Trabajo Fin de Máster

Estimación en áreas pequeñas bajo modelos de coeficientes aleatorios

Javier Fraga García

Máster en Técnicas Estadísticas

Curso 2020-2021

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación en áreas pequenas baixo modelos de coeficientes aleatorios
Título en español: Estimación en áreas pequeñas bajo modelos de coeficientes aleatorios
English title: Small area estimation under random coefficient models
Modalidad: Modalidad A
Autor/a: Javier Fraga García, Universidad de A Coruña
Director/a: María José Lombardía Cortiña, Universidad de A Coruña; María Esther Lopéz-Vizcaíno, Instituto Galego de Estadística
Breve resumen del trabajo: El objetivo principal de este trabajo es aplicar modelos mixtos de coeficientes aleatorios a la estimación en áreas pequeñas. Se hará un recorrido por la literatura estudiando posibles aplicaciones de interés en nuestro entorno. En particular, se estudiará el salario medio por ocupación laboral de acuerdo a la Encuesta de Estructura Salarial del INE.
Recomendaciones: Conocimiento y manejo de software estadístico R
Otras observaciones:

Doña María José Lombardía Cortiña, profesora de la Universidad de A Coruña, doña María Esther Lopéz-Vizcaíno, responsable del Servicio de Difusión e información del Instituto Galego de Estadística, informan que el Trabajo Fin de Máster titulado

Estimación en áreas pequeñas bajo modelos de coeficientes aleatorios

fue realizado bajo su dirección por don Javier Fraga García para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Coruña, a 10 de Diciembre de 2021.

La directora:

LOMBARDIA CORTIÑA MARIA JOSE - 33322690Z
Firmado digitalmente por LOMBARDIA CORTIÑA MARIA JOSE - 33322690Z
Fecha: 2021.12.14 15:17:18 +01'00'

Doña María José Lombardía Cortiña

La directora:

MARIA ESTHER LOPEZ VIZCAINO - 76575190W
Firmado digitalmente por MARIA ESTHER LOPEZ VIZCAINO - 76575190W
Fecha: 2021.12.14 15:21:12 +01'00'

Doña María Esther Lopéz-Vizcaíno

El autor:

FRAGA GARCIA JAVIER - 33560237V
Firmado digitalmente por FRAGA GARCIA JAVIER - 33560237V
Fecha: 2021.12.14 14:39:49 +01'00'

Don Javier Fraga García

Prefacio

El deseo de extraer cada vez una mayor cantidad de información de los datos disponibles ha llevado a un rápido aumento en la investigación e implementación de técnicas estadísticas que permitan obtener información sobre áreas pequeñas. Un área o dominio pequeño hace referencia a una subdivisión de una población (geográfica o socio-económica) para la cuál no se pueden producir estimaciones directas con un nivel de precisión aceptable (Borrelli et al., 2012). Por lo general, esto es debido a que dichos dominios no estaban contemplados en el diseño original de la muestra y es posible que existan muy pocos datos, o incluso ninguno, sobre dicha área. Ejemplos típicos de áreas pequeñas serían provincias, municipios, minorías étnicas, etc.

Sin embargo, es importante resaltar que la característica de pequeña no viene dada por el tamaño real de dicho grupo en la población sino por su tamaño en la muestra. Es decir, que si el diseño de la muestra fue pensado para realizar estimaciones sobre el conjunto del país (por ejemplo en España) una comunidad autónoma podría constituir un área pequeña si no existen suficientes observaciones de dicha comunidad como para producir estimadores directos lo suficientemente fiables.

Una primera solución para tratar con estos problemas sería intentar subsanar el problema de falta de datos en la fase del diseño muestral. Se podría aumentar el tamaño de la muestra para obtener más datos sobre los dominios sobre los que se desea hacer la estimación, no obstante, esto presenta dos inconvenientes. El primero es que dicho aumento en el tamaño supondría un aumento muy considerable en el coste del estudio y que los errores vinculados con el muestreo suelen aumentar a mayor tamaño de muestra (Aliaga, 2001). El segundo es que muchas veces el objetivo del estudio no es realizar estimaciones sobre el dominio pequeño sino que esta necesidad aparece a posteriori.

Como solución a esta situación se ha introducido la metodología de estimación en áreas pequeñas. Esta metodología se basa en la utilización de información auxiliar para subsanar el problema de la ausencia datos. Dicha información puede provenir de 3 fuentes: la misma área en otro momento del tiempo, otra área similar en el mismo momento del tiempo, o una mezcla de ambas (Rao y Molina, 2015).

La información auxiliar se puede obtener, por tanto, de diversas fuentes. Encuestas o estudios realizados para otros fines o registros gubernamentales como los censos suelen ser una buena opción (Pfeffermann, 2013). No obstante, es necesario tener en cuenta que el uso de información suplementaria obtenida a través de otras fuentes debe hacerse atendiendo a un estudio minucioso y un conocimiento amplio de la información. Esto es debido a que dicha información puede haberse utilizado con propósitos muy diversos y ello implica que la combinación de fuentes debe llevarse a cabo con prudencia (Aliaga, 2001).

Los estimadores conectan las áreas pequeñas y los datos suplementarios a través de la implementación de modelos que se utilizan para realizar un "link" entre la información disponible y la información

complementaria con la que se quiere mejorar la estimación de la característica de interés para el área pequeña. En este sentido, cabe señalar que existen dos tipos de modelos:

- **Modelos a nivel de área:** son aquellos en los que la información aportada por las covariables solamente está disponible a nivel de área.
- **Modelos a nivel de unidad:** son aquellos en los que se dispone de información sobre las covariables para todas las unidades muestrales.

Dentro de estas posibilidades existen tres modelos que son casos específicos de los modelos lineales generalizados mixtos y se caracterizan por tener en cuenta efectos fijos y aleatorios. A continuación, se citan los tres modelos y se explica brevemente cuando se formularon y en que contexto.

Modelo de errores anidados: propuesto en el año 1988 por Battese, Harter y Fuller en el contexto de la estimación de la superficie media de acres cultivados en los condados del estado de Iowa en Estados Unidos. Se recurrió a información procedente de satélites y se utilizó como información auxiliar la proporcionada por diversas encuestas. Este modelo constituye un ejemplo de modelo a nivel de unidad.

Modelo de coeficientes aleatorios: propuesto en el año 1981 por Dempser, Tubin y Tsutakawa constituye un modelo más general que el anterior. No ha sido muy utilizado en los últimos años y, por ello, constituye el objeto de estudio del presente documento.

Modelo Fay-Herriot: propuesto por Fay y Herriot en el año 1979 en el contexto de la estimación de los ingresos per capita en áreas con población inferior a los 1000 habitantes. Este es, probablemente, el modelo más usado en el contexto de la estimación en áreas pequeñas. Este modelo sería un ejemplo de modelo a nivel de área.

Como el objetivo del presente trabajo es el estudio de las estimaciones en áreas pequeñas se revisarán y explicarán con mayor detalle en el capítulo de la metodología. No obstante, para más información se remite al lector a Prasad y Rao (1990).

El modelo de errores anidados y el modelo Fay-Herriot han sido ampliamente estudiados y utilizados en los últimos años para llevar a cabo estimaciones en el contexto de las áreas pequeñas. Tanto es así, que aparecen citados con regularidad en diversos manuales sobre esta metodología elaborados por y para diversos institutos de estadística. Por citar un par de ejemplos, nos encontramos con el manual de*Metodologie di stima per piccole aree applicabili a variabili di censimento* elaborado por Borrelli et al. (2012) para el Istituto Nazionale di Statistica italiano, el manual elaborado por Eurostat (2019), el realizado por Mancho (2002) para el EUSTAT o los diversos documentos técnicos elaborados por Instituto Galego de Estatística (2009). Sin embargo, a lo largo de la revisión bibliográfica realizada para la elaboración de este documento se ha hecho patente la carencia de estudios que impliquen la utilización del modelo de coeficientes aleatorios. Teniendo en cuenta esto, y viendo que por sus características puede ser de interés en el campo de estudio de la estructura salarial, se va a evaluar la posibilidad de utilizarlo para la estimación del salario por hora. Esta variable se obtiene de los datos proporcionados por la Encuesta de Estructura Salarial (EES) llevada a cabo por el Instituto Nacional de Estadística (INE) en el contexto de los reglamentos de la Unión Europea para el estudio de la estructura salarial de los Estado Miembros. Esta encuesta se caracteriza por comprender dos operaciones, una encuesta cuatrienal y otra anual.

Encuesta cuatrienal: se realiza desde 1995 utilizando criterios y metodologías comunes con la Unión Europea con el fin último de que los resultados sean comparables y se pueda obtener una mejor

perspectiva del estado de los salarios en los países miembros. Para investigar los salarios se tienen en cuenta una gran variedad de variables sociales y económicas tales como el sexo, la rama de actividad, la antigüedad, etc.

Encuesta anual: se realiza desde el 2004 y proporciona estimaciones de la ganancia bruta anual por trabajador teniendo en cuenta cuatro variables: tipo de jornada, sexo, actividad económica y ocupación.

El objetivo del presente documento tiene, por tanto, dos objetivos fundamentales. El primero y más general es indagar en la estructura salarial y laboral española. El segundo consiste en investigar como afectan a la ganancia por hora, tomada como variable respuesta, diversas covariables como por ejemplo el *Nivel de estudios*, la *Antigüedad*, la *Edad*, la *Ocupación*, etc. El objetivo sería realizar esta estimación para las diferentes ocupaciones laborales contempladas (directores, técnicos, trabajadores cualificados, etc.). El hecho de que la encuesta haya sido recogida con el fin de realizar inferencia sobre el conjunto de la población es lo que nos lleva a utilizar la metodología de áreas pequeñas ya que las ocupaciones laborales no son un dominio especificado o planificado en el diseño de la encuesta.

La estructura en la que se dividirá el documento será la siguiente:

Capítulo 1. Datos: durante este capítulo se realizará una introducción a los datos de los que se dispone para realizar el estudio. De esta forma se realizará un análisis descriptivo para valorar como se distribuyen las variables, que información nos pueden aportar sobre la estructura salarial y laboral del país y los posibles problemas a la hora de realizar estimación sobre la variable objeto de estudio.

Capítulo 2. Metodología: este capítulo realizará una breve introducción a la metodología de áreas pequeñas por las posibilidades que estos modelos aportan para resolver los problemas encontrados y expuestos durante el capítulo anterior. Así, se realizará una introducción teórica a los modelos.

Capitulo 3. Análisis de Resultados: una vez conocidos los datos y presentada la metodología que se va a utilizar para realizar la estimación, en este capítulo se procederá al ajuste y evaluación de los modelos.

Capítulo 4. Conclusiones: se aportará en este apartado información sobre las conclusiones a las que se ha llegado y se aportaran ideas sobre futuras aplicaciones del modelo estudiado en este u otros campos de estudio.

Capítulo 1

Datos

Los datos de la Encuesta de Estructura Salarial están disponibles en la página web del INE para su descarga y utilización por parte de la población y se corresponden con el estudio elaborado para el año 2018. La base de datos consta de una muestra de 216.726 personas que, teniendo en cuenta los pesos atribuidos a cada observación (factores de elevación) representan a un total de 12.976.074 individuos.

Esta encuesta se caracteriza por recoger información de dos tipos. El primero se corresponde con información de carácter social y demográfico como el sexo, la edad, el nivel de estudios o la nacionalidad. El segundo son variables de carácter puramente económico y laboral como el tipo de contrato, la titularidad de la empresa, las bases de cotización o la categoría laboral del empleado. Se va a trabajar con un total de 15 variables consideradas de interés para el estudio, 11 de ellas categóricas y 4 numéricas. Cabe destacar que, de las variables numéricas, solamente una estaba recogida originalmente en la base de datos y las otras 3 se han construido en base a otras variables con las fórmulas aportadas por el INE para ello. Dichas variables son:

- *Salario mensual* (medida en euros): calculada teniendo en cuenta el salario base más todos los complementos percibidos.
- *Horas de trabajo en el mes de referencia*: calculada teniendo en cuenta la jornada pactada en horas y en minutos más las horas extra realizadas.
- *Salario por hora* (medida en euros): resultado de la división de las dos anteriores.

Se ha decidido utilizar la variable *Salario por hora* como variable objeto de estudio. Esta medida permite realizar una comparación entre los trabajadores independientemente del tipo de jornada o contrato que tengan. Esto la convierte en una buena forma de valorar las diferencias salariales presentes en el mercado laboral.

Una vez explicado esto y antes de empezar con el análisis descriptivo se presenta el Cuadro 1.1 donde se resumen las variables categóricas a utilizar presentando sus categorías, el número de casos en la muestra y el porcentaje de esta que representan. Resulta necesario aclarar que se ha decidido agregar parte de las categorías originales por ser demasiado exhaustivas y para aportar una mayor claridad. De esta forma en la variable *Sector económico* se han agrupado diversas categorías basándose en los grupos de la Clasificación Nacional de Actividades Económicas (CNAE-2009). En la variable *Categoría laboral* se ha decidido agrupar a los técnicos, a los empleados de oficina, a los trabajadores

de servicios, a los trabajadores cualificados, a los operadores de maquinaria y a los trabajadores no cualificados. Además cabe señalar que existía también la categoría ocupaciones militares que constaba de solamente 51 observaciones y que, teniendo en cuenta que su agrupación más natural hubiera sido con los trabajadores de los servicios de seguridad, ha quedado finalmente englobada con los trabajadores de servicios. Se ha modificado también la categorización de la variable nivel de estudios de forma que se han agrupado en estudios primarios (incluye también a las personas que no tienen dicha formación), estudios secundarios (agrupando las dos etapas de secundaria) y estudios superiores (agrupando a las personas con FP de grado superior y a los universitarios). Por último se ha retocado la variable *Edad* y se ha hecho un solo grupo que contiene a los individuos entre 16 y 29 años donde originalmente había dos.

Como se comentó previamente las variables se pueden dividir en dos grupos: aquellas que aportan una información de carácter social y demográfico, como son el *Sexo*, la *Edad*, el *Nivel de estudios* o la *Zona geográfica*, y aquellas que aportan información de carácter económico como pueden ser la *Ocupación laboral*, la *Titularidad de las empresas*, el *Mercado de actuación*, etc. En el Cuadro 1.1 se hace un resumen de todas las variables que se van a analizar y utilizar en el estudio. A continuación se va a resumir la información del cuadro.

En un primer vistazo a las variables de índole social y demográfico observamos que: en el caso del sexo, la base de datos presenta una representación equitativa de hombres y mujeres. Además, se observa que la inmensa mayoría de los trabajadores son de nacionalidad española con un 94.5 %. En relación con el nivel de estudios, se observa que el 44.3 % se concentra en la etapa de secundaria. Frente a esto, nos encontramos con la categoría de estudios superiores que es la siguiente en relevancia con un 38.8 % de las observaciones. En relación con la edad se observa que los tres grupos que presentan más representación son aquellos que comprenden a la población entre 30 y 59 años.

Por otro lado, y entrando ya en las variables de carácter económico se puede observar que es destacable la relevancia de las empresas de titularidad privada en la base de datos que suponen un 83.6 % sobre el total de las empresas que participaron en el estudio, esto remarca la importancia del sector privado en la estructura económica española. Con respecto a los mercados de actuación de dichas empresas se observa que la mayor parte se concentran en un marco Local o Regional y Nacional con casi un 80 % del total, resulta llamativo que el siguiente mercado en importancia es el mundial con un 13.5 % de las empresas operando en dicho ámbito frente al 6.39 % que operan en el marco del mercado europeo. Esta situación llama la atención teniendo en cuenta que las regulaciones del mercado único europeo deberían favorecer que las empresas operaran en dicho ámbito en detrimento del mercado mundial. Es posible que esta situación se deba a que dichas empresas no son competitivas en el contexto europeo y sí que lo son en el mundial. Se puede apreciar que la jornada a tiempo completo es claramente la mayoritaria con un 82 % de los casos, lo mismo sucede con el contrato de tipo indefinido que, con un 79.4 % de las observaciones, es el más común. Por último, en relación con la categoría laboral, se puede valorar que la representación de todas ellas es más o menos similar salvo algunas excepciones como son los directores y gerentes. También es llamativo que el sector con más representación es el de la industria manufacturera.

Una vez presentadas brevemente las variables se procede a la realización de un análisis descriptivo sobre las mismas para ver como se relacionan entre ellas, haciendo especial énfasis en su relación con la ganancia por hora que es nuestra variable objeto de estudio.

Cuadro 1.1: Tabla resumen Variables Categóricas.

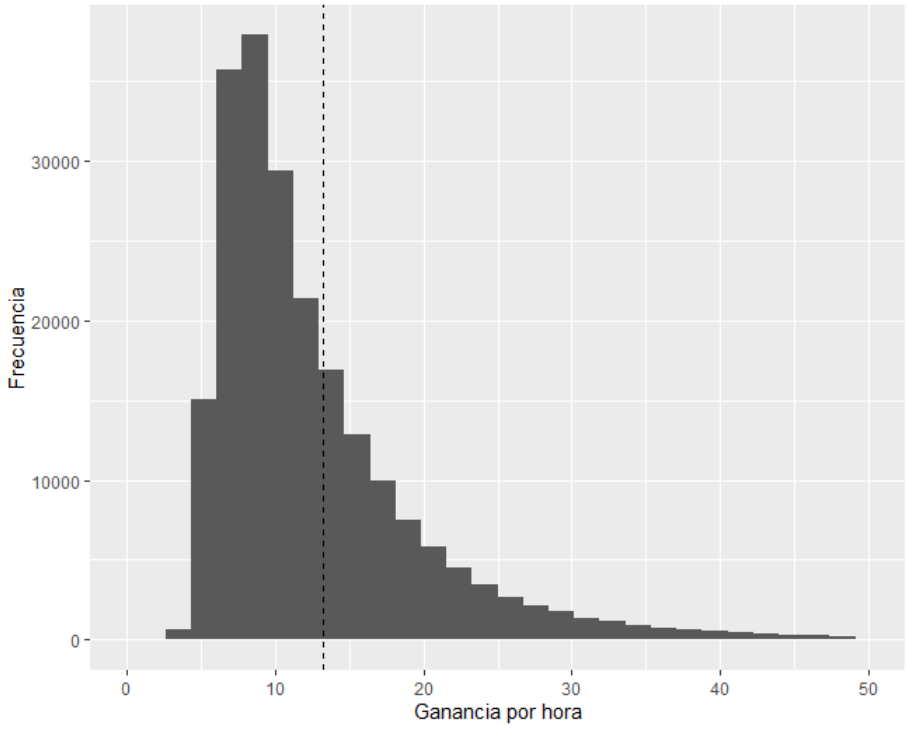
Variable	Categoría	Nº de Casos	% de Casos
Zona Geográfica	Noroeste	24.806	11.45
	Noreste	33.624	15.51
	Comunidad de Madrid	34.269	15.81
	Centro	26.428	12.19
	Este	58.852	27.16
	Sur	29.413	13.57
	Canarias	9.334	4.31
Sector de Actividad Económica	Industria manufacturera	51.158	23.60
	Construcción	12.412	6.51
	Hostelería y Comercio	14.119	11.93
	Transporte	11.226	5.18
	Información/Comunicaciones/Financieras	20.890	9.64
	Actividades científicas y Educación	23.731	10.95
	Actividades Administrativas	30.180	13.93
	Actividades sanitarias y de servicios sociales	19.250	8.88
	Otros servicios	20.312	9.37
Titularidad	Pública	35.553	16.4
	Privada	181.173	83.6
Mercado	Local/Regional	82.115	37.89
	Nacional	91.244	42.1
	Unión Europea	13.887	6.41
	Mundial	29.480	13.6
Sexo	Hombre	122.558	56.55
	Mujer	94.168	43.45
Nacionalidad	España	204.683	94.44
	Resto del Mundo	12.043	5.56
Categoría laboral	Directores y Gerentes	6.958	3.21
	Técnicos	72.072	33.25
	Empleados de oficina	30.184	13.93
	Trabajadores de servicios	32.067	14.80
	Trabajadores cualificados	29.254	13.50
	Operadores de maquinaria	19.112	8.82
	Trabajadores no cualificados	27.079	12.49
Nivel de Estudios	Primaria	36.255	16.73
	Secundaria	95.468	44.05
	Superior	85.003	39.22
Edad	De 16 a 29	21.610	9.97
	De 30 a 39	52.936	24.43
	De 40 a 49	72.439	33.42
	De 50 a 59	54.075	24.95
	Más de 59	15.666	7.23
Tipo de Jornada	Tiempo completo	177.774	82.03
	Tiempo parcial	38.952	17.97
Tipo de Contrato	Indefinido	172.292	79.5
	Temporal	44.434	20.5

1.1. Análisis Descriptivo

Teniendo en cuenta que el objetivo es realizar estimación sobre la ganancia por hora de trabajo de la población española parece razonable comenzar el análisis por esta variable. Una forma de visualizar la estructura salarial sería calculando el salario medio por hora de la población. Esto implica calcular individualmente sobre los datos cuanto gana cada persona por una hora de trabajo y calcular su media aritmética. Si atendemos a esta medida se puede decir que en España el salario medio por hora es de 13,22 euros.

