to do this, note that

$$\begin{aligned}\left\|\widehat{S}_{ww} - S_{ww}\right\| &\leq \binom{N}{2}^{-1}\sum_i\sum_{i<j}|\widehat{\omega}_{ij} - \omega_{ij}|\,\|w_{1i} - w_{1j}\|^2\\ &\leq \left(\binom{N}{2}^{-1}\sum_i\sum_{i<j}\|w_{1i} - w_{1j}\|^2\right)\max_{ij}|\widehat{\omega}_{ij} - \omega_{ij}|\,. && \text{(C.9)}\end{aligned}$$

The first term of the third line of the previous inequality is a U-statistic whose kernel has finite expectation, it converges to that expectation almost surely by the Strong Law of Large Numbers for U-statistics (see Serfling (1980), Theorem A, p. 190). We show now that under assumptions (A.1)-(A.11) $\text{máx}_{ij}\,|\widehat{\omega}_{ij}-\omega_{ij}|=o_p(1)$ and therefore (C.8) is proved. In order to show the last claim note that using a Taylor's series expansions,

$$\begin{aligned}\widehat{\omega}_{ij}-\omega_{ij} &= \left(\frac{1}{h_2^2}\right)^2 DK\left(\Delta_{1D}\right)^\top\left[(\,\widehat{z}_i-z_i\,)-(\,\widehat{z}_j-z_j\,)\right]p_ip_j \\ &+ R_{ij} \end{aligned}\tag{C.10}$$

where $DK\left(\Delta_{1D}\right)$ is a gradient function, Δ_{1D} is an intermediate values between $\widehat{z}_i-\widehat{z}_j$ and z_i-z_j and

$$R_{ij}=\left(\frac{1}{h_2^2}\right)^3\left[(\widehat{z}_i-z_i)-(\widehat{z}_j-z_j)\right]^\top H\left(\Delta_{1H}\right)\left[(\widehat{z}_i-z_i)-(\widehat{z}_j-z_j)\right]\tag{C.11}$$

$H\left(\Delta_{1H}\right)$ is the hessian functions and again Δ_{1H} is an intermediate values between $\widehat{z}_i-\widehat{z}_j$ and z_i-z_j. Then, under the assumptions 1 to 11, we show that

$$\begin{aligned}&\underset{ij}{\text{máx}}\left|DK\left(\Delta_{1D}\right)^\top\left[(\,\widehat{z}_i-z_i\,)-(\,\widehat{z}_j-z_j\,)\right]p_ip_j\right| \\ =\;& o_p(h_2^4).\end{aligned}\tag{C.12}$$

and

$$\underset{ij}{\text{máx}}\,|R_{ij}| \;=\; o_p\left(N^{-1/2}\right).\tag{C.13}$$

Following the same arguments as in Ahn y Powell (1993), p. 23, we have

$$\begin{aligned}\text{C.12}\leq 2\kappa_1h_2^{-4}\; &\left\{\sup_{w_{12}}|\widehat{p}_1(w_{12})-p_1(w_{12})|\right\}\\ \leq\;& O\left(N^{4\delta}\right)\times O_p\left(N^{-(1/3+\delta)}\right)=o_p(1).\\ \underset{ij}{\text{máx}}\,|R_{ij}| \;\leq\;& 2\kappa_0h_2^{-6}\left\{\sup_{w_{12}}|\widehat{p}_1(w_{12})-p_1(w_{12})|\right\}^2\\ \leq\;& O\left(N^{6\delta}\right)\times O_p\left(N^{-(2/3+2\delta)}\right)=o_p(N^{-1/2}),\end{aligned}$$

and eq. 4,31 is shown. Now we show 4,32 by a usual Taylor's expansion,

$$\begin{aligned}&\widehat{S}_{wu}-S_{wu}\\ &= \binom{N}{2}^{-1}\sum_i\sum_{i<j}\left(\frac{1}{h_2}\right)^2 DK\left(\Delta_{1D}\right)^\top\left[(\,\widehat{z}_i-z_i\,)-(\,\widehat{z}_j-z_{1j}\,)\right]\\ &\times p_ip_j\left(w_{1i}-w_{1j}\right)\left(u_i-u_j\right)\end{aligned}$$

