Trabajo Fin de Grado
Grado en Ingeniería de las Tecnologías Industriales

Un análisis econométrico de los rendimientos del trabajo a partir de la Muestra Continua de Vidas Laborales

Autor: Manuel Sánchez Morillo

Tutor: Fernando Núñez Hernández

Dpto. Organización Industrial y Gestión de Empresas I
Universidad de Sevilla

Sevilla, 2022

Trabajo Fin de Grado
Grado en Ingeniería de las Tecnologías Industriales

Un análisis econométrico de los rendimientos del trabajo a partir de la Muestra Continua de Vidas Laborales

Autor:

Manuel Sánchez Morillo

Tutor:

Fernando Núñez Hernández

Profesor Titular de Universidad

Dpto. Organización Industrial y Gestión de Empresas I

Universidad de Sevilla

Sevilla, 2022

Trabajo Fin de Grado: Un análisis econométrico de los rendimientos del trabajo a partir de la Muestra Continua de Vidas Laborales

Autor: Manuel Sánchez Morillo

Tutor: Fernando Núñez Hernández

El tribunal nombrado para juzgar el Trabajo arriba indicado, compuesto por los siguientes miembros:

Presidente:

Vocales:

Secretario:

Acuerdan otorgarle la calificación de:

Sevilla, 2022

El Secretario del Tribunal

Resumen

El presente Trabajo de Fin de Grado propone un análisis de los determinantes de la retribución de los trabajadores de la economía española en los años 2019 y 2020, utilizando para ello la Muestra Continua de Vidas Laborales de la Seguridad Social. Para ello, se han descrito y añadido 21 variables explicativas dentro de un modelo de regresión lineal múltiple con variables ficticias, siendo la variable endógena las rentas anuales brutas obtenidas por el trabajador. Los resultados obtenidos nos permiten cuantificar, por ejemplo, que una mujer ganó de media un 12% menos de salario que un hombre para un mismo puesto de trabajo o que existen diferencias de retribución media entre provincias de hasta el 35%. Asimismo, se observa que el sector económico donde se trabaje y la categoría de puesto realizado generan las mayores diferencias salariales, diferencias que pueden alcanzar el 80% de la retribución. El año 2020 supuso también una caída generalizada de las retribución esperada del 7%, posiblemente por la situación de pandemia originada por la COVID-19.

Abstract

This Final Degree Project proposes an analysis of the determinants of the remuneration of workers in the Spanish economy in the years 2019 and 2020, using the Continuous Sample of Working Lives of the Social Security. For this purpose, 21 explanatory variables have been described and added within a multiple linear regression model with dummy variables, the endogenous variable being the gross annual income obtained by the worker. The results obtained allow us to quantify, for example, that a woman earned on average 12% less salary than a man for the same job or that there are differences in average salary between provinces of up to 35%. It is also observed that the economic sector in which one works, and the category of the position performed generate the greatest salary differences, differences that can reach up to 80% of the salary. The year 2020 also saw a generalized drop in expected pay of 7%, possibly due to the pandemic situation caused by COVID-19.

Índice

Índice de Tablas

Índice de Figuras

1 Introducción.

Según la Organización Internacional del Trabajo (OIT), los trabajadores ocupados son todas aquellas personas de 16 años o más que, durante un período de referencia dado (semana de referencia, período actual...) declararon tener un empleo por cuenta ajena, asalariado, o ejercieron una actividad por cuenta propia. En el caso de los asalariados, el pagador y el trabajador deben responder al menos a estas dos preguntas: ¿Cuál es la cantidad de dinero que debe intercambiarse por el trabajo realizado? ¿Qué debemos tener en cuenta para llegar a esa cantidad?

La Teoría Económica ha tratado de dar respuestas a estas preguntas. A nivel teórico, el modelo clásico de naturaleza walrasiana de la oferta y la demanda de trabajo en competencia perfecta ha ido dejando paso a modelos de competencia imperfecta donde los trabajadores y los puestos son heterogéneos, siendo necesaria una negociación individual entre empresa y trabajador para formar un emparejamiento productivo y acordar un determinado salario (véanse por ejemplo los modelos de Rosen (1974) o Pissarides (2000)). En la negociación del salario, ambos lados del mercado deben valorar monetariamente la actividad realizada por el trabajador. Y dicha valoración va a depender en buena medida de los atributos del trabajador en relación con los requerimientos del puesto.

Este Trabajo de Fin de Grado pretende arrojar un poco de luz sobre los factores determinantes de las retribuciones procedentes del trabajo observadas en los años 2019 y 2020 en España. A través de la estimación de un modelo econométrico, trataremos de obtener las retribuciones medias esperadas para cada tipo de trabajador en función de sus características y de las características conocidas del puesto. Dicho de otra forma, estas predicciones estarán condicionadas, en términos estadísticos, a dicho conjunto observado de variables cualitativas y cuantitativas, quedando en el término de error del modelo todos aquellos factores inobservables que pueden afectar a la remuneración del individuo.

Para situar al trabajo dentro de su campo de investigación, hemos adoptado un marco teórico de referencia y hemos revisado una variedad de autores y artículos que tratan de explicar la fijación de precios en el mercado de trabajo, ya sea a nivel teórico o empírico. Esta revisión nos ha ayudado a seleccionar un modelo econométrico sobre los determinantes de la retribución salarial, conocido como ecuación de salarios, que está compuesto por 22 variables. Por tanto, la heterogeneidad observada e incluida dentro del modelo es abundante.

El modelo estimado es el de regresión lineal múltiple con variables ficticias. Estas variables ficticias nos han permitido codificar las variables cualitativas (también llamadas categóricas) consideradas en nuestro modelo. Para aprender esta metodología y aplicarla como es debido (a la estimación de una ecuación de salarios), hemos profundizado en los aspectos de estimación e inferencia del modelo y en el uso e interpretación de las variables ficticias o categóricas.

La información a la que se le ha aplicado la metodología procede de la Muestra Continua de Vidas Laborales (MCVL), muestra que contiene conjunto de microdatos de gran tamaño sobre individuos afiliados a la Seguridad Social. La MCVL incluye información tan diversa como el género de la persona, su salario bruto anual recibido de cada empresa o pagador o su provincia de afiliación, entre otras. Agrupando las muestras correspondientes a los años 2019 y 2020, disponemos de un total de 1.798.121 observaciones (cada observación se corresponde con una relación trabajador-pagador dentro del año) que nos proporcionarán una mejor representatividad estadística de los resultados obtenidos en la estimación; a esta muestra agrupada la llamaremos la muestra conjunta.

Los resultados obtenidos para la población estudiada nos han permitido llegar a conclusiones útiles como conocer la provincia española que retribuye de peor manera a sus trabajadores afiliados o la diferencia de salario medio entre hombres y mujeres. Además, se han acompañado al estudio de las variables del modelo una serie de predicciones de salario para distintos perfiles de trabajadores, perfiles que podrían haberse dado en España en los años 2019 y 2020. Entre los resultados obtenidos, se han visualizado ciertos indicios de discriminación entre subgrupos de trabajadores españoles; si bien, haría falta una mayor investigación sobre dichos indicios.

El trabajo se estructura como sigue: En el capítulo 2, revisaremos la literatura existente sobre los determinantes del salario de un trabajador. Después, en el capítulo 3, analizamos varias teorías que tratan de explicar la distribución de las retribuciones de los trabajadores. Tras presentarlas, nos centraremos en la Teoría del Capital Humano para fundamentar nuestro modelo. Tras ello, en el capítulo 4, se hace una descripción completa de la metodología econométrica utilizada y se definirán los supuestos necesarios para poder realizar inferencia estadística sobre la población a partir de la muestra empleada. Teniendo ya un modelo para analizar, en el capítulo 5, se procederá a describir las distintas variables explicativas del modelo y su relación con la variable endógena, el rendimiento del trabajo. En el capítulo 6, interpretaremos los distintos parámetros obtenidos de la regresión múltiple estimada, identificando de esta forma las principales variables determinantes del salario del trabajador. Finalmente, en el capítulo 7, se mostrarán las principales conclusiones del Trabajo y se comentarán las implicaciones de política económica de nuestros resultados.

2 Revisión de literatura sobre retribución de trabajadores.

Desde el final de la Segunda Guerra Mundial, el número de publicaciones con estudios econométricos ha ido aumentando de forma exponencial, debido por una parte a la gran capacidad predictiva de sus modelos, y por otra al desarrollo de las herramientas computacionales necesarias para estimar dichos modelos. El valor de los descubrimientos de la econometría ha sido varias veces reconocido con el Premio del Banco de Suecia en Ciencias Económicas en Memoria de Alfred Nobel, como el entregado a Lawrence R. Klein en 1980 o a Trygve Haavelmo en 1989.

Aun así, aquellos estudios econométricos situados en Estados Unidos no se aplicaron a Europa, mayoritariamente, hasta el principio de la década de los 80. En España se empezará a popularizar la metodología a partir de la década de los 90, de tal manera que, desde 2015 hasta hoy, se ha indexado en bases de datos bibliográficas, como Scopus, un total de artículos mayor que todos los años anteriores juntos, demostrando la gran popularidad de la práctica tanto a nivel global como en nuestro país.

Los primeros estudios y modelos econométricos centrados en la influencia de la inversión en escolarización y experiencia laboral sobre los salarios posteriores de los individuos, cuyas suposiciones se desarrollan en la Teoría del Capital Humano que explicamos en el siguiente capítulo, fueron realizados por Becker (1964), Becker y Chiswick (1966) y Mincer (1958, 1962, 1974), siendo este último el que propuso la función estándar de retribución de capital humano. Esta función se demostró representativa de la estructura salarial de la época en los EE. UU. debido a su muy buen ajuste de la realidad con las previsiones que se podían obtener de ella. Aunque el modelo estándar no ha conseguido mantenerse como modelo fiable en sus ajustes con el paso de los años (Lemieux, 2006 y Heckman *et al.*, 2003), gran cantidad de variaciones de la función original, consistentes en reformulaciones matemáticas de la variables originales o la incorporación de nuevas variables explicativas, han aparecido en artículos sobre modelado de determinantes de los salarios de trabajadores. Por ejemplo, accediendo a los artículos más citados indexados en Scopus sobre el tema, podemos destacar desde el análisis de transferencia de conocimientos de empresas multinacionales a empresas nacionales de Poole, J. P (2013), hasta estudios sobre discriminación salarial femenina en direcciones ejecutivas de empresas (Gregory-Smith *et al.*, 2014).

Sobre el caso español las publicaciones basadas y/o desarrolladas a partir de la ecuación de Mincer también han sido comunes y variadas. Así, Canal-Domínguez y Rodríguez-Gutiérrez (2008) presentan evidencias de discriminación de inmigrantes a través del estudio de Encuesta de Población Activa del Instituto Nacional de Estadística (INE). Por su parte, Lassibille (1998) analiza los salarios de los años 1990 y 1991 de los sectores público y privado, concluyendo, entre otros resultados, menor retribución por educación y experiencia en el sector público. Junto a los anteriores autores, Molina y

Montuenga (2009) establecen que para una mujer tener hijos supone una reducción media del 9% de su salario.

Debido a que los microdatos a utilizar en este trabajo proceden de la Muestra Continua de Vidas Laborales (MCVL) extraída de los registros de la Seguridad Social española, es interesante comentar que, aunque el acceso a esta información sea posible solo desde el 2004 y que esté limitado a su aceptación con anterioridad por parte de la Dirección General de Ordenación de la Seguridad Social, ya hay indexados en Scopus unas 85 publicaciones utilizando esta muestra (2022). Esto es debido a la buena representatividad de los datos de la muestra, la cual es controlada y verificada por el INE por comparaciones periódicas con la Encuesta de Población Activa.

Entre los artículos más citados en Scopus que utilizan la MCVL, encontramos el de Izquierdo *et al.* (2009) con un análisis longitudinal que concluye que la alta inmigración en España en el año 2006, al menos parcialmente, contribuyó a reducir la productividad media del país. pero que después, con el paso de los años y la asimilación del capital humano específico medio de nuestra economía, el margen productivo entre los trabajadores inmigrantes y nativos se redujo rápidamente. Otro estudio longitudinal de los datos de la muestra es el que proponen Fernández-Kranz y Rodríguez-Planas, (2011) para el periodo desde 1996-2006. Estos autores comparan el trabajo a tiempo parcial y a tiempo completo en España, observando que el trabajador a tiempo parcial recibe un menor salario una vez que se controla por variables observables y heterogeneidad inobservables; concluyendo por tanto que existe una posible discriminación a este tipo de contrato laboral en España. En otros artículos, como el de Felgueroso *et al.* (2016) se busca explicación de la reducción progresiva de los salarios por nivel educativo ocurrido en España desde 1998 hasta 2008. La conclusión al que llegan los autores es que, aunque es necesario más investigación, el grado de sobrecualificación de los trabajadores, en parte, explica una disminución salarial.

Teniendo en cuenta los cambios drásticos que ha supuesto la COVID-19 para la economía española a partir del 2020, se podría suponer que la literatura publicada ha utilizado de la MCVL para el análisis riguroso de los salarios o el trabajo en este periodo. Esto no ha sido posible debido a que hay un desfase de alrededor de 2 años hasta la posibilidad de solicitar los microdatos y por ello no hay aún estudios indexados en Scopus que hagan uso de la muestra para el análisis del impacto de la pandemia en las vidas laborales y salarios de los trabajadores de España.

La publicación que antes presentó los datos sobre los cambios que ha provocado el coronavirus en los salarios se puede considerar la Encuesta Anual de Coste Laboral (EACL) para el año 2020 publicada por el INE en 2021. De esta se obtiene que el coste total por trabajador fue un 2,1% menor que el año 2019 (31.150,20€ anuales) siendo la actividad económica con menor coste la hostelería (13.323,24€) y la que más el suministro de energía eléctrica, gas, vapor y aire acondicionado (79.544,22€). La comunidad autónoma con menor coste laboral fue Extremadura (24.062,13€) y la

que más la Comunidad de Madrid (37.124,53€). Al igual que la MCVL, hoy en día no hay literatura indexada en Scopus que se haya dedicado a analizar esta información, desde el punto de vista de los salarios.

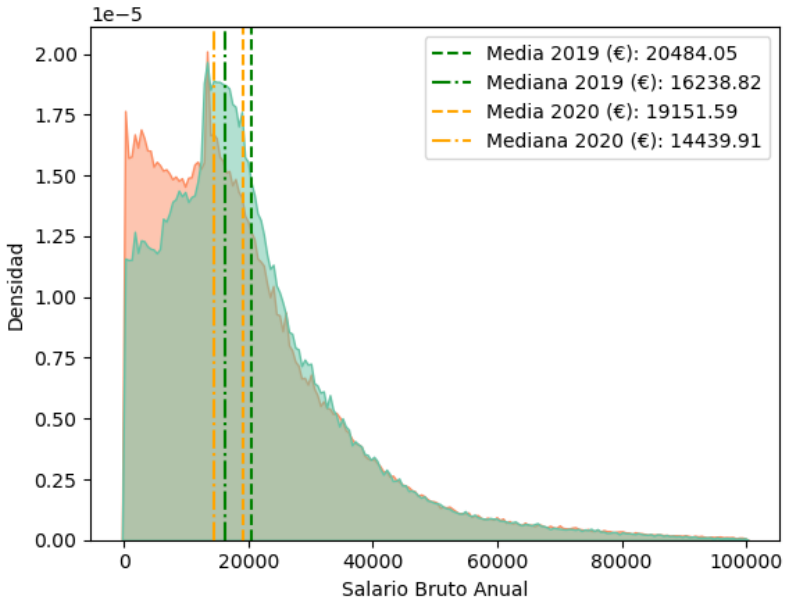
.

3 Marco teórico.

3.1 Teorías y modelos existentes.

Toda investigación econométrica está relacionada, de alguna manera, con un postulado teórico socioeconómico que busca refutar o probar basándose en la evidencia empírica. En esta práctica, como veremos a continuación, se llegan a conclusiones que son necesarias revisar y volver a comprobar en nuevas investigaciones. En esta lucha por la verdad, multitud de modelos teóricos han sido desarrollados para explicar los determinantes de las rentas de los trabajadores, pero ninguno ha conseguido explicar o predecir a la perfección las distribuciones salariales que se observan en la realidad, debido a la magnitud de los factores y variables involucrados en que una persona reciba una determinada compensación por su trabajo. Como ilustración, las retribuciones brutas anuales del año 2019 y 2020 en España se pueden ver en la Figura 3.1.

Figura 3.1: Retribuciones brutas del año 2019 y 2020 en España.



Fuente: Elaboración propia a partir de la MCVL.

Neal y Rosen (2000), en su revisión sobre las teorías de la distribución de las retribuciones, establece una serie de observaciones empíricas generales que ocurren siempre en toda distribución salarial en grandes poblaciones:

1. Las funciones de densidad de retribuciones en un periodo están sesgadas a la derecha y son leptocúrticas. Son asimétricas, presentan una cola larga a la derecha y tienen una medida de sesgo positiva (tercer momento central). Debido a esto, las retribuciones medias siempre

exceden la retribución mediana y las retribuciones más altas representan una cantidad desproporcionada de las retribuciones totales.

2. Las retribuciones cambian en gran manera al estudiar distintos grupos de trabajadores, definidos por ocupación, educación u otras atributos observados. Si estas se estudian dentro de un mismo grupo de atributos similares, también varían significativamente.
3. La dispersión de las retribuciones es mayor en trabajadores con experiencia que aquellos que acaban de iniciar su vida laboral. Un trabajador joven encontrará con una menor probabilidad salarios muy distintos al suyo, pero en cuanto aumenta la edad esta probabilidad aumenta, apreciándose mayor desigualdad en las retribuciones.

De los modelos teóricos más populares los autores de la revisión seleccionan cuatro de ellos que intentan explicar alguna o todas las observaciones descritas anteriormente. Analizamos brevemente tres de ellos en este apartado y continuaremos con el cuarto en el apartado 3.2:

Teorías estocásticas: Se centran en el problema de la "cola" derecha de las distribuciones de densidad de los salarios. La hipótesis principal de los modelos estocásticos es considerar que la retribución de una persona se genera del producto de incrementos independientes que reflejan los diversos factores que determinan la productividad de un individuo. La concepción inicial proviene de reinterpretar el Teorema Central del Límite, según el cual la distribución normal de una variable aleatoria se obtiene por la suma de componentes aleatorios independientes. Mientras estos modelos son capaces de explicar con buena precisión las rentas más altas de una distribución, no llegan a ajustarse bien a las rentas más bajas.

Modelos de selección y clasificación: Los modelos de selección, conocidos como modelos de Roy (1951). Se basan en la premisa de que según cómo los individuos se van asignando a puestos de trabajo, con intención de aumentar su salario, se irá generando una distribución salarial determinada. La formulación matemática consiste en la maximización del salario posible de cada trabajador por cada puesto de trabajo. Por su parte, los modelos de clasificación son extensiones dinámicas de los modelos de selección en los que se añade la característica de que el trabajador y las empresas no conocen de las habilidades del trabajador. En esta situación el trabajador irá conociendo su productividad con el tiempo, al igual que la empresa. La literatura plantea distintas hipótesis que definen la velocidad con la que se adquiere consciencia de las capacidades que tiene un trabajador. Al igual que con los modelos estocásticos, no consiguen ajustarse a los peores resultados de las distribuciones.