Entroncando con esto, se puede intuir que el salario por hora presentará una distribución asimétrica positiva ya que, en principio, las ganancias mínimas están acotadas por la legislación vigente o como mínimo por el 0 mientras que las ganancias máximas no tienen límite. Un estudio de la distribución del salario por hora muestra que el coeficiente de asimetría es 18,87 lo cual indica una clara asimetría positiva y que hay una cantidad considerable de datos superiores a la media. Por otro lado, el coeficiente de curtosis es 962,49 lo cual indica que estamos ante una distribución más apuntada que la gaussiana y con colas más pesadas. Esta situación se puede observar en el histograma de la Figura 1.1, donde la línea continua representa el salario medio por hora de trabajo.

Figura 1.1: Histograma del salario por hora

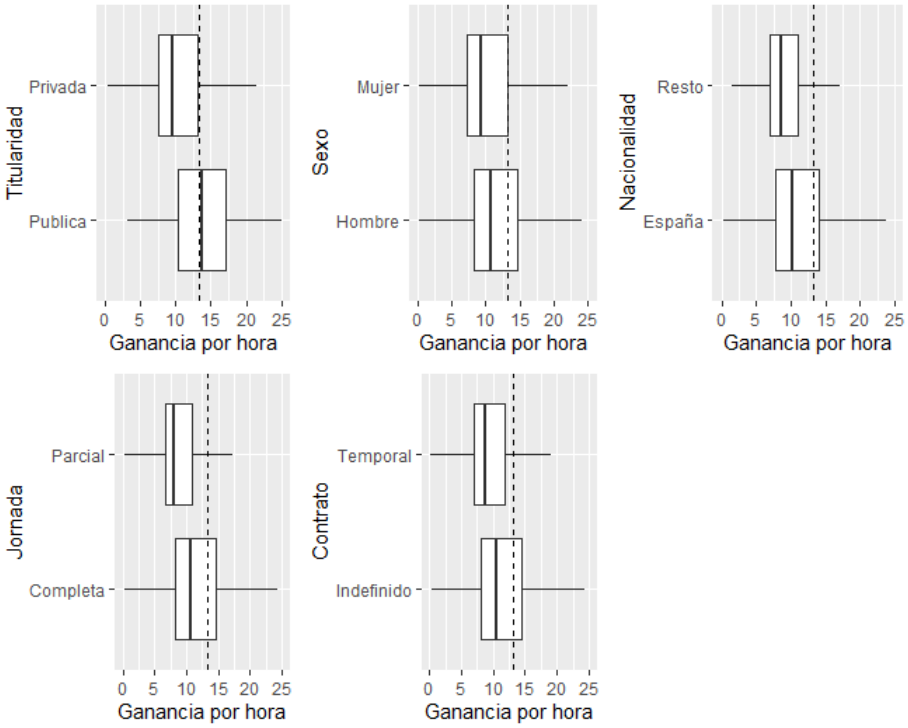


Teniendo en cuenta que una pregunta interesante a resolver es si existen diferencias significativas en las ganancias por hora en función de las demás variables un hecho importante a contrastar es si los datos provienen de una distribución normal. A la vista de lo anteriormente expuesto no parece que esto vaya a ser así, no obstante, se va a recurrir a un test de bondad de ajuste para contrastar la normalidad de los datos de forma analítica. Teniendo en cuenta la gran cantidad de datos se ha decidido recurrir al test de Kolmogorov-Smirnov con la corrección de Lilliefors. Con un p-valor casi

igual a 0 se puede concluir que para cualquiera de los niveles de significación habituales se rechaza la hipótesis de que los datos puedan provenir de una distribución normal.

Para evaluar las diferencias en las distribuciones del salario por hora se va a recurrir a gráficos de tipo boxplot. Se ha decidido recurrir a estos gráficos ya que nos permiten valorar de un modo visual tanto la distribución de la variable en las categorías (asimetría, mediana, etc) como la presencia de atípicos. Además se ha añadido a los boxplot el salario medio por hora de trabajo en forma de línea discontinua, esto nos permitirá valorar la distribución de las ganancias por hora en los grupos en relación con esta medida de referencia. Ha sido necesario eliminar de la visualización los valores atípicos porque impedían visualizar el gráfico correctamente debido a la escala. Es destacable que hay un gran número de observaciones atípicas en valores elevados de salario por hora.

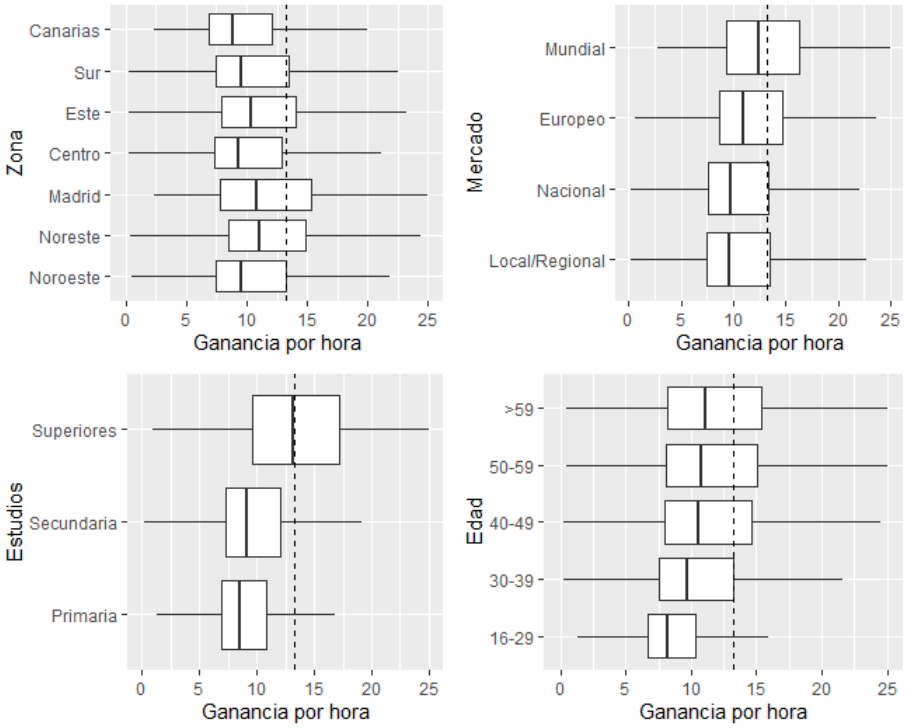
Figura 1.2: Boxplot de la ganancia por hora en función de variables con dos categorías



En la Figura 1.2 se observan diferencias en la distribución del salario por hora en función de varias de las variables. En relación con la titularidad de la empresa se puede observar que la ganancia mediana de los trabajadores del sector privado es inferior, no solo a la de los trabajadores del sector público, sino también al salario medio. Los trabajadores del sector público, sin embargo, se encuentran casi todos en valores de ganancia por hora superiores a este umbral. Las diferencias en función del sexo del individuo también parecen claras ya que se observa que la ganancia mediana de las mujeres está por debajo de la de los hombres, siendo la de estos casi igual al salario medio por hora de trabajo. Atendiendo a la nacionalidad se observa que la ganancia mediana por hora de las personas con nacionalidad española es superior a la del resto. Por último, en relación con el tipo de jornada y contrato se observa que los trabajadores con jornada completa y contrato indefinido tienen una ganancia mediana superior. Esto es destacable ya que esta medida está pensada para que las ganancias de las personas sean comparables independientemente del tipo de contrato y jornada que tengan.

En la Figura 1.3 se realiza lo mismo para variables que presentan varias categorías. Se observa la existencia de diferencias salariales por zona geográfica pero no parecen muy destacables teniendo en cuenta la escala del gráfico. Sí que es reseñable el hecho de que solamente en la Comunidad de Madrid y la zona Noreste la ganancia mediana por hora iguala el valor del salario medio por hora, en el resto de los casos se sitúan por debajo de este umbral. Con respecto a los tipos de mercado en los que operan las empresas se observa que cuanto más grande es el mercado mayores ganancias por hora presentan los trabajadores. En relación con el nivel de estudios se observa que las mayores ganancias por hora las presentan las personas con estudios superiores. Para las otras dos categorías las ganancias se encuentran por debajo del salario medio por hora y apenas se aprecian diferencias entre las dos. Esto es llamativo ya que la categoría secundaria incluye a las personas con un FP de grado medio. En relación con la edad se observa que a mayor edad mayor salario por hora parecen percibir los trabajadores.

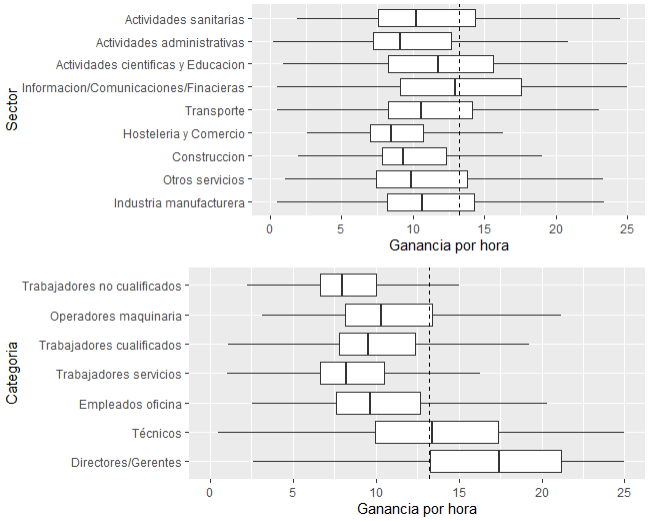
Figura 1.3: Boxplot de la ganancia por hora en función de variables con varias categorías



En la Figura 1.4 observamos las diferencias en la ganancia por hora en función del sector en el que se trabaja y de la categoría laboral a la que se pertenece. En relación con el sector se observa que los tres sectores con una ganancia por hora más elevada son el de la informática y las comunicaciones, el de los servicios financieros y el del suministro de agua. Este último es razonable teniendo en cuenta que en España este es un sector del que se hace cargo el Estado y, como ya se observó previamente, los trabajadores de las empresas públicas cobran más que los de las privadas por lo menos en media. En relación con la categoría laboral se observa que los únicos dos grupos que se sitúan por encima del umbral del salario medio por hora de trabajo son los directivos y, en parte, los técnicos. En vista de esta situación cabe preguntarse si las diferencias en la ganancia por hora en función del sector al que se pertenece son una causa o más bien una consecuencia. Es decir, cabe preguntarse si las diferencias en la ganancia por hora en determinados sectores no estarán relacionadas con la estructura económica interna de dichos sectores más que con el tipo de sector en sí. Por ejemplo, es de esperar que, por

lo menos en media las personas que trabajan en el sector financiero tengan un nivel de estudios más elevado que aquellos que trabajan en el sector de la hostelería. Además, incluso entre sectores en los que se pueda suponer que variables tales como el nivel de estudios, el tipo de jornada o incluso la categoría laboral se distribuyan uniformemente cabe preguntarse si no entran en juego otras variables, como la responsabilidad o formaciones de carácter específico que no se recogen dentro de los sistemas de enseñanza reglada, y que son derivadas de la estructura interna y de las necesidades específicas del sector que provocan claras diferencias entre los sectores en como se distribuyen las ganancias.

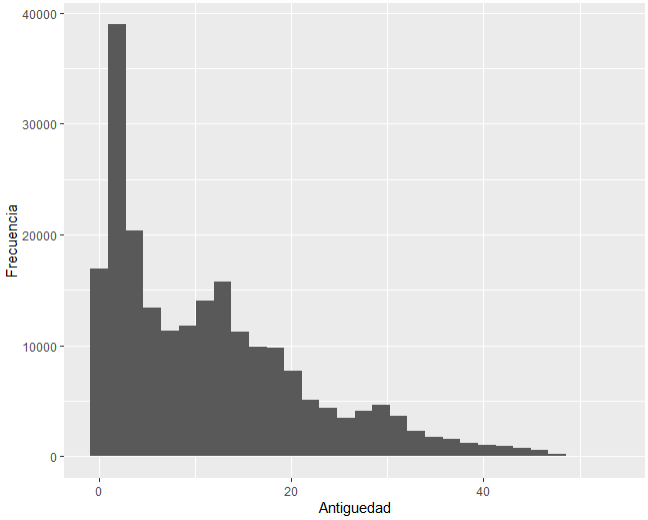
Figura 1.4: Boxplot de la ganancia por hora en función de Sector y Categoría laboral



A continuación se va a realizar un estudio sobre la relación que la ganancia por hora tiene con los años de antigüedad. Originalmente la base de datos contemplaba los años de antigüedad de los individuos en una variable y los meses de antigüedad en otra. Esta situación ha sido modificada para que una única variable contenga la mayor información posible. Así, se han dividido los meses de antigüedad de los individuos entre 12 para transformarlos a la escala de años y se han sumado a la variable años de antigüedad, de esta forma una persona que tenga una antigüedad de 5 años y 6 meses quedará recogida en la base de datos con una antigüedad de 5.5 años. Una vez explicado esto, se puede observar en la Figura 1.5 que los años presentan una distribución asimétrica hacia la derecha y que la mayor parte de las observaciones se concentran por debajo de los 20 años de antigüedad.

Con el objetivo de conocer mejor como se relacionan los años de antigüedad con la ganancia por hora se ha calculado el coeficiente de correlación entre ellas y el resultado obtenido muestra una correlación de 0.25. Esto nos indica que hay una correlación media de carácter positivo entre ellas, por lo tanto, cabe esperar que cuanta mayor antigüedad tengan los individuos mayor debe ser la ganancia por hora de trabajo. Realizando una discretización de la variable años de antigüedad en grupos de 10 años se obtiene el boxplot de la Figura 1.6 que nos permite valorar mejor como se distribuye la variable en función de los años de antigüedad. De esta manera se observa que efectivamente a mayor antigüedad mayor ganancia por hora parecen tener los grupos. No obstante, se ve un hecho curioso y es que en el último grupo, donde están agrupadas las personas que presentan una antigüedad de entre 40 y 60 años la ganancia por hora de trabajo parece sufrir un descenso, estas fluctuaciones pueden deberse al bajo tamaño muestral en esta franja de edad. Si miramos el número de personas con una antigüedad superior a los 40 años observamos que solo hay 3100 individuos que cumplen el requisito, si aumentamos el umbral a 50 años el número se reduce a 17 individuos. Esto puede llevarnos a pensar

Figura 1.5: Histograma para la variable años de antigüedad



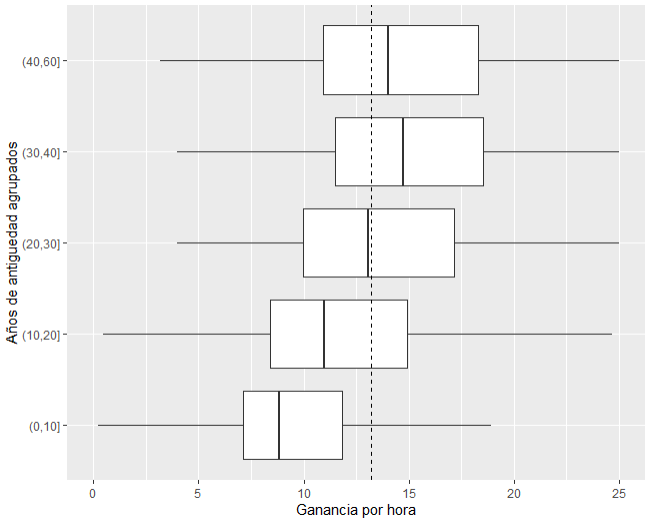
que si trabajamos directamente con la variable continúa la relación entre los años de antigüedad y la ganancia por hora no sea una relación lineal sino que incluya algún componente no paramétrico en su parte final. Con el objetivo de comprobar esto se ha realizado un ajuste mediante un modelo aditivo generalizado (GAM) que contempla la inclusión de un componente no paramétrico en la relación entre la ganancia por hora y los años de antigüedad.

El resultado se puede ver en la Figura 1.7. Se observa que la relación entre los años de antigüedad parece seguir una tendencia lineal ascendente hasta llegados los 30 años de antigüedad. En este momento la tendencia se estanca para luego decaer durante unos años y finalmente volver a ascender. A simple vista esta relación parece poco natural, no obstante, es razonable pensar que en la mayor parte de los trabajos la productividad y la capacidad de trabajo de los individuos decae con los años. De esta manera, llega un momento en que las ganancias por hora alcanzan un máximo y a partir de ahí empiezan a bajar. Esta situación es natural por ejemplo entre los conductores del sector del transporte. Por lo general los viajes que más beneficio reportan son aquellos en los que se recorren mayores distancias. Es de suponer que, eventualmente por cuestiones de edad dichos viajes tengan que evitarse o incluso dejar de hacerse lo que repercute negativamente en las ganancias por hora de los conductores. Además, también es necesario aclarar que hay pocos casos en los que la gente acumula tantos años de antigüedad debido a que ello implica comenzar a trabajar ininterrumpidamente muy joven y permanecer mucho tiempo en la misma empresa, esta escasez de muestra seguramente este, como se comentó previamente, causando las fluctuaciones en la parte final del ajuste.

Una vez analizada la relación que las variables explicativas tienen con la variable respuesta cabe preguntarse como se relacionan entre ellas las explicativas. Por ello ahora se va a proceder a realizar un análisis más exhaustivo sobre como se distribuyen los datos partiendo de lo anteriormente expuesto.