Now, by a Strong Law of Large Numbers for U-statistics (see Serfling (1980), theorem A, p. 190),

$$\begin{aligned}
&\left\|\widehat{S}_{wu} - S_{wu}\right. \\
&\quad = \binom{N}{2}^{-1} \sum_{i} \sum_{i<j} \left(\frac{1}{h_2}\right)^2 DK\left(\Delta_{1D}\right)^{\top} \left[\left(\,\widehat{z}_i - z_i\,\right) - \left(\,\widehat{z}_j - z_j\,\right)\right] \\
&\quad \times p_i p_j \left(w_{1i} - w_{1j}\right)\left(u_i - u_j\right)| \\
&\quad \leq \left\{\operatorname*{máx}_{ij} R_{ij}\right\} \binom{N}{2}^{-1} \sum_{i} \sum_{i<j} \left\|w_{1i} - w_{1j}\right\| \times \left|u_i - u_j\right\| = o_p\left(N^{-1/2}\right).
\end{aligned}$$

Thus, the normalized difference between $\widehat{S}_{wu}$ and S_{wu} is of the form

$$\begin{aligned}
\sqrt{N}\left(\widehat{S}_{wu} - S_{wu}\right) &= \frac{2}{\sqrt{N}} \sum_{i=1}^{N} r_{1N}\left(\xi_i\right)\left(\widehat{p}_{1i} - p_{1i}\right) + \frac{2}{\sqrt{N}} \sum_{i=1}^{N} r_{2N}\left(\xi_i\right)\left(\widehat{\eta}_{2i} - \eta_{2i}\right) \\
&+ \; o_p(1),
\end{aligned} \tag{C.14}$$

where

$$\begin{aligned}
r_{1N}\left(\xi_i\right) &= \frac{1}{N-1} \sum_{j=1}^{N} \left(\frac{1}{h_2}\right)^2 DK_1\left(\Delta_{1D}\right) p_i p_j \left(w_{1i} - w_{1j}\right)\left(u_i - u_j\right), \\
r_{2N}\left(\xi_i\right) &= \frac{1}{N-1} \sum_{j=1}^{N} \left(\frac{1}{h_2}\right)^2 DK_2\left(\Delta_{1D}\right) p_i p_j \left(w_{1i} - w_{1j}\right)\left(u_i - u_j\right).
\end{aligned}$$

$DK_1\left(\Delta_{1D}\right)$ and $DK_2\left(\Delta_{1D}\right)$ are the corresponding row vectors. Thus substituting and applying Lemma C.0.2 we have that

$$\begin{aligned}
\sqrt{N}\left(\widehat{S}_{wu} - S_{wu}\right) &= \frac{2}{\sqrt{N}} \sum_{i=1}^{N} \frac{r_{1N}\left(\xi_i\right)}{b_1\left(x_i\right)} \left(\widehat{t}_1\left(x_i\right) - \widehat{b}_1\left(x_i\right) p_1\left(x_i\right)\right) \\
&+ \frac{2}{\sqrt{N}} \sum_{i=1}^{N} \frac{r_{2N}\left(\xi_i\right)}{b_2\left(w_i\right)} \left(\widehat{t}_2\left(w_i\right) - \widehat{b}_2\left(w_i\right) \eta_2\left(w_{12i}\right)\right) \\
&+ \; o_p(1).
\end{aligned}$$

