UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS (BILINGÜE)

TRABAJO DE FIN DE GRADO

TÍTULO: Estudio de la ganancia media de los trabajadores españoles

AUTOR: Joaquín Álvarez López

TUTOR/ES: David Casado de Lucas

CURSO ACADÉMICO: 2016/2017

CONVOCATORIA: Junio

ÍNDICE

1. **Introducción** 2
2. **Variables y datos** 3
3. **Explicación teórica** 5
4. **Análisis de los datos.** 9
   - **4.1. Análisis individual.** 9
   - **4.2. Análisis conjunto.** 10
5. **Resultados y conclusiones** 33
6. **Bibliografía** 34

1. Introducción

¿Alguna vez os habéis preguntado qué se esconde detrás la ganancia media de un individuo? Esta pregunta es la premisa entorno a la cual se elaborará este trabajo, en el que se pretende mostrar a través del estudio de datos reales de dos Comunidades Autónomas cuáles son los factores más relevantes a la hora de determinar cuál es el salario de una persona.

Actualmente, puede parecer que las carreras profesionales han dejado de ser tan previsibles en comparación a como lo fueron en su momento, pasando a tener una mayor importancia las relaciones laborales, gracias al desarrollo exponencial del "networking" de las personas debido a las nuevas tecnologías de la información y la comunicación (TIC). Además existen muchas creencias tales como que existe un brecha salarial en función del sexo a favor del hombre, que es necesario tener mayor edad y experiencia laboral para obtener una ganancia media mas elevada o que esta variará en función de trabajar en las principales ciudades Españolas o fuera de estas. Por lo tanto, este estudio, tratará de demostrar o desmentir estas teorías a la vez que se elabora un modelo, que pueda predecir e imputar los posibles valores ausentes de la ganancia media de los trabajadores españoles.

Para analizar el problema, existen diversas técnicas estadísticas mediante las cuales se puede llevar a cabo el estudio. Para este caso, se han ideado distintos modelos de regresión múltiple por los cuales se podrán identificar las variables más relevantes a la hora elaborar el sueldo medio de un individuo. Para ello estudiaremos la variable dependiente **Y**: Ganancia media en función de las siguientes variables independientes: **X1:** Experiencia, **X2:** Educación, **X3:** Sexo, **X4:** Comunidad autónoma

Los datos seleccionados para elaborar el trabajo se encuentran disponible en las encuestas de estructuras salariales del año 2014 llevadas a cabo por el Instituto Nacional de Estadística (INE). Con la adecuación de estos datos y su posterior análisis estadístico a través de programas estadísticos, se comparan las variables mediante la elaboración de distintos modelos de regresión con la finalidad de averiguar cuales son las factores con mayor poder decisivo a la hora determinar la ganancia media de un trabajador Español.

2. Variables y datos

En un análisis más detallado del proyecto, y como ya hemos comentado anteriormente, el objetivo del estudio consistirá en averiguar cuáles son las variables que más influyen a la ganancia media del mercado laboral Español. Esta, por lo tanto será nuestra variable dependiente (**Y:** Ganancia media). Pese a que en otros estudios es común el estudio de la mediana, pues es más estable ya que resulta menos afectada por aquellos datos considerados como extremos. Para este estudio se ha considerado el estudio de la media.

Para analizar esta variable, hemos clasificado las siguientes variables independientes extraídas de la encuesta sobre la estructura salarial de 2014 elaborada por el Instituto Nacional de Estadística (INE), en la cual se han proporcionado todas y cada una de las variables como subgrupos.

Cabe destacar, que estos datos no han sido obtenidos mediante la observación de los individuos, registrando, posteriormente los valores de cada variable independiente, en su lugar, el INE ha considerado varios individuos y ha calculado su ganancia media, y por lo tanto, estos datos corresponden a variables propias de un diseño experimental. Dichas variables serán las siguientes:

- **X1:** Experiencia, variable cuantitativa que ha sido agrupada en cada uno de los siguientes segmentos:
    1: Menos de 1 año de experiencia.
    2: Entre 1 y 3 años de experiencia.
    3: Entre 4 y 9 años de experiencia.
    4: Entre 11 y 20 años de experiencia.
    5: Entre 21 y 29 años de experiencia.
    6: Mas de 30 años de experiencia.

- **X2:** Educación, variable ordinal la cual se ha codificado en función de la formación académica de los distintos individuos:
    1: Educación primaria
    2: Primera etapa de educación secundaria
    3: Segunda etapa de educación secundaria

4: Enseñanza de formación profesional de grado superior y similares

5: Diplomados universitarios y similares

6: Licenciados y similares, y doctores universitarios

- **X3:** Sexo, variable nominal con distribución dicotómica que se ha se codificará de la siguiente manera:

  0: Masculino

  1: Femenino

- **X4:** Comunidad autónoma, variable nominal cuya codificación se dará de la siguiente manera:

  0: España

  1**:** Madrid

  2: Cataluña

  Para esta variable, han sido seleccionadas aquellas comunidades autónomas con mayor población: la comunidad autónoma de Madrid y Cataluña. Las cuales compararemos con la media nacional.

- **Variables Ficticias**: Con la finalidad de estudiar el nivel de educación en mayor detalle, se han añadido una variable ficticia para cada nivel de Educación (**Ed1,** …. **Ed6**), asimismo se han generado variables ficticias para cada nivel de experiencia (**menosde1** ….**masde30**), pues de otra manera estaríamos asumiendo que la diferencia entre cada uno de los distintos niveles de educación y años de experiencia es idéntico, es decir, se asumiría que las distancias entre los distintos niveles sean iguales. Para poder analizar mejor la diferencia entre las comunidades autónomas se han incluido dos variables ficticias adicionales **Mad** para la comunidad de Madrid y **Cat** para Cataluña con la finalidad de poder analizar las regresiones condicionadas por cada comunidad de manera individual.

3. Explicación teórica

Para llevar a cabo el estudio, primero debemos elaborar un modelo adecuado para poder estudiar correctamente las variables y los factores que determinan la ganancia media en el mercado laboral, teniendo en cuenta las influencias tanto positiva como negativamente en esta. Para esto, construiremos un modelo basado en la regresión lineal múltiple (RLM)

$$\mathbf{Y_j = \beta o + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \ldots\ldots + \beta_k x_{kj} + \varepsilon_j}$$

Con la finalidad de establecer una ecuación lineal estimada que describa y prediga la variable dependiente **Y** en función de k variables independientes observadas $\mathbf{x_j}$. Asimismo estudiaremos la variación marginal de la variable dependiente **Y** debida a las variaciones de las variables independientes **X**, que se estiman por medio de los coeficientes $\boldsymbol{\beta_1}$**.** Por ultimo ε_j, es el termino de error aleatorio, el cual posee un media 0 y varianza σ^2.

Para este modelo tendremos en cuenta los siguientes supuestos sobre la variable dependiente y para el error

Supuestos de la variable explicada Y:

- Dependencia lineal $\mathbf{E(Y/X=x_{ij}) = \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \ldots\ldots + \beta_k x_{kj}}$

supuestos de la variables explicativas X_j:

- Las variables explicativas se supone que toman valores fijos, es decir, no toman valores aleatorios.

  Independencia lineal de la variables explicativas, pues ninguna puede obtenerse como en función lineal de las demás o, equivalentemente, no existen c_i tales que $\mathbf{c_1 x_{1j} + c_2 x_{2j} + c_3 x_{3j} + \ldots\ldots + c_k x_{kj} = 0}$

Supuestos del termino del error aleatorio:

- $\mathbf{E(\varepsilon_j) = 0}$, para todo j = 1,2,3…n
- $\mathbf{Var(\varepsilon_j) = E(\varepsilon^2_j) - E(\varepsilon_j)^2 = E(\varepsilon^2_j) = \sigma^2}$, para todo j = 1,2,3…n todos los errores poseen la misma varianza $\boldsymbol{\sigma^2}$**,** es decir, el modelo presentaría homocedasticidad.