Teniendo en cuenta las diferencias encontradas en función del tipo de contrato y del tipo de jornada así como del sexo, cabe preguntarse como se relacionan estas tres variables. Para ello se ha recurrido a la Figura 1.8 donde se pueden ver cual es la distribución interna de los tipos de contrato y de jornada en función del sexo. Se observa que, en relación con el tipo de contrato, no existen grandes diferencias entre los hombres y las mujeres. Sin embargo, en relación con el tipo de jornada si que se observan

Figura 1.6: Boxplot de la ganancia por hora en función de los años de antigüedad



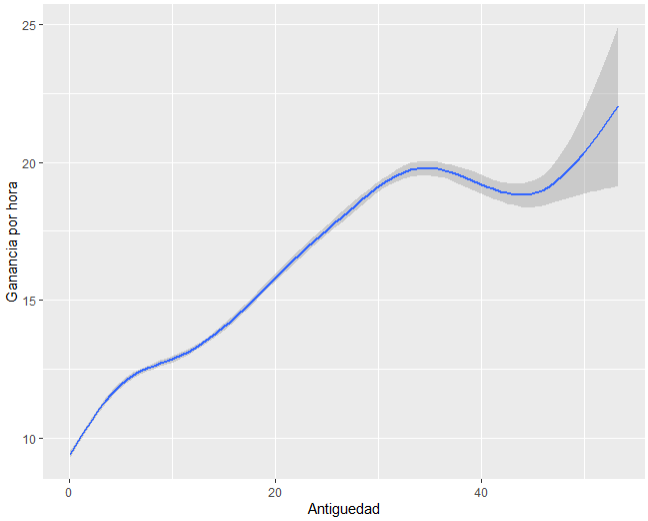
diferencias claras. De las personas que trabajan a tiempo parcial, que ya hemos observado que tienen una ganancia por hora de trabajo menor que la de los trabajadores a jornada completa, el 66.11 % son mujeres.

Hemos observado que las mujeres presentan unas ganancias por hora de trabajo inferiores a la de los varones y que también se ven afectadas por una mayor presencia en jornadas laborales de carácter parcial. Esto supone también una influencia negativa en su salario por hora. Cabe preguntarse cómo se distribuyen los niveles educativos entre los sexos. En la Figura 1.9 se puede comprobar que las mujeres presentan una mayor formación. El 43.84 % de las mujeres del estudio tienen un nivel de estudios de tipo superior frente al 36 % de los hombres. Además, el porcentaje de mujeres con un nivel educativo primario es inferior al de los hombres en 6 puntos porcentuales. Esta situación es llamativa ya que previamente se había observado que las personas con niveles educativos más elevados tienen unas ganancias por hora superiores a las demás. Sin embargo, en el caso de las mujeres parece que su mayor formación no es impedimento para sufrir una clara discriminación salarial.

En la mayor parte de los estudios sobre estructura laboral que tienen que ver con el género a parte del concepto de brecha salarial también es común el de "Techo de cristal" que supone la dificultad de las mujeres para llegar a puestos de responsabilidad a pesar de su formación o habilidades. Como en la base de datos disponemos de la variable categoría laboral podemos comprobar la diferente presencia de las mujeres y hombres en dichas categorías. Como se puede comprobar en Figura 1.10 en el caso de los directivos y gerentes solamente un 32 % son mujeres, lo cual parece ser indicativo de las dificultades que estas presentan para acceder a puestos de poder en el seno de las empresas. Algunos otros datos llamativos son que en el caso de las categorías de trabajadores de oficina y de trabajadores de servicios se observa una mayor presencia de mujeres. Esto, junto con la abrumadora mayoría de hombres en las categorías trabajadores cualificados (intimamente relacionada con la industria) y conductores de maquinaria, supone un indicativo de que ciertos patrones y roles históricos de género siguen muy presentes en el mercado laboral español.

Llegados a este punto parece interesante considerar la influencia que la categoría laboral presenta en las demás variables. Hemos comprobado que, en función de la categoría laboral a la que se per-

Figura 1.7: Relación entre ganancia por hora y años de antigüedad (Modelo GAM)



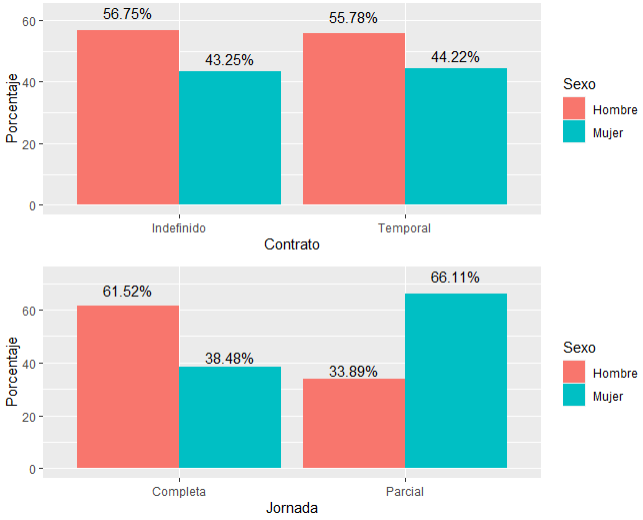
tenezca, las personas presentan unas ganancias por hora superiores o inferiores. No obstante, sería interesante comprobar si, por ejemplo, los salarios por hora de los directivos se distribuyen de igual forma independientemente del sexo o del nivel de estudios. También hemos valorado que los años de antigüedad presentan una relación con la ganancia por hora que en sus primera etapas es de carácter lineal, posteriormente se estanca en un máximo y luego comienza a bajar pero, una pregunta relevante a responder es si esta situación es igual para todas las categorías laborales, es decir, cabe preguntarse si la pendiente de la recta es igual para todos ya que el intercepto ha quedado patente que no debido a que hay categorías laborales en las que la ganancia por hora es superior. Si se comprueba que la categoría laboral ejerce influencia en las demás variables sería necesario considerar dicha variable como un efecto aleatorio a la hora de hacer estimación sobre la ganancia por hora en un modelo de regresión. Esto nos obligaría a recurrir a la metodología de los modelos lineales mixtos.

En primer lugar, se va a comprobar si las diferencias salariales en función de la categoría laboral afectan también a las personas según su género, es decir, si podemos encontrar diferencias en la ganancia por hora dentro de las propias categorías laborales. En la Figura 1.11 se puede ver que dentro de las mismas categorías laborales encontramos diferencias en función del género del trabajador. Se puede ver que en todas las categorías existen diferencias más o menos claras siendo las de los directivos y las de los trabajadores cualificados las más destacables. Además, resulta llamativo el alargamiento de la distribución en las categorías de los directivos y de los técnicos. Esto nos da a entender que, a pesar de ser los grupos que presentan una mayor ganancia por hora, también presentan una mayor variabilidad.

En segundo lugar, en la Figura 1.12 se evalúan las diferencias que existen por grupo de edad y categoría laboral. Como se puede ver la situación es similar a la hallada en el caso del género. Parece haber diferencias claras siendo estas más marcadas en las categorías de los directores y los técnicos. Es destacable, no obstante, que en el caso de los trabajadores no cualificados parece que no existen diferencias reseñables. Además, se corrobora lo anteriormente estudiado en relación con la edad, a mayor edad mayor ganancia por hora.

En tercer lugar, se ha realizado la misma comprobación teniendo en cuenta el nivel de estudios. En la Figura 1.13 se repite la misma situación anterior. Se observa que existen diferencias dentro de

Figura 1.8: Gráficos de barras. Distribución interna de los tipos de contrato y jornada por género



las categorías laborales por el nivel de estudios del trabajador y se distribuyen de forma similar a las dos variables anteriores, es decir, son diferencias más reseñables en el caso de las categorías de los directores y técnicos, donde el impacto de un nivel de estudios superior es más elevado pero también existe una mayor variabilidad en la ganancia por hora.

A la vista de lo expuesto hasta este momento parece razonable pensar que la variable categoría laboral influye en las demás variables respuesta. De este modo, ante la posibilidad de realizar un modelo de regresión para ajustar posibles estimaciones sobre la ganancia por hora parece adecuado valorar la variable categoría laboral como un posible efecto aleatorio. Por lo expuesto hasta ahora, es posible pensar en un modelo con un intercepto aleatorio derivado de que las distintas rectas de regresión podrían tener orígenes distintos en función de la categoría laboral. Por ejemplo, a raíz de lo observado hasta el momento, parece que los directivos parten de una ganancia por hora superior a la de los demás trabajadores. No obstante, cabe preguntarse si la influencia de esta variable será solamente sobre el origen de la regresión o puede afectar también a su pendiente. Como previamente se analizó la relación entre los años de antigüedad y la ganancia por hora a través de un modelo GAM vamos a comprobar la misma relación en función de la categoría laboral. A la vista de la Figura 1.14 parece que tanto el origen de la recta como su pendiente se ven afectadas por la categoría laboral. De esta manera parece que, por ejemplo, los directivos no llegan a sufrir una caída en sus ganancias por hora a partir de los 20 años de antigüedad sino que dichas ganancias se estabilizan y dejan de crecer con tanta velocidad. Sin embargo, en el caso por ejemplo de los técnicos se aprecia el mismo patrón que cuando se realizaba el análisis de forma conjunta.

A la vista de estos resultados cabe pensar en la necesidad de recurrir a modelos lineales mixtos para realizar la estimación. Más concretamente a modelos que contemplen la inclusión de efectos aleatorios que afecten tanto al intercepto como a la pendiente del modelo. Además, teniendo en cuenta que las categorías laborales no eran dominios contemplados para realizar estimación en el diseño original de la encuesta nos veremos obligados a recurrir a la metodología para áreas pequeñas. En el próximo capítulo se va a hacer un repaso a dicha metodología y se va a proponer un modelo que nos permita realizar estimaciones para la ganancia por hora teniendo en cuenta la situación expuesta a lo largo de este apartado.

Figura 1.9: Gráficos de barras. Nivel de estudios en función del sexo

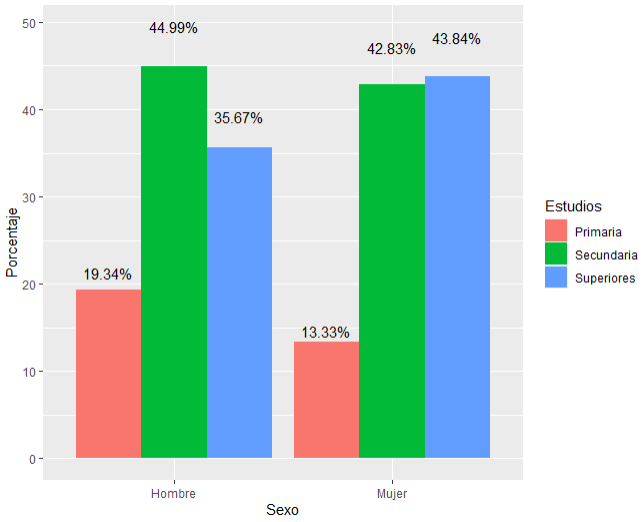


Figura 1.10: Gráficos de barras. Sexo en función de la categoría laboral

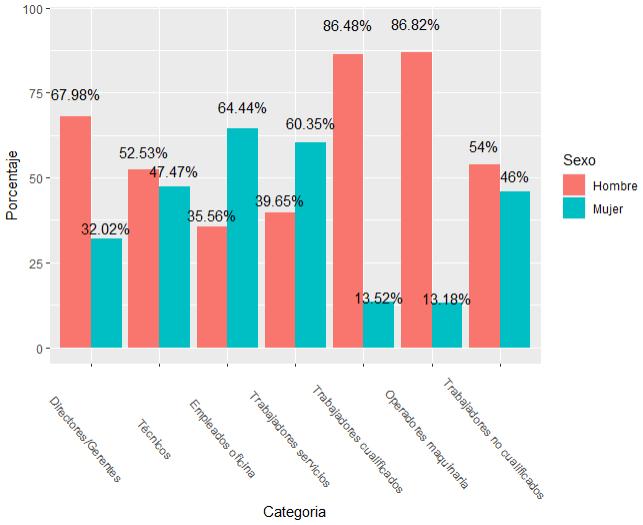


Figura 1.11: Boxplots de la ganancia por hora por categoría laboral y sexo

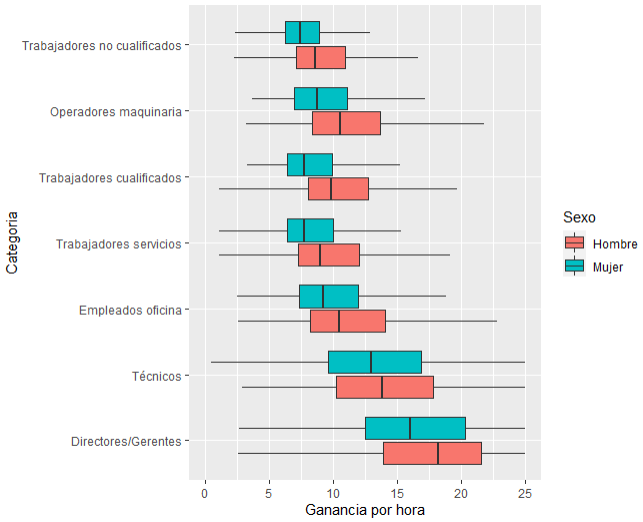


Figura 1.12: Boxplots de la ganancia por hora por categoría laboral y edad

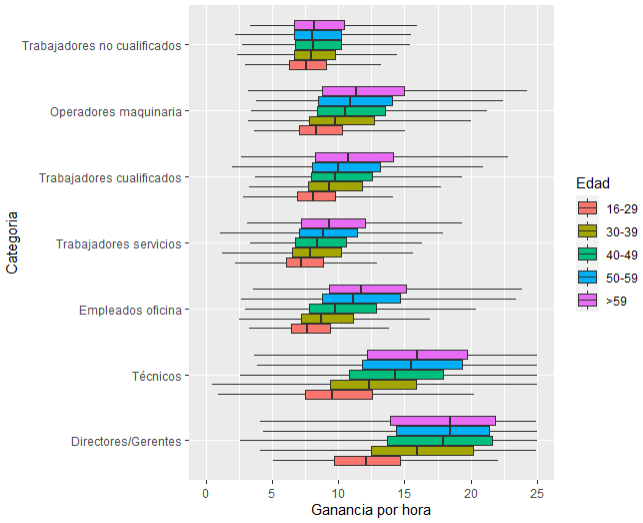


Figura 1.13: Boxplots de la ganancia por hora por categoría laboral y nivel de estudios

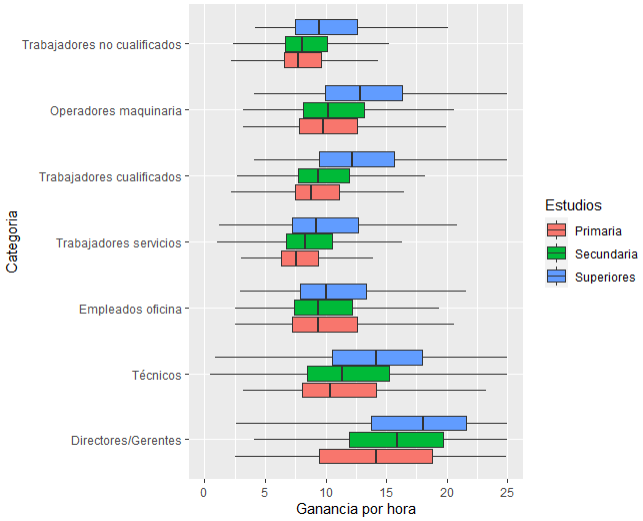
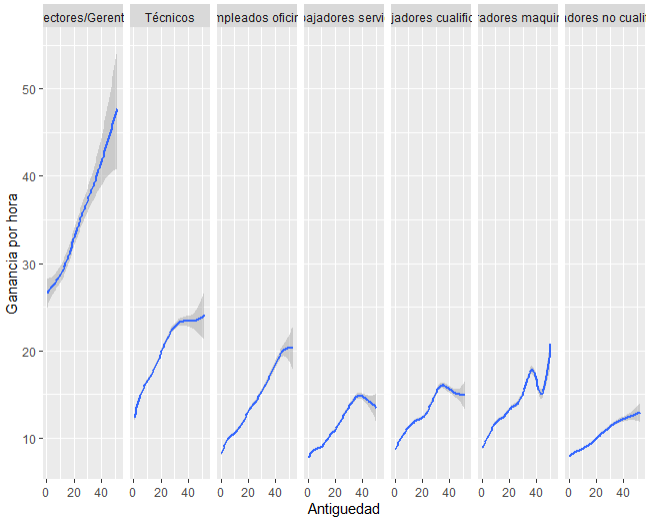


Figura 1.14: Modelos GAM para la ganancia por hora en función de los años de antigüedad. Separación por categoría laboral



Capítulo 2

Metodología

Ahora que se tiene un conocimiento más amplio de los datos recogidos en la Encuesta de Estructura Salarial elaborada por el INE resulta necesario realizar una introducción a la metodología para el estudio de estos datos. La idea es buscar algún modelo que nos permita poner en relación la información aportada por las variables independientes con la variable dependiente. Por tanto, nuestro objetivo es hallar un modo de poner en relación las variables explicativas previamente analizadas con la variable *Salario por hora* con el fin de llevar a cabo estimaciones sobre esta métrica. Los métodos de estimación basados en modelos asumen un modelo para la muestra y a través de este modelo se relaciona la información disponible con la información auxiliar. Esto es de vital importancia ya que, como apunta Pfeffermann (2013), incluso los modelos más complejos pueden no producir estimaciones o predicciones especialmente precisas en ausencia de un tamaño muestral elevado o de covariables con una capacidad predictiva elevada. Algunos de los modelos más utilizados para llevar a cabo esta asociación son los modelos lineales. Gracias a su facilidad de interpretación y a su buen rendimiento en multitud de estudios con datos reales se han convertido en modelos muy usados en diversos ámbitos. El modelo al que se va a recurrir para la realización de las estimaciones en este trabajo es el modelo de coeficientes aleatorios. Está basado en los modelos lineales mixtos y es una metodología propuesta para el estudio de relaciones entre variables en el contexto de las áreas pequeñas. No obstante, a lo largo del capítulo del análisis de resultados se va a proceder mediante un proceso iterativo en el cual los modelos irán aumentando su complejidad al tiempo que se busca la realización de un mejor ajuste de los datos. Se partirá de un modelo de regresión múltiple en el que todos los efectos serán fijos, luego se hará un ajuste en el que se valorará la modelización de la variable *Categoria laboral* como un efecto aleatorio que afecte al intercepto del modelo y posteriormente se valorará que dicha variable afecte tanto al intercepto como a la pendiente del modelo. La forma de proceder mencionada hace necesaria una introducción a los modelos lineales para su mejor conocimiento.

Antes de comenzar con la presentación de los modelos es importante introducir la notación que se va a utilizar. Como se ha explicado, se dispone de una base de datos proporcionada por el INE en la que nos basaremos para realizar estimaciones sobre el conjunto de la población española sobre la variable de interés. Por tanto, consideremos que tenemos una población U de tamaño N que se corresponde con el conjunto de la población trabajadora española. Esta población está dividida en M áreas $U_1 \cup ... \cup U_M$, en nuestro caso estas áreas son las ocupaciones laborales de los individuos, con N_d unidades en el área d. Por tanto, $\sum_{d=1}^{M} N_d = N$. Suponiendo que disponemos de una muestra con $m < M$ áreas y dejemos que el total de la muestra sea $s = s_1 \cup ... \cup s_m$ donde el s_d de tamaño n_d es la muestra observada para la d área muestreada. Por tanto, bajo el mismo razonamiento que antes $\sum_{d=1}^{m} n_d = n$. El objetivo es estimar θ_d. En nuestro caso el objetivo es estimar el salario medio por

hora de la población española para cada ocupación laboral, es decir, $\theta_d = \bar{Y}_d = \sum_{j=1}^{N_d} y_{dj}/N_d$, donde j representa a cada unidad perteneciente al área d. Partiendo de esto, tenemos que y_d define los valores de la variable respuesta observados en la muestra mientras que x_d define los correspondientes valores de las covariables en cada área.