Teoría de la agencia: Esta teoría se edifica a partir de la suposición de que las empresas y los trabajadores suelen tener intereses contrapuestos, de tal manera que la empresa exige un nivel de productividad alto y el trabajador tiende a evitar tales objetivos. Para solucionar este problema, se analizan en la teoría distintas mecánicas de compensación que intentan alinear al trabajador con los

intereses de la empresa. De esta manera se llega a la conclusión de que los trabajadores no se les retribuye acorde a una determinada productividad asegurada por la competencia, si no que la productividad de los individuos vendrá determinada por la retribución que esté dispuesta a asignarles la empresa. Los resultados de la revisión concluyen que, aunque este modelo puede llegar a explicar de buena manera ciertas retribuciones medias entre sectores de la economía, no se puede llegar a conseguir hacer una descripción completa de la distribución general de los salarios.

Es importante remarcar que ninguno de los modelos anteriores ni el siguiente que veremos a continuación llega a explicar por qué las empresas tienen distintos tipos de demandas y elasticidades por la variedad de habilidades que puede llegar a tener un trabajador. Debido a que esto no es despreciable, la literatura empírica intenta también responder a la necesidad de incluir la demanda de trabajo en los modelos retributivos, pero para no superar el alcance del trabajo no trataremos este tipo de modelos. En el siguiente apartado, describiremos con más profundidad la teoría que servirá como base para formular el modelo econométrico utilizado en este trabajo: la teoría del capital humano.

3.2 El modelo de capital humano.

Según Neal y Rosen (2000), la idea central de teoría moderna del capital humano fue formulada por Adam Smith en su obra *La Riqueza de las Naciones* (1776), en el que se expone que al igual que debe esperarse que una máquina reemplace el capital invertido en ella por una determinada tasa de beneficios hasta su depreciación de un hombre educado en un determinado tiempo y trabajo se debe esperar lo mismo. Por tanto, un trabajador que haya invertido "trabajo y tiempo" en su educación tendrá una retribución mayor que aquel que no haya realizado esa inversión. Este supuesto obliga que haya siempre una cierta desigualdad en la retribución de salarios de una economía, debido a que se debe retribuir de mejor manera aquellos puestos laborales que necesiten de mayor "inversión educativa" respecto a aquellos que no. La teoría moderna del capital humano se centra en como varían los salarios según los costes de la educación, la experiencia y el entrenamiento profesional de los trabajadores. Mincer (1958) fue de los primeros autores en formalizar las ideas de Adam Smith que se ocupan de la desigualdad salarial y de los requisitos de experiencia productiva en una ocupación laboral.

Neal y Rosen (2000) presentan este ejemplo para explicar el modelo más básico de Capital Humano: Supongamos un trabajador que quiere maximizar su retribución y que tiene a su disposición dos ofertas de trabajo distintas, una necesita d periodos de entrenamiento y paga un salario de W_0 por periodo y otra requiere $d+s$ períodos, pero paga un salario por período de W_1. El trabajador trabaja

en su vida hasta *n* periodos y actualiza los salarios futuros según una tasa *ρ.* El trabajador no tiene preferencia de puesto si los dos puestos le retribuyen el mismo salario en su vida, tal que:

$$\int_{d}^{n} W_0 e^{-\rho t}\, dt = \int_{d+s}^{n} W_1 e^{-\rho t}\, dt \tag{3.1}$$

La tasa que permite la indiferencia entre dos ocupaciones es:

$$k(d,s,n) = \frac{W_1}{W_0} = \frac{e^{-\rho n} - e^{-\rho d}}{e^{-\rho n} - e^{-\rho(d+s)}} \tag{3.2}$$

La tasa *k* mide el precio relativo de demanda de la ocupación que necesita de educación previa. Aumentar el tiempo de *s* postpone la entrada en el mercado laboral y aumenta las ganancias, entonces *k* debe aumentar para mantener la indiferencia. Si la vida laboral del individuo aumenta, implica un menor ingreso por período, y aunque la retribución total no aumente, *k* se reduce. Si se considera que *n* va hasta el infinito, *k* solo es dependiente de s (Mincer, 1974):

$$\ln W_1 = \ln W_0 + \rho s \tag{3.3}$$

En la práctica se puede observar para cualquier individuo *i* su salario y número de años de educación, pero no se puede observar W_0 por lo que Mincer propone una nueva formulación:

$$\ln W_i = \alpha + \rho s_i + \varepsilon_i \tag{3.4}$$

siendo ε_i un término de error (Cahuc *et al.*, 2014) que expresa la heterogeneidad de los individuos y los coeficientes α y ρ términos a estimar. El estimador de mínimos cuadrados ordinarios de la tasa de retribución ρ sería insesgado si el resto de los términos fueran independientes. Esto no es posible ya que las capacidades individuales medidas por ε_i influyen en el tiempo necesario para educarse s_i, por tanto, se debe establecer ρ como sesgado.

Para mejorar el ajuste del modelo propuesto inicial, Mincer asume que es posible obtener capital humano después de la educación formal a través de experiencia profesional, principalmente al inicio de su vida laboral, a la vez que considera que la acumulación de capital humano decrece de forma lineal con el tiempo pasado desde la salida de la educación reglada. Añadiendo estos supuestos a la fórmula original se obtiene la ecuación de salario estándar de Mincer:

$$\ln W(s+x) = \ln W(0) + \rho s + \rho_x t_0 x - \rho_x \left(\frac{t_0}{2T}\right) x^2 \tag{3.5}$$

siendo $W(s+x)$ la retribución por periodo pasado un tiempo *s* de educación y *x* el tiempo de experiencia laboral. Los parámetros ρ y ρ_x se consideran las tasas de retribución de la educación y la experiencia laboral respectivamente. Por su parte, t_0 es el momento temporal en el que se salió de la educación reglada y T el tiempo que ha pasado desde entonces. Este modelo de Mincer es relativamente simple debido a que se sustenta en base a una serie de hipótesis muy restrictivas que pueden llegar a sesgar los resultados. Brevemente, estas suposiciones son:

1. La tasa de retribución por un año más de educación es independiente de la duración de los estudios.
2. El coste de añadir un año más de educación reglada es proporcional al salario.
3. La duración de la vida laboral se considera lo suficientemente larga, según lo visto en la ecuación (3.3).
4. La duración de la vida laboral es independiente de la duración de la educación reglada.
5. Los salarios son funciones multiplicativas respecto a la experiencia y a la educación.

Es muy probable que estas hipótesis carezcan de validez. Por ejemplo, el coste de la educación no se puede suponer únicamente como ingresos no adquiridos por el hecho de no trabajar de forma indirecta aumentando *s*, concentrando mayor retribución en menor tiempo, sino también costes directos como costes psicológicos, préstamos estudiantiles o tasas de matrícula.

Heckman *et al.* (2008) proponen calcular los rendimientos de la educación desde una interpretación menos restrictiva de la realidad que las suposiciones anteriores de Mincer. Asumen que la edad de jubilación puede depender de la duración de su educación reglada, por lo que el final del trabajo no se produce a unos *n* periodos fijos, sino a los *n*+*s* periodos. Además, introducen la influencia de tasas de matrícula al modelo y calculan ρ a través de los valores del censo estadounidense. Para llegar a calcular la ecuación final obtienen información empírica, como que los trabajadores pasan 47 años de media trabajando independientemente de sus estudios (n=47) o información sobre tasas de matrícula e impuestos. Con esto se obtiene una nueva reformulación de la función de Mincer (3.4) que vimos anteriormente, tratando de estimar las retribuciones logarítmicas según los distintos perfiles educativos de trabajadores. Posteriormente, utilizando el mismo razonamiento que ya analizamos en la ecuación (3.1), obtienen el parámetro ρ que genera la indiferencia entre los perfiles censales. La función utilizada para la estimación que utilizan los autores es:

$$\ln W = \alpha + \beta s + \delta x - \gamma x^2 + \varepsilon \qquad (3.6)$$

Siendo W la retribución del periodo x de trabajo, α el término constante de retribución logarítmica, s una variable categórica que codifica niveles educativos progresivamente mayores. El parámetro δ es la tasa de remuneración según el aumento de periodos realizados y el parámetro γ, al contrario, se define como la tasa de reducción cuadrática de la retribución. Por último, ε es el término de error.

En el capítulo 4 entraremos más en profundidad en la estimación de modelos como este. Lo remarcable aquí es que se define s como una variable categórica capaz de introducir nueva información a la ecuación, mientras se mantiene la variable continua de la experiencia laboral. Los parámetros de la ecuación deben ser posteriormente estimados: las tasas de retribución β, δ y γ, el

término de error ε y el término constante α. Esta modificación de la fórmula obtiene mejores ajustes que las anteriores ecuaciones analizadas y llega a conclusiones interesantes, como que los aumentos de salarios debidos a educación reglada son mayores al terminar la educación secundaria y los estudios superiores con respecto a otros niveles educativos.

Heckman *et al.* (2008) demuestran en su trabajo como los supuestos de Mincer llevan a grandes sesgos al analizar las variables que influyen en las retribuciones de los trabajadores. Aun así, también abren nuevos caminos hacia futura investigación al relajar las hipótesis iniciales de la teoría del capital humano y obtener ajustes más cercanos a la realidad empírica, como veremos que realizaron varios autores en el siguiente apartado.

3.3 Desarrollos de la teoría del capital humano.

De las hipótesis simplificadas de Mincer al comportamiento humano frente a la escolarización y la adquisición de experiencia productiva hay muchos campos por explorar. Desde el renacimiento de la teoría por Becker (1964) y las formulaciones de Mincer, gran cantidad de autores han enriquecido la literatura con apreciaciones, nuevas hipótesis y descartes de anteriores suposiciones tras analizarse empíricamente. Para ilustrar las múltiples posibilidades que se pueden tener en cuenta en un modelado, mostraré distintas interpretaciones publicadas.

Weiss (1986) presenta que, si se tiene en cuenta que las horas trabajadas y la retribución por horas varían durante una vida laboral, normalmente aumentando desde el inicio de la vida profesional hasta llegar a su máximo antes de la jubilación. Añadiendo el número de horas trabajadas al modelo se pueden analizar, para distintos perfiles laborales, las elecciones de consumo, ocio e inversión en capital humano, teniendo en cuenta la posibilidad de instrucción mientras se trabaja. Por ello, habría una intención de aumentar el capital humano, trabajando más, al inicio de la vida laboral para después disfrutarlo con una reducción de horas de trabajo años después. Teniendo en cuenta también que una mayor inversión en capital humano puede suponer una reducción de riesgo en las posibles retribuciones futuras, esta suposición adquiere aún más fuerza.

Spence (1973) recupera la idea antes vista de que los estudios reglados mejoran empíricamente las retribuciones salariales y se dedica a comprender las posibles relaciones causales de esta interacción. El autor llega a la conclusión de que la educación reglada actúa de "filtro" para la selección de individuos en ocupaciones laborales, sin llegar a afectar la eficiencia productiva, considerándola esta última una característica inherente al individuo. Esta estaría determinada por factores relativamente independientes de la educación tales como el entorno familiar, historia personal, etc. Las personas más capaces en su vida activa general se hacen entonces también las más capaces en su educación y según esto aumentar los años de estudio no tendría relevancia, sino el

"probar" al individuo en los distintos "filtros" que suponen entornos como el colegio, instituto o universidad. De esta manera, un título educativo supone una herramienta del trabajador para señalar sus cualidades intrínsecas individuales. A esta teoría se le conoce como la teoría de la señalización y aunque nace de unas premisas parecidas a la Teoría del Capital Humano, sus conclusiones son muy distintas. Una de ellas es que, aunque el capital humano siempre se beneficia de un aumento de los años de entrenamiento, Spence afirma que los trabajadores tienden a sobreeducarse, mucho más lejos del estándar de lo socialmente eficiente, con el objetivo de destacar frente a los empleadores.

Un modelo basado en la teoría de la señalización tiene en cuenta esta secuencia de hechos:

1. Los trabajadores optan a un nivel de educación *s.*
2. Las empresas entran en el mercado laboral libremente, observan las señales *s* y hacen ofertas simultaneas a los trabajadores.
3. Los trabajadores aceptan o rechazan las ofertas.

En el modelo propuesto, se muestra el rol de la educación de forma muy negativa, representándola como una mera forma de seleccionar a los trabajadores según su eficiencia, sin mejorar la asignación de recursos en la economía. Este problema que se intenta solucionar en nuevas iteraciones de la fórmula. Además, la suposición de la existencia de sobreeducación ha sido debatida por otros autores, como Weiss (1983), en la que se remarca algo que sucede en la realidad: las empresas suelen contratar a estudiantes que aún no han terminado sus estudios. De esta manera, las suposiciones de Spence se ponen en duda y deben de reformularse en nuevos términos. De este modo se llega hasta el punto de contradecir las posturas de la ineficiencia económica del sistema educativo, al resaltar Weiss que, según lo que observa, son realmente los individuos con menos cualidades los que suelen estudiar niveles educativos superiores.

Card (1999), Blundell *et al.* (2005) y Blundell y Costa Dias (2009) ahondan en lo que es conocido como el problema de selección en econometría. En los artículos publicados se intenta explicar la relación causal entre el tiempo de entrenamiento, sea en escolarización o experiencia laboral, y las retribuciones. Los autores entonces hablan de lo que se conoce como el sesgo de habilidad, en los que las estimaciones salariales obtenidas vistas anteriormente estarían sobreestimadas. Esto se debe a que las estimaciones de capital humano y señalización entienden a los individuos más productivos como los que además estudian más años, pero que tal vez las retribuciones no vengan por el tiempo de preparación, sino por las capacidades individuales de trabajo. La literatura ha intentado en múltiples ocasiones probar o refutar esta hipótesis de distintas maneras, hasta en algunos casos llegar a estudiar el comportamiento de gemelos monocigóticos en sus vidas educativas, pero sin llegar a comprobar si realmente estamos ante un caso de sesgo o no de forma conclusa.

Ante las distintas teorías propuestas y para concluir, hay que recuperar lo comentado en la revisión de literatura de este trabajo: la falta de inclusión de la demanda empresarial de trabajadores y las características de esta en los modelos. Debido a que estas teorías se centran en las cualidades productivas de los trabajadores en una población y sus retribuciones heterogéneas, obteniendo innegables buenos resultados, se suele dejar de lado el aspecto importante de la demanda empresarial que suele ser, como complicación añadida al estudio, difícil de adquirir para el investigador medio. Abowd *et al.* (1999) publican un artículo basado en una muestra longitudinal de un millón de trabajadores franceses empleados en más de 500 empresas. Los microdatos utilizados incorporan información detallada de las firmas estudiadas. Los resultados presentados son:

- La heterogeneidad de variables personales y de las empresas es un determinante importantes de las retribuciones de un trabajador, aunque las personales influyen en mayor medida.
- Las variables personales son capaces de explicar el 90% de las retribuciones interempresariales y el 75% de las retribuciones según el tamaño de la empresa. Las características empresariales explican menos que las personales.
- Las empresas que contratan a trabajadores que solicitan mucho más salario que la media suelen ser más productivas, pero no las más rentables. Además, dependen más de capitales mayores de la empresa y de trabajo muy cualificado.
- Aquellas empresas que pagan salarios generalmente mayores a la media del sector son más productivas y rentables. Estas también son más dependientes del capital de la empresa, pero no necesitan de trabajo muy cualificado.

Ante estos resultados, la demanda de las empresas, caracterizada por las distintas variables que conforman el tejido productivo de un país, se ve como determinante menor de los salarios, pero significativo e influyente. Desde la fecha de publicación hay que tener en cuenta que este artículo se ha convertido en uno de los más citados sobre el análisis de cómo influye la demanda empresarial en los salarios, indicando que ha abierto nuevos frentes de investigación. Además, están llegando sets de datos cada vez más completos al acceso de más investigadores, que ayudan a que se estén debatiendo nuevas teorías y modelos aún más complejos.

4 Metodología.

Para la explicación de la metodología utilizada en este trabajo, la regresión múltiple con variables ficticias, nos basaremos en los capítulos que explican el fundamento de la regresión múltiple de los manuales de Wooldridge (2012) y Gujarati (2015).

4.1 Regresión múltiple: Estimación.

El modelo de regresión múltiple es actualmente la herramienta estadística más ampliamente usada para análisis empírico en economía y en otras ciencias sociales. Esto es debido a que, si se tiene una muestra de sección transversal de una población desconocida, podemos utilizar la información contenida en ella para construir modelos que expliquen las relaciones que subyacen entre las características de la población. Este modelo puede ser luego verificado para comprobar que su representación de la población desconocida de referencia es estadísticamente correcta. La obtención de los parámetros que definen a un modelo a través de la muestra es lo que llamaremos la estimación del modelo muestral de regresión múltiple. Tras haber probado que nuestra estimación es representativa de la población, trasladaremos los resultados muestrales a toda la población. En este momento hablamos de la inferencia estadística del modelo.

Los modelos de regresión múltiple, concretamente, pretenden explicar el comportamiento de una variable dependiente y, o también conocida como regresando o variable endógena, utilizando la información de un conjunto de variables independientes $\{x_1, x_2, \dots, x_k\}$, también denominadas regresores, covariables o variables exógenas o explicativas. El modelo de regresión lineal múltiple aplicado a una población estadística determinada puede escribirse como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots + \beta_k x_k + u \qquad (4.1)$$

siendo β_0 el término constante de la ecuación, midiendo valor esperado de y en el caso de que todas las variables explicativas sean 0, β_k el parámetro o estimador asociado a la variable x_k y u el término de error, el cual tiene naturaleza aleatoria y, por tanto, sigue una determinada distribución de probabilidad. Los parámetros β son llamados en conjunto coeficientes de regresión. El término de error de la observación, o perturbación aleatoria, contiene factores distintos de x_k que no son observados directamente por el investigador y que también afectan a la variable endógena.

Dicho error mide la distancia entre el valor observado de y y su valor esperado condicionado a las variables explicativas, por lo que la ecuación (4.1) puede escribirse tal que:

$$y = E(y|x) + u \tag{4.2}$$

$E(y|x)$ recibe el nombre de función de regresión poblacional (FRP) y al estimarla a partir de la muestra, ya que no conocemos la población completa, pasa a llamarse función de regresión muestral (FRM):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ \ldots + \hat{\beta}_k x_k \tag{4.3}$$

en la que los parámetros poblaciones de regresión pasarían a ser sus estimaciones muestrales. La técnica para estimar los parámetros del modelo es mínimos cuadrados ordinarios (MCO), que consiste en minimizar la suma de las desviaciones al cuadrado de la FRM respecto a los valores reales de *y,* para las *n* distintas observaciones:

$$\sum_{i=1}^{n} \left(y_i - \widehat{\beta_k} x_{ik}\right)^2 \tag{4.4}$$

A esta ecuación también se le denomina la suma de cuadrados de residuos del modelo estimado (SCR). Asimismo, a las desviaciones, $\widehat{u_i} = y_i - \widehat{y_i}$, se le denominan residuos y se consideran al cuadrado para evitar compensaciones entre valores positivos y negativos en el proceso de minimización de la SCR. Si el residuo es mayor que 0, la variable dependiente ha sido subestimada y si es menor que 0 ha sido sobreestimada. Si el modelo está bien estimado, la media de los residuos debe ser 0 y la covarianza entre los residuos y las variables independientes debe ser nula, como veremos más adelante.