- **Cov ($\varepsilon_j\varepsilon_h$) = E($\varepsilon_j\varepsilon_h$) - E($\varepsilon_j$)E($\varepsilon_h$) = E($\varepsilon_j\varepsilon_h$) = 0**, para todo j,h= 1,2,3….n. errores no tienen correlación
- **$\varepsilon_j \sim N(0, \sigma^2)$,** normalidad del error, la cual permite aplicar inferencia estadística sobre los coeficientes.
- Puesto que los errores son independiente de cada variable explicativa: **E(ε_j/x_{ij}) = E(ε_j) = 0** para todo i,j= 1,2,3….n

Para poder estudiar el modelo correctamente, analizaremos la variabilidad total (STC), puede dividirse en dos componentes, uno SCR, la variabilidad explicada por la pendiente de la ecuación y otro SCE, la desviación sin explicar de los puntos respecto a la ecuación. Realizaremos un ajuste del modelo mediante el método de mínimos cuadrados ordinarios mediante el cual, minimizaremos la desviación sin explicar: la suma de los cuadrados de los residuos (SCE), siendo el error la diferencia entre y_i e $\overline{y}$ gráficamente se interpretaría como la distancia en vertical de cada error a la recta.

STC= SCR + SCE

STC = $\sum_i^n (y_j - \overline{y})^2$

SCR= $\sum_i^n (\overline{y}_j - \overline{y})^2$

SCE = $\sum_i^n e_j^2$

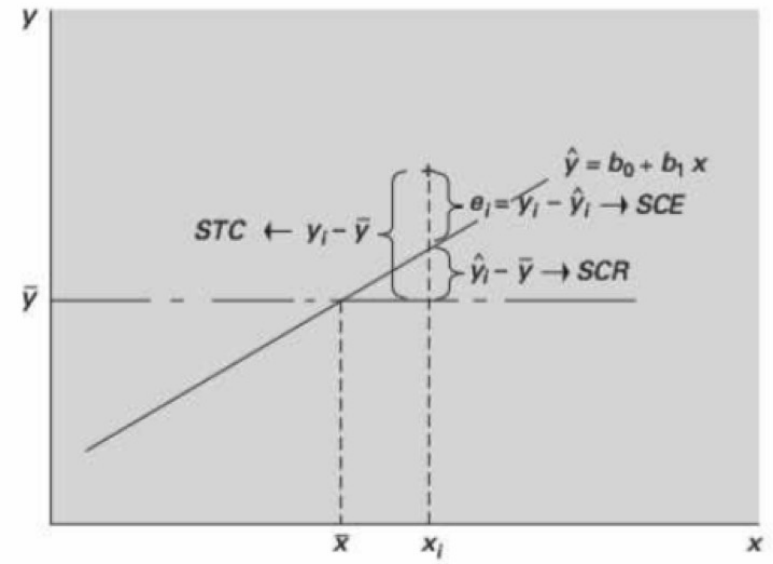


Figura 1: descomposición de la variabilidad, caso K=1, Fuente: Newbold

De esta manera, se consigue hallar el hiperplano que mejor represente un conjunto de puntos en el espacio.

Con la aplicación de proceso iterativo trataremos de encontrar un modelo sencillo y explicativo mediante una estrategia de selección de variables y un criterio de ajuste global para estudiar el conjunto de los coeficientes y variables independientes.

Analizaremos el ajuste de los términos de la expresión anteriormente mencionados mediante la diagnosis por tablas estadísticas, las cuales obtendremos mediante el uso de el software estadístico R y Gretl. De este modo, observaremos el ajuste global de cada coeficiente β_i de cada una de las variables explicativas $\mathbf{X_i}$ de nuestro modelo de regresión. Nos apoyaremos en distintas herramientas estadísticas para analizar el ajuste global y el ajuste estadístico Para cada uno de los modelos :

$$\mathbf{y_j = \beta o + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \ldots\ldots + \beta_k x_{kj} + \varepsilon_j}$$

Durante la diagnosis analizaremos:

A. Para cada uno de los términos del modelo:

a) El coeficiente estimado $\boldsymbol{b_j = \widehat{b}_i}$

b) error estándar de la estimación anterior

c) El p-valor (interpretado en último termino como el apoyo que los datos dan a la hipótesis nula), nos servirá para determinar la validez de las distintas variables para los distintos niveles de significación establecidos:

$$\mathbf{H_0:}\ \beta_j = \mathbf{0}$$

$$\mathbf{H_0:}\ \beta_j \neq \mathbf{0}$$

Se rechaza H_0 si

$$\left(\frac{b_j - 0}{S_{bj}}\right) > \mathbf{t_{n-k-1,\,\alpha/2}} \qquad \text{o} \qquad \left(\frac{b_j - 0}{S_{bj}}\right) < \mathbf{-\,t_{n-k-1,\,\alpha/2}}$$

B. Para el modelo global ajustado:

a) El coeficiente de determinación R^2 de la regresión es la proporción de la variabilidad maestral total explicada por la regresión

$$\mathbf{R^2} = \frac{SCR}{STC} = \mathbf{1} - \frac{SCR}{STC}$$

y se deduce que $0 \leq R^2 \leq 1$

b) El R^2 ajustado, que se empleará con la finalidad de tener en cuenta el hecho de que las variables independientes irrelevantes provocan una pequeña reducción de la suma de los cuadrados de los errores. Por lo que nos permitirá comparar mejor los modelos de regresión múltiple que tengan distinto número de variable, lo cual nos llevaría a incluir variables adiciónales de manera continua.

$$\overline{R}^2 = \frac{\frac{SCE}{n-K-1}}{\frac{STC}{n-1}}$$

c) El p-valor (interpretado en último termino como el apoyo que los datos dan a la hipótesis nula), es una herramienta para probar la validez del conjunto de variables para los distintos niveles de significación establecidos:

$$\mathbf{H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0}$$

$$\mathbf{H_1 :} \text{ al menos un } \mathbf{\beta_j \neq 0}$$

Se rechaza H_0 si $\frac{\frac{SCR}{K}}{S_e^2} > \mathbf{F_{k,n-k-1,\alpha}}$

Donde $F_{k,n-k-1,\alpha}$ es el valor crítico para el que $P(F_{k,n-k-1} > F_{k,n-k-1,\alpha}) = \alpha$

Finalmente, y una vez encontremos un buen modelo, mediante la interpretación del $\overline{R}^2$ analizaremos el porcentaje de variabilidad que cada modelo es capaz de explicar. De cada coeficiente estimado b_j, interpretaremos, aplicando *"*ceteris paribus*"* (cláusula mediante la cual, si todas las X_j son fijas para j = 1,2,3 … k) b_j se interpretará como la cantidad que aumenta **y** para el aumento en unidad de $\mathbf{x_1}$

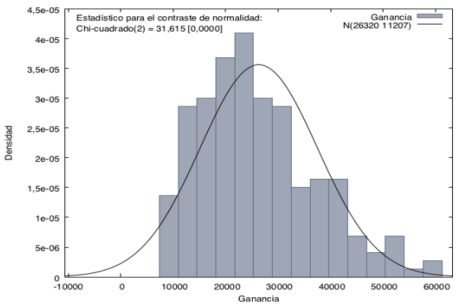
4. Análisis de los datos.

Las variables empleadas en el estudio han sido organizadas mediante Microsoft Excel para poder clasificar cada grupo de individuos en función de su ganancia media anual, formato en el cual se extrajeron los datos, obteniendo un total de 216 grupos de los cuales 12 son valores ausentes. Por esto, se han omitido en el estudio con la finalidad de simplificar el análisis y evitar caer en soluciones erróneas, mediante la omisión por defecto de los programas estadísticos empleados para el posterior análisis, R y GRETL

4.1. Análisis individual.

En primer lugar, comenzamos analizando detenidamente nuestra variable dependiente **Y**: Ganancia media. Con este análisis, observaremos como se comporta individual o marginalmente, sin añadir ninguna variable dependiente. Para este análisis nos centraremos exclusivamente en la ganancia media pues es la variable principal obtenida de la base de datos del INE, el resto de variables han sido obtenidas de forma agrupada, además son de carácter nominal y ordinal, por lo que carecen de interés para un análisis individual.

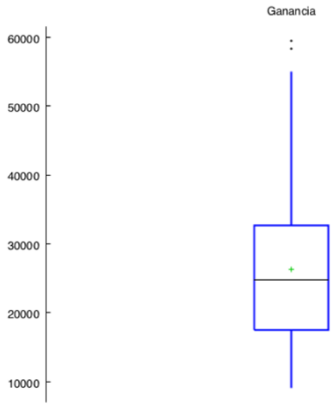
Distribución de la ganancia media.