Para la elaboración de este capítulo se ha recurrido a varias fuentes, pero las principales son: el artículo de Prasad y Rao de 1990, el artículo de Pfeffermann del 2013, el libro escrito por Rao y Molina en el 2015 y el libro sobre modelos lineales mixtos de Badiella y Sánchez del 2011. A ellos se remite al lector para una mayor información sobre los temas en adelante tratados.

2.1. Modelos Lineales

Los modelos lineales se basan en la suposición de que el valor esperado de la variable respuesta se puede expresar como una combinación lineal de las variables explicativas (Badiella y Sánchez, 2011). Esto da lugar a un modelo de muy fácil interpretación y que se ha usado de forma común en muy diversos ámbitos. El modelo presentaría la siguiente forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon \tag{2.1}$$

siendo y la variable respuesta sobre la cuál se desea hacer la predicción o estimación y $(x_1, x_2 ... x_p)$ el conjunto de variables explicativas con el cuál se desea realizar la predicción. Además, nos encontramos con el coeficiente β_0 que representa el valor que tomaría la variable respuesta en ausencia de cualquier otra información, constituye el intercepto del modelo, y con el vector formado por $(\beta_1, \beta_2 ... \beta_p)$ que representan el incremento medio en la variable y al aumentar en una unidad el valor de la covariable que los acompaña dejando igual el resto, estos coeficientes se interpretan como la pendiente del modelo. Por último, ϵ es un componente de error derivado del muestreo que se suele asumir con media 0 y varianza desconocida σ^2.

Este modelo ha sido ampliamente utilizado gracias a su facilidad de interpretación y su versatilidad en los más diversos campos: medicina, economía, sociología, biología, etc. A modo de ejemplo, pensemos en el caso que nos atañe. Deseamos hacer estimación sobre el salario por hora de la población y para ello queremos utilizar como variable explicativa los años de antigüedad. Por tanto, podemos ajustar el siguiente modelo lineal:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

donde y es el salario por hora, x_1 son los años de antigüedad y deseamos estimar β_0 y β_1. β_0 sería el origen de la recta de regresión, es decir, el salario por hora del que se parte con 0 años de antigüedad y β_1 sería la pendiente del modelo, es decir, el aumento o decremento que se produce en el salario por hora al aumentar en una unidad los años de antigüedad. Se puede observar que, a partir de un modelo muy sencillo, se puede modelizar de forma muy intuitiva la relación entre dos variables. Por comodidad en las siguientes páginas denominaremos a este modelo como "Modelo 0".

Los modelos lineales más comunes y conocidos probablemente sean el modelo de regresión lineal (tanto simple como múltiple) y el modelo de análisis de la varianza. En el modelo de regresión lineal múltiple una serie de variables cuantitativas se utilizan para estimar una variable objetivo también cuantitativa mientras que en el caso del análisis de la varianza la variable respuesta (cuantitativa)

se explica a través de la inclusión de un conjunto de variables cualitativas también llamadas factores (Badiella y Sánchez, 2011).

2.1.1. Modelo lineal mixto

El modelo lineal mixto constituye una generalización del modelo lineal general. Esta generalización se basa en la idea de que las observaciones analizadas pueden no ser independientes o presentar una variabilidad heterogénea (Badiella y Sánchez, 2011). Lo más común para ajustar este tipo de modelos es contemplar la presencia de factores fijos y aleatorios simultáneamente. Es necesario recordar que un factor es una variable de carácter cualitativo que presenta varias categorías.

- Factor fijo: sería aquel en el cuál las categorías suponen el conjunto de todos los niveles observables, o bien, dichas categorías han sido fijadas por el investigador previamente.
- Factor aleatorio: es aquel cuyos niveles son una muestra de todos los posibles niveles reales presentes en la población, de esta forma la variable respuesta se observa para cada categoría que constituye el factor aleatorio y se espera que haya homogeneidad entre los elementos dentro de una misma categoría y heterogeneidad entre las distintas categorías.

Es habitual asumir que en los modelos lineales con factores aleatorios existen dos componentes de variabilidad claramente diferenciados. En primer lugar, nos encontraríamos con la variabilidad entre factores, que se suele asumir que se corresponde con los denominados efectos aleatorios. Es habitual asumir que estos efectos aleatorios tienen media 0 y varianza desconocida. En segundo lugar, nos encontramos con la variabilidad entre las observaciones dentro de los propios factores, esta es una variabilidad experimental y se corresponde con el ϵ en los modelos lineales presentados con anterioridad. Teniendo esto en cuenta se puede formular el siguiente modelo:

$$y_{dj} = \beta_0 + \beta_1 x_{1dj} + ... + \beta_p x_{pdj} + v_d + \epsilon_j$$

siendo $j = 1, ..., n$ el indicador del individuo y $d = 1, ..., D$ el indicador del nivel del factor aleatorio. Los coeficientes β presentan la misma interpretación que en los modelos presentados con anterioridad al igual que ϵ_j que constituye la variabilidad entre los individuos observados y es un componente de error relacionado con la aleatoriedad propia del muestreo y la experimentación. Sin embargo, la novedad de este modelo la encontramos con el componente v_d que es el encargado de explicar la variabilidad entre los distintos niveles del factor aleatorio. Este componente se supone generalmente con media 0 y varianza desconocida, es normal asumir que $v_d \sim \mathcal{N}(0, \sigma_v)$.

Este modelo puede dar lugar a varias situaciones:

1 Modelos con intercepto aleatorio: este modelo se basa en la suposición de que, debido a la presencia de un efecto aleatorio, los distintos niveles de estudio del factor aleatorio presentan una recta con un origen distinto.

2 Modelos de pendiente aleatoria: en este caso, el efecto aleatorio afecta a todas las covariables que ayudan a estimar la variable respuesta de forma que modifican la pendiente del modelo. Esto provoca en la práctica la aparición de una recta distinta para cada grupo.

3 Modelos con intercepto y pendiente aleatoria: este modelo supone una combinación de los dos mencionados previamente y consiste en suponer que los niveles de factor aleatorio provocan cambios tanto en el origen como en la pendiente de la recta ajustada.

Estos modelos aportan una gran flexibilidad para ajustar relaciones entre variables donde es fácil suponer la existencia de una mayor variabilidad de la esperada debido a simples errores de medición o a la aleatoriedad inherente a la investigación y el muestreo. Y en esta situación nos encontramos en el estudio de las áreas pequeñas. Por ello, partiendo de los modelos lineales mixtos se han propuesto varias metodologías para realizar estimaciones sobre variables de interés en el contexto de las áreas pequeñas.

2.2. Modelos de áreas pequeñas

En esta sección se va a introducir la metodología de áreas pequeñas. Como se ha comentado en la sección anterior, estos modelos constituyen una caso particular dentro de los modelos lineales mixtos. La utilización de estos modelos es recomendable cuando lo que se pretende es hacer estimación sobre un parámetro de interés medido en relación con un dominio no planificado previamente en el diseño del estudio. En este caso nos encontramos nosotros ya que el objetivo del presente documento es realizar estimaciones sobre el salario medio por hora dependiendo de la ocupación de los indiviudos. Por su diseño, la Encuesta de Estructura Salarial permite realizar estimaciones directas fiables sobre las condiciones laborales del conjunto de la población española pero no sobre las ocupaciones de los individuos. Esta situación nos obliga a recurrir a la metodología de áreas pequeñas.

Cuando se habla de áreas pequeñas, existen dos familias de modelos que dependen de la disponibilidad de la información auxiliar. Encontramos modelos a nivel de unidad, aquellos en los que la información está disponible para todos los datos de la muestra, y modelos a nivel de área, aquellos en los que la información está agregada por áreas y no la tenemos disponible para los individuos. Los modelos a nivel de área son ampliamente utilizados debido a que no siempre se dispone de información a nivel de unidad para realizar las estimaciones, el más conocido es el modelo Fay-Herriot. No obstante, en nuestro caso tenemos disponible la información para todas las unidades muestreadas por lo que recurriremos a alguno de los modelos de unidad.

2.2.1. Modelo de errores anidados

El primer modelo que nos puede ser útil es el denominado *modelo de errores anidados*. Este modelo fue utilizado por Battese, Harter y Fuller (1988) en el contexto de la estimación de la media de superficie cultivada por condados en Iowa. Para ello recurrieron a información proporcionada por un satélite y por encuestas. Las encuestas fueron realizadas a los agricultores de las distintas zonas de cultivo y la estimación por satélite se realizó mediante el análisis de píxeles de superficie cultivada mostrados en las imágenes, se estima que cada píxel se corresponde con 0.45 hectáreas. Partiendo de esta situación, los autores supusieron que la información se podía relacionar a través del siguiente modelo:

$$y_{dj} = x'_{dj}\beta + v_d + e_{dj}, \quad d = 1, ..., D, \ j = 1, ..., n_d \tag{2.2}$$

donde y_{dj} es la característica de interés en la j-ésima unidad muestreada de la d área, con D siendo el número total de áreas muestreadas. x'_{dj} es un vector $k - dimensional$, siendo k el número de variables,

correspondiente a los valores auxiliares que, como se puede observar, aportan información disponible para todas las unidades muestreadas, por ello se considera que este modelo es un modelo a nivel de unidad. β es un vector $k - dimensional$ de parámetros desconocidos, v_d es el error que explicaría la diferencia entre las distintas áreas y que se supone independiente e identicamente distribuido a $\mathcal{N}(0, \sigma_v)$ y e_{dj} sería el componente de error del muestreo que se asume independiente e identicamente distribuido a $\mathcal{N}(0, \sigma_e)$. Cabe señalar también que v_d y e_{dj} son independientes entre sí. Bajo este modelo, como explica Pfeffermann (2013), la verdadera media de las áreas viene dada por $\bar{Y}_d = \bar{X_d}'\beta + v_d + \bar{\epsilon}_d$, sin embargo como para N_d grandes tenemos que $\bar{\epsilon}_d = \sum_{j=1}^{N_d} \epsilon_{dj}/N_d \cong 0$ entonces las medias suelen definirse como $\theta_d = \bar{X}_d'\beta + v_d = E(\bar{Y}_d|v_d)$

Se puede ver que este modelo es, en esencia, un modelo lineal mixto de intercepto aleatorio. El modelo se basa en suponer que cada condado presenta una componente de variabilidad añadida que provoca cambios en el origen de las rectas de regresión. Este modelo resulta muy intuitivo ya que valora la posibilidad de que cada condado presente en origen superficies cultivadas distintas.

Se pueden encontrar otros ejemplos de uso de los modelos mixtos de intercepto aleatorio fuera de las áreas pequeñas. Por ejemplo, en el año 1964, Potthoff y Roy realizaron un estudio con el objetivo de conocer las curvas de crecimiento en la distancia desde la hipófisis hasta la fisura pterigomaxilar (medida en milímetros mediante rayos X) en un grupo de niños de 8, 10, 12 y 14 años. El objetivo de los investigadores era valorar posibles diferencias en las curvas de desarrollo de los sujetos en función de su sexo. No obstante, un análisis descriptivo mostraba que, si se seleccionaba solo a los varones o a las mujeres para una representación de cómo evolucionaba el crecimiento, se observaba que las diferencias en el desarrollo a lo largo de los años no eran especialmente pronunciadas. Sin embargo, las diferencias iniciales sí que resultaban claras. Por ello, tenía sentido valorar la existencia de un efecto aleatorio asociado a los individuos que afectaba al intercepto del modelo mientras la pendiente se podía asumir igual para los distintos sujetos.

Este planteamiento puede resultar de utilidad en el contexto que nos atañe ya que, como se observó en el Capítulo 1, se observa que los individuos presentan salarios por hora diferentes en función de su ocupación. Esto puede llevarnos a pensar que sería adecuado tomar esta variable como un efecto aleatorio y evaluar la posibilidad de ajustar un modelo lineal en el que el intercepto de la recta sea distinto en función de la ocupación del individuo. De esta manera podríamos valorar la posibilidad de que, por ejemplo, los directivos partan de un salario por hora más elevado que los técnicos independientemente del resto de variables explicativas.

No obstante, existen críticas hacia el modelo (2.2). Estas se basan en que las rectas de regresión asignadas por el modelo a los distintos dominios son en la práctica paralelas ya que permiten al intercepto variar pero no permiten a la pendiente variar entre dominios (Hobza y Morales, 2012). Esta situación no parece razonable a la vista de las observaciones realizadas en el Capítulo 1. Se había observado que, por ejemplo, los años de antigüedad tienen un efecto distinto sobre la ganancia por hora dependiendo de la ocupación de los individuos. Esto nos indica que posiblemente la ocupación laboral esté afectando no solamente al origen de la recta sino también a la pendiente. Para estas circunstancias se ha planteado la posibilidad de usar un modelo ligeramente distinto al modelo (2.2). En adelante consideraremos a este modelo como nuestro "Modelo 1".

2.2.2. Modelo de coeficientes aleatorios

Este modelo fue planteado en el año 1981 por Dempser, Rubin y Tsutakawa. Estos autores propusieron un modelo más general basado en el modelo de errores anidados. Además, Hobza y Morales publican un trabajo en el 2012 donde explican que el modelo propuesto por Dempser et al (1981)

podría ser más adecuado en ciertas situaciones en las que la suposición de que la variación entre dominios solamente afecta al intercepto del modelo sigue pareciendo demasiado rígida a la vista de los datos. Esto es debido a que los dominios pequeños pueden estar afectando de un modo desconocido a la pendiente del modelo. Una mayor flexibilidad en el ajuste de los datos se conseguiría formulando un modelo donde los coeficientes ajustados β presentan una componente aleatoria y dando lugar al llamado modelo de coeficientes aleatorios. El modelo viene dado por:

$$y_{dj} = \sum_{k=0}^{p} \beta_k x_{kdj} + \sum_{k=0}^{p} v_{kd} x_{kdj} + e_{dj} \tag{2.3}$$

donde y_{dj} es la $j-$ésima observación del dominio d, x_{kdj} son las variables auxiliares y β_k son los parámetros desconocidos de la regresión. Además, los coeficientes aleatorios $v_{kd} \sim \mathcal{N}(0, \sigma_v)$ y los errores aleatorios $e_{dj} \sim \mathcal{N}(0, \sigma_e)$ son independientes entre sí, $d = 1, ..., D$, $j = 1, ..., n_d$, $k = 1, ..., p$, $v_{0d} \sim \mathcal{N}(0, \sigma_0)$ son independientes de e_{dj}'s, $E(v_{0d_1} v_{kd_2}) = 0$ si $d_1 \neq d_2$ y $E(v_{0d} v_{kd}) = \tau_k$, $d = 1, ..., D$, $k = 1, ..., p$. La correlación entre v_{0d} y la parte aleatoria del $k-$ésimo parámetro de la regresión, v_{kd}, dentro del área d es modelada por las medias de la covarianza τ_k. En nuestro caso, no parece necesario asumir correlación entre los coeficientes de regresión del modelo por lo que en el próximo capítulo ajustaremos un modelo restringido por $\tau_k = 0$, $k = 1, ..., p$. Otra forma de denotar este modelo sería mediante su versión matricial, dada por:

$$y = X\beta + \sum_{k=0}^{p} Z_k v_k + e \tag{2.4}$$

donde $n = \sum_{d=1}^{D}$, $\beta = \beta_{(p+1)\times 1}$, $y = col_{1 \leq d \leq D}(y_d)$, $y_d = col_{1 \leq j \leq n_d}(y_{dj})$, $e = col_{1 \leq d \leq D}(e_d)$, $e_d = col_{1 \leq j \leq n_d}(e_{dj})$, $v_k = col_{1 \leq d \leq D}(v_{kd})$, $X = col_{1 \leq d \leq D}(X_d)$, $X_d = col^t_{0 \leq k \leq p}(x_{k,n_d})$, $x_{kn_d} = col_{1 \leq j \leq n_d}(x_{kdj})$, $Z_k = diag_{1 \leq d \leq D}(x_{k,n_d})$.

Esto supone en la práctica la inclusión de una pendiente aleatoria dentro del modelo. Existe escasa bibliografía sobre la aplicación de este modelo en el campo de las áreas pequeñas. No obstante, los modelos lineales mixtos son ampliamente utilizados en estudios de datos longitudinales. Aquellos en los que una misma medida se repite en el tiempo para una misma unidad muestral. En estos casos es habitual suponer que dicha unidad muestral supone un efecto aleatorio sobre la variable respuesta. En Diggle et al. (2002), por ejemplo, encontramos mediciones del crecimiento de una especie de abetos que han estado expuestos a distintos niveles de ozono. Un análisis de los datos muestra que, tomando como variable respuesta el tamaño del árbol, se observa que aunque el ozono afecta de la misma manera en un principio al tamaño de los árboles, su desarrollo a lo largo de los días de exposición es diverso. Por ello se valora un modelo con efectos aleatorios a la pendiente de la recta y no al origen. Otro ejemplo lo encontramos en Belenky et al (2003), aquí los autores estudiaban los tiempos de reacción de un grupo de soldados sometidos a privación del sueño durante varios días. Un análisis descriptivo de la relación entre el tiempo de reacción y el tiempo de estudio transcurrido mostraba que cada soldado presentaba, ya originalmente, un tiempo de reacción distinto. Esto podía llevar a plantear que el individuo era un factor aleatorio y que tenía un efecto aleatorio sobre el intercepto de la recta que relaciona el tiempo de reacción con el tiempo transcurrido en privación del sueño. No obstante, también se observaba que la pendiente de la recta era distinta para cada caso, habiendo soldados cuyos tiempos de reacción presentaban mayores o menores reducciones para los mismos tiempos de privación del sueño. Por ello, se considera que un modo adecuado para analizar la relación entre el tiempo de reacción y los días transcurridos en privación del sueño sería ajustar un modelo lineal mixto con un factor aleatorio (los individuos) y un efecto aleatorio que influye tanto al intercepto como a los coeficientes del modelo.

Estos estudios presentan similitudes con nuestro caso. En el Capítulo 1 se observó que, como mínimo, los años de antigüedad presentan pendientes distintas en función de la ocupación del individuo. De esta forma se veía que, por ejemplo, la pendiente de los directivos era más pronunciada. Ello implica que en esta categoría no solo se parte de un salario más elevado (intercepto distinto) sino que también se presenta una lógica de acumulación del capital diversa en la que el salario por hora aumenta más rápido que en otras categorías a medida que aumentas la antigüedad. Teniendo esto en cuenta, parece razonable pensar que la ocupación está afectando, no solo a la variable respuesta, sino también a las variables explicativas y resulta intuitivo valorar el modelo de pendientes aleatorias como modelo a ajustar. Así podríamos valorar el siguiente modelo:

$$y_{dj} = \beta_d x_{dj} + e_{dj},$$
$$j = 1, ..., n_d,\ d = 1, ..., D$$

donde y_{dj} sería el salario por hora del individuo j perteneciente a la ocupación d, con $d = 1, ..., 7$. x_{dj} sería un vector de variables explicativas medidas en el individuo $j-$ésimo perteneciente a la ocupación d. β_d sería un vector de coeficientes ajustados para la ocupación d y e_{dj} sería como en el modelo de errores anidados un componente de error de muestreo. Teniendo en cuenta que β_d se puede descomponer en $\beta + v_d$, siendo v_d el componente de aleatoriedad derivado de pertenecer a la ocupación d y que se puede asumir independiente e idénticamente distribuido conforme a $\mathcal{N}(0, \sigma_v)$, se puede decir que el objetivo del modelo es estimar β y σ_v^2.