Now, we treat each of the previous terms as in the proof of Theorem 3.1. (ii), pp. 25-28 in Ahn y Powell (1993) and we obtain,

$$\begin{aligned}
\widehat{S}_{wu} - S_{wu} &= \frac{2}{N} \sum_{i=1}^{N} \rho_{1i}^2 \gamma_i D\mu_1(p_{1i}, \eta_{2i}) \\
&\quad \times \left(w_{1i} - \phi(p_{1i}, \eta_{2i})\right)\left[p_i - p_{1i}\right]\left[s_i - \eta_{2i}\right] \\
&+ \; o_p\left(N^{-1/2}\right),
\end{aligned}$$

where

$$D\mu_1(p_{1i}, \eta_{2i}) = \partial\mu_1\left(p_{1i}, \eta_{2i}\right)/\partial p_{1i} + \partial\mu_1\left(p_{1i}, \eta_{2i}\right)/\partial\eta_{2i}. \tag{C.15}$$

Bibliografía

Bibliografía

R. Aaberge. “Choosing measures of inequality for empirical applications”. Discussion Papers, 158 in *Research Department of Statistics Norway.* 1995

R. Aaberge, U. Colombino, y S. Strom. “Labor supply in italy: An empirical analysis of joint household decisions, with taxes and quantity constraints”. *Journal of Applied Econometrics*, 14(4):403–422, 1999.

H. Ahn y C. Manski. “Distribution theory for the analysis of binary choice under uncertainty with nonparametric estimation of expectations”. *Woorking paper. Social System Research Institute, University of Wisconsin-Madison.*, 8913, 1989.

H. Ahn y J. Powell. “Semiparametric estimation of censored selection models with a nonparametric selection mechanism”. *Journal of Econometrics.*, 58:3–29, 1993.

R. Alaez y M. Ullibarri. “Diferencias salariales entre los sectores público y privado por género, escolaridad y edad. el caso de espaa”. *ICE. Tribuna de Economía.*, 789:117–138, 2001.

T. Amemiya. *Advanced Econometrics.* Harvard University Press, Cambridge, Massachusets, 1985.

D. Andrews y M. Schafgans. “Semiparametric estimation of the intercept of a sample selection model”. 1998.

J. Angrist y A. Krueger. “Instrumental variables and the search for identification: From supply and demand to natural experiments”. *Journal of Economic Perspectives.*, 15:69–85, 2001.

A. Aradillas-Lopez, E. Honor, y J. Powell. Pairwise difference estimation with nonparametric control variables*. *International Economic Review*, 48(4):1119–1158, 2007.

M. Arellano y S. Bonhomme. “Quantile selection models with an application to understanding changes in wage inequality”. *Econometrica.*, 85:1–28, 2017.

W. Arulampalam, A. Booth, y M. Bryan. “Is there a glass ceiling over europe? exploring the gender pay gap across the wage”. *ILR Review*, 60:163–186, 2007.

G. Becker. “Human capital”, 1964. New York: Columbia university press.

M. Ben-Akiva y SR. Lerman. "Discrete choice analysis: Theory and application to travel demand", 1985. The MIT Press, Cambridge/Massachusetts.

W. Bielby y J. Baron. "Men and women at work: Sex segregation and statistical discrimination". *American Journal of Sociology.*, 91:759–799, 1986.

F. Blau y L. Kahn. "Changes in the labor supply behavior of married women: 1980-2000". *Journal of Labor Economics.*, 25:393–438, 2007.

A. Blinder. "Wage discrimination: Reduced form and structural variables". *Journal of Human Resourcess.*, 8:436–455, 1973.

R. Blundell y J. Powell. "Endogeneity in nonparametric and semiparametric regression models". *In M. Dewatripont, L. P. Hansen and S. J. Turnovsky (eds.) Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress.*, II, 2003.

C. Boot, D. Deeg, T. Abma, K. Rijs, S. van der Pas, T. van Tilburg, y A. van der Beek. Predictors of having paid work in older workers with and without chronic disease: A 3-year prospective cohort study. *Journal of Occupational Rehabilitation*, 24(3):563–572, 2014.

V. Burris y A. Wharton. "Sex segregation in the u.s. labor force". *Review of Radical Political Economics.*, 14:43–56, 1982.