Media	26320,
Mediana	24824,
Mínimo	9147,1
Máximo	59470,
Desviación típica	11207,
C.V.	0,42580
Asimetría	0,70598
Exc. de curtosis	-0,034554
Percentil del 5%	10986,
Percentil del 95%	48113,
Rango intercuartílico	15208,
Observaciones ausentes	12

Figura1: Distribución de la ganancia media, fuente: elaboración propia

La ganancia media española, se sitúa sobre los 26320,0 € anuales, siendo la mayor ganancia obtenida 54956 € anuales y la menor de 9147,1 € anuales. La mediana se sitúa en 24824,0 € pues se vera afectada en menor medida por aquellos valores considerados extremos. La desviación típica, resulta ser de prácticamente el 50% del valor medio, pues sin la adición de ninguna variables independientes, no contamos con ninguna explicación para estas variabilidades.



Atendiendo a los resultados obtenidos en los datos estadísticos principales de nuestra variable dependiente podemos observar que existe una gran dispersión, de la ganancia media de los individuos. Esto se puede observar en el grafico de caja donde se ve claramente la extensión de los cuartiles primero y último en comparación con el resto.

Figura 2: Gráfico de caja de la distribución de la ganancia media, fuente: elaboración propia.

4.2. Análisis conjunto.

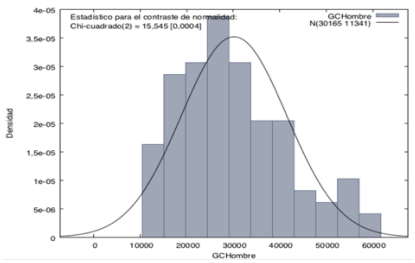
Una vez analizada nuestra variable dependiente, estudiaremos cómo se transforma su distribución a medida que se van considerando las distintas variables independientes seleccionadas para nuestro estudio y veremos cómo se relacionan entre sí.

A través de estos análisis se obtendrán información cualitativa útil para un mejor entendimiento de nuestras variables y para ajustar los distintos modelos de regresión.

i. Distribuciones condicionadas

Distribución de la ganancia media de los hombres y las mujeres

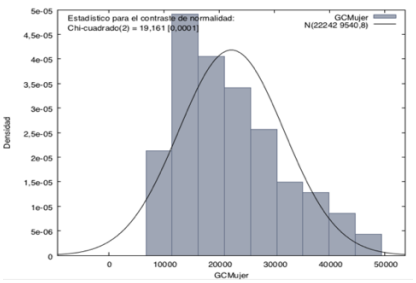
Distribución de la ganancia media de los hombres



Media	30165,
Mediana	28352,
Mínimo	12807,
Máximo	59470,
Desviación típica	11341,
C.V.	0,37598
Asimetría	0,64746
Exc. de curtosis	-0,28152
Percentil del 5%	14686,
Percentil del 95%	53680,
Rango intercuartílico	15596,
Observaciones ausentes	111

Figura 3: Distribución de la ganancia media de los hombres, fuente: elaboración propia

Distribución de la ganancia media de las mujeres



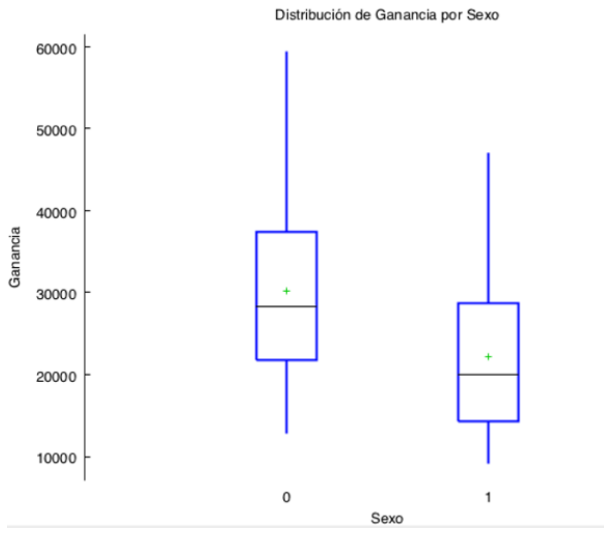
Media	22242,
Mediana	20031,
Mínimo	9147,1
Máximo	47015,
Desviación típica	9540,8
C.V.	0,42895
Asimetría	0,69728
Exc. de curtosis	-0,32616
Percentil del 5%	10149,
Percentil del 95%	42419,
Rango intercuartílico	14373,
Observaciones ausentes	117

Figura 4: Distribución de la ganancia media de las mujeres, fuente: elaboración propia

Tras observar ambas distribuciones podemos sospechar, que en la actualidad, sigue existiendo una discriminación salarial condicionada al sexo del trabajador. Como ya hemos visto previamente, la ganancia media española se sitúa sobre los 26320,0 € anuales. Ahora bien, la ganancia media anual de los hombres aumenta en su media un total de 3845 € hasta situarse en una ganancia anual media de 30165 €.

Por otro lado, la ganancia media anual de las mujeres desciende en 4078 € situándose en los 22242 € euros de ganancia media anual. Tras analizar las figura 3 y 4 podríamos deducir que la variable independiente **X3**, es decir el Sexo, supone en la ganancia media anual una disminución de un 27% para las mujeres, con una diferencia monetaria total de 7923 €, por lo que posteriormente en la regresión, esta variable debería tener una β_j negativa.

Por otra parte, observando la desviación típica de las ganancias medias, podemos observar como esta, es considerablemente mayor en los hombres con un valor de 11341 €, en comparación con las mujeres siendo esta de 9540,8 €, lo cual supone una variabilidad un 18% superior en los ganancia media de los hombres.



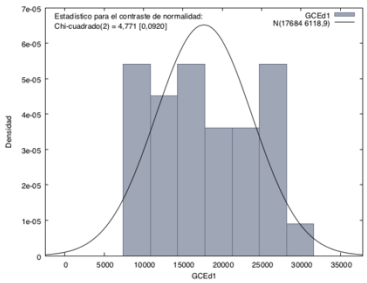
Para poder observar esta variabilidad con mayor detalle, se ha elaborado una gráfica de caja con separación de factores por la variable independiente a analizar: el Sexo.

Figura 5: distribución de la ganancia media con factor de separación por sexo, fuente: elaboración propia

A través de esta, podemos observar que efectivamente la mediana de los hombres es considerablemente mayor que la de las mujeres por lo que esta variable se espera que sea significativa a la hora de elaborar los distintos modelos, pero además, se observa como la distancia entre los cuartiles es mucho mas amplia en los hombres a medida que la ganancia media aumenta.

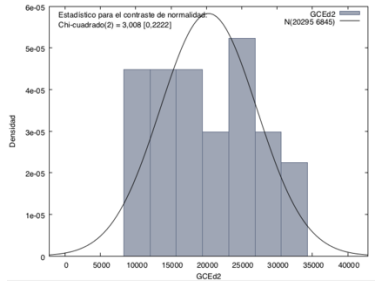
Distribución de la ganancia media condicionada a la educación

Otra de las variables que usualmente se ha considerado relevantes a la hora de determinar la ganancia media de un trabajador, es la variable independiente **X2**: Educación. Para poder estudiar en qué medida esta es importante, se han analizado individualmente cada uno de los distintos niveles de estudio que se han tenido en cuenta por el INE a la hora de reunir los datos:



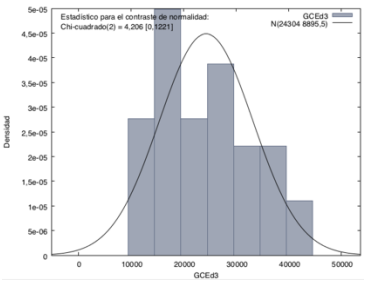
Media	17684,
Mediana	15990,
Mínimo	9147,1
Máximo	29928,
Desviación típica	6118,9
C.V.	0,34602
Asimetría	0,32923
Exc. de curtosis	−1,1397
Percentil del 5%	9392,4
Percentil del 95%	28073,
Rango intercuartílico	11259,
Observaciones ausentes	184

figura 6: distribución de la ganancia media condicionado a la educación en la educación primaria, fuente: elaboración propia