Entre las principales ventajas de este modelo está el que, por lo menos a priori, parece que es capaz de ajustar los datos conforme a lo que se observa en la realidad. En cuestiones de índole económica las variables suelen estar interrelacionadas y la suposición de que existe una única recta de regresión que modelice la relación entre el salario por hora y las demás variables parece poco razonable. Sin embargo, es necesario tener en cuenta que como desventaja el modelo reviste una mayor complejidad que dificulta su ajuste para conjuntos de datos grandes desde un punto de vista computacional y, además, el nivel de interpretabilidad del modelo se reduce en comparación con un modelo de regresión múltiple. Por este motivo, se hace necesario valorar su rendimiento sobre los datos y valorar si la mejora en el ajuste (en caso de existir) justifica la aplicación de un modelo de estas características. En la próxima sección haremos referencia a este modelo como el "Modelo 2".

2.3. Estimación de los modelos

En esta sección procederemos a realizar una explicación de la manera de estimar los parámetros del modelo y el parámetro objetivo, en nuestro caso un promedio.

2.3.1. Modelo 0

Este modelo se detalló en la Sección 2.1 veáse la ecuación (2.1). El objetivo es estimar $\hat{\beta}_0$ y $\hat{\beta}_1$ para los parámetros β_0 y β_1 que nos permiten obtener la recta de regresión dada por:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

que mejor se ajuste a los datos que tenemos disponibles.

Partiendo de este punto sabemos que la diferencia entre cada observación y_j y su estimación $\hat{y}_j$ dada por:

$e_j = y_j - \hat{y}_j$

se conoce como residuo y nuestro objetivo será minimizar la suma de los cuadrados de dichos residuos, es decir:

$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 = \sum_{j=1}^{n} (y_j - (\hat{\beta}_0 + \hat{\beta}_1 x))^2$

partiendo de esto, y gracias al cálculo diferencial podemos obtener que:

$\hat{\beta}_1 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{Cov(x,y)}{Var(x)}$

una vez obtenido el coeficiente $\hat{\beta}_1$ podemos obtener $\hat{\beta}_0$ de la siguiente manera:

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$

Para más información puede consultarse Sheater (2009) o Badiella y Sánchez (2011).

2.3.2. Modelo 1

El modelo dado por (2.2) puede escribirse también de la siguiente forma:

$$Y = X\beta + Zv + e$$

donde Y es el vector de observaciones de la variable objetivo, β es un vector de coeficientes de las variables explicativas o efectos fijos, $v \sim \mathcal{N}(0, \sum_v)$ es el vector de coeficientes aleatorios, que tiene una matriz de covarianzas $\sum_v = \sigma_u^2 I_D$, donde $I_D = diag(1, ..., 1)_{D \times D}$, X es una matriz de incidencia con elementos conocidos, $Z = diag(1_{N_d}; d = 1, ..., D)$ es una matriz diagonal y $e \sim \mathcal{N}(0, \sum_e)$ es el vector de errores aleatorios. Bajo esta notación la varianza viene dada por

$$Var(Y) = Z\Sigma_v Z^t + \Sigma_e = V.$$

En el proceso de inferencia tenemos una muestra de una población determinada de modo que $n < N$. Dejemos a s denotar al conjunto de unidades de dicha población que han sido seleccionadas en la muestra y a r denotar al conjunto de unidades no seleccionadas para la muestra. Las restricciones para el área d de n, N, s y S son n_d, N_d, s_d y S_d respectivamente. Teniendo esto en cuenta, tenemos que

$$V = \begin{pmatrix} V_s & V_{sr} \\ V_{rs} & V_r \end{pmatrix}$$

donde $V_s = Z_s \Sigma_v Z_s^t + \Sigma_{es}$, $V_r = Z_r \Sigma_v Z_r^t + \Sigma_{er}$ y $V_{rs} = Z_r \Sigma_{vr} Z_s^t$.

En base a esto podemos obtener una estimación para los coeficientes β y para el componente v. El mejor estimador empírico insesgado (BLUE) de β vendría dado por

$$\hat{\beta}_B = (X_s^t V_s^{-1} X_s)^{-1} X_s^t V_s^{-1} Y_s$$

mientras que el BLUP de v vendría dado por

$$\hat{v}_B = \sigma_v^2 Z_s^t V_s^{-1} (Y_s - X_s \hat{\beta}_B)$$

Los desarrollos para este modelo pueden consultarse en González-Manteiga et al (2008), Rao y Molina (2015) y Morales et al (2021).

2.3.3. Modelo 2

El modelo 2 viene dado por (2.3) y, como se comentó anteriormente, el objetivo es obtener estimaciones para los coeficientes β y para σ_v^2. La variabilidad del modelo viene dada por:

$$V = var(y) = V_e + \sum_{k=0}^{p} Z_k V_{vk} Z_k^t + \sum_{k=1}^{p} Z_0 V_{0k} Z_k^t + \sum_{k=1}^{p} Z_k V_{k0} Z_0^t = diag_{1 \leq d \leq}(V_d)$$

donde V_e es la matriz de covarianzas de los vectores de errores aleatorios e_d, V_{v_k} es la matriz de covarianzas para los vectores de regresores aleatorios v_k con $k = 0, 1, ..., p$, $Z_k = diag_{1 \leq d \leq D}(x_{k,n_d})$, $V_{0k} = cov(v_0, v_k) = E(v_0 v_k^t) = \tau_k I_D$ y:

$$V_d = \sigma_e^2 + \sum_{k=0}^{p} \sigma_k^2 x_{k,n_d} x_{k,n_d}^t + \sum_{k=1}^{p} \tau_k x_{0,n_d} x_{k,n_d}^t + \sum_{k=1}^{p} \tau_k x_{k,n_d} x_{0,n_d}^t, \quad d = 1, ..., D.$$

Para el ajuste del modelo es útil considerar los parámetros alternativos

$$\sigma^2 = \sigma_e^2, \quad \varphi_k = \frac{\sigma_k^2}{\sigma_e^2}, \quad k = 0, 1, ..., p, \quad \emptyset_k = \frac{\tau_k}{\sigma_e^2}, \quad k = 1, ...p.$$

Dejemos a $\varphi = (\sigma^2, \varphi_0, \varphi_1, ..., \varphi_p, \emptyset_1, ..., \emptyset_p)$ ser el vector de componentes de la varianza con $\sigma^2 > 0$, $\varphi_0 > 0$ $\varphi_k > 0$ y $\emptyset_k \in \mathcal{R}$ para $k = 1, ...p$. Denotemos $v = col_{0 \leq k \leq p}(v_k)$, la correspondiente matriz de covarianzas

$$V_v = var(v) = \begin{pmatrix} V_{v_0} & col^t_{1 \leq k \leq p}(V_{0k}) \\ col_{1 \leq k \leq p}(V_{0k}) & diag_{1 \leq k \leq p}(V_{vk}) \end{pmatrix}$$

Si φ es conocida, entonces el mejor estimador lineal insesgado (BLUE) para $\beta = (\beta_0, \beta_1, ..., \beta_p)^t$ y el mejor predictor lineal insesgado (BLUP) para v vienen dados por

$$\hat{\beta} = (X^t V^{-1} X)^{-1} X^t V^{-1} y \quad \text{y} \quad \hat{v} = V_v Z^t V^{-1} (y - X\hat{\beta}).$$

Como en la práctica no es posible conocer φ entonces se recurre al llamado "mejor estimador lineal insesgado empírico" (EBLUE) y al "mejor predictor lineal insesgado empírico" (EBLUP). Para su obtención solamente es necesario sustituir en la expresión anterior los componentes de la varianza con sus estimadores, de forma que la expresión quedaría de la siguiente manera:

$$\hat{\beta} = (X^t \hat{V}^{-1} X)^{-1} X^t \hat{V}^{-1} y \quad \text{y} \quad \hat{v} = \hat{V}_v Z^t \hat{V}^{-1} (y - X\hat{\beta}).$$

2.4. Predicción de la media

Ahora se procede a indicar como calcular el parámetro de interés. Como se explicó en secciones previas disponemos de una población N. De esta población se extrajo una muestra n con n_d elementos en el área d, $n = \sum_{d=1}^{D} n_d$. Entonces, sin pérdidad de generalidad, podemos reordenar la población del siguiente modo $y = (y_s^t, y_r^t)$, donde y_s constituye un vector de n elementos observados mientras y_r es un vector de tamaño $N - n$ de elementos no observados. En lo que sigue, y al igual que en la sección anterior, se utilizará el índice s para los elementos de la muestra y el índice r para el resto de la población. Teniendo en cuenta esta notación podemos escribir la matriz de covarianzas V de la siguiente manera.

$$V = var(y] = \begin{pmatrix} V_{ss} & V_{sr} \\ V_{rs} & V_{rr} \end{pmatrix}$$

donde $V_{ss} = var(y_s)$ es la matriz de covarianzas de los elementos observados, $V_{rr} = var(y_r)$ es la matriz de covarianzas de los elementos no observados y $v_{rs} = cov(y_r, y_s)$.

Estamos interesados en la estimación (en base a la muestra y_s) de la media $\bar{Y}_d$ del dominio d que puede ser expresada de la siguiente manera

$$\bar{Y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj} = a_s^t y_s + a_r^t y_r$$

donde $a^t = (1/N_d)(\mathbf{0}_{N_1}^t, \ldots, \mathbf{0}_{N_{d-1}}^t, \mathbf{1}_{N_d}^t, \mathbf{0}_{N_{d+1}}^t, \ldots, \mathbf{0}_{N_D}^t)$, $\mathbf{0}_m^t = (0, \ldots, 0)_{1\times m}$ y $\mathbf{1}_m^t = (1, \ldots, 1)_{1\times m}$. Además,

$$a_s^t = \frac{1}{N_D}(\mathbf{0}_{n_1}^t, \ldots, \mathbf{0}_{n_{d-1}}^t, \mathbf{1}_{n_d}^t, \mathbf{0}_{n_{d+1}}^t, \ldots, \mathbf{0}_{n_D}^t)$$

y

$$a_r^t = \frac{1}{N_D}(\mathbf{0}_{N_1-n_1}^t, \ldots, \mathbf{0}_{N_{d-1}-n_{d-1}}^t, \mathbf{1}_{N_d-n_d}^t, \mathbf{0}_{N_{d+1}}^t, \ldots, \mathbf{0}_{N_D-n_D}^t)$$

Como $\bar{Y}_d$ es un parámetro lineal, el predictor que minimiza el error cuadrático medio es el BLUP. Entonces, basándonos en las estimaciones obtenidas con anterioridad para los coeficientes β y para v podemos deducir que

$$\hat{\bar{Y}}_d^{blup} = a_s^t y_s + a_r^t [X_r \hat{\beta} + Z_r \hat{v}].$$

Como ya se comentó en la sección previa, cabe suponer que en la práctica no vamos a poder ajustar el BLUP por lo que generalmente se recurre al EBLUP. Solamente es necesario sustituir las varianzas de la expresión anterior por sus correspondientes estimadores. Para más información sobre estos desarrollos puede consultarse Hobza y Morales (2012) y Pfeffermann (2013). Hay que señalar que, en nuestro caso, no podemos calcular este estimador debido a que no disponemos de toda la información auxiliar en la población sino solo en la muestra. Por tanto, proponemos usar como estimador alternativo uno basado en la propuesta de Prasad y Rao (1990) que vendría dado por:

$$\hat{\bar{Y}}_d^{eblup} = a_s^t \; X_s \hat{\beta} + a_s^t \; Z_s \hat{v} \tag{2.5}$$

Capítulo 3

Análisis de Resultados

En este capítulo se va a aplicar la metodología expuesta en el Capítulo 2 a un caso de estudio con datos reales. El objetivo es estimar el salario medio por hora de la población española. Se va a recurrir a los datos recabados por el INE y que fueron presentados y analizados en el Capítulo 1. Nuestra variable objetivo será, por tanto, el *Salario por hora* y el resto de variables presentadas en el Cuadro 1.1 serán tomadas como variables explicativas. Para llevar a cabo este estudio se va a realizar un proceso iterativo en el que partiremos del modelo más simple, el dado por (2.1), que nos dará una primera aproximación de como se relacionan las variables y posteriormente aumentaremos en complejidad ajustando el modelo dado por la ecuación (2.2) y el modelo dado por (2.3). Para el ajuste de los modelos mixtos se tomará la variable *Ocupación laboral* como variable de efecto aleatorio. En el Capítulo 1 se observó que dicha variable podía estar afectando a las demás variables, por ejemplo en la Figura 1.6 y en la Figura 1.14 se observó que podía estar afectando a la variable *Antigüedad* y que podría ser razonable incluirla como efecto aleatorio.

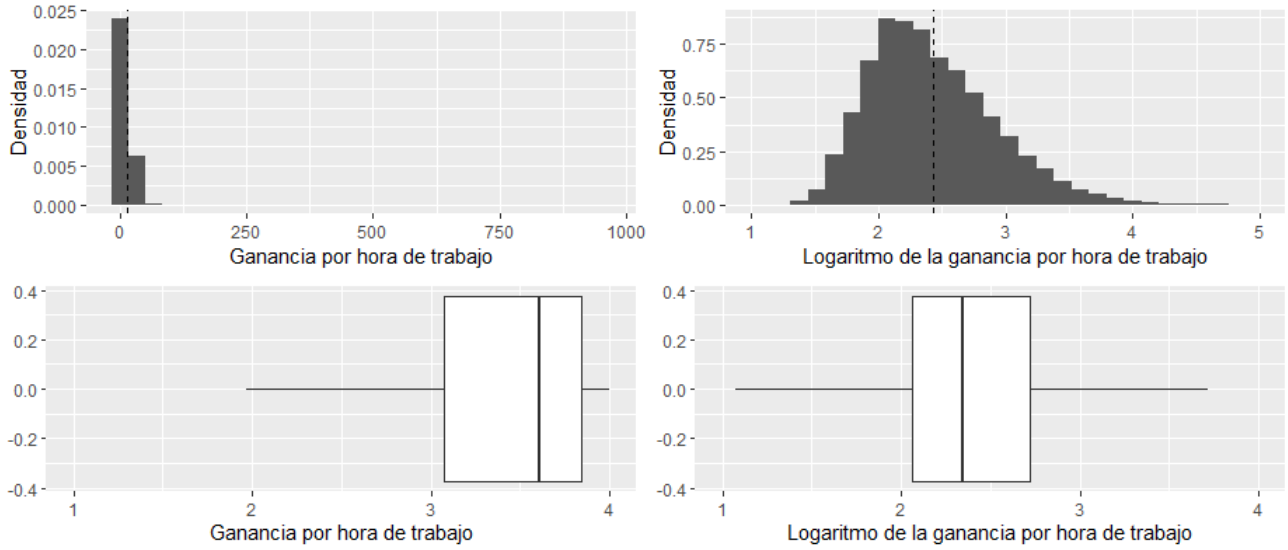
Para la realización de este capítulo se ha recurrido al paquete básico de R (*stats*) para el ajuste del modelo de regresión lineal múltiple de efectos fijos y a los paquetes *nlme* y *lme4* que permiten trabajar con modelos lineales mixtos de una forma más cómoda. Además, se usará el paquete *lmerTest* que complementa al paquete *lme4* y permite obtener de un modo sencillo, entre otras cosas, los niveles de significación de los coeficientes ajustados.

3.1. Transformación de la variable respuesta y las variables explicativas

En el Capítulo 1 se observó que la variable respuesta presenta una fuerte asimetría positiva y una elevada curtosis. Para un modelo de regresión se hace necesario intentar corregir esta situación y conseguir la normalidad de la variable sobre la que se desea hacer predicción. En estos casos lo más habitual es proceder a su transformación mediante la aplicación del logaritmo neperiano y valorar si con esto se consigue la normalidad deseada. Se ha aplicado este procedimiento y los resultados parecen satisfactorios. En la Figura 3.1 se presentan la variable original y los resultados de la transformación. Parece que el problema de la asimetría ya se ha corregido en gran medida si bien aún se observa una ligera asimetría positiva. En el boxplot para el logaritmo de la ganancia por hora de trabajo se ve con claridad que aún permanece una ligera asimetría pero que el problema parece resuelto en su mayor

parte. Además, se ha decidido aplicar el test de Lilliefors para contrastar la normalidad de la variable y se ha obtenido que, aún con la transformación logarítmica, la variable sigue sin presentar normalidad, hay que tener en cuenta que se está trabajando con unos 220000 datos lo cuál puede afectar a la potencia del test. No obstante, a nivel gráfico sí que se observa una distribución acampanada que se corresponde más o menos correctamente con la normalidad. Teniendo esto en cuenta, se ha decidido trabajar con la variable transformada y por tanto los modelos se ajustarán tomando el logaritmo de la ganancia por hora de trabajo como variable respuesta. Esta situación debe tenerse en cuenta a la hora de interpretar los resultados obtenidos al ajustar los modelos.

Figura 3.1: Resultados de la transformación de la variable respuesta.



Además, a la hora de ajustar los modelos se ha detectado que algunas variables presentan categorías con coeficientes no significativos. Por ello, y con el fin de posibilitar la posterior comparación entre modelos para elegir el que mejor ajuste los datos se ha decidido agrupar dichas categorías. Las categorías agrupadas son la categoría Canarias de la variable *Zona Geográfica* que pasará a unirse a la categoría Sur bajo el nombre de "Sur y Canarias" y las categorías Actividades sanitarias y Otros servicios de la variable *Sector de Actividad económica* que se unirán a la categoría actividades administrativas bajo el nombre de "Actividades administrativas, sanitarias y otros". Durante el ajuste de los modelos algunas variables se tomarán como variables de referencia y no tendrán coeficiente asociado en los cuadros de resultados ya que formarán parte del caso base. Se presenta en el Cuadro 3.1 un resumen de dichas variables para facilitar la posterior interpretación de los modelos presentados.

Cuadro 3.1: Tabla resumen categorías de referencia de las variables categóricas.

Variable	Categoría de referencia
Zona Geográfica	Noroeste
Sector de Actividad Económica	Industria Manufacturera
Titularidad	Pública
Mercado	Local/Regional
Sexo	Hombre
Nacionalidad	Española
Categoría laboral	Directores y Gerentes
Nivel de Estudios	Primarios
Edad	De 16 a 29
Tipo de Jornada	Tiempo Completo
Tipo de Contrato	Temporal

3.2. Modelo de regresión lineal múltiple (Modelo 0)

En primer lugar, se va a proceder a ajustar un modelo de regresión lineal que contemple la inclusión de todas las variables tenidas en cuenta. Son un total de 13 debido a la necesidad de descartar como variables explicativas el *Salario Mensual* y el *Número de horas trabajadas* ya que fueron utilizadas para construir la variable *Salario por hora* y no tendría sentido su inclusión. A continuación, se presentan en el Cuadro 3.2 los resultados obtenidos. El modelo resulta significativo y es capaz de explicar un 47 % de la varianza. Además, todas las variables resultan significativas para los niveles de significación habituales.