Para una regresión, si queremos comprobar la bondad del ajuste obtenido, es útil utilizar el coeficiente de determinación (R^2). Este coeficiente que siempre estará entre 0 y 1, mide si los residuos han sido, de media, de poco o mucho valor. En el caso de R^2 cercano a 0, podemos concluir que nuestro modelo no es capaz de explicar la variable endógena con precisión a partir de variables independientes seleccionadas. Al contrario, valores cercanos a 1, nos informa de que nuestro modelo sí cumple. No existe un valor correcto de R^2 en la literatura con el que descartar o aceptar unos resultados. Sin embargo, mejorar los valores de R^2 que otros autores pueden haber conseguido para estudios parecidos al nuestro puede ser indicador de buenos resultados, aunque no sea el único que debemos tener en cuenta. Para calcular el coeficiente de determinación, definimos tres sumas de cuadrados que usamos en su cálculo:

Tabla 4.1: El coeficiente de determinación y sus componentes.

Suma de Cuadrados Totales	$\text{SCT} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = n\,Var(y_i)$	SCT=SCE+SCR
Suma de Cuadrados Explicada (por el modelo estimado)	$\text{SCE} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = n\,Var(\hat{y}_i)$	SCE=SCT-SCR
Suma de Cuadrados de Residuos (del modelo estimado)	$\text{SCR} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{u}_i^{\,2}$	SCR=SCT-SCE
Coeficiente de determinación	$R^2 = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}}$	

Fuente: Elaboración propia.

Otro aspecto importante del modelo de regresión es la interpretación de los coeficientes. Los regresores estimados se interpretan como efectos parciales en la estimación. Si, por ejemplo, tomamos una FRM de tamaño *k*=2 y consideramos un incremento de $\hat{y}$, tenemos que:

$$\Delta\hat{y} = \widehat{\beta_1}\Delta x_1 + \widehat{\beta_2}\Delta x_2 \qquad (4.5)$$

Si mantenemos con un incremento nulo una de las dos variables, el incremento de la variable dependiente será igual a la variable independiente restante únicamente: $\Delta\hat{y} = \widehat{\beta_1}\Delta x_1$ o $\Delta\hat{y} = \widehat{\beta_2}\Delta x_2$. Esta interpretación de derivada parcial es uno de los motivos por el que los modelos de regresión múltiple son considerados de gran utilidad. Cada parámetro de las variables explicativas indica el aumento de $\hat{y}$ al aumentar una unidad x, mientras se mantienen el resto de las variables fijas. En el momento que trabajemos con otras formas funcionales, más adelante, comentaremos sus interpretaciones específicas de parámetros.

Para la obtención de buenos resultados en datos de sección cruzada, es necesario que se cumplan ciertas suposiciones sobre el modelo lineal planteado, el residuo obtenido y la muestra transversal a utilizar. Estas condiciones son conocidas como los 5 supuestos de Gauss-Markov para la obtención de Mejores Estimadores Lineales Insesgados (MELI):

4.1.1 Linealidad en los parámetros.

Supuesto 1*:* El modelo puede ser escrito como la ecuación (4.1).

Es importante remarcar que la linealidad se asume únicamente en los parámetros β_k y no a las covariables $\{x_1, x_2, \dots, x_k\}$. Dicha linealidad en los parámetros impone la existencia de una relación lineal entre la variable endógena y las variables exógenas.

Este supuesto no es realista en algunas aplicaciones económicas, como puede ser el salario y la educación que analizamos en el anterior capítulo. Para superar esta limitación, se crean relaciones no lineales entre variables, pero manteniendo la linealidad en los parámetros. De esta manera, algunas relaciones no lineales en los parámetros se pueden linealizar transformando sus variables, tal que sus variables transformadas si puedan ser explicadas por un modelo lineal. A continuación, indicamos un resumen de formas funcionales para posibles modelos lineales y como son sus pendientes.

Tabla 4.2: Resumen de formas funcionales lineales en los parámetros.

Modelo	Modelo no lineal (linealizable)	Forma lineal	Interpretación
Nivel-Nivel	-	$y = \beta_0 + \beta_1 x + u$	$\beta_1 = \frac{\Delta \mathrm{y}}{\Delta \mathrm{x}}$
Nivel-Cuadrado	-	$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$	$\frac{\partial \mathrm{y}}{\partial \mathrm{x}} = \beta_1 + 2\beta_2 x$
Log-Log	$y = e^{\beta_0} X^{\beta_1} e^u$	$\ln y = \beta_0 + \beta_1 \ln x + u$	$\beta_1 = elasticidad = \frac{\Delta(\%)\mathrm{y}}{\Delta(\%)\mathrm{x}}$
Log-Nivel	$y = e^{\beta_0} e^{\beta_1 x} e^u$	$\ln y = \beta_0 + \beta_1 x + u$	$\beta_1 = semielast. = \frac{\Delta(\%)\mathrm{y}}{\Delta \mathrm{x}}$
Nivel-Log	$e^y = e^{\beta_0} X^{\beta_1} e^u$	$y = \beta_0 + \beta_1 \ln x + u$	$\beta_1 = semielast. = \frac{\Delta \mathrm{y}}{\Delta(\%)\mathrm{x}}$

Fuente: Elaboración propia.

4.1.2 Muestreo aleatorio.

Supuesto 2*:* Tenemos una muestra aleatoria simple de *n* observaciones, $\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i): i = 1, 2, \dots, n\}$, que siguen el modelo establecido en el supuesto anterior para cada una de ellas.

Según este supuesto, todas las observaciones de la muestra deben contener información para todas las variables del modelo y mantener la misma FRP en todas las observaciones.

4.1.3 No existe colinealidad perfecta.

Supuesto 3: En la muestra, al igual que en la población, ninguna de las variables independientes es constante (salvo la propia constante de regresión) y no hay relaciones lineales exactas entre ellas.

Si una variable es una combinación lineal exacta de otras variables independientes el modelo, éste experimenta de colinealidad perfecta y no se puede estimar por MCO. Puede existir cierta correlación entre las variables explicativas, pero esta nunca puede ser perfecta o demasiado elevada. Que exista una elevada correlación entre covariables supone un problema de estimación, ya que aumenta la varianza de los coeficientes estimados dificultando la inferencia sobre la población.

4.1.4 Media condicional nula.

Supuesto 4: El error *u* tiene un valor esperado de 0 dado cualquier valor de sus variables independientes:

$$E(u|x_1, x_2, \dots, x_k) = 0 \tag{4.6}$$

Esta propiedad se conoce también como exogeneidad del término de error y consiste en que las covariables no deben contener información sobre los valores del término de error de cada unidad muestral. Esto permite, junto al supuesto 2, que la covarianza entre las variables independientes y el error sea también 0. La explicación de este supuesto es que cualquier variable no observada debe estar incorrelada con cualquiera de las variables independientes del modelo. En el caso de no cumplirse la exogeneidad hablamos de su caso contrario: la endogeneidad. Pueden ocurrir casos de endogeneidad, por ejemplo, al dejar inobservadas variables relevantes para el modelo, llegando entonces a estimadores sesgados e inconsistentes.

4.1.5 Homocedasticidad.

Supuesto 5: El error *u* tiene la misma varianza dado cualquier valor de las variables independientes:

$$Var(u|x_1, x_2, \dots, x_k) = \sigma^2 \tag{4.7}$$

Este supuesto significa que la varianza del error es igual en toda combinación de variables explicativas que se utilicen. Si este supuesto falla, el modelo presenta heterocedasticidad. Si se cumple, todas las variables explicativas deberán de tener una distribución de densidad de probabilidad igual. De la ecuación (4.7), obtenemos que la varianza del estimador de los parámetros es:

$$Var(\widehat{\beta}_j) = \frac{\sigma^2}{\mathrm{SCT}_j(1 - R_j^2)} \tag{4.8}$$

4.1.6 Teorema de Gauss-Markov.

Teorema: Teniendo en cuenta las cinco suposiciones anteriores, $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_k}$ son los mejores estimadores lineales insesgados (MELI) de $\beta_0, \beta_1, \dots, \beta_k$ respectivamente.

Debido a este teorema y a que nos vamos a mover en el entorno de modelos estadísticos lineales, podemos usar la técnica de estimación de MCO. En los siguientes apartados entraremos más en profundidad en el análisis de inferencia de los MELI propuestos.

4.2 Regresión múltiple: Inferencia.

Según vimos anteriormente, bajo los supuestos 1 a 4 de Gauss-Markov podemos considerar los estimadores de MCO insesgados y añadiendo el supuesto 5 podemos medir sus varianzas. El teorema nos asegura que estos estimadores, si se dan bajo los 5 supuestos, son los MELI. Dicho esto, para realizar inferencia estadística es necesario conocer no solo la insesgadez y varianza del estimador, sino toda su distribución muestral, por lo que nos vemos obligados a establecer un nuevo supuesto:

4.2.1 Normalidad.

Supuesto 6: El error del modelo poblacional u debe cumplir los supuestos 4 y 5 de Gauss-Markov y además tener una distribución normal de media cero y varianza σ^2:

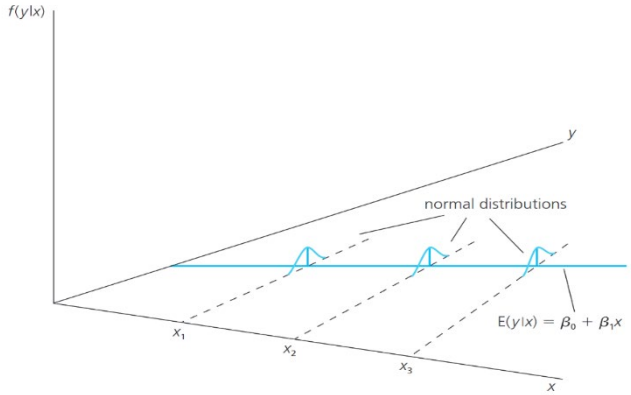
$$u \sim \text{Normal}(0, \sigma^2) \tag{4.9}$$

Este supuesto refuerza los de Gauss-Markov para que los errores no presenten distintas formas de distribución, fijándolas a la normal. Para regresiones transversales, se le llama al conjunto de las seis suposiciones anteriores suposiciones del modelo clásico lineal (MCL). Bajo estas suposiciones los estimadores de MCO se pueden considerar estimadores insesgados de varianza mínima tanto entre los estimadores lineales como entre los no lineales. Gracias a esto, no habrá ningún estimador con menor varianza bajo estas suposiciones. Una forma de resumir los supuestos del MCL es mediante la siguiente:

$$\text{y|x} \sim \text{Normal}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots + \beta_k x_k, \sigma^2) \tag{4.10}$$

Consecuencia de esta ecuación, la variable aleatoria *y*, condicionada a $\{x_1, x_2, \ldots, x_k\}$, tiene una distribución normal con media fijada en cada x_k y una varianza constante. Esto obliga a que el punto definido por la media de cada variable explicativa esté siempre en la línea de regresión. Si modelamos una única variable independiente, la ecuación (4.10) define la Figura 4.1.

Figura 4.1: Distribución normal homocedástica con una variable independiente.



Fuente: Wooldridge (2012)

Como se explicó en el apartado 4.1.1, es posible realizar transformaciones de las variables de modelos no lineales (en los parámetros) para lograr que las variables transformadas sí se relacionen de forma lineal. Por ejemplo, en este trabajo utilizaremos las transformaciones log-log entre la variable endógena y varias variables independientes. Asimismo, el logaritmo de la variable endógena puede mostrar una distribución más próxima a la Normal que el propio nivel de la variable como veremos más adelante en el capítulo 5, y eso siempre facilita la inferencia estadística.

Con el supuesto de normalidad asumido, debemos contrastar, a partir de la muestra, las hipótesis de significancia estadística de los parámetros de la FRP (ecuación 4.1). Como nunca podremos conocer los valores reales de una determinada β_j, ya que no conocemos la población, lo que haremos es hipotetizar sobre sus valores y utilizar la muestra para aceptar o no dichas hipótesis. Para ello, debemos definir:

4.2.2 Distribución *t* para los coeficientes estandarizados.

Bajo los supuestos del MCL:

$$\frac{(\widehat{\beta_J} - \beta_j)}{se(\widehat{\beta_J})} \sim t_{n-k-1} = t_{gl} \qquad (4.11)$$

donde *k+1* es el número de parámetros desconocidos del modelo, *n* el tamaño de la muestra utilizada, *n-k-1* los grados de libertad del modelo (*gl*) y *se* el error estándar del estimador (*standard error*).

Este teorema es importante ya que permite que contrastemos distintas hipótesis sobre β_j. La que nosotros utilizaremos en nuestro trabajo es el contraste de hipótesis bilateral, cuya hipótesis nula e hipótesis alternativa, respectivamente, son:

$$H_0: \beta_j = a_j \quad H_1: \beta_j \neq a_j \tag{4.12}$$

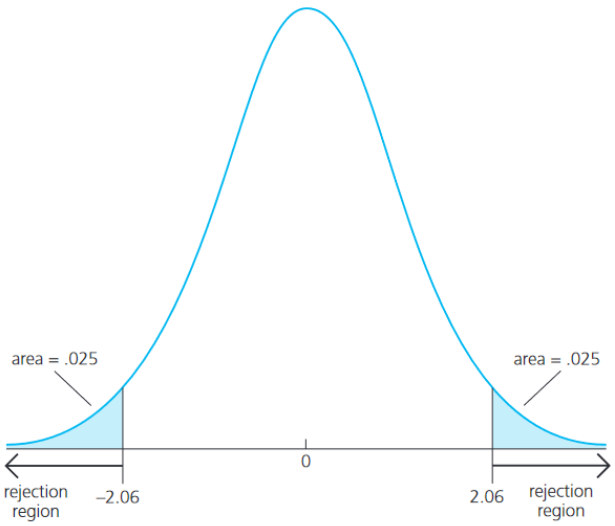
donde a_j es el valor hipotetizado de β_j. Con ello, obtenemos su estadístico utilizando ecuación (4.11):

$$t = \frac{(\hat{\beta}_j - a_j)}{se(\hat{\beta}_j)} \tag{4.13}$$

Este estadístico *t* representa la distancia de las desviaciones entre el valor estimado y el valor de la hipótesis nula si los representáramos en sus funciones de densidad de probabilidad. Unos valores menores del estadístico *t* conllevan una mayor probabilidad de que ocurra la hipótesis nula y, al contrario, unos valores mayores provocan una menor probabilidad.

Aun así, necesitamos establecer un criterio sobre el estadístico para saber si podemos descartar o no la hipótesis nula. Para ello, establecemos un valor límite de probabilidad acumulada en las colas de la distribución *t* llamado valor crítico *c*, que consideraremos como la probabilidad de que erremos si no rechazamos H_0. Este valor crítico se calcula estableciendo un porcentaje arbitrario de observaciones de la muestra, distribuida según t_{gl}, que consideraremos que rechazan la hipótesis nula: el nivel de significancia (α), que en la práctica econométrica es del 5%. Para cuando los grados de libertad de un modelo sean 25 y se quiera rechazar $H_0: \beta_j = 0$, la Figura 4.2 sirve para visualizar la distribución y las regiones críticas.

Figura 4.2: Regiones críticas bilaterales para una significancia del 5% y 25 grados de libertad.



Fuente: Wooldridge (2012)

Debido a que trabajaremos con dos colas, si el valor crítico es menor que el valor absoluto del estadístico *t* calculado, rechazamos la hipótesis nula a favor de la hipótesis alternativa. En este caso hablamos de que "$\widehat{\beta_j}$es estadísticamente diferente de a_j" para el nivel de significancia utilizado. Si $a_j = 0$, hablaremos de que "x_j es estadísticamente significativo" para el α utilizado.

La necesidad de establecer un nivel de significancia antes de realizar un contraste *t* supone un problema, ya que, dependiendo del α arbitrariamente seleccionado, los resultados que se puedan obtener son potencialmente distintos. Además, es importante remarcar que no existe un nivel de significancia que podamos considerar correcto. Un α puede llegar a esconder información útil sobre los resultados de un contraste estadístico, como, por ejemplo, en el caso de un investigador que quiera obligar a la prueba a rechazar la hipótesis nula aumentando el nivel de significancia. Para solucionar este problema, es necesario normalmente hacerse esta pregunta: Dado el valor observado del estadístico *t*, ¿cuál es el menor nivel de significancia con el que la hipótesis nula se rechazaría?

La respuesta de la pregunta se llama p-valor. Este es igual al nivel de significancia del contraste cuando usamos el valor del estadístico *t,* calculado como valor crítico de la prueba. Como el p-valor es una probabilidad, su valor se comprende siempre entre 0 y 1. El p-valor de cada variable explicativa de nuestro trabajo se calculará automáticamente con cada regresión usando nuestro software estadístico, tal que $P(|T| > |t|)$, donde T es una variable aleatoria *t* distribuida. El p-valor resume la fuerza o debilidad de la evidencia empírica contra la hipótesis nula. Una interpretación útil es considerar el p-valor como la probabilidad de observar un estadístico *t* como el nuestro si la hipótesis nula es verdadera. Por lo tanto, esto significa que valores pequeños de p-valor proveen evidencia contra H_0. En nuestro trabajo, descartaremos la hipótesis nula cuando el p-valor tome valores menores que α = 0,05, como se suele hacer en la literatura econométrica, considerando significativos aquellos estimadores en los que ocurra.

Otra forma de contrastar las hipótesis anteriores es construir intervalos de confianza (IC) para el parámetro poblacional β_j. Sabiendo que $(\widehat{\beta_j} - \beta_j)/se(\widehat{\beta_j})$ tiene una distribución *t* con *n-k-1* grados de libertad, para los límites inferiores y superiores de un intervalo de confianza del 95%, necesitamos resolver:

$$\widehat{\beta_j} \pm c * se(\widehat{\beta_j}) \tag{4.14}$$

donde la constante *c* es el percentil número 97.5 en la distribución t_{n-k-1}. Dadas las hipótesis de las ecuaciones (4.12), podremos rechazar la hipótesis nula a favor de la alternativa si, para el nivel de significancia utilizado, a_j no está dentro del intervalo de confianza. Igual que en el p-valor, el software estadístico utilizado calculará automáticamente cada uno de los IC.