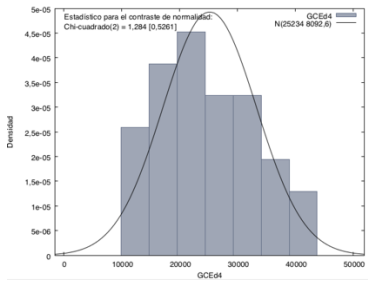
Media	20295,
Mediana	19605,
Mínimo	10149,
Máximo	32471,
Desviación típica	6845,0
C.V.	0,33727
Asimetría	0,15983
Exc. de curtosis	−1,1187
Percentil del 5%	10149,
Percentil del 95%	31890,
Rango intercuartílico	11130,
Observaciones ausentes	180

figura 7: distribución de la ganancia media condicionado a la educación en las primeras etapas de la ESO, fuente: elaboración propia



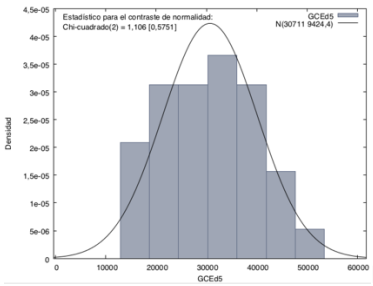
Media	24304,
Mediana	23678,
Mínimo	11962,
Máximo	42080,
Desviación típica	8895,5
C.V.	0,36601
Asimetría	0,36699
Exc. de curtosis	−0,97942
Percentil del 5%	12341,
Percentil del 95%	40920,
Rango intercuartílico	13691,
Observaciones ausentes	180

figura 8: distribución de la ganancia media condicionado a la educación en las segundas etapas de la ESO, fuente: elaboración propia



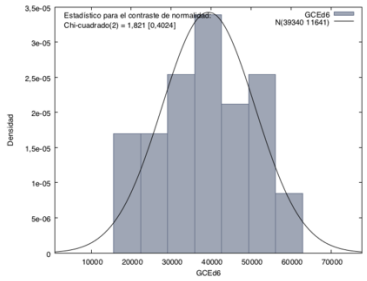
Media	25234,
Mediana	24118,
Mínimo	12369,
Máximo	41363,
Desviación típica	8092,6
C.V.	0,32070
Asimetría	0,21437
Exc. de curtosis	−0,87326
Percentil del 5%	13054,
Percentil del 95%	40365,
Rango intercuartílico	13088,
Observaciones ausentes	184

figura 9: distribución de la ganancia media condicionado a la educación en grados superiores y similares, fuente: elaboración propia



Media	30711,
Mediana	30317,
Mínimo	15739,
Máximo	50592,
Desviación típica	9424,4
C.V.	0,30687
Asimetría	0,17337
Exc. de curtosis	-0,87322
Percentil del 5%	16866,
Percentil del 95%	47487,
Rango intercuartílico	14171,
Observaciones ausentes	183

figura 10: distribución de la ganancia media condicionado a la educación de diplomados universitarios y similares , fuente: elaboración propia



Media	39340,
Mediana	41661,
Mínimo	18995,
Máximo	59470,
Desviación típica	11641,
C.V.	0,29591
Asimetría	-0,10085
Exc. de curtosis	-1,0298
Percentil del 5%	19613,
Percentil del 95%	58507,
Rango intercuartílico	17139,
Observaciones ausentes	181

figura 11: distribución de la ganancia media condicionado a la educación de licenciados y similares y doctores universitarios, fuente: elaboración propia

Observando las resultados obtenidos, vemos cómo la ganancia media evoluciona en nivel de estudios: la educación primaria, con un media 17684€ hasta llegar al ultimo nivel considerado: licenciados y similares y doctores universitarios, con una ganancia media de 39340€. Lo cual supone un aumento total del 45%.

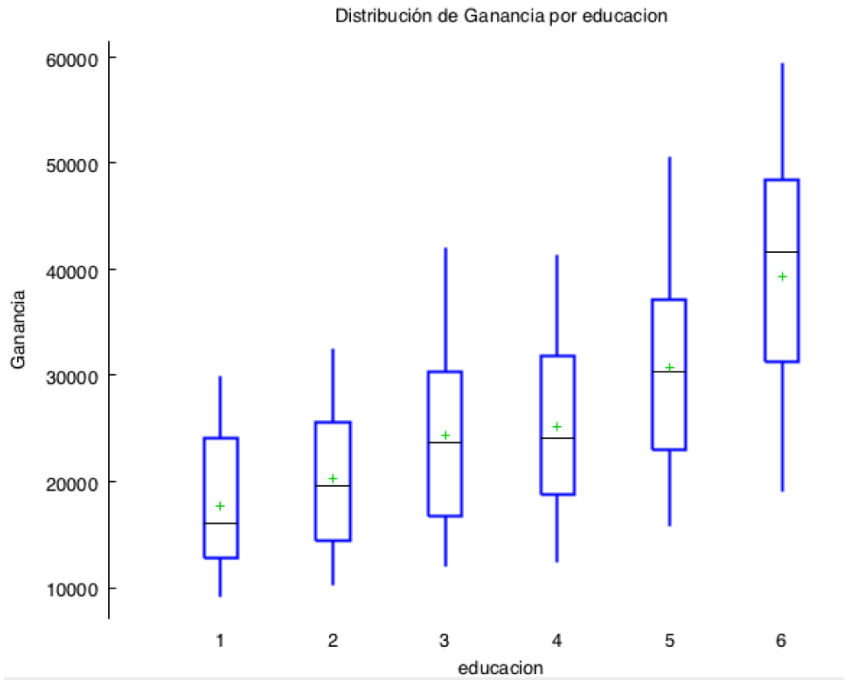


figura 12: Gráfico de caja con factor de separación por niveles de educación, fuente: elaboración propia

Para poder observar en mayor detalle la evolución tanto de la media como de la mediana se ha desarrollado un gráfico de cajas, donde se incluye la ganancia media condicionada

a la educación. A simple vista se puede observar como tanto la media como la mediana aumentan a cada uno de los distintos niveles considerados para el estudio.

Como conclusión, se podría decir que esta variable resultaría significativa al elaborar un modelo de regresión, puesto que influye positivamente y de manera significativa en la ganancia media de los trabajadores a medida que esta crece.

i. Regresiones

Tras analizar las distribuciones condicionadas, se realizarán distintas regresiones, tanto simples como múltiples, para estudiar cómo la ganancia media varia en función de las distintas variables independientes, así como con las distintas combinaciones posibles entre ellas, hasta encontrar el modelo que mejor se ajuste a nuestro objetivo.

En primer lugar, es importante mencionar que a consecuencia del tipo de datos proporcionado por el INE no realizaremos un análisis de la correlación puesto que como hemos mencionado anteriormente, nuestras variables independientes son básicamente factores propios de un análisis experimental, por lo que no tendrán ningún tipo de relación entre ellas. Por esto mismo, tampoco realizaremos gráficos de dispersión para cada par de variables, pues los datos únicamente se mueven a través del eje Y, es decir, variarán únicamente en su ganancia media.

Seguidamente se calculará la matriz de correlación, donde, a simple vista podemos observar que: Cov (x_j, x_h) = E($x_j x_h$) - E(x_j)E(x_h) = E($x_j x_h$) = 0 para todo j,h=1, 2, …, n) es decir, no existe correlación en ninguna de las variables independientes(por las características mencionadas en el párrafo anterior).

```
Coeficientes de correlación, usando las observaciones 1 - 216
valor crítico al 5% (a dos colas) = 0,1335 para n = 216

            Sexo     experiencia       educacion       Comunidad
          1,0000          0,0000          0,0000          0,0000 Sexo
                          1,0000          0,0000          0,0000 experiencia
                                          1,0000          0,0000 educacion
                                                          1,0000 Comunidad
```

figura 13: Matriz de correlación, fuente: elaboración propia

Regresiones simples

En las siguientes regresiones, puesto a que anteriormente ha sido demostrado por distintos autores que estas variables no son suficientes para explicar la ganancia media, se ajustarán las siguientes regresiones mostrando únicamente la información relevante, para la posterior evaluación de la evolución de los coeficientes estimados al añadir variables nuevas.

a) Ganancia media con respecto a la experiencia

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p
const	11949,4	1374,06	8,696	1,19e-15 ***
experiencia	4267,27	365,891	11,66	2,29e-24 ***

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	1,52e+10	D.T. de la regresión	8684,999
R-cuadrado	0,402400	R-cuadrado corregido	0,399441
F(1, 202)	136,0187	Valor p (de F)	2,29e-24
Log-verosimilitud	-2138,606	Criterio de Akaike	4281,213
Criterio de Schwarz	4287,849	Crit. de Hannan-Quinn	4283,897