En este caso la interpretación del modelo se hace teniendo en cuenta que los coeficientes nos indican las diferencias observadas por pasar de la categoría de referencia a las otras. Así, por ejemplo, se observa que pasar de trabajar en una empresa pública a una privada (manteniéndose iguales el resto de categorías) supone un descenso en el salario por hora, mientras que pasar de la categoría estudios primarios a estudios secundarios supone un aumento. Una vez seleccionado el modelo se procede a hacer una diagnosis para comprobar si se respetan los supuestos básicos de linealidad, homocedasticidad, normalidad e independencia de los residuos y, de esta manera ver si es correcta su aplicación.

Con respecto a la hipótesis de linealidad y homocedasticidad en la Figura 3.2 se presentan los residuos del modelo frente a los valores ajustados. Se comprueba que la nube de puntos no presenta ninguna tendencia por lo que cabe deducir que no existen problemas con estas hipótesis. Además, cabe señalar que si el supuesto de independencia se violara se observaría algún patrón en este gráfico por lo que podemos suponer que no existen problemas tampoco con dicho supuesto.

En la Figura 3.3 se observa un gráfico cuantil-cuantil que nos muestra que la hipótesis de normalidad falla debido a la presencia de atípicos en las colas de la distribución. Si nos fijáramos solamente en el sector central de dicha distribución los puntos parecen mantenerse bastante pegados a la línea teórica y la normalidad es perfectamente asumible. Para comprobar esta suposición se ha realizado un histograma de los residuos que se presenta también en la Figura 3.3 y que nos sirve para valorar

Cuadro 3.2: Resultados ajuste del Modelo 0.

Variable	Categoría	Coeficiente	t-valor	p-valor
Intercepto		2.767	369.930	0
Zona	Noreste	0.097	31.690	0
	Madrid	0.108	34.715	0
	Centro	-0.014	-4.525	0
	Este	0.083	29.592	0
	Sur y Canarias	0.017	5.820	0
Sector	Construccion	-0.029	-10.936	0
	Hostelería y Comercio	0.016	4.321	0
	Transporte	-0.017	-5.403	0
	Información/Comunicaciones/Finacieras	0.044	11.055	0
	Actividades científicas y Educación	0.035	10.580	0
	Actividades administrativas, sanitarias y otros	-0.081	-24.474	0
Titularidad	Privada	-0.152	-58.327	0
Mercado	Nacional	0.040	20.964	0
	Europeo	0.115	32.232	0
	Mundial	0.160	57.514	0
Sexo	Mujer	-0.147	-83.797	0
Nacionalidad	Resto	0.029	8.342	0
Categoria	Técnicos	-0.453	-96.959	0
	Empleados oficina	-0.702	139.095	0
	Trabajadores servicios	-0.732	138.249	0
	Trabajadores cualificados	-0.700	132.621	0
	Operadores maquinaria	-0.669	120.981	0
	Trabajadores no cualificados	-0.774	144.289	0
Nivel de Estudios	Secundaria	0.046	19.619	0
	Superiores	0.253	86.833	0
Antigüedad		0.012	123.991	0
Tipo de Jornada	Parcial	-0.040	-17.856	0
Contrato	Temporal	-0.023	-10.885	0
Edad	de 30 a 39	0.094	31.099	0
	de 40 a 49	0.141	46.932	0
	de 50 a 59	0.145	43.445	0
	más de 59	0.139	31.640	0

la distribución de los residuos del modelo. Se ve que la distribución presenta una ligera asimetría positiva. Sin embargo, si nos centramos solo en el sector central de la distribución se observa una forma claramente acampanada cuya normalidad es asumible.

Figura 3.2: Gráfico de residuos del Modelo 0

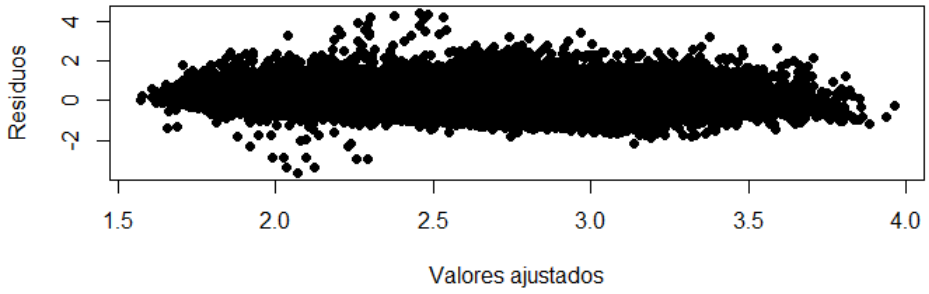
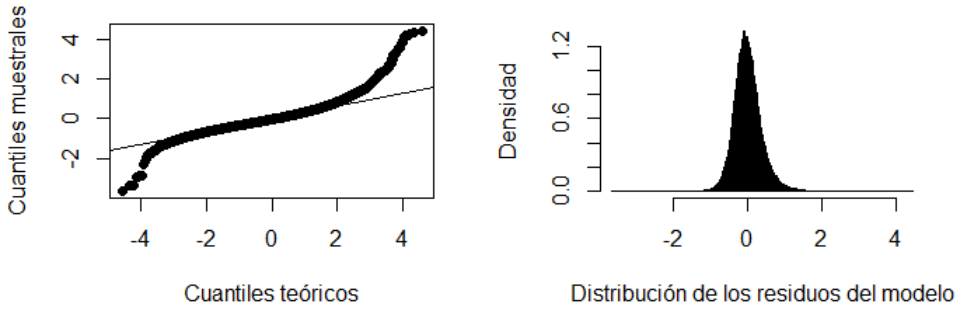


Figura 3.3: Gráfico QQ e histograma de los residuos (Modelo 0)



3.3. Modelo lineal con efecto aleatorio (Modelo 1)

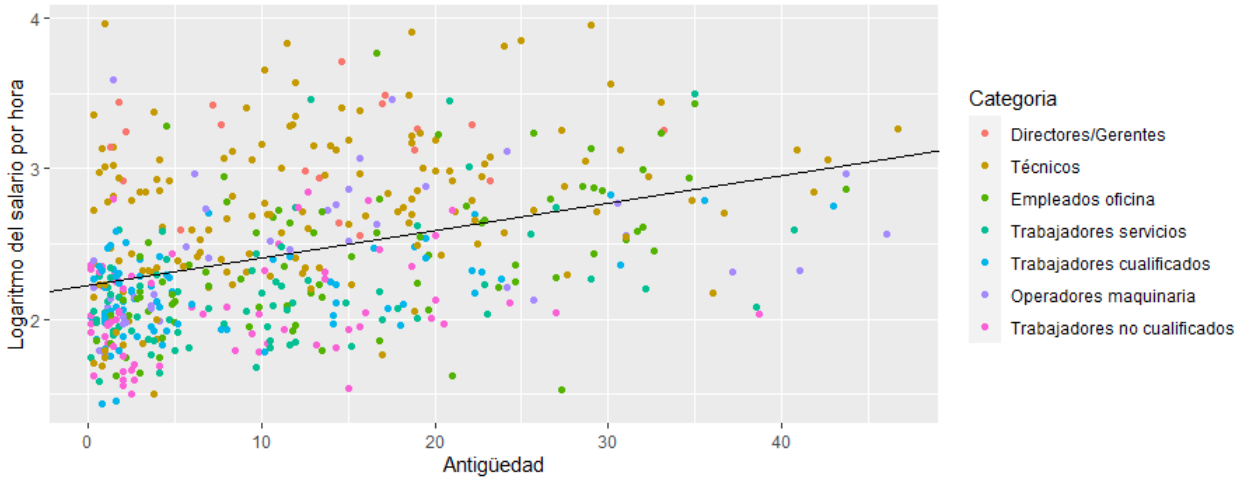
Se ha observado que la ocupación de los individuos se puede incluir como efecto aleatorio en el modelo. Esto nos indicaría que existe un componente de variabilidad asociado a la pertenencia a una u otra ocupación. En el Capítulo 1 se valoró que, por ejemplo, la distribución del salario por hora varía dependiendo de la categoría laboral en la que se encuentren englobados los individuos, de forma que un directivo parte de un salario por hora superior. Esto nos llevó a pensar que podría ser correcto valorar un modelo que tome la ocupación como efecto aleatorio y cuantifique la variabilidad extra que la pertenencia a dicha categoría incluye en el modelo.

Con el fin de entender mejor el funcionamiento de este modelo pensemos en una situación menos

compleja. Si quisiésemos ajustar un modelo de regresión lineal simple que intentase estimar los valores del logaritmo del salario medio por hora en base a los años de antigüedad de los individuos el resultado sería una recta de regresión como la que se observa en la Figura 3.4 (para la mejor visualización de los datos se ha seleccionado una muestra aleatoria de 500 casos del conjunto de datos original). Como se puede ver, se trata de una única recta que debe ajustar los valores de todos los individuos independientemente de su ocupación. En general, la recta parece ajustar bien los datos en su conjunto. Sin embargo, hay observaciones para las que el error de predicción es muy elevado, sobre todo entre los directores y gerentes. En la Figura 3.5 se ajusta un modelo lineal mixto de la forma dada por la ecuación (2.2) que toma la ocupación laboral de los individuos como un efecto aleatorio que afecta al origen de la recta. Se puede observar como cada ocupación laboral presenta una recta de regresión propia, todas ellas paralelas entre sí. Además, se ha añadido la recta ajustada con el modelo de regresión lineal simple con una línea discontinua para que se puedan comparar los dos ajustes.

En la Figura 3.6 se puede ver el mismo ajuste pero dividido por ocupación. De esta manera podemos valorar el ajuste de cada recta de regresión con respecto a las observaciones de cada categoría, una vez más, el ajuste de regresión lineal simple se representa con una línea discontinua.

Figura 3.4: Logaritmo del salario por hora frente a Años de Antigüedad (Modelo 0)



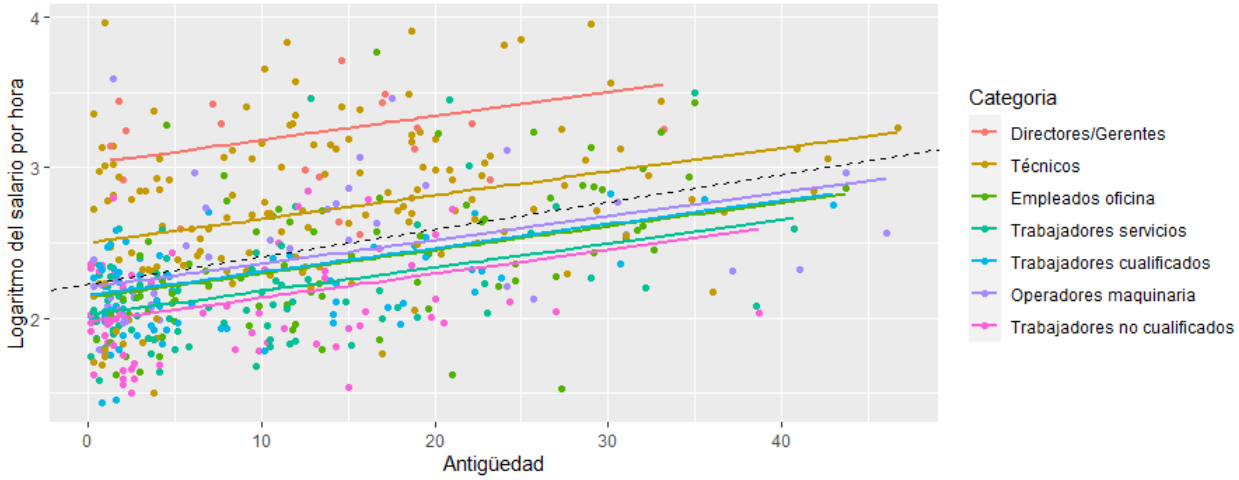
Como se puede ver, hay categorías laborales para las que el ajuste de regresión lineal simple parece bastante adecuado. Este es el caso de los trabajadores cualificados o los operadores de maquinaria. Sin embargo, para los directores y gerentes o los trabajadores cualificados parece que el modelo de intercepto aleatorio consigue un ajuste más realista de los datos.

Una vez explicado en mejor medida el modelo y presentado su funcionamiento de forma gráfica, se procede a ajustar este modelo teniendo en cuenta todas las variables explicativas utilizadas en la sección anterior. Los resultados para los efectos fijos del modelo se presentan en el Cuadro 3.3 y muestra que todas las categorías de las variables resultan ser significativas. En el Cuadro 3.5 se presentan las estimaciones para los interceptos aleatorios del modelo. Como se explicó en el Capítulo anterior dichos interceptos siguen una distribución $\mathcal{N}(0, \sigma_v)$, por tanto, estos valores son las estimaciones obtenidas para los valores de $\hat{v}$ indicados en la Sección 2.4. En el Cuadro 3.4 se presentan las estimaciones de la varianza de los efectos aleatorios tanto sobre el intercepto como sobre los residuos.

Cuadro 3.3: Resultados ajuste del Modelo 1.

Variable	Categoría	Coeficiente	t-valor	p-valor
Intercepto		2.1910	21.122	0
Zona	Noreste	0.0979	31.690	0
	Madrid	0.1087	34.716	0
	Centro	-0.0148	-4.525	0
	Este	0.0830	29.593	0
	Sur y Canarias	0.0175	5.820	0
Sector	Construcción	-0.0292	-10.935	0
	Hostelería y Comercio	0.0160	4.322	0
	Transporte	-0.0175	-5.403	0
	Información/Comunicaciones/Finacieras	0.0445	11.056	0
	Actividades científicas y Educación	0.0358	10.585	0
	Actividades administrativas, sanitarias y otros	-0.0815	-24.472	0
Titularidad	Privada	-0.1522	-58.326	0
Mercado	Nacional	0.0403	20.964	0
	Europeo	0.1154	32.232	0
	Mundial	0.1600	57.515	0
Sexo	Mujer	-0.1477	-83.798	0
Nacionalidad	Resto	0.0297	8.342	0
Nivel de Estudios	Secundaria	0.0466	19.622	0
	Superiores	0.2537	86.845	0
Antigüedad		0.0127	123.992	0
Tipo de Jornada	Parcial	-0.0403	-17.858	0
Contrato	Temporal	-0.0240	-10.886	0
Edad	de 30 a 39	0.0944	31.099	0
	de 40 a 49	0.1418	46.933	0
	de 50 a 59	0.1453	43.446	0
	más de 59	0.1395	31.641	0

Figura 3.5: Logaritmo del salario por hora frente a Años de Antigüedad (Modelo 1)



Cuadro 3.4: Estimación de la variabilidad de los efectos aleatorios.

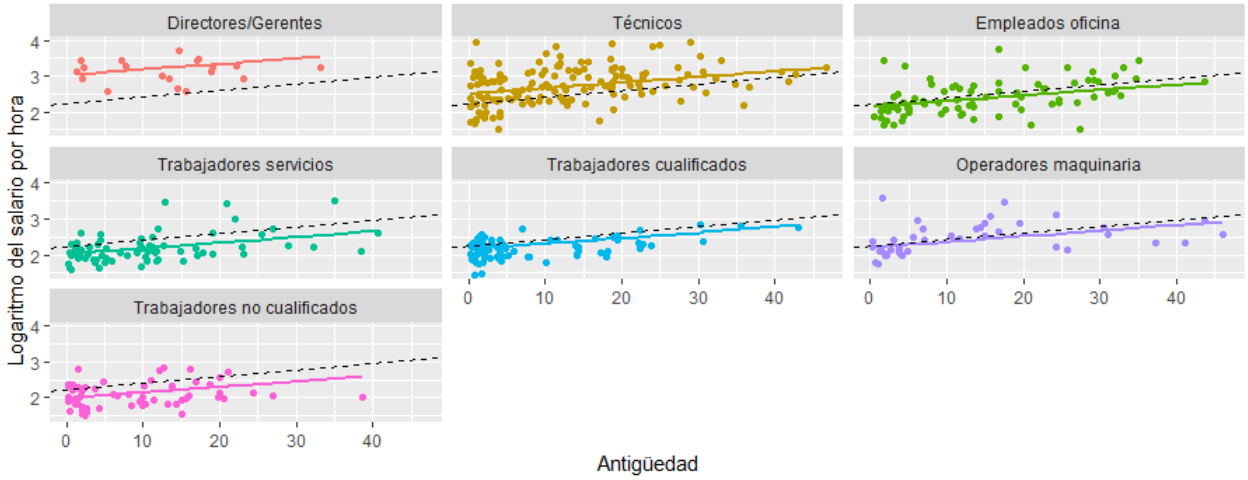
Grupo	Varianza
Intercepto	0.07512
Residuos	0.13556

Cuadro 3.5: Estimación de los interceptos aleatorios.

Ocupación	Intercepto aleatorio
Directores/Gerentes	0.576
Técnicos	0.122
Empleados oficina	-0.126
Trabajadores servicios	-0.155
Trabajadores cualificados	-0.123
Operadores maquinaria	-0.093
Trabajadores no cualificados	-0.198

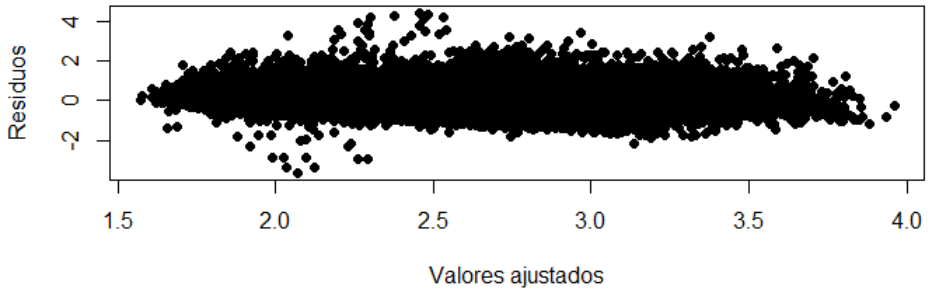
Ahora conviene realizar una diagnosis del modelo para comprobar que no hay problemas de incumplimiento de los supuestos de los modelos lineales. Para ello se puede recurrir al gráfico de residuos frente a valores ajustados y al gráfico QQ para valorar la normalidad. En la Figura 3.7 se observa que no existen patrones en la nube de puntos lo que nos indica que no existen problemas de heterocedasticidad o dependencia. En relación con el supuesto de normalidad, se observa una vez más que dicho supuesto falla debido a la presencia de observaciones con valores extremos en las colas de la

Figura 3.6: Ajuste del Modelo 1 (visualización en rejilla)



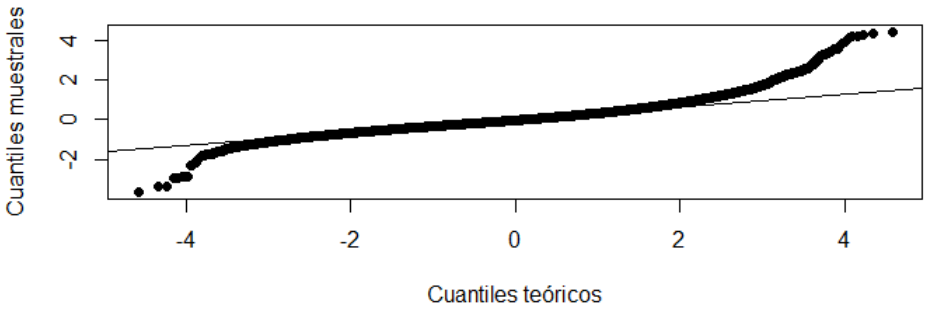
distribución. En la Figura 3.8 se ve que los puntos se ajustan bien a la línea teórica en el centro de la distribución mientras que falla en las colas.