Habiendo revisado la forma de comprobar si una de nuestras variables independientes no tiene efecto en nuestra variable dependiente, nos interesa comprobar el mismo caso, pero para la totalidad

de nuestro modelo. Debemos descartar la hipótesis nula de que todo los regresores sean iguales a 0, a lo que llamaremos el modelo restringido, a favor del modelo en los que los regresores son distintos de 0, que lo trataremos como el modelo no restringido. Para ello, definimos el estadístico *F* tal que:

$$F \equiv \frac{\frac{(\mathrm{SCR}_r - \mathrm{SCR}_{nr})}{k}}{\frac{\mathrm{SCR}_{ur}}{n-k-1}} \tag{4.15}$$

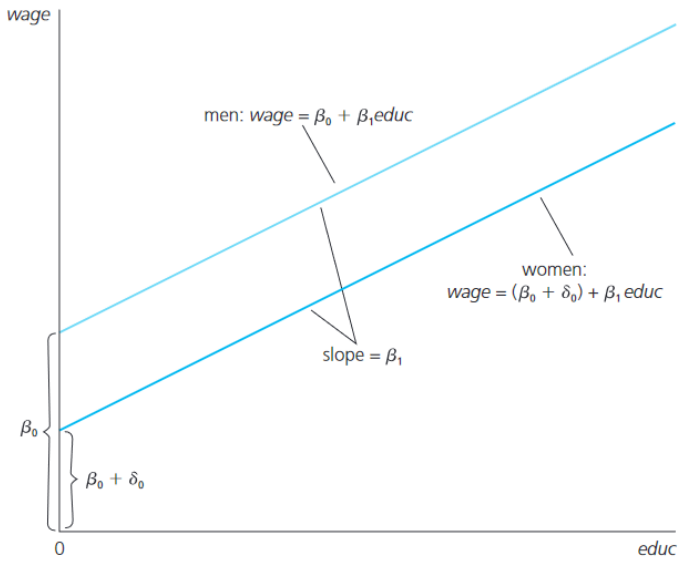
donde SCR_r es la suma de cuadrados de los residuos del modelo restringido y SCR_{nr} la suma de cuadrados de los residuos del modelo no restringido. Como $F \sim F_{k,n-k-1}$, podemos descartar la hipótesis nula si el *F* calculado supera el percentil crítico $F_{k,n-k-1,1-\alpha}$ y de esta manera probar la validez global de nuestro modelo. Al igual que con los anteriores indicadores, el programa utilizado nos devolverá un *F* con cada regresión.

4.3 Regresión múltiple con variables ficticias.

Al querer introducir en un modelo información cualitativa es necesario codificarla de alguna manera. Una forma de hacerlo es establecer variables binarias para cada cualidad que queramos estudiar, como, por ejemplo, tratar el género masculino como la ausencia de género femenino, siendo en el caso de los hombres la variable de valor 0 y en el caso de las mujeres de valor 1. En econometría este tipo de variables son conocidas como variables categóricas o también llamadas "*dummy*" en inglés. Al introducir una variable categórica en un modelo, lo que ocurre es un desplazamiento del término constante en la estimación según una determinada cantidad y por cada una de las variables categóricas presentes. La cantidad desplazada por una variable categórica está determinada por el valor de su parámetro, que llamaremos para el caso de variables categóricas δ.

Por ejemplo, en un modelo de salarios (*wage*) cuyas variables explicativas son los años de estudio de un trabajador (*educ*) y su género (*female*) se visualizaría según la Figura 4.3:

Figura 4.3: Gráfico de modelo salarial con una variable categórica



Fuente: Wooldridge (2012)

Este ejemplo se interpretaría que las mujeres, de media, tienen un salario con una reducción de δ_0 frente al de un hombre, para el mismo tiempo estudiado.

El motivo por el que introducimos una única variable categórica de género en vez de dos, como podría ser *female* y *male*, es para no crear colinealidad perfecta, ya que la suma de las variables que modelan una misma propiedad cualitativa es siempre igual a 1. A esta situación se le conoce como la trampa de las variables categóricas. Por ello, debemos de establecer un grupo base con el que hacer comparaciones para una determinada propiedad. En el ejemplo anterior sería el género femenino. Es posible obtener modelos que permitan que distintas variables categóricas interactúen entre sí, pero, debido a que no ha sido la metodología aplicada, no lo trataremos en este trabajo.

Si quisiéramos probar, utilizando el modelo de ejemplo, que las mujeres sufren discriminación salarial deberíamos de estimar el modelo, realizar el contraste t en la variable explicativa δ_0 y comprobar si se rechaza la hipótesis nula. El modelo no nos permite establecer si el efecto producido por la variable cualitativa es causal o no, solo podemos garantizar que existe una diferencia en la variable aleatoria en el caso de poseer o no una determinada propiedad cualitativa.

Habiendo explicado como introducir variables cualitativas al modelo, es necesario entender cómo interpretarlas en el momento que trabajemos con distintas formas funcionales. Trabajando en nivel-nivel, como definimos en la tabla 4.2, podemos considerar el parámetro *dummie* como una desviación absoluta δ cuando su variable explicativa es igual a 1. Sin embargo, nuestro modelo presenta formas funcionales de log-log y log-nivel, como veremos más adelante. En el momento que trabajemos log-log debemos interpretar los parámetros como una desviación porcentual de la variable

dependiente entre la existencia de una propiedad o no. En estos casos, hablamos de una diferencia de $(100 * \delta)\%$ en la variable aleatoria entre observaciones. Si trabajamos con modelos log-nivel la interpretación es levemente más compleja y debemos realizar una transformación del parámetro para interpretarlo como porcentaje:

$$100 * \left[\exp\left(\widehat{\beta}_j\right) - 1\right] \tag{4.16}$$

Aplicando esta ecuación llegamos a poder realizar la misma interpretación que en el caso log-log. Hay que tener en cuenta que el signo del regresor es necesario para un resultado correcto.

Las interpretaciones anteriores son también parecidas para el caso de tratar con variables no categóricas. En el momento que existan relaciones log-log hablaremos de que duplicar la variable independiente o hacerla 0, siendo estos casos un incremento del 100% de su valor, provocará que la variable aleatoria aumente o disminuya respectivamente un $\beta\%$ su valor. En el caso de que existan relaciones log-nivel, la ecuación (4.16) expresa el porcentaje que se incrementa la variable dependiente si se incrementa una unidad la variable independiente.

En el siguiente capítulo aplicaremos lo comentado en apartado 4.1 estudiando la MCVL en profundidad y revisando que esta cumple las suposiciones del MCL. Después de esto, en el capítulo 6, utilizaremos lo revisado en el apartado 4.2 para hacer un diagnóstico de los resultados de la inferencia del modelo que proponemos e interpretar los resultados obtenidos.

5 La Muestra Continua de Vidas Laborales.

5.1 Origen de la muestra.

La Muestra Continua de Vidas Laborales (MCVL) con datos fiscales, tal y cómo nos explican en la guía del contenido publicada en el año 2021 autores de la misma (Seguridad Social), es un extracto de datos individuales anonimados, procedentes de las bases de datos de la Seguridad Social, a los que se añaden otros que se toman del Padrón Continuo Municipal y del resumen anual de retenciones e ingresos a cuenta del IRPF de la Agencia Estatal de Administración Tributaria (AEAT).

En nuestro caso, los datos corresponden a una muestra longitudinal de personas seleccionadas al azar que fueron afiliados a la Seguridad Social durante los años 2019 y 2020. De estas personas, se incluyen tanto datos sobre su relación con la Seguridad social en dicho año como datos históricos de la vida laboral, si están presentes en los registros informatizados. Esta información se pone a disposición de investigadores en forma de microdatos: datos en bruto (sin apenas modificaciones) y completamente desagregados; es decir, los datos se presentan en una forma muy semejante a los datos almacenados en las bases de datos de donde proceden.

La MCVL se estructura en seis tablas, correspondientes a distintas áreas de información: personas, afiliación (vida laboral), bases de cotización, pensiones, convivientes y datos fiscales (IRPF). Este trabajo se basa principalmente en la información contenida en la segunda y sexta tabla, la de las vidas laborales y la del módulo fiscal, aunque se ha complementado con ciertos datos del resto. En la muestra conjunta realizada (para trabajar los datos) consideramos conjuntamente las muestras de 2019 y 2020; por el carácter longitudinal de la muestra, la mayor parte de los individuos se encuentran en ambas muestras.

Los criterios para incluir a una persona en la MCVL son:

1. Que esta disponga de un documento identificador de persona física (IPF). Este IPF deberá estar después incluido en un conjunto de 4 millones de números seleccionados al azar entre los primeros 100 millones de números naturales. La selección de números se mantiene para todas las ediciones de la MCVL
2. El individuo debe haber formado parte de la población de referencia el año considerado. La población de referencia contempla a trabajadores afiliados durante al menos un día a la Seguridad Social o pensionistas

Al exigir estas dos condiciones simultánea e independientemente se garantiza que el 4% de todas las personas que forman parte de los afiliados estén seleccionados al azar. La muestra es siempre verificada para que sea representativa de la población por contrastes estadísticos.

Las personas saldrán o no de la población de referencia dependiendo de sus circunstancias personales como, por ejemplo, estar en paro, tener empleo… pero el estar entre los 4 millones de números seleccionados es una criterio que se cumplirá, en la mayoría de los casos, de forma permanente, ya que el número de Documento Nacional de Identidad (DNI) se asigna una vez en la vida. En el raro caso de que haya cambios del número de documento, como en el momento en el que un extranjero adquiera la nacionalidad, en el que se le asigna un DNI sustituyendo al NIE, este puede salirse o no de la MCVL. Dicho esto, los autores de la MCVL establecen una serie de conclusiones que se deben tener en cuenta al utilizar la muestra:

1. La MCVL se concibe como una muestra de personas con idéntica probabilidad de ser seleccionadas. Aunque se incluyan datos de otros elementos, como los empleadores, la MCVL no puede considerarse una muestra de otras unidades estadísticas distintas.
2. Cada edición de la MCVL es representativa sólo de la composición de la población de referencia a lo largo del año de referencia. La MCVL no permite extraer conclusiones sobre la situación de la población de referencia en años anteriores.
3. La persona que no cumpla la primera condición nunca entrará en la MCVL, salvo que cambie su número de identificador.
4. La permanencia en sucesivas ediciones de la MCVL de personas que no se han ausentado de la población de referencia durante más de un año, permite observar su evolución en periodos amplios de tiempo.

5.2 Descripción de la muestra.

La muestra conjunta realizada de la muestra (años 2019 y 2020) está conformada por 1.798.121 observaciones de episodios en los que se le ha retribuido una cantidad dineraria o en especie a una persona con un determinado identificador IPF. Un único IPF puede recibir durante un año múltiples retribuciones por distintos conceptos (salarios, prestaciones, pensiones, etc.). En nuestro Trabajo Fin de Grado, nos centramos en los episodios laborales y en la retribución obtenida en dichos episodios; esto es, prestamos especial interés a la retribución económica del IPF por trabajar para una determinada persona física o jurídica, definida en la muestra por un valor Identificador de la Persona Jurídica (IPJ), durante un determinado tiempo. Existen en la muestra conjunta 647.800 IPF distintos para el año 2019 y otros 670.461 para el año 2020.

Con intención de describir en profundidad la agrupación muestral utilizada y comprobar las distribuciones de las diferentes variables de la muestra, en los siguientes apartados estudiaremos los 22 atributos, o variables, de la muestra conjunta que se han utilizado para la inferencia del modelo de regresión múltiple. De este conjunto de variables, 15 son cualitativas y se modelarán como

variables categóricas dentro del modelo econométrico. El resto son variables cuantitativas, que como veremos en el siguiente apartado, a algunas de ellas se les han hecho transformaciones matemáticas para obtener un mayor ajuste en el modelo estadístico.

Para ilustrar las variables de la muestra conjunta, se ha realizado una estimación núcleo (*kernel* en inglés) de sus distribuciones de densidades de probabilidad. Los autores de la metodología, Rosenblatt (1956) y Parzen (1962), definen la ecuación de la función de densidad estimada para un conjunto de datos $(x_1, x_2, \dots, x_n)$ independientes e idénticamente distribuidos como:

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(u) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{5.1}$$

Donde *K* es el núcleo utilizado, una función no negativa integrable, *h* es un parámetro de suavizado llamado ancho de banda. En nuestro caso, el núcleo utilizado es el gaussiano, tal que:

$$K(u) = \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}} \tag{5.2}$$

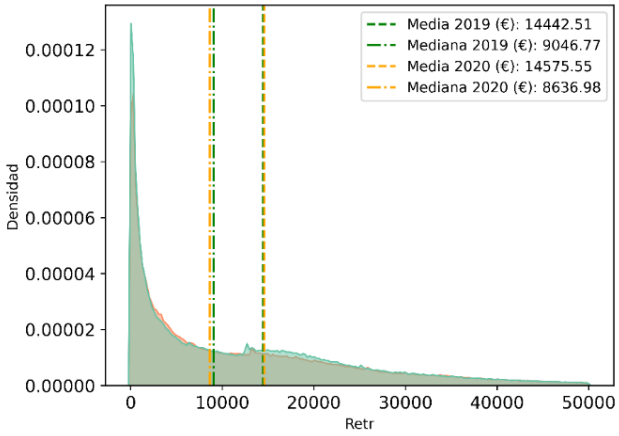
Para la estimación se suele utilizar un ancho de banda *h* que sea capaz de minimizar el error de estimación, pero para la ilustración de las figuras de este trabajo se ha seleccionado una serie de anchos de bandas para cada distribución. Con ello, se mantendrá una correcta visibilidad de las distribuciones de datos. Por ejemplo, en el caso de las variables categóricas, utilizaremos una estimación de densidad para cada propiedad, las acumularemos y normalizaremos a uno para presentar la probabilidad de observar las propiedades analizadas para cada tramo retributivo.

5.2.1 Retribuciones brutas percibidas por episodio.

La MCVL únicamente nos presenta los datos de retribuciones totales del trabajo de un IPF para un IPJ que hayan ocurrido durante un año. Esta forma de recopilar la información nos impide hacer un estudio individualizado para cada contrato de una persona con una determinada empresa. Por ello, el enfoque aplicado al estudio es el de analizar el total de las retribuciones que gana un IPF trabajando para un IPJ, que podrá haber ocurrido en uno o varios contratos. En el caso de que existan varios contratos, medidos como salidas y entradas en la afiliación, estudiaremos las propiedades cualitativas del contrato que más tiempo duró durante el año.

El valor que vamos a considerar como retribución por episodio es la suma de la parte dineraria y la valoración en euros de la parte en especie: *Retr*. Este valor es la retribución bruta obtenida antes de aplicar retenciones de cualquier tipo, por lo que hay que tener en cuenta que cada persona de la muestra, según su situación personal, disfrutará de una retribución neta distinta. Habiendo definido los episodios retributivos de la muestra conjunta, podemos ver su distribución de probabilidad en la Figura 5.1.

Figura 5.1: Distribución de densidad de salarios hasta 50.000€. (2019 y 2020)

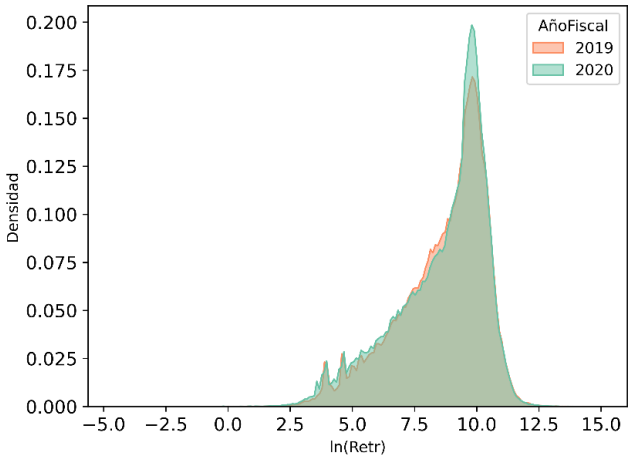


Fuente: Elaboración propia a partir de la MCVL

Los resultados obtenidos en la figura anterior están recortados hasta los 50.000 € para una mejor visualización de las diferencias entre 2019 y 2020, perdiéndose por ello tan solo 70.051 observaciones (el 3,89% del total de la muestra) que presentan valores muy elevados. La "desigualdad" de las observaciones es tal que los episodios con mayor retribución en 2019 y 2020 fueron, respectivamente, de 3.300.132,5 € y de 3.390.059 €.

Debido a las características de la distribución, se ha aplicado la transformación logarítmica natural a los datos de la muestra para acercarla todo lo posible a la normal y mejorar los resultados de la estimación e inferencia del modelo final. La distribución obtenida completa, que será la variable aleatoria de nuestro modelo, *ln(Retr)*, queda como se representa en la Figura 5.2.

Figura 5.2: Distribución de densidad de salarios logarítmicos por episodio. (2019 y 2020)

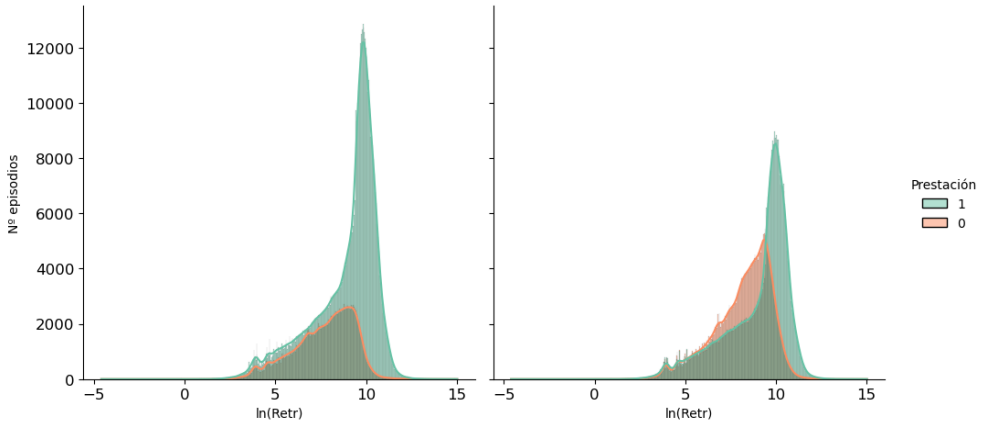


Fuente: Elaboración propia a partir de la MCVL

5.2.2 Retribuciones no salariales.

Son variables categóricas que indican que una persona ha tenido asociadas algún tipo de retribución no salarial. Estas pueden ser pensiones de viudedad, discapacidad o jubilación (*Pensión*), prestaciones como la de desempleo, becas o extraordinarias públicas (*Prestación*) o rentas de actividades económicas (*RAE*), como pueden ser ciertas actividades artísticas, intelectuales, agrarias o exentas de gravamen. En la muestra conjunta realizada para 2019 y 2020, 790.499 observaciones tienen una o varias retribuciones no salariales, aun siendo el receptor un trabajador afiliado a la Seguridad Social. De estas, presentan variables *Pensión, Prestación o RAE* positivas respectivamente, 49.846, 637.630 y 195.181 observaciones. Si visualizamos (Figura 5.3) los datos de prestación para los años 2019 y 2020 de forma desagregada vemos un gran aumento de prestaciones en el año 2020, probablemente determinado por la cantidad de ERTEs ejecutados con el inicio de la pandemia.

Figura 5.3: Histograma de frecuencias de retribuciones logarítmicas separados según prestación. (2019 y 2020)



Fuente: Elaboración propia a partir de la MCVL

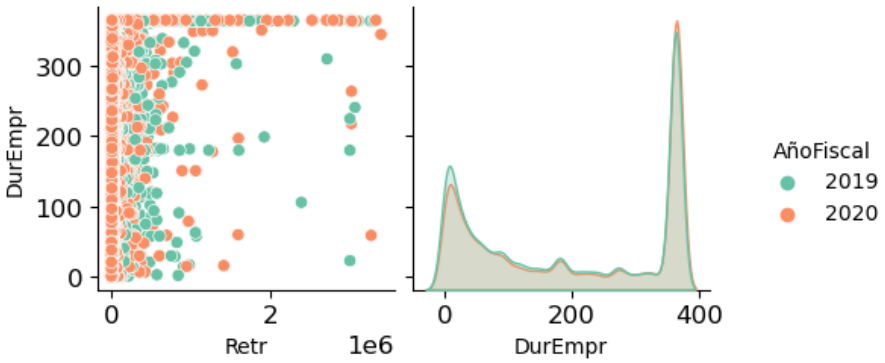
5.2.3 Variables de duración.