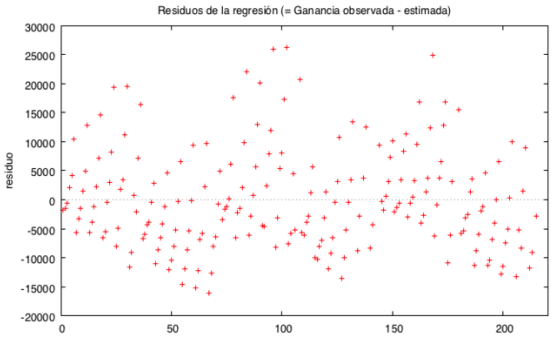


figura 14:Modelo MCO ganancia media con respecto a la experiencia, fuente: elaboración propia

Observando este modelo donde:

$$\textbf{Ganancia media} = \beta o + \beta_1 \text{Experiencia} + \varepsilon$$

La ganancia media de los individuos siendo la ganancia base $\boldsymbol{\beta o}$:11949,4€, la cual aumentará positivamente $\boldsymbol{\beta_1}$: 4267,22€ en función de cada uno de los distintos niveles de experiencia en los que se encuentre el individuo.

Atendiendo a la desviación típica de la $\boldsymbol{\beta_1}$ este modelo, la cual es significativamente menor que su coeficiente, posee un valor de 365,891€ por lo que los resultados de este modelo parece, tienen cierto grado de precisión. Si observamos la $\boldsymbol{\overline{R}^2}$, observamos que la variable **X1**: Experiencia explica prácticamente el 40% de la variabilidad de la ganancia media del modelo.

Finalmente, para contrastar la hipótesis $\mathbf{H_0}$**:** $\beta_j = \mathbf{0}$ frente $\mathbf{H_0}$**:** $\beta_j \neq \mathbf{0}$, atenderemos a estadístico T del modelo así como a su p-valor, los cuales nos llevan a rechazar la

hipótesis de nula de que la variable dependiente **X1**: Experiencia, no sea significativa, Es decir, existen indicios de que será necesaria incluir esta variable para elaborar nuestro modelo.

b) Ganancia media con respecto a la educación

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p
const	12136,1	1418,26	8,557	2,90e-15 ***
educacion	4046,91	363,999	11,12	1,03e-22 ***

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	1,58e+10	D.T. de la regresión	8848,966
R-cuadrado	0,379622	R-cuadrado corregido	0,376551
F(1, 202)	123,6081	Valor p (de F)	1,03e-22
Log-verosimilitud	-2142,422	Criterio de Akaike	4288,844
Criterio de Schwarz	4295,480	Crit. de Hannan-Quinn	4291,528

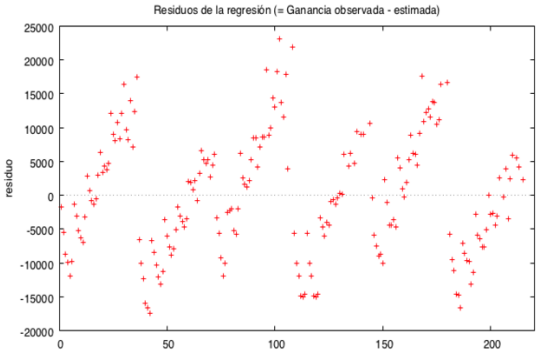


figura 15:Modelo MCO ganancia media con respecto a la educación, fuente: elaboración propia

Observando este modelo donde:

$$\textbf{Ganancia media} = \beta o + \beta_1 \text{Educación} + \varepsilon$$

La ganancia media de los individuos siendo la ganancia base $\boldsymbol{\beta o}$:12136,1€, la cual aumentará positivamente $\boldsymbol{\beta_1}$: 4046,91€ en función de cada uno de los distintos niveles de experiencia en los que se encuentre el individuo.

Atendiendo a la desviación típica de la $\boldsymbol{\beta_1}$ este modelo, la cual es significativamente menor que su coeficiente, posee un valor de 363,999€ por lo que los resultados de este modelo parece, tienen cierto grado de precisión. Si observamos la $\boldsymbol{\overline{R}^2}$, observamos que la variable **X2**: Educación explica prácticamente el 38% de los errores del modelo.

Finalmente, para contrastar la hipótesis $\mathbf{H_0: \beta_j = 0}$ frente $\mathbf{H_0: \beta_j \neq 0}$, atenderemos a estadístico T del modelo así como a su p-valor, los cuales nos llevan a rechazar la hipótesis de nula de que la variable dependiente **X2**: Educación, no sea significativa. Como sucedía con la variable anterior, existen indicios de que será necesaria incluir esta variable para elaborar nuestro modelo.

Tras analizar las variables **X1**: Experiencia y **X2:** Educación se observa como en el gráfico de los residuos existe cierta estructura, lo cual indica la necesidad de incluir nuevas variables para poder explicar la ganancia media con mayor detalle.

Regresiones múltiples

a) Ganancia media respecto a experiencia y educación

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	−2327,35	1120,58	−2,077	0,0391	**
experiencia	4280,73	220,236	19,44	4,77e−48	***
educacion	4060,44	215,038	18,88	2,07e−46	***

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	5,49e+09	D.T. de la regresión	5227,631
R-cuadrado	0,784560	R-cuadrado corregido	0,782417
F(2, 201)	365,9878	Valor p (de F)	9,98e−68
Log-verosimilitud	−2034,542	Criterio de Akaike	4075,084
Criterio de Schwarz	4085,038	Crit. de Hannan-Quinn	4079,111

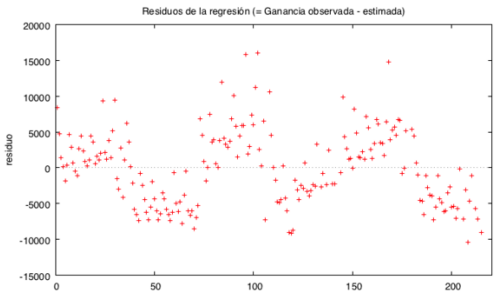


figura 16:Modelo MCO ganancia media con respecto a la experiencia y la educación , fuente: elaboración propia

Observando este modelo donde:

$$\textbf{Ganancia media} = \beta o + \beta_1 \text{Experiencia} + \beta_2 \text{Educación} + \varepsilon$$

La ganancia media de los individuos siendo la ganancia base β_o: -2327,35€, la cual aumentará positivamente β_1: 4280,73€ en función de cada uno de los distintos niveles de experiencia en los que se encuentre el individuo, así como esta aumentara positivamente β_2: 4060,44€ en función a que nivel de educación el individuo llegue.

Atendiendo a la desviación típica de la β_1 y β_2 de este modelo, son significativamente menor que su respectivos coeficientes, con un valor de 220,236€ y 215,038€ por lo que los resultados de este modelo parece, que tienen cierto grado de precisión. Si observamos la $\overline{R}^2$, observamos que las variables **X1**: Experiencia y **X2:** Educación, explican el 78% de la variabilidad de la ganancia media del modelo.. Por otro lado si atendemos a los criterios de Akaike (AIC) y de Schwarz (BIC) los cuales ofrecen una estimación relativa de la información perdida en la elaboración del modelo de 4079,111€ y 4085,038€ respectivamente.

Para contrastar la hipótesis $\mathbf{H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0}$ frente a $\mathbf{H_1 : al\ menos\ un\ \beta_j \neq 0}$, atenderemos al estadístico F del modelo así como a su p-valor, los cuales nos llevan a rechazar la hipótesis nula de que el conjunto de variables independientes **X1**: Experiencia y **X2**: Educación, no sean significativo.

Para estudiar la normalidad de los residuos, hemos realizado un grafico Q-Q. En el cual se aprecia, como la mayoría de los residuos caen bastante alineados con respecto a la media a excepción de los puntos extremos tanto inferiores como superiores lo cual indica un problema de normalidad. Además, mediante un test de normalidad de los residuos cuyo estadístico para el contraste de normalidad obtenido es de Chi-Cuadrado(2) = 7,950[0,0188] lo cual vuelve a indicar un problema de normalidad, únicamente aceptado al 1% de significación.