Figura 3.7: Residuos frente a valores ajustados (Modelo 1)



Una vez comprobado que el modelo cumple los supuestos básicos y es válido se procede a comprobar si la inclusión del efecto aleatorio es relevante para el análisis o, por el contrario, el modelo de regresión lineal múltiple de efectos fijos resulta suficiente para explicar la relación entre variables. Para llevar a cabo esta comprobación se va a recurrir a la función *ranova* propia del paquete *lme4*. Dicha función recibe por parámetro un modelo lineal mixto y comprueba cada efecto aleatorio mediante un proceso de reducción. Es decir, en nuestro caso ajustaría un modelo de regresión lineal sin el intercepto aleatorio y lo comparará con el modelo de intercepto aleatorio. El resultado será una tabla de tipo ANOVA en la que nos indicará si debemos aceptar el modelo más simple o por el contrario el modelo lineal mixto aporta una mayor información. Los resultados se presentan en el Cuadro 3.6. La hipótesis nula sobre la que trabaja esta función es que un modelo es significativamente peor que el otro, por tanto, un p-valor

Figura 3.8: Gráfico cuantil-cuantil de los residuos (Modelo 1)



inferior a los niveles de significación habituales nos indica que dicho modelo es significativamente peor que el otro y se debe rechazar su uso. En este caso, debemos rechazar el uso del modelo lineal de efectos fijos ya que su nivel de significación nos indica que es peor que el de intercepto aleatorio.

Cuadro 3.6: Resultados comparación Modelo 0 y Modelo 1

Efecto aleatorio	Verosimilitud	AIC	p-valor
Sin efecto aleatorio	-104875	209806	0
Ocupación	-91131	182320	

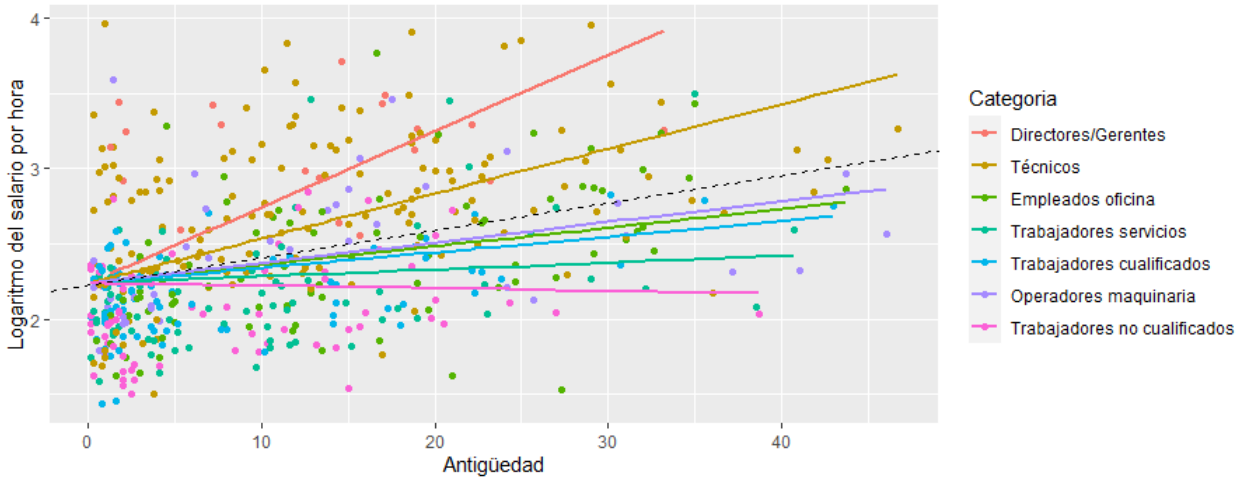
3.4. Modelo de pendientes aleatorias (Modelo 2)

En esta sección se pretende ajustar un modelo de pendientes aleatorias que nos permita representar de un mejor modo aquello que sucede en la realidad. Como se comentó previamente, un modelo que toma la ocupación del individuo como efecto aleatorio supone una mejora en relación con un modelo lineal estándar. No obstante, en cuestiones de índole socioeconómica parece aún demasiado rígido para representar correctamente la realidad. En relación con esto, se comprobó en el Capítulo 1 que, por ejemplo, la variable *Antigüedad* presenta evoluciones distintas en función de la ocupación que tiene el individuo. Esto nos llevó a plantear la teoría de que podía ser adecuado ajustar dicha variable con un efecto aleatorio que provoca cambios en su pendiente. En relación con el resto de variables, cabe pensar que pueden verse afectadas también por la ocupación del individuo. De esta manera, si miramos una vez más, por ejemplo, la Figura 1.3 vemos que la distribución del salario por hora es distinta en función del nivel de estudios que se tiene y de la ocupación a la que se dedica el individuo. Se observa, por ejemplo, que las diferencias entre los empleados de oficina son relativamente pequeñas mientras que entre los directivos o los técnicos son más destacables. Esta situación nos puede hacer sospechar que esta variable también se está viendo afectada por la ocupación del individuo.

Con el fin de que el modelo propuesto se entienda mejor se va a proceder igual que en la sección anterior. Se va a ajustar un modelo de regresión lineal simple que explique el salario por hora en

función de los años de antigüedad de los individuos y después se va a realizar un ajuste que tome la ocupación de los individuos como efecto aleatorio. Sin embargo, en este caso se va a suponer que el intercepto es igual para todas las ocupaciones y que la variación se produce en la pendiente de las rectas. Los resultados se presentan en la Figura 3.9.

Figura 3.9: Logaritmo del salario por hora frente a Años de Antigüedad (Modelo 2)



Para una visualización del ajuste de forma independiente con respecto a las observaciones de cada categoría se presenta la Figura 3.10. Podemos observar que, por ejemplo, los directores y los gerentes presentan una pendiente mucho más pronunciada que las demás ocupaciones. Además, los directores, los técnicos, los trabajadores de servicios y los trabajadores no cualificados son los que presentan unas mayores diferencias con respecto al ajuste realizado sin tener en cuenta efectos aleatorios. A la vista de los datos, parece razonable pensar que el ajuste realizado por el modelo de pendiente aleatoria es más realista ya que la variabilidad a mayores aportada por la ocupación laboral del individuo le permite ajustar en mejor medida el salario por hora para valores elevados de la variable *Antigüedad*.

Una vez presentado el modelo para un ajuste con una única variable explicativa se procede a realizar el ajuste tomando todas las variables de la base de datos utilizadas hasta el momento. La variable *Antigüedad* será modelada como una variable con un efecto aleatorio derivado de la pertenencia a una u otra ocupación laboral mientras que el resto de variables serán tomadas como efectos fijos. Los resultados se presentan en el Cuadro 3.7 y se observa que todas las variables resultan significativas. En el Cuadro 3.8 se presentan los valores de las varianzas del efecto aleatorio que supone la *Categoría laboral* sobre la *Antigüedad* y la varianza de los residuos. Por último, se presenta también el Cuadro 3.9 que contiene las estimaciones de las pendientes aleatorias del modelo. Al igual que en el caso anterior, estos valores son las estimaciones obtenidas para los valores de $\hat{v}$ indicados en la sección 2.4.

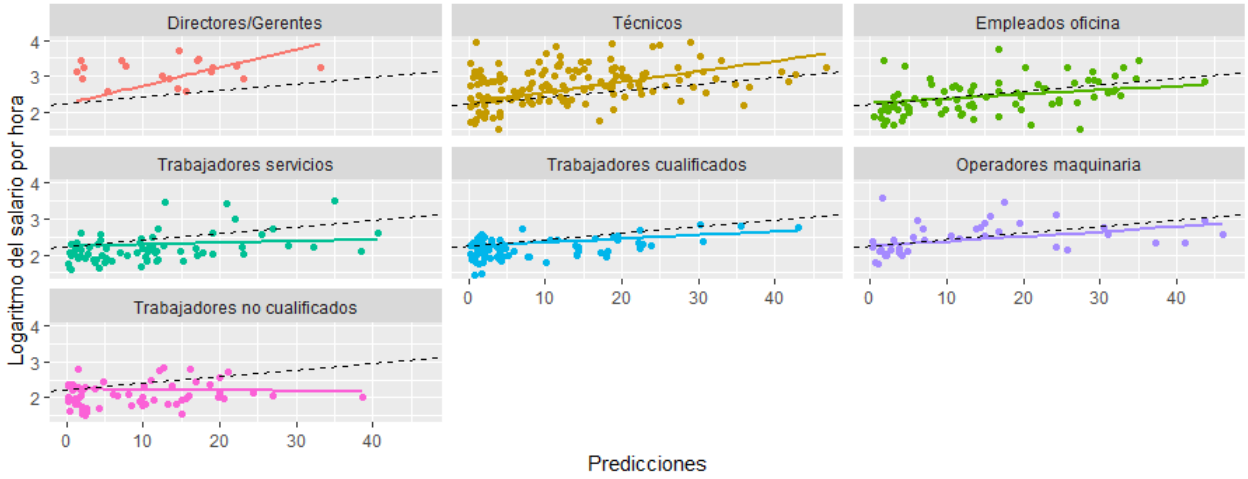
Cuadro 3.8: Estimación de las varianzas del efecto aleatorio (Modelo 2).

Grupo	Varianza
Categoria | Antigüedad	0.0001421
Residuos	0.1419090

Cuadro 3.7: Resultados ajuste del Modelo 2.

Variable	Categoría	Coeficiente	t-valor	p-valor
Intercepto		2.1150	377.484	0
Zona	Noreste	0.0976	30.907	0
	Madrid	0.1201	37.541	0
	Centro	-0.0162	-4.862	0
	Este	0.0884	30.823	0
	Sur y Canarias	0.0159	5.178	0
Sector	Construcción	-0.0252	-9.507	0
	Hostelería y Comercio	0.0112	2.965	0
	Transporte	-0.0288	-9.025	0
	Información/Comunicaciones/Finacieras	0.0382	9.390	0
	Actividades científicas y Educación	0.0663	19.590	0
	Actividades administrativas, sanitarias y otros	-0.0434	-13.101	0
Titularidad	Privada	-0.1585	-59.306	0
Mercado	Nacional	0.0421	21.404	0
	Europeo	0.1232	33.638	0
	Mundial	0.1714	60.311	0
Sexo	Mujer	-0.1562	-88.183	0
Nacionalidad	Resto	0.0126	3.478	0
Nivel de Estudios	Secundaria	0.0559	23.235	0
	Superiores	0.3339	120.270	0
Antigüedad		0.0139	3.077	0.02
Tipo de Jornada	Parcial	-0.0485	-21.124	0
Contrato	Temporal	-0.0413	-18.327	0
Edad	de 30 a 39	0.1036	33.373	0
	de 40 a 49	0.1479	47.895	0
	de 50 a 59	0.1459	42.681	0
	más de 59	0.1484	32.927	0

Figura 3.10: Ajuste del Modelo 2 (visualización en rejilla)



Cuadro 3.9: Estimación de las pendientes aleatorias.

Categoría	Pendientes aleatorias
Directores/Gerentes	0.0245
Técnicos	0.0058
Empleados oficina	-0.0045
Trabajadores servicios	-0.0067
Trabajadores cualificados	-0.0056
Operadores maquinaria	-0.0030
Trabajadores no cualificados	-0.0104

Una vez ajustado el modelo podemos realizar una diagnosis para valorar si se cumplen los supuestos básicos. En la Figura 3.11 se observa que no existen patrones en los residuos frente a los valores ajustados con lo cual no parece haber problemas de heterocedasticidad o dependencia de las observaciones. Además, en la Figura 3.12 se presenta un gráfico cuantil-cuantil que muestra que la hipótesis de normalidad falla debido a los atípicos presentes en las colas de la distribución.

Al igual que en el caso anterior, conviene comparar el modelo obtenido mediante la inclusión de las pendientes aleatorias con el modelo de regresión lineal múltiple para ver si en este caso también se puede suponer que el ajuste es mejor. Una vez más, se recurre a la función *ranova* para llevar a cabo el cálculo. Los resultados se presentan en el Cuadro 3.10. Los resultados muestran que el ajuste con el efecto aleatorio es más adecuado para explicar la relación entre las variables.

Figura 3.11: Residuos frente a valores ajustados (Modelo 2)

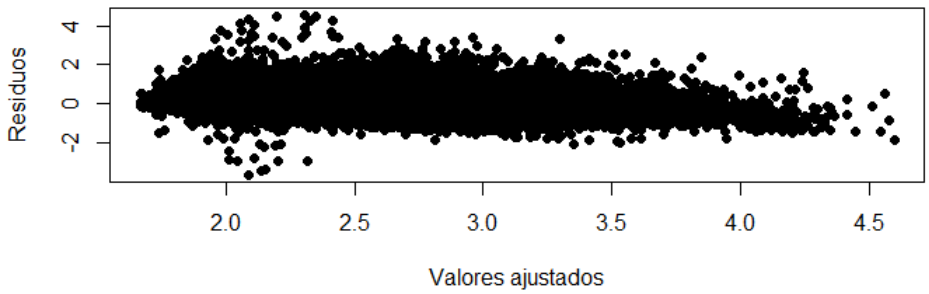
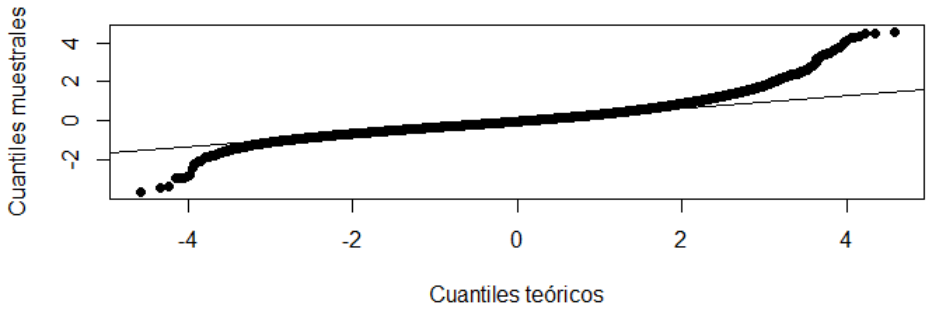


Figura 3.12: Gráfico cuantil-cuantil de los residuos (Modelo 2)



Cuadro 3.10: Resultados comparación Modelo 0 y Modelo 2

Efecto aleatorio	Verosimilitud	AIC	p-valor
Sin efecto aleatorio	-104875	209806	0
Antigüedad | Categoría	-96088	192234	

En este caso no podemos comparar los modelos 1 y 2 mediante el uso de la función *ranova* ya que uno no supone una reducción del otro. Son dos modelos que parten de suposiciones distintas acerca de la relación de las variables y como afecta la variable *Categoría laboral* al resto. En un caso suponemos que solamente afecta al origen de la recta y en el otro suponemos que afecta a la pendiente de la recta. En los dos casos dicha suposición resulta más adecuada que suponer un modelo de regresión lineal con todas las variables tomadas como efectos fijos. Sin embargo, parece que los dos modelos aportan una

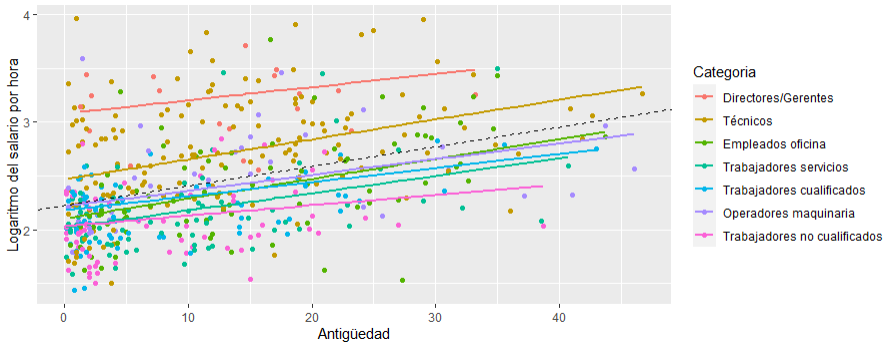
mayor información que el modelo lineal estándar pero aún parecen demasiado rígidos para explicar la realidad de los datos socioeconómicos que estamos analizando.

3.5. Modelo de intercepto y pendientes aleatorias (Modelo 2.1)

Si nos fijamos en las Figuras 3.5 y 3.9 vemos que, en el primer caso, parece que el intercepto aleatorio ayuda a ajustar mejor los datos en el origen de la recta mientras que para valores altos de la variable *Antigüedad* parece ser demasiado rígido. En el segundo, por el contrario, vemos que el modelo es capaz de ajustar bien los datos para valores altos de la variable *Antigüedad* pero que falla al suponer un mismo origen para las rectas. A la vista de esta situación se propone ajustar un nuevo modelo basado en (2.3) suponiendo que el efecto aleatorio dado por la ocupación de los individuos afecta tanto a la pendiente como al intercepto del modelo.

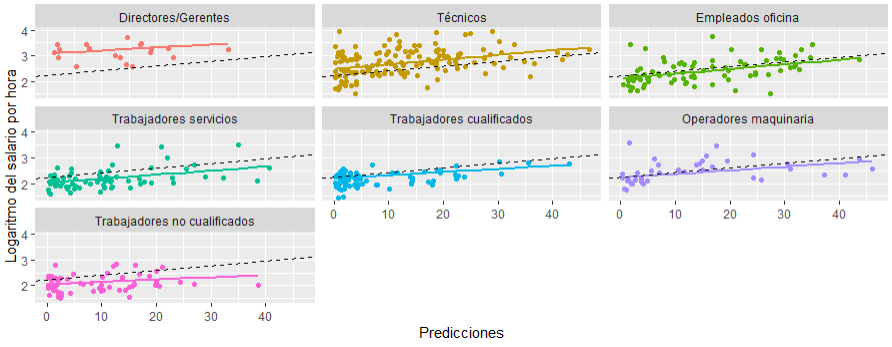
Al igual que en secciones anteriores se va a presentar el ajuste del modelo de forma gráfica con un ejemplo más sencillo, suponiendo solo un modelo que explique el salario por hora en función de los años de antigüedad. En la Figura 3.13 se puede ver el ajuste del modelo. Se puede ver que cada recta parte de un intercepto distinto, lo cual permite un ajuste mejor a los datos en el origen de la recta y al mismo tiempo se observan variaciones en las pendientes de las distintas rectas, sobre todo para valores altos de la variable *Antigüedad*. De esta manera conseguimos lo que parece una aproximación más realista a la relación entre las dos variables. En la Figura 3.14 se observa el ajuste de cada recta en relación con las observaciones de cada ocupación laboral.