G. Chamberlain. "Quantile regression, censoring, and the structure of wages". *In: Sims, C., (ed.), Advances in Econometrics: Sixth World Congress. Econometric Society Monograph.*, 1, 1994.

V. Chernozhukov, I. Fernandez, y B. Melly. "Inference on counterfactual distributions". *Econometrica.*, 81:2205–2268, 2013.

D. Cotter, J. Hermsen, y R. Vanneman. "Gender inequality at work". *New York: Russell Sage; Washington, D.C.: Population Reference Bureau*, 2004.

J. Dagsvik. "Discrete and continuous choice, max-stable processes, and independence from irrelevant attributes". *Econometrica*, 62(5):1179–1205, 1994.

J. Dagsvik y S Strom. "Sectoral labor supply, choice restrictions and functional form", 2004. Discussion Papers, 388 in *Research Department of Statistics Norway.*

J. Dagsvik y S. Strom. "Sectoral labor supply, choice restrictions and functional form". *Journal of Applied Econometrics*, 21(6):803–826, 2006.

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

M.L. Di Tommaso, S. Strom, y E.M. Saether. "Nurses wanted: Is the job too hard or is the wage too low". *Journal of Health Economics*, 28:748–757, 2009.

E. Duguet y C. Le-Clainche. "Une valuation de límpact de lámnagement des conditions de travail sur la reprise du travail aprs un cancer". *Working Papers halshs-00966861, HAL.*, 2014.

I. Fernandez, J. Rodriguez-Poo, y S. Sperlich. "A note on the parametric three step estimator in structural labor supply models". *Economic Letters*, 74:31–41, 2001.

I. Garcia-Perez y J. Jimeno. "Public sector wage gaps in spanish regions.". *Manchester School*, 75:501–531, 2007.

R. Giordano, D. Depalo, M. Coutinho, B. Eugene, E. Papapetrou, J. Perez, L. Reiss, y M. Roter. "The public sector pay gap in a selection of euro area countries". *Working paper series. European Central Bank*, 1406, 2011.

B. Greve. "The labor market situations of disabled people in european countries and implementation of employment policies; a summary of evidence from country reports and research studies.". *Report prepared for the Academic Network of European Disability Experts (ANED)*, 2009.

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

B. E. Hansen. "Uniform convergence rates for kernel estimation with dependent data". *Econometric Theory*, 24:726–748, 2008.

J. Hausman y D. McFadden. "Specification test for the multinomial logit model". *Econometrica*, 52(5):1219–1240, 1984.

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

J. Heckman. "Dummy endogenous variables in a simultaneous equation system". *Econometrica*, 46:931–959, 1978.

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

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

J. Jacobs. "Women's entry into management: Trends in earnings, authority, and values among salaried managers". *Administrative Science Quarterly.*, 37:282–301, 1992.

B. Jann. "Standard errors for the blinder-oaxaca decomposition". *German Stata Users Group Meeting.*, 3rd, 2005.

L. Kahn. “The impact of employment protection mandates on demographic temporary employment patterns: International microeconomic evidence”. *Economic Journal.*, 117:333–356, 2007.

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

R. Koenker y G. Bassett. “Regression quantiles”. *Econometrica*, 46:33–50, 1978.

P. Langley, J. Tornero, C. Margarit, C. Prez, A. Tejedor, y M. Ruiz. The association of pain with labor force participation, absenteeism, and presenteeism in spain. *Journal of Medical Economics*, 14(6):835–845, 2011.

G. Livermore, D. Stapleton, y D. Wittemburg. “The economics of policies and programs affecting the employment of people with disabilities”, 2000. Discussion Papers, 5 in *http://www.ilr.cornell.edu/RRTC/papers.html.*

J. Machado y J. Mata. “Counterfactual decomposition of changes in wage distributions using quantile regression”. *Journal of Applied Econometrics.*, 20:445–465, 2005.

H. Mandel. “Occupational mobility of american women: Compositional and structural changes, 1980 2007”. *Research in Social Stratification and Mobility*, 30:5–16, 2012.