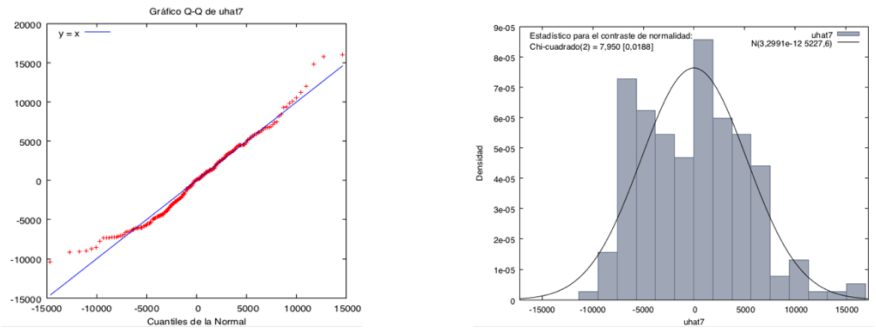


figura 17: Contraste de la normalidad sobre el modelo MCO a, fuente: elaboración propia

Para estudiar la heterocedasticidad del modelo en primer lugar se realizara un contraste de Breusch-Pagan donde estudiaremos la varianza estimada de los residuos en la regresión

```
Suma de cuadrados explicada = 36,7031

Estadístico de contraste: LM = 18,351539,
con valor p  = P(Chi-cuadrado(2) > 18,351539) = 0,000104
```

figura 18: Contraste de Breusch-Pagan sobre el modelo MCO a, fuente: elaboración propia

Analizando el contraste, vemos que nuestro p-valor devuelve un resultado de 0,000104 por lo que rechazaremos la homocedasticiad de los residuos. Fijándonos en el gráfico de los residuos, podemos observar que la cantidad de datos atípicos es mínima y poco significativos para la elaboración del modelo. Por otro lado de dispersión de los residuos parece seguir una tendencia, la cual se corregirá posteriormente mediante la inclusión de nuevas variables.

Para finalizar se ha estudiado la estabilidad estructural del modelo mediante un contraste de Chow en el individuo que se encuentra en el punto medio de nuestra muestra, así como un estudio de la especificación del modelo mediante un contraste de Ramsey (RESET) sobre todas las variables. Ambos contrastes devolvieron un p-valor de prácticamente de 0, por lo que no apoyan nuestra hipótesis nula de significación conjunta, lo cual indica que pudiese ser necesario la inclusión de nuevas variables para un modelo óptimo.

b) Ganancia media respecto a experiencia, educación y sexo

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	1531,00	791,230	1,935	0,0544	*
experiencia	4193,25	147,912	28,35	5,51e-72	***
educacion	4112,61	144,357	28,49	2,51e-72	***
Sexo	-7720,25	491,971	-15,69	1,05e-36	***

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	2,46e+09	D.T. de la regresión	3508,422
R-cuadrado	0,903445	R-cuadrado corregido	0,901997
F(3, 200)	623,7883	Valor p (de F)	3,2e-101
Log-verosimilitud	-1952,680	Criterio de Akaike	3913,359
Criterio de Schwarz	3926,632	Crit. de Hannan-Quinn	3918,728

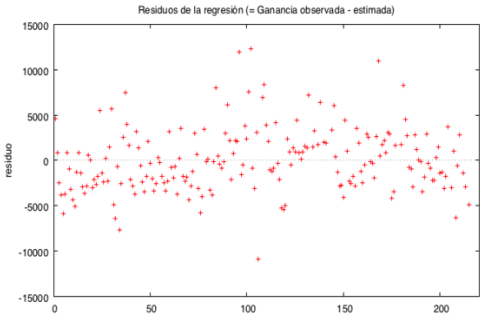


figura 19:Modelo MCO ganancia media con respecto a la experiencia educación y sexo, fuente: elaboración propia

Observando este modelo donde:

$$\textbf{Ganancia media} = \beta o + \beta_1 \text{Experiencia} + \beta_2 \text{Educación} + \beta_3 \text{Sexo} + \varepsilon$$

La ganancia media de los individuos siendo la ganancia base β_o: 1531,00€, la cual aumentará positivamente β_1: 4193,25€ en función de cada uno de los distintos niveles de experiencia en los que se encuentre el individuo, así como esta aumentara positivamente β_2: 4112,61€ en función a que nivel de educación el individuo llegue. Por otro lado la variable independiente del sexo tiene un efecto negativo β_3: -7720,25€ sobre la ganancia media siendo el individuo una mujer.

Atendiendo a la desviación típica de la β_1, β_2 y β_3 de este modelo, son significativamente menor que su respectivos coeficientes, con un valor de 147,912€, 144,357€ y 491,971€ siendo esta ultima considerablemente mayor que las demás, debido a la variabilidad estudiada anteriormente en la distribución de la ganancia media condiciona a los hombres. Pese a esto, los resultados de este modelo parece ser, que tienen cierto grado de precisión. Si observamos la $\overline{R}^2$, observamos que las variables **X1**: Experiencia, **X2:** Educación y

X3: Sexo, explican el 90% de la variabilidad de la ganancia media del modelo. Por otro lado si atendemos a los criterios de Akaike (AIC) y de Schwarz (BIC) los cuales ofrecen una estimación relativa de la información perdida en la elaboración del modelo de 3918,728€ y 3926,632€ respectivamente, por lo que mediante la inclusión de la variable independiente **X3:** Sexo estamos perdiendo menos información.

Para contrastar la hipótesis $\mathbf{H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0}$ frente a $\mathbf{H_1 : al\ menos\ un\ \beta_j \neq 0}$, atenderemos al estadístico F del modelo así como a su p-valor, los cuales nos llevan a rechazar la hipótesis nula de que el conjunto de variables independientes **X1**: Experiencia, **X2**: Educación y **X3:** Sexo no sean significativo.

Para estudiar la normalidad de los residuos, hemos realizado un grafico Q-Q. En el cual se aprecia, como la mayoría de los residuos caen bastante alineados con respecto a la media a excepción de los puntos extremos tanto inferiores como superiores lo cual supone nuevamente un problema de normalidad. Además, mediante un test de normalidad de los residuos cuyo estadístico para el contraste de normalidad obtenido es de Chi-Cuadrado(2) = 13,529[0,0012] el cual nuevamente plantea problemas con respecto a la normalidad del modelo.

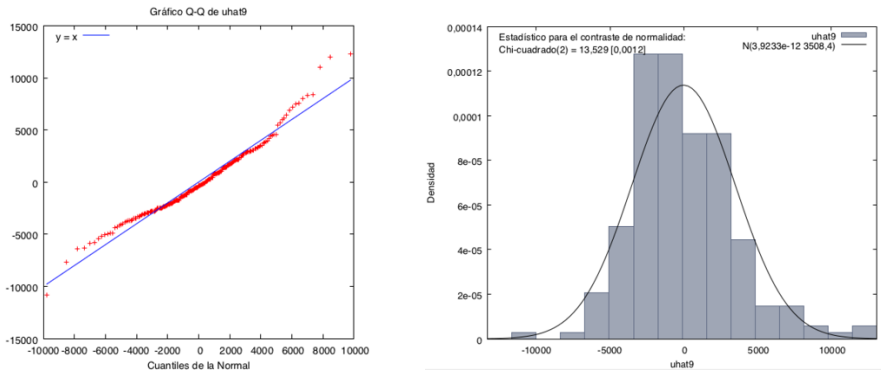


figura 20: Contraste de la normalidad sobre el modelo MCO b, fuente: elaboración propia

Para estudiar la heterocedasticidad del modelo en primer lugar se realizara un contraste de Breusch-Pagan donde estudiaremos la varianza estimada de los residuos en la regresión

```
Suma de cuadrados explicada = 45,1204

Estadístico de contraste: LM = 22,560220,
con valor p  = P(Chi-cuadrado(3) > 22,560220) = 0,000050
```

figura 21: Contraste de Breusch-Pagan sobre el modelo MCO b, fuente: elaboración propia

Analizando el contraste, vemos que nuestro p-valor devuelve un resultado de 0,00050 rechazaremos la homocedasticiad de los residuos.

Fijándonos en el gráfico de los residuos, podemos observar que la cantidad de datos atípicos es mínima y poco significativos para la elaboración del modelo. Por otro lado de dispersión de los residuos al añadir la variable **X3**: Sexo corrige el patrón obtenido en el modelo anterior, por lo que el modelo parece ajustarse de una manera mas adecuada.