Este subgrupo está conformado por tres variables de duración: la duración en días en relación laboral con el IPJ dentro del año, en uno o varios contratos (*DurEmpr*), la edad del individuo en el año fiscal de la observación (*Edad*) y el tiempo en años que hace que está en contacto con el pagador, esto es, el tiempo transcurrido desde que el individuo recibió un primer pago del pagador hasta el año muestral considerado (*1erContacto)*.

En el caso de duraciones en la empresa donde el IPF ha estado a trabajando a tiempo parcial, se ha utilizado la variable de coeficiente a tiempo parcial que adjunta la MCVL para normalizar dicha duración a una jornada de 8 horas estándar dentro de *DurEmpr*. En la Figura 5.4 podemos ver como las observaciones que han durado un año entero han sido las que han llegado a tener mejores

retribuciones respecto a las demás. Además, las observaciones de un año entero son las más comunes de encontrar, aunque los episodios de alrededor de un mes han sido los segundos más probables.

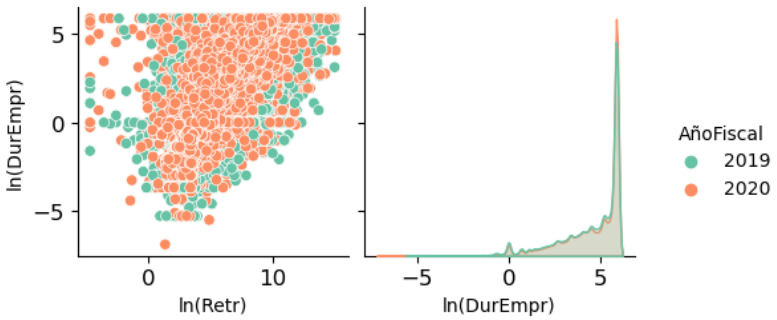
Figura 5.4: Duración de la observación y salario



Fuente: Elaboración propia a partir de la MCVL.

Como hemos comentado en otros casos, en la muestra se aprecian dispersiones y distribuciones no normalizadas que no nos permiten realizar inferencia con un buen ajuste. Por ello, se ha realizado la transformación logarítmica, obteniendo resultados más normalizados (Figura 5.5):

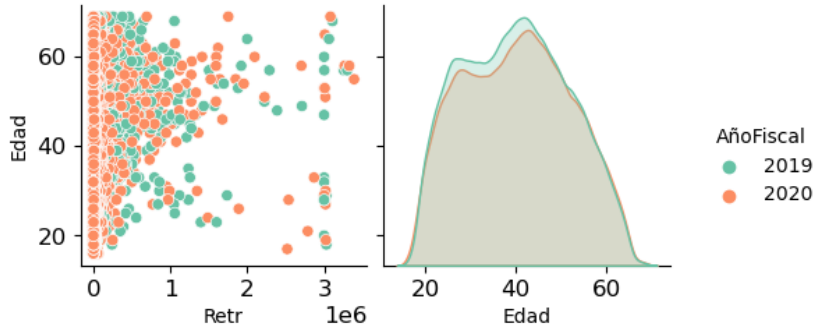
Figura 5.5: Duración logarítmica de la observación y salario bruto logarítmico.



Fuente: Elaboración propia a partir de la MCVL.

La distribución de las edades de los IPF observada (Figura 5.6) es similar a la de las edades de la población española trabajadora, donde, de media, la edad de cada observación de la muestra es elevada, siendo de 41,42 años en 2019 y de 41,29 en 2020. También hay que comentar que la muestra está relativamente distribuida según una normal. Aun así, si lo comparamos con las retribuciones obtenidas, vemos claramente que, salvo alguna excepción, al envejecer un IPF los ingresos aumentan y se hacen más dispersos. Esta forma de distribución se comentó al tratar la teoría del capital humano.

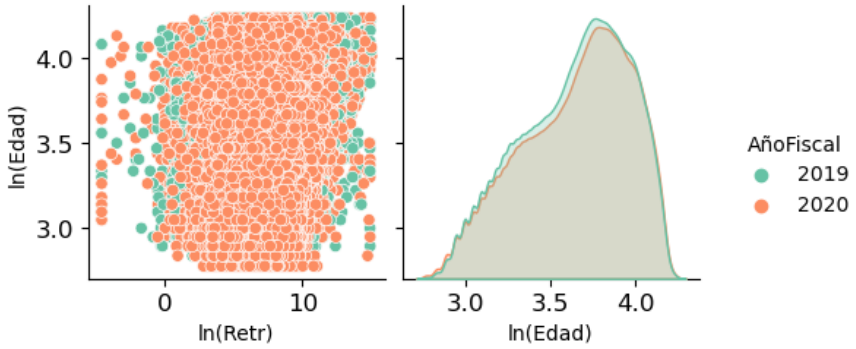
Figura 5.6: Edad del IPF de la observación y salario bruto.



Fuente: Elaboración propia a partir de la MCVL.

Igual que con la variable *DurEmpr,* se ha aplicado una transformación logarítmica a los datos (Figura 5.7). No se ha aplicado la transformación cuadrática de Mincer ya que el ajuste logarítmico daba mejores resultados de estimación e inferencia.

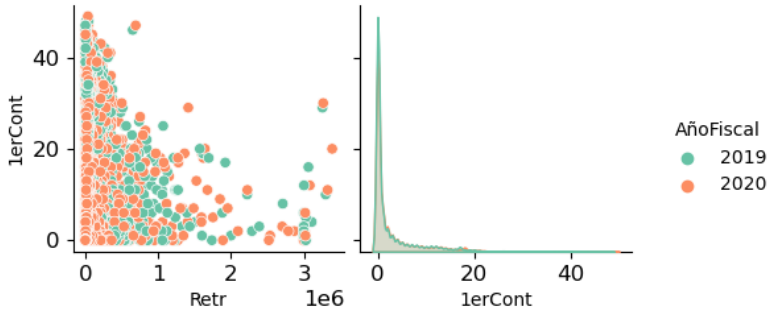
Figura 5.7: Edad logarítmica del IPF de la observación y salario bruto logarítmico.



Fuente: Elaboración propia a partir de la MCVL.

En el caso de la variable *1erContacto* no hay que confundir este dato con el tiempo trabajado en años para un IPJ, ya que *1erContacto* refleja la posibilidad de que una persona haya terminado su relación productiva con un IPJ una o más veces, para después haber vuelto a tener relación laboral en el año fiscal de estudio. En la Figura 5.8, si comparamos las retribuciones y el tiempo que se conoce al IPF, vemos que estar más tiempo conociendo a una empresa no supone mejorar la retribución con seguridad. Las retribuciones más altas se dieron entre los IPF que entraban por primera vez a trabajar para un IPJ. En cuanto a la distribución de probabilidad, vemos que lo más común es trabajar para IPJ nuevas.

Figura 5.8: Años desde el primer contacto con el IPJ y salario bruto.



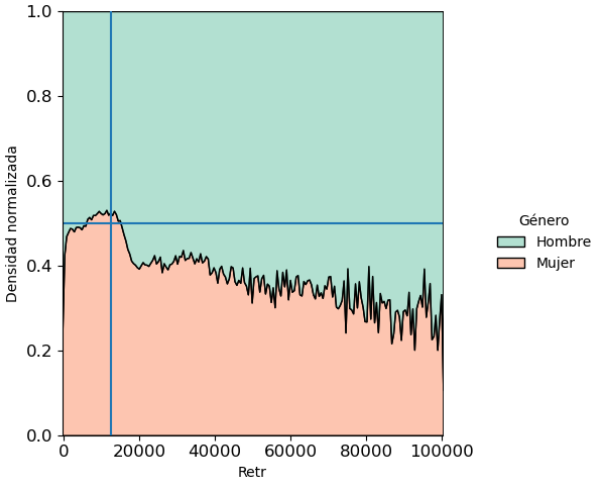
Fuente: Elaboración propia a partir de la MCVL.

Para ajustar la distribución de *1erContacto* a la normal se ha considerado añadirla al modelo de forma cuadrática. Las implicaciones de haberla añadido de esta manera se discutirán en el siguiente capítulo.

5.2.4 Género.

La variable categórica que controla la propiedad de que una observación sea de una mujer o un hombre se ha denominado *Género*. En el año 2019 se registran 417.507 observaciones para mujeres y 500.641 para hombres. En el año 2020 ocurren 401.073 episodios de mujeres y 478.903 de hombres. Si queremos comprobar la probabilidad de encontrar un determinado género para una determinada retribución, podemos utilizar el año 2019 como base al ser este más estable al no haber ocurrido la pandemia. En la Figura 5.9 vemos como, en la muestra, habrá solo más posibilidad de encontrar mayoría femenina de observaciones alrededor del salario mínimo interprofesional de 2019, 12.600 €.

Figura 5.9: Retribuciones según género. (2019)

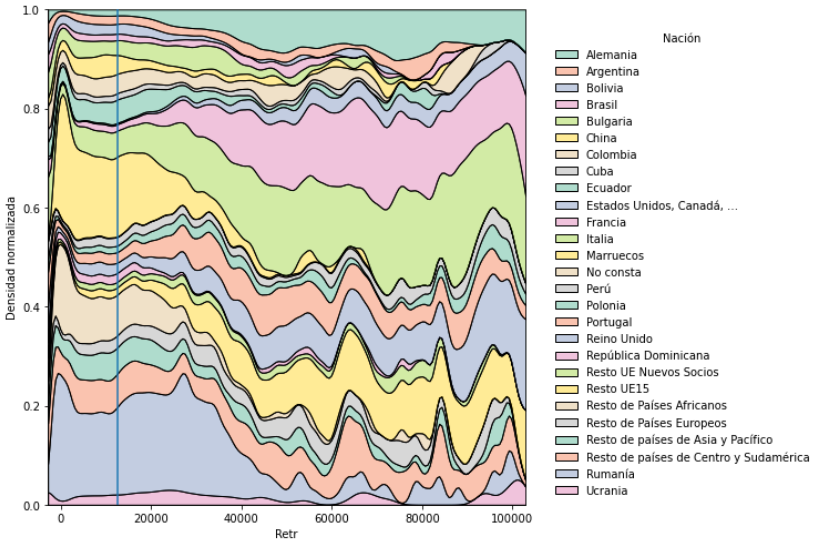


Fuente: Elaboración propia a partir de la MCVL.

5.2.5 Nacionalidad y provincia de afiliación.

Propiedades geográficas de la muestra son la nacionalidad del IPF, *Nación*, y la provincia donde está afiliado ese determinado IPF, *Prov.* En el caso de las retribuciones según nacionalidad, se es capaz de ver que para todos los tramos retributivos la probabilidad de encontrar españoles es muy alta. En el caso de tramos retributivos menores a los 30.000€ brutos, los IPF de nacionalidad española son aproximadamente el 80% de los episodios y para mayores rentas suponen el 90% de probabilidad. Si visualizamos (Figura 5.10) las retribuciones desagregadas por nacionalidad distinta a la española en 2019, podemos ver que extranjeros son más probables de encontrarse en cada uno de los distintos tramos estudiados. Esta distribución es similar también a la de 2020. Las nacionalidades distintas a la española más comunes de observar son los IPF procedentes de Rumanía, seguidos por Marruecos y otros países africanos.

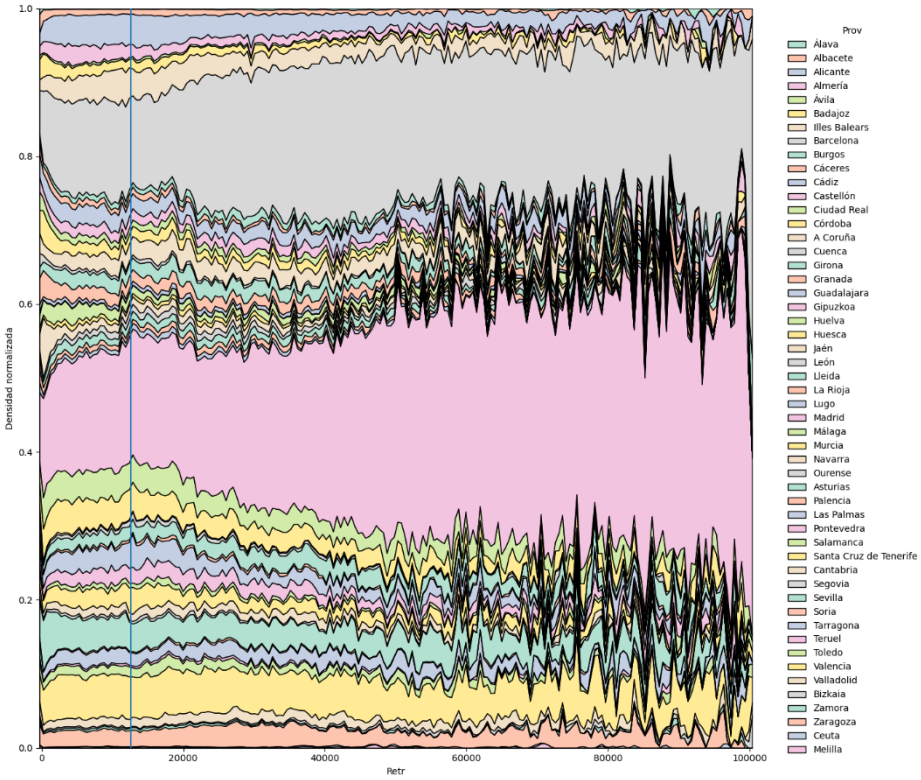
Figura 5.10: Retribuciones según nacionalidad distinta a la española. (2019)



Fuente: Elaboración propia a partir de la MCVL.

En el caso de la provincia de afiliación del IPF, las proporciones de los datos son similares para los años 2019 y 2020. En la Figura 5.11 comprobamos como las capitales del tejido productivo español, Madrid (157.102 observaciones) y Barcelona (124.961), son las dos provincias con mayor número de observaciones, seguidas por Valencia (50.478), Sevilla (42.000) y Murcia (33.299).

Figura 5.11: Retribuciones según provincia de afiliación. (2019)



Fuente: Elaboración propia a partir de la MCVL.

Es interesante ver como a partir de los 40.000 € anuales es significativamente más probable ver retribuciones más altas en Madrid que en Barcelona. Si contamos la provincia de afiliación de las retribuciones superiores a 100.000€, en 2019, encontramos 2.284 observaciones (1,45% de las totales) para el caso de Madrid y 1.116 (0,89%) para el caso de Barcelona, muy lejos de la tercera opción de Valencia de 162 (0,32%) o Sevilla con 96 (0,22%). Esto podría significar que hay una mayor intención por parte de las altas retribuciones por tributar en Madrid, aunque no entraremos en mayor profundidad sobre la causalidad.

5.2.6 Nivel de estudios.

La variable *Estudios* contiene los últimos estudios reglados que ha terminado el IPF de la observación. Los resultados de 2019 no son muy distintos de la tendencia que ya se comentó en la revisión teórica del capital humano, siendo los niveles superiores de estudios mejor retribuidos de media. Los resultados de 2019 y 2020 son también similares. Los niveles de estudios observados más frecuentes en 2019 son: Primaria y bachiller elemental; FP básica (206.505), Bachiller (101.926) y Licenciado o Grado Universitario (101.046).

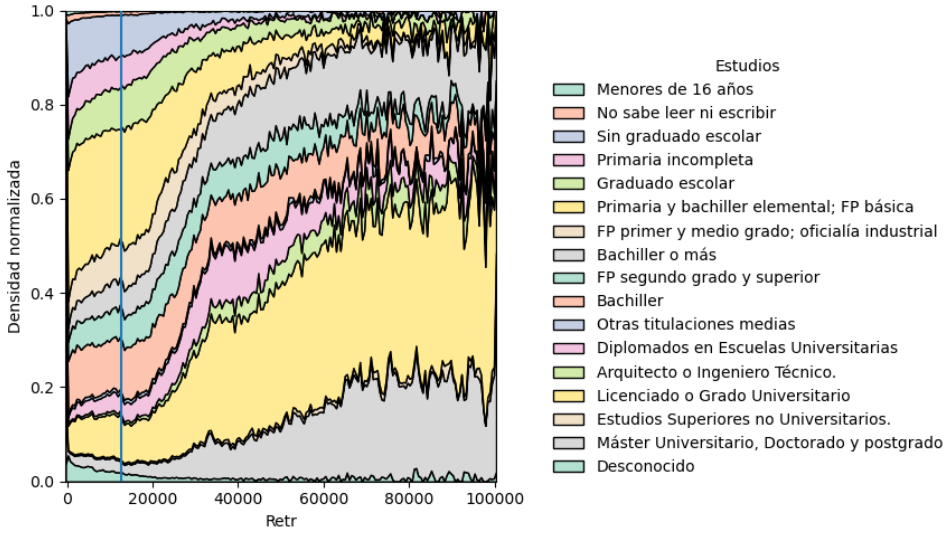
Teniendo en cuenta esto, es interesante ver que en la Figura 5.12 existe un desplazamiento de mayores densidades hacia la derecha, aumentando el salario en el momento de ir creciendo los niveles

educativos. Aunque, también es importante ver que siguen existiendo probabilidades significativas de encontrar estudios, como el de bachiller, compitiendo con unas retribuciones similares a los de másteres universitarios.

Si contamos las observaciones según estudios para retribuciones mayores de los 100.000€ en 2019 encontramos que la observación más común, con diferencia, es la de Licenciado o Grado Universitario (2.193), lejos de los siguientes estudios: Máster (654) y Bachiller o más (587).

Los salarios menores que el salario mínimo de 2019 fueron observados en mayor medida para Primaria y bachiller elemental; FP básica (139.347), Sin graduado escolar (61.795) y Bachiller (59.722).

Figura 5.12: Retribuciones según nivel de estudios. (2019)



Fuente: Elaboración propia a partir de la MCVL.

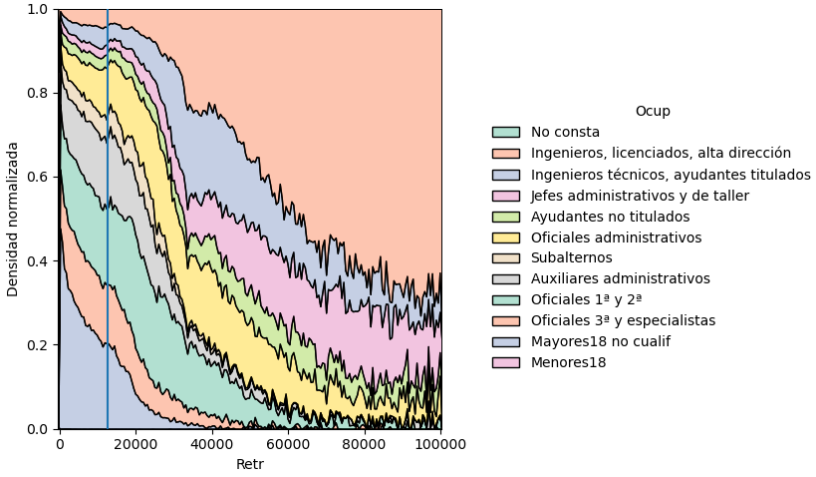
5.2.7 Tipo de ocupación y sector productivo.