H. Mandel. “Up the down staircase: Women’s upward mobility and the wage penalty for occupational feminization. 1970-2007”. *Social Forces*, 91:1183–1207, 2013.

H. Mandel y M. Semyonov. “Gender pay gap and employment sector: Sources of earnings disparities in the u.s., 1970-2010”. *Demography*, 51:1597–1618, 2014.

R. Meenan, P. Gertman, J. Mason, y R. Dunaif. “The arthritis impact measurement scales: further investigation of a health status instrument”. *Arth. Rheum*, 25:1048–1053, 1982.

B. Melly. “Estimation of counterfactual distributions using quantile regression.”. *University of St. Gallen, Discussion Paper.*, 2006.

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

J. Mincer y S. Polacheck. “Family investments in human capital: Earnings of women”. *Journal of Political Economy*, 82(5):76–108, 1974.

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

E. Nadaraya. “On nonparametric estimates of density functions and regression curves”. *Theory of Applied Probability*, 10:186–190, 1965.

R. Oaxaca. “Male-female wage differentials in urban labor markets”. *International Economic Review.*, 14:693–709, 1973.

G. Parodi y D. Scuilli. “Disability in italian households: income, poverty and labour market participation.”. *Applied Economics*, 40(20):2615–2630, 2008.

L. Petersen, T. adn Morgan. “Separate and unequal: Occupationestablishment segregation and the gender wage gap”. *American Journal of Sociology.*, 101:329–365, 1995.

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

J. L. Powell. *Semiparametric estimation of censored selection models*, págs. 165–196. Cambridge University Press, 2001.

J. L. Powell, J. H. Stock, y T. M. Stoker. Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430, 1989.

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

P. Rosenbaum y D. Rubin. “The central role of the propensity score in observational studies for causal effects”. *Biometrika*, 70:41–55, 1983.

E.M. Saether. “Nurses’labor supply with an endogenous choice of care level and shift type : a nested discrete choice model with nonlinear income”, 2004. University of Oslo, Norway.

R. J. Serfling. *Approximation theorems of mathematical statistics.* John Wiley & Sons, Inc., New York, 1980. Wiley Series in Probability and Mathematical Statistics.

D.M. Shapiro y M. Stelcner. “Canadian publicprivate sector earnings differentials, 19701980”. *Industrial Relations*, 28:72–81, 1989.

S. Smith. “Government wage differentials by sex”. *The Journal of Human Resources.*, 11(2):185–199, 1976.

Moon-Gi Suh. Determinants of female labor force participation in south korea: Tracing out the u-shaped curve by economic growth. *Social Indicators Research*, 131(1):255–269, 2017.

K. Train. “Discrete choice methods with simulation”, 2003. Cambridge University Press.

D. Treiman y H. Hartmann. “Women, work and wages: Equal pay for jobs of equal value”. *Washington, D.C.: National Academy Press.*, 1981.

A. Van Soest. “Discrete choice models of family labor supply”. *Journal of Human Resources*, 30:63–88, 1995.

F. Vella. "A simple estimator for models with censored endogenous regressors." *International Economic Review.*, 34:441–457, 1993.

F. Vella. "Estimating models with sample selection bias: A survey." *Journal of Human Resources.*, 33:127–169, 1998.

E. Vytlacil y N. Yildiz. "Dummy endogenous variables in weakly separable models". *Econometrica.*, 75:757–779, 2007.

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

K. Weeden. “Profiles of change: Sex segregation in the united states, 1910-2000”. *In M. Charles and D. B. Grusky (Eds.), Occupational ghettos: The worldwide segregation of women and men.*, págs. 131–178, 2004.

J. Wooldridge. "Control function methods in applied econometrics". *Journal of Human resources.*, 50:420–445, 2015.

X. Zhang, X. Zhao, y A. Harris. "Chronic diseases and labour force participation in australia". *Journal of health economics*, 28 1:91–108, 2009.