Para finalizar se ha estudiado la estabilidad estructural del modelo mediante un contraste de Chow en el individuo que se encuentra en el punto medio de nuestra muestra, así como un estudio de la especificación del modelo mediante un contraste de Ramsey (RESET) sobre todas las variables . Ambos contrastes devolvieron un p-valor de prácticamente de 0, por lo que no apoyan nuestra hipótesis nula de significación

C) Ganancia media respecto a los distintos niveles de educación, la experiencia y el sexo

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	7475,42	721,216	10,37	2,37e-20	***
experiencia	4173,33	130,508	31,98	1,05e-79	***
Ed2	2047,47	751,386	2,725	0,0070	***
Ed3	6056,12	751,386	8,060	7,37e-14	***
Ed4	7920,32	772,016	10,26	4,85e-20	***
Ed5	13527,7	766,769	17,64	2,45e-42	***
Ed6	21280,3	755,842	28,15	8,49e-71	***
Sexo	-7668,35	433,413	-17,69	1,73e-42	***

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	1,87e+09	D.T. de la regresión	3087,558
R-cuadrado	0,926717	R-cuadrado corregido	0,924099
F(7, 196)	354,0780	Valor p (de F)	1,5e-107
Log-verosimilitud	-1924,551	Criterio de Akaike	3865,101
Criterio de Schwarz	3891,646	Crit. de Hannan-Quinn	3875,839

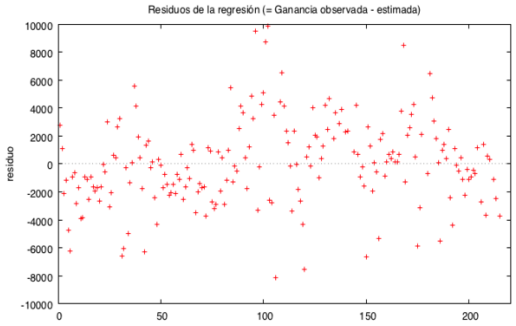


figura 22:Modelo MCO ganancia media con respecto a la experiencia, los distintos niveles de educación y el sexo, fuente: elaboración propia

Observando este modelo donde:

Ganancia media = $\beta o + \beta_1 Experiencia + \beta_2 Ed2 + \beta_3 Ed3 + \beta_4 Ed4 + \beta_5 Ed5 + \beta_6 Ed6 + \beta_7 Sexo + \varepsilon$

Para poder estudiar con mayor detalle como afectan los distintos niveles de educación de un individuo en su ganancia media como el efecto positivo que genera en cada nivel de educación, en lugar de una media fija, se ha construido este modelo, donde observamos la ganancia base de β_0: 7475,42€, la cual aumentará positivamente en función de los años de experiencia de nuestro individuo en β_1: 4173,33€ y también aumentara positivamente según el nivel de educación de nuestro individuo β_2 .... β_6. Por otro lado la variable

independiente del sexo tiene un efecto negativo β_7: -7668,35€ sobre la ganancia media siendo el individuo una mujer.

Atendiendo a la desviación típica de cada una de nuestras $\beta_2 \dots \beta_6$ de este modelo, son significativamente menores que sus respectivos coeficientes, aunque son bastante mayores que las desviaciones típicas obtenidas hasta ahora, lo cual podría indicar que estos coeficientes no son 100% fiables. Finalmente la desviación típica de la β_7 es menor que el resto, aunque considerablemente alta, nuevamente debido a la variabilidad observada en el análisis de la distribución de la ganancia media condicionada al sexo. Si observamos la $\overline{R}^2$, observamos que las variables **X1**: Experiencia, **X2:** Educación y **X3:** Sexo, explican el 92% de la variabilidad de la ganancia media del modelo. Por otro lado si atendemos a los criterios de Akaike (AIC) y de Schwarz (BIC) los cuales ofrecen una estimación relativa de la información perdida en la elaboración del modelo de 3875,839€ y 3891,646€ respectivamente.

Finalmente para contrastar la hipótesis $\mathbf{H_o : \beta_1 = \beta_2 = \dots = \beta_k = 0}$ frente a $\mathbf{H_1}$ **: al menos un** $\beta_j \neq 0$, atenderemos al estadístico F del modelo así como a su p-valor, los cuales nos llevan a rechazar la hipótesis nula de que el conjunto de variables independientes **X1**: Experiencia, **X2**: Educación y **X3:** Sexo no sean significativo.

D) Ganancia media respecto a los distintos intervalos de experiencia, la educación y el sexo

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	5964,58	784,515	7,603	1,17e-12	***
de1a3	3028,09	803,168	3,770	0,0002	***
de4a10	8340,84	803,168	10,38	2,08e-20	***
de11a20	13244,0	803,168	16,49	6,87e-39	***
de21a29	17828,0	821,741	21,70	5,33e-54	***
masde30	19136,7	868,105	22,04	5,77e-55	***
educacion	4093,54	140,332	29,17	3,09e-73	***
Sexo	-7731,66	477,991	-16,18	6,09e-38	***

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	2,28e+09	D.T. de la regresión	3407,554
R-cuadrado	0,910739	R-cuadrado corregido	0,907551
F(7, 196)	285,6872	Valor p (de F)	3,47e-99
Log-verosimilitud	-1944,668	Criterio de Akaike	3905,336
Criterio de Schwarz	3931,881	Crit. de Hannan-Quinn	3916,074

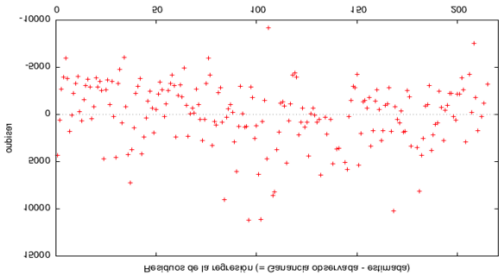


figura 23:Modelo MCO ganancia media con respecto a los distintos años de experiencia, la educación y el sexo, fuente: elaboración propia

Observando este modelo donde:

$$\textbf{Ganancia media} = \beta o + \beta_1 de1a3 + \beta_2 de4a10 + \beta_3 de11a20 + \beta_4 de21a29 + \beta_5 masde30 + \beta_6 Experiencia + \beta_7 Sexo + \varepsilon$$

Para poder estudiar con mayor detalle como afectan los distintos años de experiencia de un individuo en su ganancia media como el efecto positivo que genera en rango de experiencia, en lugar de una media fija, se ha construido este modelo, donde observamos la ganancia base β_0: 5964,58€, la cual aumentará positivamente en función del nivel de educación de nuestro individuo β_1 .... β_5. Por otro lado las variable independientes del nivel de educación aumentara la ganancia media en β_6: 4093,54€. Finalmente la variable independiente del sexo tiene un efecto negativo β_7: -7731,66€ sobre la ganancia media siendo el individuo una mujer.

Atendiendo a la desviación típica de cada una de nuestras β_1 ... β_5 de este modelo, son significativamente menores en comparación a sus respectivos coeficientes, aunque son bastante mayores que las desviaciones típicas obtenidas hasta ahora, lo cual podría indicar que estos coeficientes no son 100% fiables. Finalmente la desviación típica de la β_7 es menor que el resto, aunque considerablemente alto, nuevamente debido a la variabilidad observada en el análisis de la distribución de la ganancia media de los hombres. Si observamos la $\overline{R}^2$, observamos que las variables **X1**: Experiencia, **X2:** Educación y **X3:** Sexo, explican el 90% de la variabilidad de la ganancia media del modelo. Por otro lado si atendemos a los criterios de Akaike (AIC) y de Schwarz (BIC) los cuales ofrecen una estimación relativa de la información perdida en la elaboración del modelo de 3905,336€ y 3931,881€ respectivamente.

Finalmente para contrastar la hipótesis $\mathbf{H_o : \beta_1 = \beta_2 = \ldots = \beta_k = 0}$ frente a $\mathbf{H_1}$ **: al menos un** $\beta_j \neq 0$, atenderemos al p-valor del estadístico F del modelo así como a su p-valor, los cuales nos llevan a rechazar la hipótesis de nula de que las variable independiente **X1**: Experiencia, **X2**: Educación y **X3:** Sexo no sean significativas en su conjunto.