Las variables cualitativas que definen la categoría de la ocupación de la observación y el sector productivo donde se está realizando son, respectivamente, *Ocup* y *SAct*. Teniendo en cuenta que la estructura laboral de 2020 es similar a la de 2019, en la representación de las retribuciones por ocupación de 2019 (Figura 5.13) y al contabilizar los datos vemos cuatro rangos de retribuciones aproximadas que producen diferencias cualitativas en la distribución de la muestra:

1. El rango [0-12.600] está dominado principalmente por personal no cualificado, con 181.942 observaciones (34,38% del total de observaciones para ese rango), seguido por Oficiales 1ª y 2ª (16,98%). El resto de las ocupaciones tiene mucha menor presencia que la dominante.

2. En el rango [12.600-33.500] la ocupación más común fue la de Oficiales 1ª y 2ª con 65.541 observaciones (22,25%), seguido de cerca por los oficiales administrativos, con 45.828 observaciones (15,56%). Los episodios ocurren de forma relativamente proporcionada.
3. En el rango [33.500-72.000] hay un crecimiento constante de la ocupación de Ingenieros, licenciados, alta dirección, hasta llegar a las 25.029 observaciones (31,41%). También hay un aumento menor de ingenieros técnicos, ayudantes titulados, hasta llegar a las 14.702 observaciones (18,45%), mientras que se reduce la presencia de oficiales de 1ª, 2ª y oficiales de 3ª y especialistas.
4. El rango [>72.000] está desproporcionadamente dominado por ingenieros, licenciados, alta dirección, teniendo 8.524 observaciones (65,95%) seguidos por jefes administrativos y de taller con 1.886 (14,59%).

Figura 5.13: Retribuciones según tipo de ocupación. (2019)



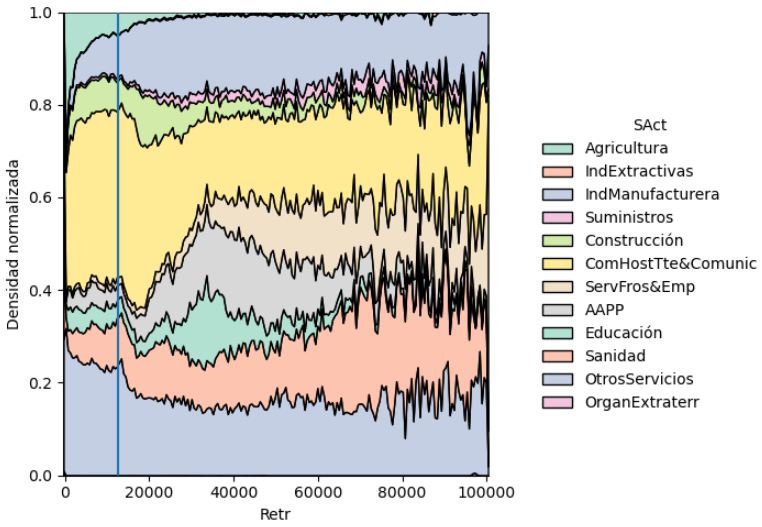
Fuente: Elaboración propia a partir de la MCVL.

La actividad sectorial de la observación, *SAct*, también se puede dividir en 4 sectores según los rangos de retribución establecidos anteriormente (Figura 5.14). Para el análisis, representamos las actividades económicas según agrupaciones de la Clasificación Nacional de Actividades Económicas (CNAE 09):

1. Entre [0-12600] dominan ComHostTte&Comunic (comercio al por mayor, hostelería, transporte y comunicaciones) con 175.933 observaciones (33,14%) y OtrosServicios con 140.944 (26,55%), teniendo la siguiente actividad, agricultura, unas 75.362 observaciones (14,19%).
2. Entre [12.600-33.500] la agricultura se reduce y cambian las distribuciones del resto de actividades. La más común, ComHostTte&Comunic, se reduce a 93.481 observaciones (31,73 %). OtrosServicios, la segunda, también se reduce a 51.980 (17,64%) y aparece como tercera más frecuente IndManufacturera con 42.262 (14,34%).

3. Entre [33.500-72000] las probabilidades se hacen más proporcionales y las diferencias entre actividades serán menores. ComHostTte&Comunic se mantiene como la más común con 14.719 episodios (18,47%), IndManufacturera la segunda con 12.816 (16,08%) y OtrosServicios tercera con 12.105 (15,19%).
4. En el rango de [>72.000] pierden significancia educación y AAPP, volviendo a cambiarse las distribuciones retributivas. ComHostTte&Comunic adquiere más importancia con 3.275 observaciones (25,34%), OtrosServicios sigue con 2.603 (20,14%) y Sanidad entra tercera con 2.021 (16,64%).

Figura 5.14: Retribuciones según sector. (2019)

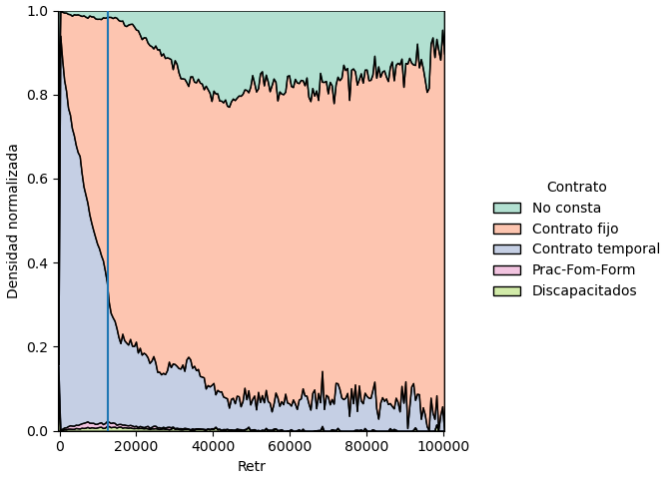


Fuente: Elaboración propia a partir de la MCVL.

5.2.8 Tipo de contrato.

En la MCVL existe información sobre el tipo de contrato de un episodio, pero, aunque es significativa, también es problemática de analizar. Esta información se obtiene de los datos asignados a un Código de Cuenta de Cotización (CCC) que contiene a colectivos de trabajadores afiliados de un IPJ, clasificados según diversas propiedades de contrato. Esta variable, que nosotros hemos añadido en el modelo como *Contrato*, no tiene por qué contener datos siempre, estando algunos marcados como "No consta". Esto se debe a que los empleadores no tienen obligación de suministrarla en muchos casos como, por ejemplo, empleados del hogar, funcionarios o autónomos entre otros. Por ello, gran parte de la muestra queda sin información, ya que "No consta" aparece 38.210 veces (4,16%) en 2019 y de forma similar en 2020. En la Figura 5.15 vemos la distribución de tipos de contratos registrados.

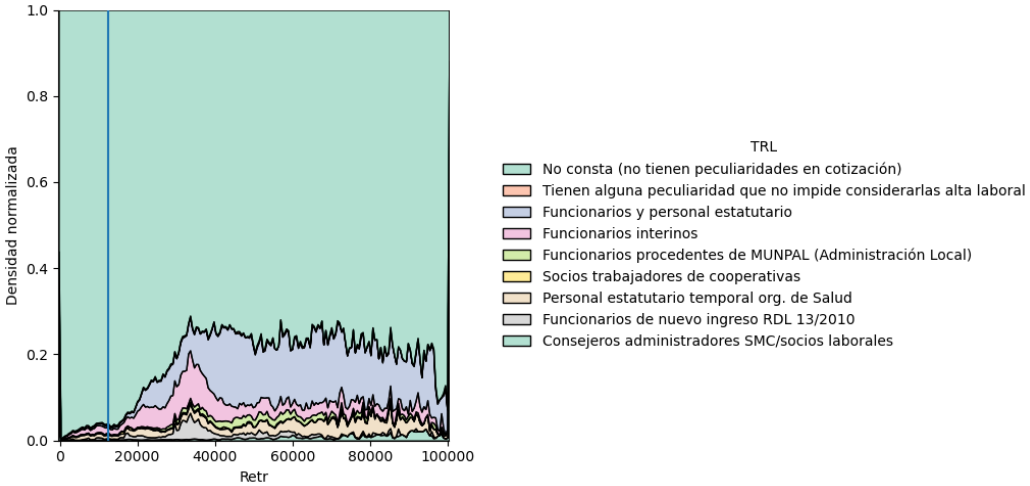
Figura 5.15: Retribuciones según tipo de contrato. (2019)



Fuente: Elaboración propia a partir de la MCVL.

Para paliar esta peculiaridad de la variable *Contrato*, se añade la variable *TRL* para complementar las propiedades que se quedan sin observar. Esta última variable cualitativa contiene datos sobre el tipo de relación laboral del IPF con el IPJ, solo si esta es especial, y se suele registrar en los casos en los que *Contrato* haya sido "No consta". Que no conste *Contrato* en una observación no obliga a que *TRL* contenga valor y viceversa, pero sí es bastante común que, al no aparecer la primera, aparezca la segunda cumplimentada. En 2019, por ejemplo, en los *Contrato* que no constan, solo 326 de 38.210 observaciones quedaron sin rellenar en *TRL* o fueron marcados como que no había peculiaridades en cotización. La distribución de *TRL* (Figura 5.16) se puede comprobar similar a la de la propiedad "No consta" de *Contrato*.

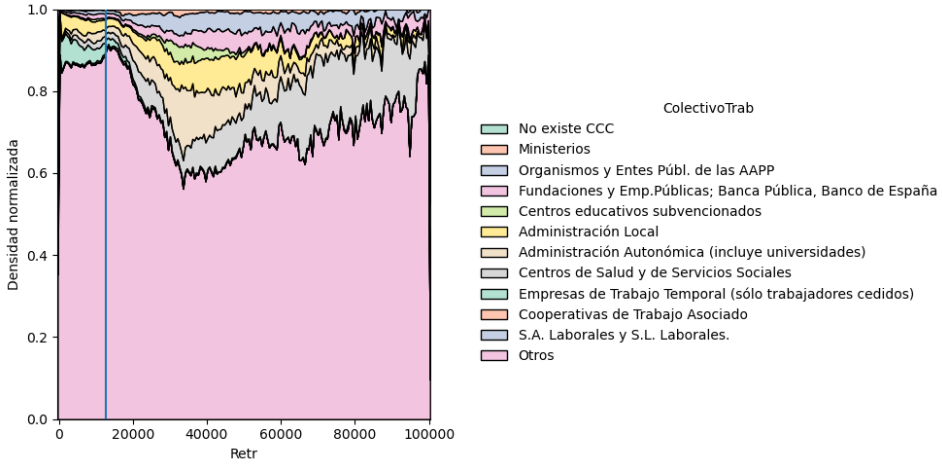
Figura 5.16: Retribuciones según relación laboral especial. (2019)



Fuente: Elaboración propia a partir de la MCVL.

Otra variable que se introduce en el modelo para sacar conclusiones sobre *Contrato* y *TRL* es la variable *ColectivoTrab*. Esta contiene el colectivo de trabajadores para el que se creó el CCC al que está asignado el IPF de la observación. De forma similar a *TRL*, si *Contrato* es "No consta", *ColectivoTrab* solo deja sin desglosar 1.433 observaciones de 38.210. En la Figura 5.17 podemos ver los distintos colectivos presentes en la variable.

Figura 5.17: Retribuciones según colectivo asignado al CCC. (2019)



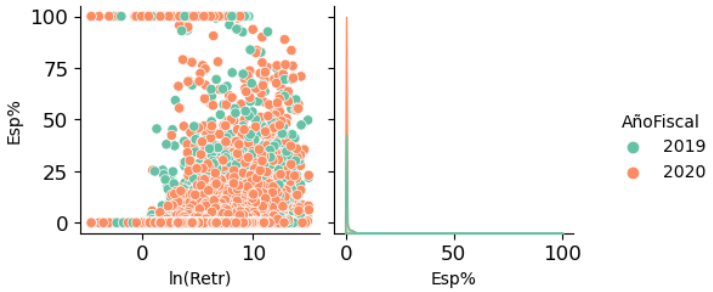
Fuente: Elaboración propia a partir de la MCVL.

5.2.9 Otras propiedades cuantitativas del contrato.

Este subgrupo incluye variables cuantitativas menos relevantes que se han incluido al modelo para comprobar cuánto afectan en las retribuciones de un episodio. *Esp%* mide el porcentaje de la retribución que ha sido en especie, *NºContr* contiene el número de contrataciones que se han hecho en ese determinado IPJ y *Tam* es el valor del número de empleados que están recibiendo retribuciones desde un CCC asociado a un IPJ, que se introduce en el modelo con una transformación logarítmica.

Si visualizamos lo datos del porcentaje de retribución en especie (Figura 5.18) podemos ver que la gran mayoría de los episodios se sitúan cercanos a un 0% de pago en especie, teniendo solamente el 14,64% de los episodios totales de la muestra conjunta valores distintos de 0.

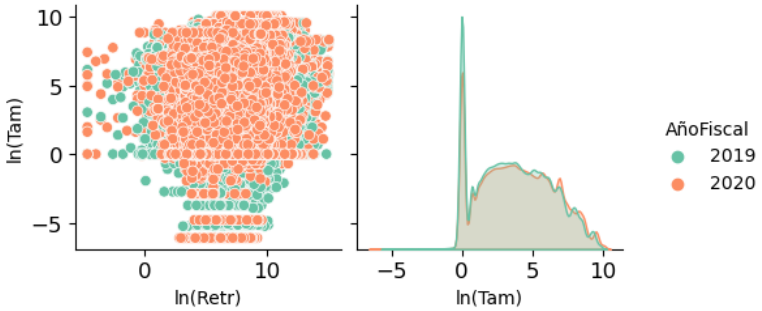
Figura 5.18: Porcentaje en especie observación y salario bruto logarítmico.



Fuente: Elaboración propia a partir de la MCVL.

Si hacemos lo mismo para el número de empleados asociado a las CCC de un IPJ (Figura 5.19) nos encontramos una situación parecida a la anterior variable: el tamaño logarítmico de 0 es lo más común. Aun así, en el momento de que nos separamos de CCC de una única persona, existe una distribución relativamente normal. El 15,07% de los episodios están asociados a un IPJ que solo está retribuyendo a una persona.

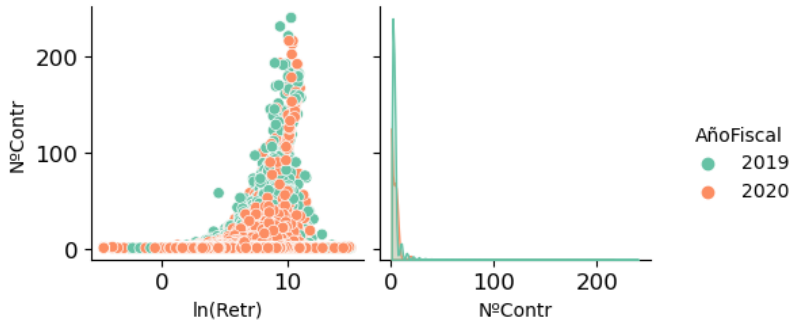
Figura 5.19: Número de empleados logarítmico y salario bruto logarítmico.



Fuente: Elaboración propia a partir de la MCVL.

Similar a los casos anteriores, en el momento que visualizamos el número de contratos de una persona en un año para un IPJ (Figura 5.20) encontramos distribuciones parecidas. De todos los episodios de la muestra utilizada, solo el 17,51% presenta IPF con más de un contrato asociado a un IPJ.

Figura 5.20: Número de contratos en la observación y salario bruto logarítmico.



Fuente: Elaboración propia a partir de la MCVL.

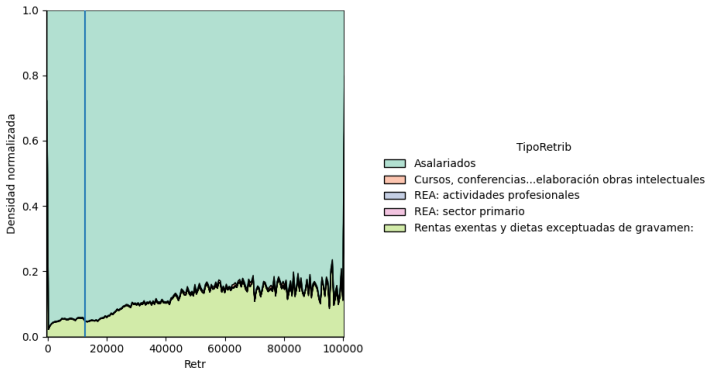
5.2.10 Otras propiedades cualitativas del contrato.

En este subgrupo encontramos *TipoRetrib*, variable cualitativa que se ha querido añadir al modelo para cuantificar las diferencias entre asalariados y distintos tipos de rentas registradas en la MCVL. También *TipoEmpleador*, con el objetivo de determinar cómo afecta que una persona este

recibiendo retribución por parte de una persona física, extranjera o persona jurídica. Las distribuciones obtenidas de 2019 son similares a las de 2020.

En la Figura 5.21 podemos ver como el tipo de retribución más común de encontrar en la muestra conjunta es el de asalariados. Solo el 5,99% de los episodios de la muestra han sido rentas distintas a un salario.

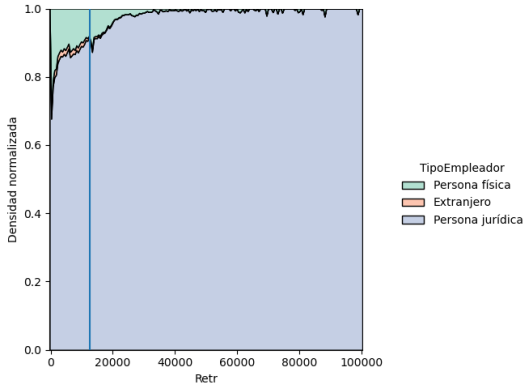
Figura 5.21: Retribuciones según asalariados, trabajo intelectual o rentas. (2019)



Fuente: Elaboración propia a partir de la MCVL.

Al comprobar la información de los tipos de empleadores del episodio (Figura 5.22) vemos que una parte importante de los episodios de menor retribución al salario mínimo son realizados para personas físicas, aunque la gran mayoría presente sean las personas jurídicas. El 12,20% de los episodios son realizados para IPJ distintos a una persona jurídica.

Figura 5.22: Retribuciones según tipo de empleador. (2019)



Fuente: Elaboración propia a partir de la MCVL.