Figura 3.13: Logaritmo del salario por hora frente a Años de Antigüedad (Modelo 2.1)



Una vez presentado el modelo para un ajuste con una única variable explicativa se procede a realizar el ajuste tomando todas las variables de la base de datos utilizadas hasta el momento. La variable *Antigüedad* será modelada como una variable con un efecto aleatorio derivado de la pertenencia a una u otra ocupación laboral mientras que el resto de variables serán tomadas como efectos fijos. Los resultados se presentan en el Cuadro 3.11. Se puede ver que, en este caso, todos los coeficientes han resultado ser significativos para cualquiera de los niveles de significación de referencia habituales. En el Cuadro 3.12 se presentan las estimaciones de la varianza asociadas a los coeficientes aleatorios introducidos en el modelo. En este caso, tenemos una estimación de la varianza asociada a la pendiente del modelo, una estimación de la varianza asociada al intercepto del modelo y otra asociada a los residuos. Por último, en el Cuadro 3.13 se presentan las estimaciones para los coeficientes aleatorios.

Figura 3.14: Ajuste del Modelo 2.1 (visualización en rejilla)



Al igual que en secciones anteriores, estos valores se corresponden con los valores $\hat{v}$ indicados en la Sección 2.4.

Cuadro 3.12: Estimación de los componentes de la varianza de los efectos aleatorios (Modelo 2.1).

Grupo	Varianza
Intercepto	0.06647
Categoria | Antigüedad	0.00001
Residuos	0.13480

Cuadro 3.13: Estimación de los coeficientes aleatorios.

Categoría	Intercepto aleatorio	Pendientes aleatorias
Directores/Gerentes	0.5833	-0.00022
Técnicos	0.0786	0.00367
Empleados oficina	-0.1615	0.00297
Trabajadores servicios	-0.1443	-0.00094
Trabajadores cualificados	-0.1015	-0.00173
Operadores maquinaria	-0.0921	0.00014
Trabajadores no cualificados	-0.1622	-0.00389

Al igual que en las secciones previas, se procede a realizar una diagnosis del modelo para valorar si se respetan los supuestos básicos de los modelos lineales. En la Figura 3.15 se presentan los residuos del modelo ajustado. Como se puede ver, no existen patrones en la nube de puntos que pudiesen indicar posibles problemas de heterocedasticidad o dependencia de las observaciones. Además, en la Figura 3.16 se observa que, al igual que en casos anteriores, la normalidad de los residuos es asumible a excepción de en las colas de la distribución.

Cuadro 3.11: Resultados ajuste del Modelo 2.1.

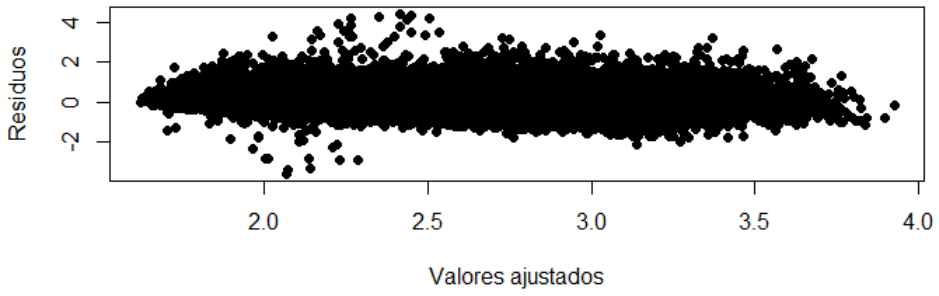
Variable	Categoría	Coeficiente	t-valor	p-valor
Intercepto		2.198	22.522	0
Zona	Noreste	0.098	31.907	0
	Madrid	0.110	35.240	0
	Centro	-0.015	-4.610	0
	Este	0.083	29.787	0
	Sur y Canarias	0.016	5.614	0
Sector	Construcción	-0.027	-10.456	0
	Hostelería y Comercio	0.009	2.623	0
	Transporte	-0.019	-5.880	0
	Información/Comunicaciones/Finacieras	0.044	10.979	0
	Actividades cientificas y Educación	0.034	10.072	0
	Actividades administrativas, sanitarias y otros	-0.075	-22.781	0
Titularidad	Privada	-0.148	-56.834	0
Mercado	Nacional	0.040	21.227	0
	Europeo	0.118	33.141	0
	Mundial	0.161	58.295	0
Sexo	Mujer	-0.146	-83.428	0
Nacionalidad	Resto	0.023	6.490	0
Nivel de Estudios	Secundaria	0.045	19.087	0
	Superiores	0.254	87.406	0
Antigüedad		0.011	11.806	0
Tipo de Jornada	Parcial	-0.038	-16.890	0
Contrato	Temporal	-0.028	-12.911	0
Edad	de 30 a 39	0.094	31.280	0
	de 40 a 49	0.139	45.991	0
	de 50 a 59	0.140	42.013	0
	más de 59	0.136	31.002	0

Una vez valorado que no existen problemas de incumplimiento de los supuestos básicos y el modelo es adecuado se procede a utilizar la función *ranova* para contrastar si el modelo es mejor que uno que solamente tenga en cuenta un efecto aleatorio o, por el contrario, es adecuado asumir que la variable *Categoría laboral* está afectando tanto al intercepto como a la variable *Antigüedad* y, por tanto, es correcto utilizar un modelo lineal mixto de coeficientes aleatorios. Los resultados de la prueba se presentan en el Cuadro 3.14. El resultado del test muestra que hay evidencias para afirmar que el modelo de intercepto y pendiente aleatoria es significativamente mejor que un modelo que solo comtemple la inclusión de un intercepto o una pendiente aleatoria por separado.

Cuadro 3.14: Resultados comparación Modelo 1 y 2 contra Modelo 2.1.

Efecto aleatorio	Verosimilitud	AIC	p-valor
Intercepto o pendiente aleatorias	-91131	182320	0
Intercepto y pendiente aleatoria	-90574	181211	

Figura 3.15: Residuos frente a valores ajustados (Modelo 2.1)

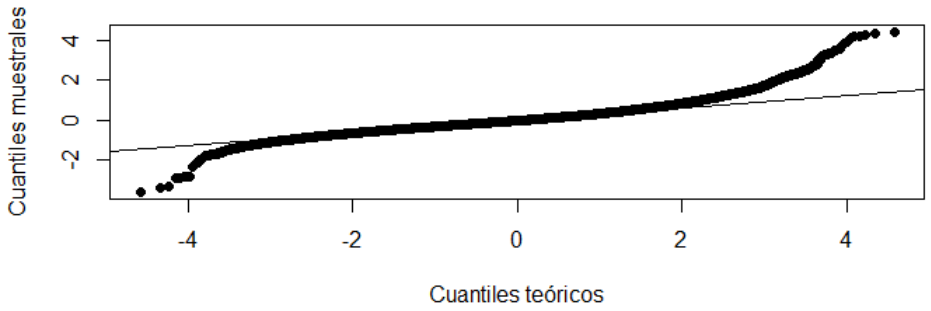


Una vez obtenido el modelo final y con los resultados presentados en los Cuadros 3.11 y 3.13 se puede obtener la estimación de la media para cada área pequeña (categoría laboral) basándonos en la ecuación (2.5).

Finalmente, se analizarán brevemente los resultados que nos proporciona el modelo sobre la relación entre nuestra variable objetivo y las covariables. No obstante, es necesario recordar que los coeficientes proporcionados en todos los cuadros y figuras son para el logaritmo de la variable *Salario por hora* ya que su transformación fue necesaria para posibilitar los análisis. Los efectos fijos del modelo nos indican que:

- Todas las zonas geográficas, a excepción de la zona centro, suponen un aumento de la ganancia por hora con respecto a la categoría de referencia (Noroeste). Este resultado era hasta cierto punto esperable y concuerda con los sistemas productivos y económicos presentes en el país. Destaca claramente el coeficiente asociado a Madrid que es el área donde se produce una mayor ganancia por hora manteniendo iguales el resto de covariables.

Figura 3.16: Gráfico cuantil-cuantil de los residuos (Modelo 2.1)



- Con respecto a la variable *Sector* se observa que el sector en el que hay una mayor ganancia por hora es el de la información, las comunicaciones y las actividades financieras.
- Se observa que el paso de una empresa pública a una privada supone una reducción de la ganancia por hora.
- En relación con el tipo de mercado observamos que, cuanto mayor es el tamaño del mercado en el que opera la empresa, mayor es la ganancia por hora de sus trabajadores.
- El coeficiente asociado a la variable *Sexo* nos muestra que, manteniéndose igual el resto de variables, las mujeres presentan una ganancia por hora menor que los hombres.
- El contrato temporal y la jornada parcial también suponen menores ganancias por hora para los trabajadores que se encuentran en estas condiciones.
- En relación con la edad, se observa que el grupo que presenta una mayor ganancia por hora es el grupo de personas entre 50 y 59 años.

Los resultados obtenidos para los efectos fijos parecen razonables y no parece haber motivo para pensar en posibles fallos del modelo. En relación con los efectos aleatorios observamos que las variaciones en el intercepto del modelo son más pronunciadas que en las pendientes. En los interceptos observamos que los directores y gerentes, así como los técnicos, presentan mayores ganancias en origen que el resto de categorías. Es llamativo el hecho de que, en el caso de los directores y gerentes, a medida que avanzan los años y ganan antigüedad su pendiente es decreciente e inversamente proporcional mientras que en el caso de los técnicos hay una relación creciente. Una vez más, los resultados del modelo parecen corroborar lo observado durante el análisis exploratorio de los datos llevado a cabo en el Capítulo 1 y no parece haber motivos para dudar de buen ajuste a los datos.

Capítulo 4

Conclusiones

Durante el presente documento se ha tratado de dar respuesta a la necesidad de realizar estimaciones sobre un parámetro de interés en aquellos casos en los que, o bien el tamaño muestral es insuficiente, o bien el hecho de que originalmente la muestra no estuviera pensada para proporcionar dicha estimación, nos impide utilizar metodologías más comunes para la obtención de resultados lo suficientemente precisos y rigurosos. La metodología para lograr este objetivo fue la metodología de áreas pequeñas (SAE) que recurre a la utilización de modelos lineales mixtos para mejorar las estimaciones en aquellos dominios o áreas para los cuáles los métodos de estimación tradicionales no producirían buenos resultados.

Cuando comenzamos la elaboración del trabajo nos encontramos con que los modelos lineales mixtos son ampliamente utilizados en el campo de los datos longitudinales o en estudios de carácter médico pero existe poca bibliografía sobre su uso en otros campos. El planteamiento del que partimos se basaba en la creencia de que, si consideramos que los individuos son lo suficientemente diversos en sí mismos como para que sea necesaria la inclusión de elementos de variabilidad extra en los modelos lineales estándar, parece razonable pensar que en escenarios de carácter social y económico también podría resultar de interés. Al fin y al cabo, cada país o región tiene características económicas y poblacionales que la hacen distinta de las de su entorno, en ocasiones tanto que no parece que pertenezcan a la misma zona económica. Sin embargo, a pesar del interés inherente que tendría realizar un análisis de las diferencias en diversos indicadores económicos que existen entre países o incluso entre comunidades autónomas finalmente nos decidimos por un problema similar pero ligeramente diferente. Intentar estimar el salario por hora de la población española en función de su ocupación laboral. Durante el Capítulo 1 se observó que variables como el sexo o la ocupación laboral implican diferencias en lo que los individuos ganan por una hora de trabajo, esta situación hacía fácil pensar que la probabilidad de que la ocupación laboral o el sexo tengan algún tipo de efecto aleatorio sobre la ganancia o incluso sobre alguna de las covariables fuera perfectamente razonable. En la decisión final de usar la ocupación laboral como área pequeña en lugar del género influyó el hecho de que ya existiera un ejemplo de uso de esta metodología para realizar una estimación de la brecha salarial, fue realizado por Lombardía, López-Vizcaíno y Rueda en 2020 y a él se remite al lector por su interés en relación con esta metodología y por los resultados obtenidos y presentados en dicho trabajo.

Una vez escogido el posible efecto aleatorio la siguiente cuestión a responder era que modelo utilizar. Nuestros datos contaban con información desagregada a nivel de individuo, es decir, teníamos disponible toda la información para todas las unidades muestrales, eso descartaba aquellos modelos a nivel de área como el Fay-Herriot y nos dejaba la opción de recurrir a un modelo de errores anidados

o a un modelo de coeficientes aleatorios. Teniendo en cuenta que el modelo de coeficientes aleatorios se puede combinar con el modelo de errores anidados, se decidió que era posible realizar un estudio polietápico que ajustara varios modelos, los comparara entre sí y valorara si el ajuste era mejor o peor al añadir efectos aleatorios. Por ello, se decidió partir de un modelo de regresión lineal múltiple con todos los efectos fijos e ir aumentando el número de efectos aleatorios incluídos en el modelo. De esta forma, se ajustaron dos modelos con un efecto aleatorio que afectara o bien al intercepto del modelo o bien a la pendiente y, finalmente, un modelo con un efecto aleatorio que afectara tanto a la pendiente como al intercepto.

Los resultados mostraron que el modelo de intercepto aleatorio supuso una mejora clara con respecto al modelo de regresión lineal múltiple con efectos fijos. Esto se vio tanto gráficamente, en el ajuste con una única variable, como en los resultados de la función *ranova* y en los índices AIC obtenidos. Esto indica que la inclusión de un efecto aleatorio en el origen de la recta mejora la capacidad del modelo para ajustar los datos y corrobora lo observado en el Capítulo 1 en el que se veía que los directivos partían de salarios por hora más elevados que los técnicos o los trabajadores cualificados aunque mantuviéramos igual otras variables como el nivel de estudios. Con respecto al modelo con un coeficiente aleatorio los resultados mostraron que era significativamente mejor que el modelo de regresión lineal múltiple estándar. Sin embargo, por lo menos en la parte del ajuste gráfico con una sola variable parecía que sí era capaz de ajustar mejor los datos en valores altos de antigüedad de los individuos mientras que parecía claramente insuficiente a la hora de pronosticar los orígenes de la recta. Posteriormente, el ajuste del modelo que valoraba la inclusión tanto de un intercepto como de una pendiente aleatoria demostró, tras la aplicación de la función *ranova*, ser significativamente mejor que los modelos que solo valoraban un tipo de coeficiente aleatorio. A la vista de estos resultados, se concluye que la metodología de áreas pequeñas supone una buena opción a la hora de mejorar las estimaciones para la ganancia por hora en función de la ocupación laboral.

A la luz de los resultados obtenidos cabe preguntarse por las futuras aplicaciones o líneas de investigación que se abren para la utilización de modelos de coeficientes aleatorios. Pero antes es necesario poner sobre la mesa algunas limitaciones que la utilización de este modelo conlleva:

- Reducción de la interpretabilidad de los resultados: una de las grandes ventajas de los modelos lineales es su sencillez de interpretación y es posible que esa sea una de las razones por las que son ampliamente utilizados en ámbitos muy diversos, no solo en investigación sino también en entornos empresariales. A pesar de que esta metodología nos permite mejorar los resultados obtenidos en la estimación del parámetro de interés su interpretación es más compleja y se hace menos intuitiva.

- Aumento de los costos computacionales: los modelos de regresión son modelos de muy fácil ajuste y no requieren de grandes prestaciones a nivel computacional. El paso de un modelo de regresión lineal múltiple a un modelo lineal mixto de estas carácterísticas puede aumentar el tiempo de ajuste de los modelos desde segundos a minutos dependiendo el número de efectos aleatorios a ajustar o del número de datos disponible.

- Dificultad para evaluar los modelos: como se explicó en el Capítulo 2 este tipo de modelos no se pueden evaluar mediante la máxima verosimilitud como los modelos lineales estándar. En su lugar es necesario recurrir a la máxima verosimilitud restringida (REML). Esto implica que ciertas funciones y paquetes de uso habitual en el software R no sean adecuados para este tipo de modelos.

- Escasez de herramientas: entroncando con el punto anterior, actualmente en R solamente hay disponibles dos paquetes que ofrecen posiblidades de ajustar este tipo de modelos (*nlme* y *lme4*). La falta de un uso más generalizado de estos paquetes así como la falta de herramientas similares

en otros lenguajes de uso común como Python hacen que sea difícil comenzar a manejar estos modelos más allá de un ámbito académico.

Como futuras líneas de investigación y trabajos a realizar se proponen los siguientes:

- Buscar una forma de deshacer la tranformación logarítmica de los modelos ajustados para obtener predicciones para la ganancia por hora que se pudieran interpretar directamente.
- Realizar el cálculo del MSE de los estimadores presentados. Dicho cálculo se podría realizar mediante desarrollos analíticos o mediante técnicas de remuestreo.
- Buscar nuevos campos de estudio y aplicación de los modelos de áreas pequeñas. Otros estudios de carácter económico o quizás estudios relacionados con el campo del marketing podrían ser opciones interesantes.

Bibliografía

[1] Aliaga, A. (2001). Métodos de estimación para áreas pequeñas y una aplicación a la prevalencia anticonceptiva[ponencia]. Conferencia conjunta IAOS-AFSA, Addis Abeba, Etiopía.

[2] Badiella, LL. y Sánchez, J.A. (2011). Modelos Mixtos con R. Barcelona: Servei d´Estatistica Aplicada UAB.

[3] Belenky, G. Wesensten, N. Thorne, D. Thomas, M. Sing, H. Redmond, D. Russo, M. y Balkin, T. (2003). Patterns of performance degradation and restoration during sleep restriction and subsequent recovery: a sleep dose-response study. Journal of Sleep Research 12, 1-12.

[4] Borrelli, F., Carbonetti, G., De Felice, L. y Solari, F. (2012). Metodologie di stima per piccole aree applicabili a variabili di censimento. Istituto Nazionale di Statistica. Istat working papers, 3.

[5] Diggle, P. Heagerty, P. Liang, K. y Zeger, S. (2002). Analysis of Longitudinal Data. 2ªed. UK: Oxford.

[6] González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008) Bootstrap mean squared error of a small-area EBLUP, Journal of Statistical Computation and Simulation, 78:5, 443 - 462.

[7] Lombardía, M., López-Vizcaíno, E. y Rueda, C. (2020). A new approach to gender pay gap decomposition by economic activity.

[8] Mancho, J. (2002). Técnicas de estimación en áreas pequeñas. Eustat.

[9] Morales, J., Esteban, M., Pérez, A. y Hobza, T. (2021). A Course on Small Area Estimation and Mixed Models. 1ªed. Springer.

[10] Münnich, R. Burgard, J.P., Ertz, F., Lenau, S. Manecke, J. y Merkle, H. (2019). Guidelines on small area estimation for city statistics and other functional geographies. Luxemburgo. Eurostat.

[11] Hobza, T. y Morales, D. (2012). Small area estimation under random regression coeficient models. Journal of Statistical Computation and Simulation, 83:11, 2160-2177. Recuperado de http://dx.doi.org/10.1080/00949655.2012.684094

[12] IGE. (2009). DOCUMENTOS TÉCNICOS. Estimación de áreas pequenas: o ingreso medio mensual por comarca nos fogares galegos.

[13] Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. Statistical Science, 28(1), 40-68, DOI: 10.1214/12-STS395

[14] Potthoff, R. y Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika, 51(3-4), 313.

[15] Prasad, N. G. N. y Rao, J. N. K. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. Journal of the American Statistical Association, 85(409).

[16] Rao, J. N. K. y Molina, I. (2015). Small Area Estimation. 2ªed. USA: Willey.

[17] Sheather, S.J. (2009). A modern approach to regression with R. Springer.