6 Estimación de la Ecuación de Salarios de la Economía Española. Años 2019 y 2020.

6.1 Modelo utilizado y diagnóstico.

Después de analizar distintas opciones para formular la ecuación de salarios, el modelo de regresión múltiple finalmente seleccionado que se ha aplicado para estudiar las retribuciones de los afiliados de las observaciones de la muestra conjunta (procedente de la MCVL) es el siguiente:

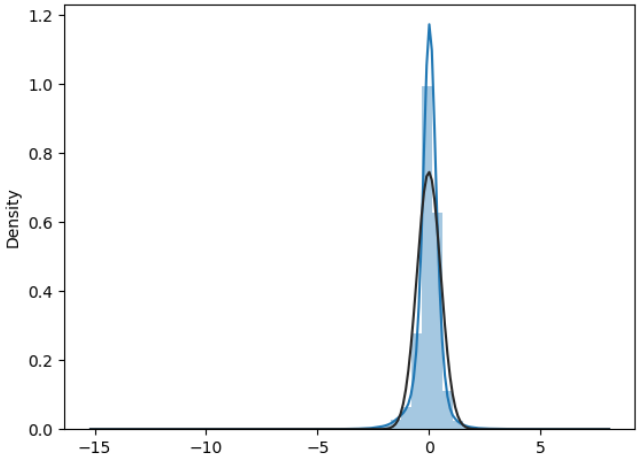
$$\begin{aligned}\ln(Retr) = {} & \beta_0 + \beta_1\, AñoFiscal + \beta_2\, ln(DurEmpr) + \beta_3\, Pensión \\ & + \beta_4\, Prestación + \beta_5\, RAE + \beta_6\, TipoRetrib + \beta_7\, Género \\ & + \beta_8\, ln(Edad) + \beta_9\, Nación + \beta_{10}\, Esp\% + \beta_{11}\, NºContr \\ & + \beta_{12}\, 1erCont + \beta_{13}\, 1erCont^2 + \beta_{14}\, TipoEmpleador \\ & + \beta_{15} \ln(Tam) + \beta_{16}\, Ocup + \beta_{17}\, SAct + \beta_{18}\, TRL \\ & + \beta_{19}\, Prov + \beta_{20}\, Contrato + \beta_{21}\, Estudios \\ & + \beta_{22}\, ColectivoTrab + u\end{aligned} \tag{6.1}$$

Siendo cada una de las variables:

a. Retribución obtenida por el trabajo de una persona durante una relación laboral con un pagador ($Retr$)
b. El año fiscal de la retribución ($AñoFiscal$)
c. Duración en días de la relación laboral con el pagador durante el año ($DurEmpr$)
d. Si se recibe pensión, prestación o RAE durante el trabajo ($Pensión, Prestación, RAE$)
e. Tipo de retribución: asalariado u otro tipo de rentas ($TipoRetrib$)
f. Género, edad y nacionalidad de la persona ($Género, Edad\ y\ Nación$)
g. Porcentaje que representa la retribución en especie ($Esp\%$)
h. Número de contratos realizados dentro de la empresa en el año considerado ($NºContr$)
i. Tiempo en años desde el primer contacto con el pagador ($1erCon$)
j. Persona jurídica, física o extranjera del pagador ($TipoEmpleador$)
k. Tamaño en personas de la cuenta de cotización del pagador (Tam)
l. Categoría de la ocupación realizada ($Ocup$)
m. Sector económico donde se realiza el trabajo ($SAct$)
n. Tipo de relación laboral (TRL)
o. Provincia de afiliación ($Prov$)
p. Tipo de contrato ($Contrato$)
q. Nivel de estudios ($Estudios$)
r. Colectivo de trabajadores especial ($ColectivoTrab$)

Tras la regresión del modelo utilizando MCO, los residuos de la ecuación (6.1) han sido revisados para comprobar que mantengan una distribución normal. La distribución obtenida y una distribución normal N(0,1) ajustada de comparación pueden visualizarse en la Figura 6.1.

Figura 6.1: Distribución de los residuos comparada con la distribución normal estándar.



Fuente: Elaboración propia a partir de los datos.

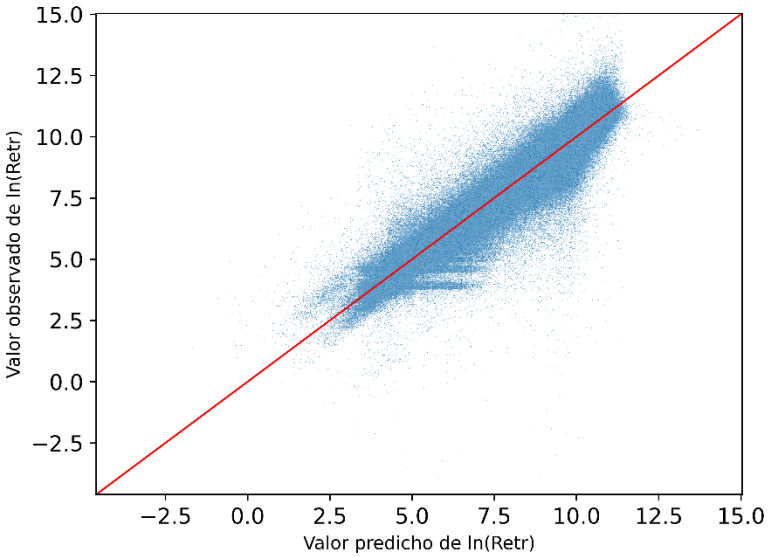
Podemos ver que los residuos siguen una distribución parecida a la normal, pero es necesario que probemos que los residuos siguen el supuesto 6 de normalidad que definimos en la metodología. Para ello, utilizaremos la prueba de Jarque-Bera, que es la más popular para hacer contraste de hipótesis sobre la normalidad muestral. La fórmula para la prueba (Gujarati, 2015) se define como:

$$JB = n\left[\frac{S^2}{6} + \frac{(K-3)^2}{24}\right] \sim \chi_2^2 \qquad (6.2)$$

donde *n* es el tamaño de la muestra, *S* el coeficiente de asimetría (tercer momento respecto de la media) y *K* el coeficiente de curtosis (cuarto momento respecto de la media). Para que el residuo fuese perfectamente normal debería de darse que *S*=0 y *K*=3, que consideraremos que es la hipótesis nula del contraste. Como está demostrado que este valor sigue la distribución χ^2 con dos grados de libertad, podemos calcular el p-valor del estadístico JB. El p-valor obtenido a través del programa estadístico utilizado es menor de 0,01 para la hipótesis alternativa distinto a la normal, por lo que somos capaces aceptar la hipótesis nula de normalidad de los residuos.

El R^2 obtenido de la regresión ha sido de 0,911 (91,1%) y el p-valor del estadístico F es menor de 0,01, indicándonos que nuestra estimación del modelo tiene un buen ajuste a la muestra conjunta de la MCVL utilizada y que se puede considerar estadísticamente valido para estudiar las retribuciones en nuestro país en los años de referencia.

Figura 6.2: Residuos asociados a cada observación y predicción de la muestra.



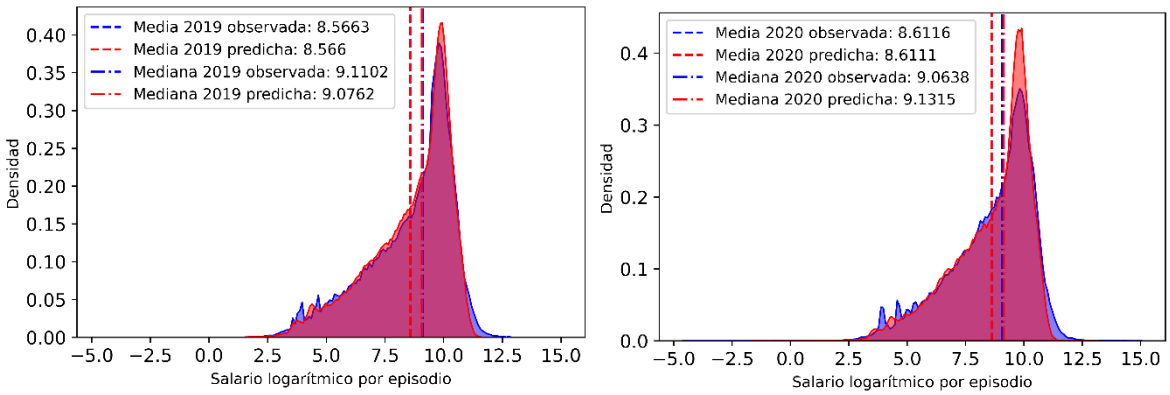
Fuente: Elaboración propia a partir de los datos.

Debido a la cantidad de variables categóricas utilizadas y a la multitud de propiedades que miden, el resultado final del cálculo de MCO nos devuelve 158 parámetros estimados, incluyendo la constante, acompañados con su debida información sobre el error estándar del coeficiente, estadístico *t-student*, p-valor y sus intervalos de confianza. De estos 158 parámetros, 11 han sido no significativos (en las variables de nacionalidades y provincias), 4 tienen un p-valor en el rango [0,05-0,01], 5 están el rango [0,01-0,001] y 138 han tenido un p-valor menor que 0,001, estos últimos serían significativamente distintos de 0 con un nivel de confianza del 99,9%. En el siguiente apartado interpretaremos los coeficientes estimados para entender cómo se comportaron los sueldos en España en los años 2019 y 2020.

6.2 Resultados obtenidos.

La distribución de la retribución logarítmica por episodio predicha para los años 2019 y 2020 utilizando la información de las variables independientes contenida en la muestra, como hemos dicho, ha sido buena, pero no perfecta. Si analizamos las predicciones frente a las retribuciones observadas para los años 2019 y 2020 (Figura 6.3) podemos ver como la media se mantiene significativamente igual en los dos casos, pero la mediana predicha varia frente a la observada. Existen pequeñas diferencias en las colas izquierda y derecha de ambas comparaciones, junto a los valores de mayor densidad de la distribución.

Figura 6.3: Salarios logarítmicos observados y predichos por el modelo (2019 y 2020).



Fuente: Elaboración propia a partir de los datos.

Esto es prueba de que la metodología MCO utilizada ha sido realizada correctamente. A continuación, analizaremos el efecto sobre las retribuciones medias de las variables explicativas de nuestro modelo. Las variables no observadas, información contenida en los residuos, explicarían las diferencias entre los valores observados y predichos de la Figura 6.3 si se añadiesen al modelo tras observase como variables independientes.

6.2.1 Interpretación de los parámetros significativos.

1. El resultado obtenido para *AñoFiscal* 2020 es un valor de -0,0709 para su parámetro, con un p-valor menor de 0,001. Utilizando la ecuación (4.16) comprobamos que las retribuciones brutas en 2020 fueron un 6,84% menores en media que las de 2019. Este resultado se debe, muy probablemente, a la crisis económica de la COVID-19, que azotó la economía del país y del tejido económico global. Por ejemplo, a igualdad del resto de factores, esta diferencia entre los dos años fiscales supuso una caída de media del 61,6 € al mes para personas que ganasen el salario mínimo mensual de 2019 (900 €) o una reducción de 136,89 € para otras que ganaran 2.000 € mensuales, de acuerdo con nuestro modelo.

2. En el caso de las retribuciones no salariales, en los dos años de referencia modelados, aquellas personas que tuviesen asociado a sus episodios laborales algún tipo de pensión, prestación o renta les supuso una reducción de sus retribuciones del 8,95%, 7,67% y 4,24% respectivamente.

3. Las variables de duración utilizadas han sido altamente significativas, pudiendo comprobarse que los contratos a jornada completa que duran un año son retribuidos un 81,81% más de media que un contrato de medio año a jornada completa o a media jornada durante un año. En general, al comparar entre dos contratos, por cada aumento del 1% de la duración de uno frente al otro, supondrá un aumento del 0,86% de la retribución esperada.

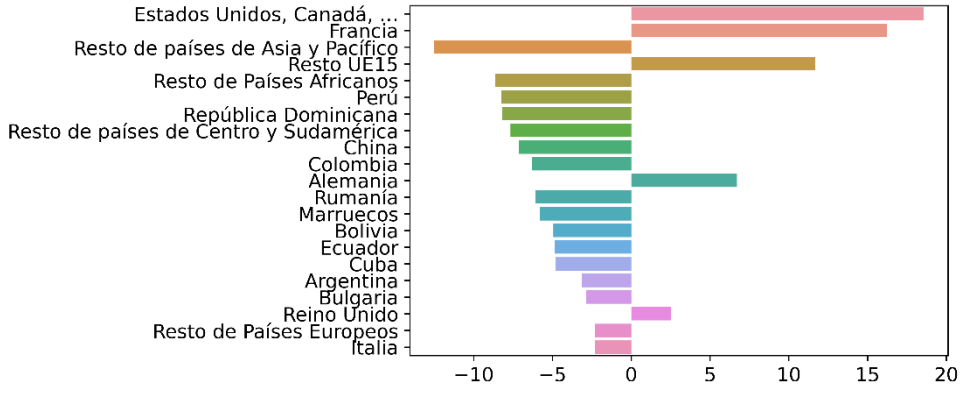
Una edad mayor del trabajador también es determinante en un aumento del salario, pero en menor manera. Para un mismo puesto, una persona de 65 años cobrará un 25,90% más que un joven de 18 años. Una persona de 40 años, al mismo tiempo, ganará un 15,41% que otra de 18. De forma general, un 1% más de edad supone un incremento de renta del 0,1794%.

La variable *1erContacto* entra en el modelo de forma cuadrática. Los coeficientes obtenidos muestran una relación positiva pero decreciente entre esta duración (tiempo que hace que conocemos al pagador) y la retribución en logaritmo. A partir de los 28 años de haber tenido contacto con una empresa la tendencia de la retribución será de decrecimiento. Nótese que no necesariamente el trabajador ha tenido que estar los 28 años trabajando en la empresa; aunque si ha tenido relaciones contractuales con la empresa en esos años. Desde el primer contacto con una empresa hasta los 28 años existe un aumento del 33,28% de las retribuciones, pero el aumento por año se va ralentizando hasta llegar al máximo. A partir de entonces, si pasasen 10 años, existiría una reducción de las retribuciones del 3,97%.

4. Para los años 2019 y 2020, la retribución media de las mujeres fue un 11,66% menor que la de los hombres, para una misma ocupación. Con este dato, el modelo nos está avisando de la existencia de una brecha salarial significativa entre hombres y mujeres en el intervalo temporal objeto de estudio.

5. Para las nacionalidades de los trabajadores de España, podemos observar en la Figura 6.4 las diferencias medias de salario de los extranjeros en comparación a los españoles. Los trabajadores estadounidenses y canadienses son los mejor retribuidos, 18,56% más que los españoles, y los trabajadores asiáticos o del pacífico son los que peor se retribuyen, 12,5% menos.

Figura 6.4: Diferencias en porcentaje de retribución según nacionalidad (Base España).

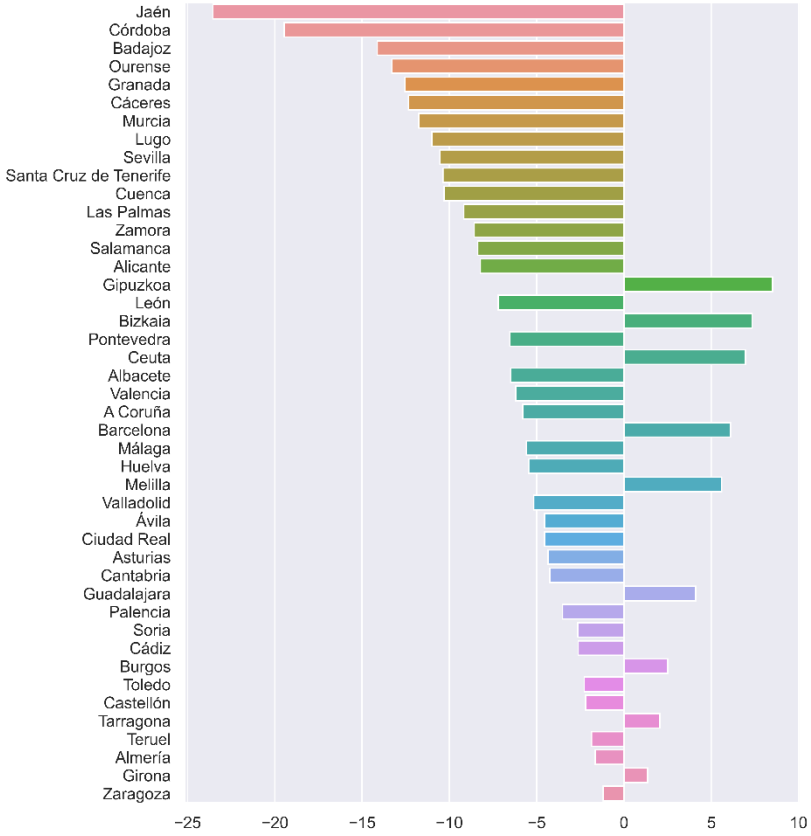


Fuente: Elaboración propia a partir de los datos.

Para el caso de estudiar los resultados según la afiliación a provincias del país, podemos hacer la misma gráfica de torbellino que en el caso anterior (Figura 6.5) utilizando como referencia los

asalariados de Madrid. De las 44 provincias significativas para estudiar distintas de Madrid, vemos que los trabajadores madrileños solo son superados en retribución de media por 9 provincias, mientras el resto reciben retribuciones menores. Los trabajadores jiennenses son los peor remunerados, con un 26,86% de diferencia respecto a Madrid y los trabajadores guipuzcoanos los mejor retribuidos, con un 8,52% bruto más. De las 10 provincias peor remuneradas de media, cuatro son andaluzas y Sevilla es la novena en el ranking de provincias peor retribuidas.

Figura 6.5: Diferencias en porcentaje de retribución según provincia de afiliación (Base Madrid).

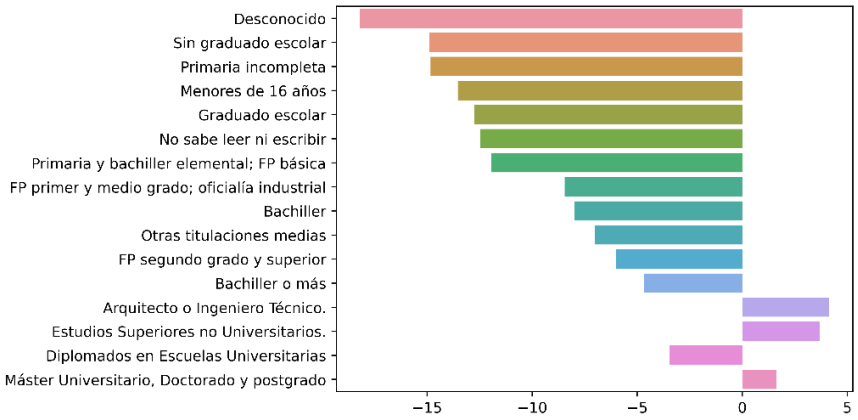


Fuente: Elaboración propia a partir de los datos.

6. Las diferencias de salarios según el nivel de estudios, utilizando de base lo estudios de licenciatura o grado universitario, no son muy distintos a lo esperado tras el análisis de la teoría del capital humano (Figura 6.6). Podemos establecer que, de media, las mejores retribuciones fueron obtenidas por los titulados en arquitectura o ingenierías técnicas, sin poder diferenciar entre los dos, obteniendo un 4,15% más que un graduado o licenciado. Entre graduados y licenciados también se han considerado los ingenieros superiores. Las peores retribuciones son para personas que no tienen ninguna información registrada en la Seguridad social sobre sus estudios, con un 18,19% menos de retribución frente a un graduado o licenciado. Los siguientes peor retribuidos son las personas sin graduado escolar o primaria incompleta, con un 14,90% menos de salario que un graduado

universitario o licenciado. Si analizamos las observaciones de la muestra conjunta para las personas sin información sobre sus estudios podemos obtener que el 76,02% de ellos tienen una nacionalidad distinta a la española, por lo que hay que tener en cuenta esta relación para entender el porcentaje obtenido de diferencias.

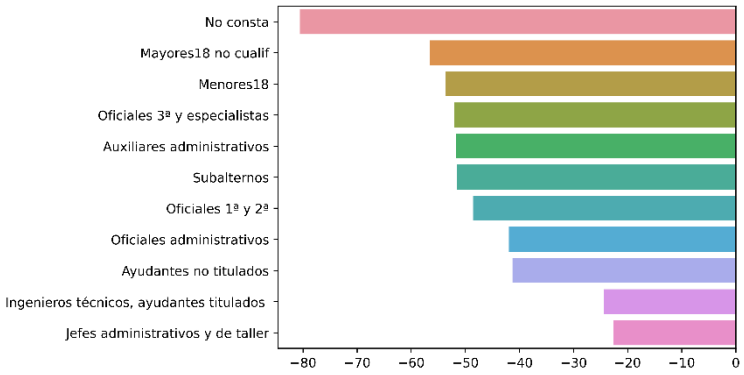
Figura 6.6: Diferencias en porcentaje de retribución según estudios (Base graduados universitarios o licenciados).



Fuente: Elaboración propia a partir de los datos.

7. Si seguimos por las categorías de ocupación. *Ocup*, nos encontramos (Figura 6.7) una de las mayores diferencias de retribución medias de las variables del modelo: No consta. Que no haya información rellenada por el empleador sobre la categoría de la actividad realizada por el IPF supone una reducción frente a la categoría base de un 80,57%. También vemos que la categoría base de ingenieros, licenciados y alta dirección es la mejor retribuida de todas las otras opciones posibles, habiendo grandes diferencias porcentuales entre los grupos de estudio.

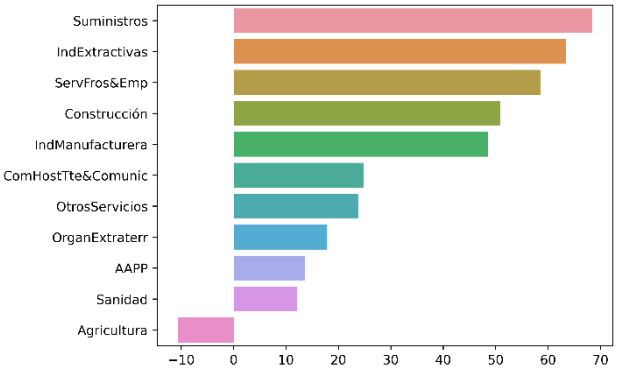
Figura 6.7: Diferencias en porcentaje de retribución según categoría de ocupación (Base ingenieros, licenciados y alta dirección).



Fuente: Elaboración propia a partir de los datos.

El sector económico español mejor retribuido fue el de suministros de energía eléctrica, de agua, gestión de residuos…, con una diferencia del 68,42% frente al educativo (Figura 6.8). El sector educativo, base del modelo, solo tuvo una actividad estudiada con peor retribución, la agricultura con un 10,67% menos de retribución media.

Figura 6.8: Diferencias en porcentaje de retribución según sector (Base educación).



Fuente: Elaboración propia a partir de los datos.

8. Los resultados según tipos de contratos descritos nos avisa de una posible discriminación entre las distintas opciones. Al usar contrato fijo como base podemos concluir que los contratos de discapacidad, solo por esta característica y no por otro contexto laboral, ganan de media un 11,97% menos. En el caso de los contratos temporales también vemos una reducción del 6,94% que no debería ocurrir ya que estamos analizando para mismas duraciones en la empresa. Los contratos de practica o formativos son los peor retribuidos, con un 22,40% menos que uno fijo.

En el caso de relaciones laborales especiales, principalmente representado por trabajadores del sector público, resultados interesantes obtenidos es que los funcionarios son el grupo mejor retribuido, siendo, por ejemplo, los funcionarios interinos retribuidos un 21,71% menos de media en comparación. El personal estatuario temporal de salud también recibe un 24,56% menos retribución que un funcionario.

De los colectivos especiales estudiados ($ColectivoTrab$), también es interesante comentar que, si utilizamos de base el colectivo que incluye universidades españolas, los trabajadores de empresas de trabajo temporal ganan un 11,74% menos, pero, por otra parte, los centros educativos subvencionados ganan un 15% más que el personal de administración autonómica y universidades.

9. Analizamos ahora las tres variables cuantitativas *Esp%*, *NºContr* y *Tam* definidas en el anterior capítulo. Se ha obtenido que un incremento de un 1% en *Esp%* y *Tam*, supone, respectivamente, un incremento retributivo del 0,006% y del 0,03%. Mientras que el primer resultado se puede considerar despreciable por la poca influencia que supone un incremento en la retribución, el pasar de, por ejemplo, una cuenta de cotización de 20 trabajadores a otra de 200 supone un aumento

de la retribución media del 7,89%. De *NºContr* obtenemos que tener cinco contratos con un determinado IPJ supuso un 7,32% mayor retribución que solo tener uno. Dicho de otra manera, el número de contratos con un determinado IPJ en uno supuso un aumento retributivo del 1,79% de media. Esto puede ser causado por aumentos de salario al firmar mejores contratos dentro de la misma empresa.

10. Si el tipo de retribución fueron rentas del sector primario u otras activades profesionales, estas llegaron a ser de media, respectivamente, un 169,39% y 115,99% más que un salario en 2019 y 2020. Por otra parte, haber recibido retribución por una persona jurídica o IPJ extranjeros, supuso haber recibido de media un 13,72% o 5,70% más que si el pagador hubiese sido una persona física.

6.2.2 Predicciones de retribución media por tipo de perfil.

El modelo obtenido nos permite calcular la retribución media esperada de un trabajador, en los años 2019 y 2020, dadas sus variables. De esta manera, somos capaces de sacar conclusiones sobre los salarios en los años estudiados para perfiles específicos de trabajadores que nos interesen, además de poder comparar el efecto de los cambios de varias covariables de un perfil a otro. A continuación, se describen un conjunto de posibles trabajadores que podrían haberse afiliado a la Seguridad Social en los años analizados, mostrando sus retribuciones medias esperadas. Además, a cada valor obtenido se le acompaña con unos ingresos netos esperados aproximados, según el porcentaje de retenciones de IRP y Seguridad Social de cada año, obtenidos utilizando la calculadora online de retenciones que publica el ministerio de hacienda y función pública cada año:

a) Joven sin experiencia de 22 años recién graduado en ingeniera, que encuentra trabajo a jornada completa con contrato en prácticas en una mediana empresa de Sevilla del sector automovilístico (365 días de 2019): 30.681.53 € brutos (1.783,48 € netos mensuales en 14 pagas).
b) Ingeniero industrial con estudios especializados madrileño de 50 años, que lleva 20 años trabajando para una gran empresa del sector de suministro eléctrico (365 días de 2020): 80.659.83 € brutos (3.958.67 € netos mensuales en 14 pagas).
c) Temporera de 30 años marroquí sin estudios, que va a trabajar durante un mes en la campaña de recolección del limón de Murcia (2020): 706 € netos.
d) Obrero de la construcción de Barcelona con 35 años y estudios de formación profesional básica, que lleva 5 años trabajando para una empresa con contrato fijo (365 días de 2020): 19.425.38 € brutos (1.424,52 € mensuales netos en 12 pagas).
e) Enfermera de atención primaria con 45 años de Soria, que trabaja para la sanidad pública como personal estatuario temporal (365 días de 2020): 29.315,38 € brutos (1.715,16 € netos mensuales en 14 pagas).

f) Profesora interina doctorada con 33 años, que trabaja a media jornada en la universidad de Salamanca (365 días de 2019): 18.502,58 € brutos (1.183,37 € netos mensuales en 14 pagas).

Esta información viene a confirmar que nuestro modelo estimado puede verse como un sistema de información que orienta al buscador de empleo sobre la retribución esperada de acuerdo con sus características, lo cual le puede otorgar cierta ventaja en la negociación salarial.

7 Conclusiones.

En este Trabajo de Fin de Grado hemos identificado y cuantificado distintos determinantes de los salarios de los trabajadores de nuestro país a través de un modelo econométrico aplicado a una muestra de microdatos de empleo para los años 2019 y 2020. Para poder disponer de información relevante sobre la remuneración del trabajo en España y sus diferentes regiones, hemos recurrido a la Muestra Continua de Vidas Laborales (MCVL) y a una de las herramientas econométricas más utilizadas en este campo, la regresión lineal múltiple con variables ficticias. Con esta base empírica y metodológica, pretendemos arrojar luz sobre un proceso, el de fijación de precios en el mercado de trabajo, que no siempre produce resultados deseables desde un punto de vista social.

Para dar sustrato teórico al ejercicio empírico propuesto, hemos desarrollado las teorías del capital humano y de la señalización (teorías que no son excluyentes, sino complementarias); estos modelos teóricos pretenden explicar la relación entre la educación del individuo y los beneficios privados y sociales de dicha inversión en educación. Asimismo, hemos revisado literatura económica que trata fundamentalmente de explicar la distribución salarial observada en la economía.

El siguiente paso de nuestra investigación ha sido revisar la metodología necesaria para realizar inferencia estadística a través de la estimación de un modelo econométrico muestral. Dada la naturaleza categórica de algunos regresores, hemos empleado algunas variables ficticias (*dummies*) en la estimación. La estimación por Mínimos Cuadrados Ordinarios del modelo nos permite, al menos dos tipos de análisis: (1) la interpretación de los coeficientes estimados, algunos de los cuales apuntan hacia la existencia de discriminación o de segregación en el mercado de trabajo (por motivos de género o nacionalidad, por ejemplo); y (2) hemos sido capaces de obtener las retribuciones medias esperadas para los años 2019 y 2020 para distintos contextos laborales que vienen definidos por las variables explicativas del modelo de salarios propuesto; el modelo permite por tanto predecir el salario de un trabajador "virtual" (no observado en la muestra) creado por nosotros.

En lo que se refiere a los datos empleados, hemos considerado de forma conjunta los datos de la MCVL de 2019 y 2020 dentro del modelo, introduciendo para ello una variable ficticia de año. Nótese que la MCVL es una muestra longitudinal, por lo que la mayoría de los trabajadores que aparecen en la MCVL de 2019 también son observados en la MCVL de 2020. Los datos han sido preparados para disponer de la retribución bruta total obtenida por el trabajo de una persona para un pagador determinado durante el año completo; esta variable (expresada en logaritmos) será la variable dependiente del modelo econométrico. A partir de esta muestra conjunta, se ha descrito cada una de las variables independientes del modelo, ilustrando su relación con la variable dependiente.

Los resultados de nuestro modelo, definidos por sus parámetros estimados, han sido significativos para 147 variables de 158, siendo las que no algunas provincias de afiliación o nacionalidades. Gracias al elevado número de las observaciones de la muestra y al modelo propuesto, se ha conseguido llegar a un R^2 de 0,911 (91,1%). Las interpretaciones obtenidas se consideran, por tanto, representativas de las retribuciones medias españolas de los años 2019 y 2020. De estas, las variables cualitativas que más diferencias provocan son el sector económico donde se trabaje y la categoría de la ocupación realizada. Dicho de otra forma, el trabajador que quiera maximizar a corto plazo su salario esperado debería orientarse hacia puestos vacantes de grupos de ocupación elevados y situados en sectores productivos bien retribuidos como el de suministros energéticos o industrias extractivas. El trabajador también tiene la opción de aumentar sus estudios o de trasladarse a provincias con mejor salario, como sería Ceuta o Barcelona, con la inversión económica y de tiempo que esto supone. Aun así, mientras las diferencias en el caso de traslado de provincias pueden suponer, como máximo, un aumento de alrededor del 30% de su retribución, una mayor demanda laboral, representada por una mayor cantidad de vacantes en los sectores y categorías comentados, puede suponerle, aproximadamente en el mejor de los casos, un aumento del 80%.

Lo comentado hasta ahora supone aumentos posibles medios de retribución respecto a las variables que el trabajador es capaz de cambiar o invertir en mejorar. Por otro lado, variables del individuo como su nacionalidad, edad, género o discapacidad, según los resultados, pueden suponer en las diferencias de su retribución tanto como podrían ser sus estudios o los años trabajados para un pagador. En el caso de las mujeres es especialmente preocupante, ya que salvo que exista alguna variable no observada que influya en que las mujeres estén asignadas a ocupaciones peor retribuidas de forma generalizada por toda España, significaría una diferencia de retribución del 11,66% por el único motivo del género.

Los resultados obtenidos en nuestro estudio nos confirman que tanto atributos del trabajador, como el nivel de estudios o la experiencia laboral, como atributos del puesto, como la ubicación geográfica del centro de trabajo o el sector de actividad de la empresa, son determinantes significativos del salario de un trabajador español. Obsérvese que disponemos de información de los dos lados del mercado. Gracias a ello, hemos sido capaces de obtener una cuantificación de las variables implícitas de la oferta y de la demanda laboral.

Desde el punto de vista de la política económica, nuestro modelo tiene al menos dos lecturas. Por un lado, muestra que en el mercado de trabajo existe un importante grado de segmentación laboral, de manera que ciertos sectores de actividad y ciertas regiones retribuyen mejor a sus trabajadores, posiblemente por las condiciones de oferta y demanda en dichos segmentos de mercado. Esto puede ayudar al estudiante universitario a orientar mejor su formación académica. Digamos que nuestra estimación contiene en sí misma un sistema de información para la formación y la búsqueda de empleo. Por otro lado, nuestros resultados apuntan a diferencias laborales que pueden tener que

ver menos con las condiciones del mercado y más con la existencia de segregación, o incluso discriminación, en el mercado laboral español, ya sea por razón de género o de nacionalidad.

En próximas líneas de investigación se ve necesario aumentar el número de años de la muestra para poder comprobar si los resultados han sido solo significativos para los años 2019 y 2020 o para periodos mayores de tiempo. También debe completarse los datos de las comunidades autónomas del País Vasco y Navarra, debido a sus regímenes fiscales especiales, utilizando el fichero de cotizaciones de la MCVL para los datos retributivos. Por otra parte, es necesario indagar en profundidad para descartar o probar discriminación por género, nacionalidad u otras variables del trabajador, comprobando la relación causal de las diferencias de rentas observadas.

8 Referencia bibliográfica.

Abowd, J. M., Kramarz, F., & Margolis, D. N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, 67(2), 251-333. https://doi.org/10.1111/1468-0262.00020

Becker, G. S (1964) *Human capital: a theoretical analysis with special reference to education.* New York: Columbia University Press para NBER.

Becker, G. S. & Chiswick, B. R. (1966) Education and the distribution of earnings. En *American Economic Review, Proceedings*, 56:358-369.

Blundell, R., Dearden, L., & Sianesi, B. (2005). Evaluating the effect of education on earnings: Models, methods and results from the National Child Development Survey. *Journal of the Royal Statistical Society Series* A, 168(3), 473–512.

Blundell, R., & Costa Dias, M. (2009). Alternative approaches to evaluation in empirical microeconomics. *Journal of Human Resources*, 44(3), 565–640.

Cahuc, P., Carcillo, S., & Zylberberg, A. (2014). *Labor economics* (Second Edition). MIT Press.

Canal-Domínguez, J. F., & Rodríguez-Gutiérrez, C. (2008). Analysis of wage differences between native and immigrant workers in Spain. *Spanish Economic Review*, 10(2), 109-134. https://doi.org/10.1007/s10108-007-9033-3

Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (vol. 3A, chap. 30). Amsterdam: Elsevier Science.

Fernández-Kranz, D., & Rodríguez-Planas, N. (2011). The part-time pay penalty in a segmented labor market. *Labour Economics*, 18(5), 591-606. https://doi.org/10.1016/j.labeco.2011.01.001

Felgueroso, F., Hidalgo-Pérez, M., & Jiménez-Martín, S. (2016). The Puzzling Fall of the Wage Skill Premium in Spain: The Fall of Wage Skill Premium. *The Manchester School, 84*(3), 390-435. https://doi.org/10.1111/manc.12116

Gregory-Smith, I., Main, B. G. M., & O'Reilly, C. A. (2014). Appointments, Pay and Performance in UK Boardrooms by Gender. *The Economic Journal*, 124(574), F109-F128. https://doi.org/10.1111/ecoj.12102

Gujarati, D. N. (2015). *Econometrics by example* (2. ed). Palgrave Macmillan.

Heckman, J., Lochner, L., & Todd, P. (2003). *Fifty Years of Mincer Earnings Regressions* (N.º w9732; p. w9732). National Bureau of Economic Research. https://doi.org/10.3386/w9732

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2008). Earnings Functions and Rates of Return. *Journal of Human Capital*, 2(1), 1-31. https://doi.org/10.1086/587037

Izquierdo, M., Lacuesta, A., & Vegas, R. (2009). Assimilation of immigrants in Spain: A longitudinal analysis. *Labour Economics*, 16(6), 669-678. https://doi.org/10.1016/j.labeco.2009.08.011

Lassibille, G. (1998). Wage gaps Between the public and private sectors in Spain. *Economics of Education Review,* 17(1), 83-92. https://doi.org/10.1016/S0272-7757(97)00012-5

Lemieux, T. (2006). The ''Mincer equation'' Thirty Years After Schooling, Experience, and Earnings. En S. I. Grossbard (Ed.), *Jacob Mincer: A pioneer of modern labor economics* (pp. 127-145). Springer US.

Mincer, J. (1958) Investment in human capital and personal income distribution. En *Journal of Political Economy*:281-302.

Mincer, J. (1962) On-the-job training: costs, returns and some implications. En *Journal of Political Economy*, 70(5):50-79.

Mincer, J. & Polachek, S. (1974) Family investments in human capital: earnings of women. En *Journal of Political Economy* (Supplement), 82:S76-S108.

Molina, J. A., & Montuenga, V. M. (2009). The Motherhood Wage Penalty in Spain. *Journal of Family and Economic Issues*, 30(3), 237-251. https://doi.org/10.1007/s10834-009-9153-z

Neal, D., & Rosen, S. (2000). Chapter 7 Theories of the distribution of earnings. En *Handbook of Income Distribution* (Vol. 1, pp. 379-427). Elsevier. https://doi.org/10.1016/S1574-0056(00)80010-X

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), 1065-1076. https://doi.org/10.1214/aoms/1177704472

Pissarides, C.A. (2000). *Equilibrium Unemployment Theory*. Cambridge (Mass.): The MIT Press.

Poole, J. P. (2013). Knowledge Transfers from Multinational to Domestic Firms: Evidence from Worker Mobility. *Review of Economics and Statistics*, 95(2), 393-406. https://doi.org/10.1162/REST_a_00258

Rosen. S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy*, 82(1), 34-55.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27*(3), 832-837. https://doi.org/10.1214/aoms/1177728190

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, 87, 355–374.

Weiss, A. (1983). A sorting-cum-learning model of education. *Journal of Political Economy*, 91, 420–442.

Weiss, Y. (1986). The determination of life-cycle earnings: A survey. In O. Ashenfelter & R. Layard (Eds.), *Handbook of labor economics* (vol. 1, chap. 11, pp. 603–640). Amsterdam: Elsevier Science.

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach*. (5. ed). South-Western Cengage Learning.