E) Ganancia media por sexo en función de la educación

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	14868,9	1722,48	8,632	1,88e-15	***
educacion	4412,30	445,167	9,912	4,31e-19	***
Sexo	-5960,86	2518,25	-2,367	0,0189	**
EDUCxSexo	-651,377	645,947	-1,008	0,3145	

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	1,23e+10	D.T. de la regresión	7839,661
R-cuadrado	0,517892	R-cuadrado corregido	0,510660
F(3, 200)	71,61495	Valor p (de F)	1,69e-31
Log-verosimilitud	-2116,702	Criterio de Akaike	4241,403
Criterio de Schwarz	4254,676	Crit. de Hannan-Quinn	4246,772

figura 24:Modelo MCO ganancia media por sexo en función de la experiencia, fuente: elaboración propia

Observando este modelo donde:

$$\textbf{Ganancia media} = \beta o + \beta_1 \text{Educacion} + \beta_2 \text{Sexo} + \beta_3 \text{Educacion*Sexo} + \varepsilon$$

Observando los resultados obtenidos, se aprecia como la evolución de la ganancia media afecta por igual a todos los trabajadores, solo es afectada por el sexo partiendo del primer nivel de educación en un aumento único, el cual se mantendrá constante en los distintos niveles de educación.

F) Ganancia media por sexo en función de la experiencia

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	15376,7	1750,81	8,783	7,20e-16	***
experiencia	4313,24	458,658	9,404	1,26e-17	***
Sexo	-6476,36	2497,30	-2,593	0,0102	**
EXPxSexo	-273,932	665,911	-0,4114	0,6812	

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	1,24e+10	D.T. de la regresión	7887,238
R-cuadrado	0,512023	R-cuadrado corregido	0,504703
F(3, 200)	69,95170	Valor p (de F)	5,63e-31
Log-verosimilitud	-2117,936	Criterio de Akaike	4243,872
Criterio de Schwarz	4257,144	Crit. de Hannan-Quinn	4249,241

figura 25:Modelo MCO ganancia media por sexo con respecto a la experiencia, fuente: elaboración propia

Observando este modelo donde:

$$\textbf{Ganancia media} = \beta o + \beta_1 \text{Experiencia} + \beta_2 \text{Sexo} + \beta_3 \text{Experiencia*Sexo} + \varepsilon$$

Observando los resultados obtenidos, se aprecia como la evolución de la ganancia media afecta por igual a todos los trabajadores a medida que acumulan años de experiencia. solo es afectada por el sexo partiendo del primer año de experiencia en un aumento único, el cual se mantendrá constante el resto de los años..

Modelos finales:

Tras estudiar las distintas posibilidades que nos ofrecen nuestras variables independientes, observar como interactúan y completan el modelo entre ellas, se elaborarán dos modelos considerando nuestra ultima variable independiente **X4:** Comunidad autónoma.

En primer lugar consideraremos cada uno de los distintos niveles de educación y experiencia como una variable independiente más, pues no se pretende considerar el hecho de alcanzar cada uno de los niveles con la misma facilidad. En segundo lugar se consideraran tanto los distintos niveles de educación como los años de experiencia como una variable independiente única **X1**: Experiencia, **X2**: Educación

A) Ganancia media respecto a los distintos intervalos de experiencia, niveles de educación, el sexo y la comunidad autónoma.

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	11233,7	738,221	15,22	9,03e-35	***
de1a3	3028,09	684,659	4,423	1,63e-05	***
de4a10	8340,84	684,659	12,18	1,24e-25	***
de11a20	13244,0	684,659	19,34	7,14e-47	***
de21a29	17879,9	700,978	25,51	2,03e-63	***
masde30	19059,5	743,654	25,63	9,97e-64	***
Ed2	2026,10	707,489	2,864	0,0047	***
Ed3	6034,75	707,489	8,530	4,45e-15	***
Ed4	7826,63	726,685	10,77	1,93e-21	***
Ed5	13376,3	722,264	18,52	1,64e-44	***
Ed6	21229,7	711,471	29,84	7,88e-74	***
Sexo	-7663,98	407,972	-18,79	2,82e-45	***
Comunidad	608,854	248,340	2,452	0,0151	**

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	1,61e+09	D.T. de la regresión	2904,764
R-cuadrado	0,936792	R-cuadrado corregido	0,932820
F(12, 191)	235,8959	Valor p (de F)	1,7e-107
Log-verosimilitud	-1909,465	Criterio de Akaike	3844,930
Criterio de Schwarz	3888,066	Crit. de Hannan-Quinn	3862,379

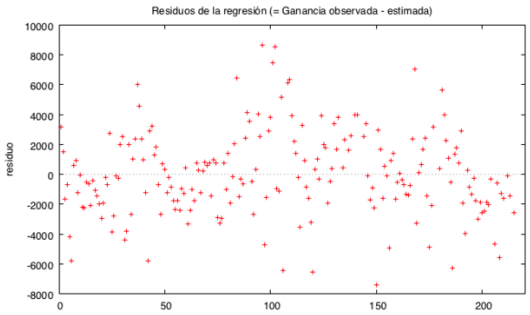


figura 26:Modelo MCO ganancia media con respecto a los distintos años de experiencia, niveles de educación, el sexo y la comunidad autónoma, fuente: elaboración propia

Observando este modelo donde:

$$\textbf{Ganancia media} = \beta_o + \beta_1 de1a3 + \beta_2 de4a10 + \beta_3 de11a20 + \beta_4 de21a29 + \beta_5 masde30 + \beta_6 Ed2 + \beta_7 Ed3 + \beta_8 Ed4 + \beta_8 Ed5 + \beta_{10} Ed6\beta_7 + \beta_{11} Sexo + \beta_{12} Comunidad\ autónoma + \varepsilon$$

Tras elaborar el modelo, a simple viste se observa como con la inclusión de las variables ficticias, se tiene en cuenta el aumento de la ganancia media correspondiente a cada uno de los distintos niveles de nuestras variables **X1:** Experiencia y **X2:** Educación de una forma mas detallada, a diferencia de los modelos anteriores en los cuales se otorga un aumento fijo en la ganancia media por cada nivel.

La ganancia media de los individuos siendo la ganancia base **βo**: 5964,58€, la cual aumentará positivamente en función del nivel de educación de nuestro individuo β_1 .... β_{10}. Por otro lado la variable independiente del sexo tiene un efecto negativo $\beta_{11:}$ -7663,66€ sobre la ganancia media siendo el individuo una mujer. Finalmente el hecho de que nuestro individuo pertenezca a la comunidad autónoma de Madrid o a Cataluña aumentara su ganancia media en $\beta_{12:}$ 608,854€

Atendiendo a la desviación típica de cada una de nuestras β_i de este modelo, son significativamente menores en comparación a sus respectivos coeficientes, son bastante mayores que las desviaciones típicas obtenidas hasta ahora, lo cual podría indicar que estos coeficientes no son 100% fiables. Finalmente la desviación típica de la β_7 es menor que el resto, aunque considerablemente alto, nuevamente debido a la variabilidad observada en el análisis de la distribución de la ganancia media de los hombres. Si observamos la $\overline{R}^2$, observamos que las variables **X1**: Experiencia, **X2:** Educación, **X3:** Sexo y **X4:** Comunidad autónoma explican el 93% de la variabilidad de la ganancia media del modelo, siendo este el valor mas alto obtenido. Por otro lado si atendemos a los criterios de Akaike (AIC) y de Schwarz (BIC) los cuales ofrecen una estimación relativa de la información perdida en la elaboración del modelo de 3844,930€ y 3888,066€ respectivamente, por lo que en este modelo, aunque muy levemente, se pierde menos información en comparación con el resto de modelos.

Para contrastar la hipótesis $\mathbf{H_o : \beta_1 = \beta_2 = ... = \beta_k = 0}$ frente a $\mathbf{H_1 :}$ **al menos un** $\beta_j \neq 0$, atenderemos al p-valor del estadístico F del modelo así como a su p-valor, los cuales nos llevan a rechazar la hipótesis de nula de que las variable independiente **X1**: Experiencia, **X2**: Educación, **X3:** Sexo y **X4:** Comunidad autónoma no sean significativas en su conjunto.

Para estudiar la normalidad de los residuos, hemos realizado un grafico Q-Q. En el cual se aprecia, como la mayoría de los residuos caen muy alineados con respecto a la media a excepción de los puntos extremos tanto inferiores como superiores lo cual podría suponer un problema de normalidad. Sin embargo, mediante un test de normalidad de los residuos cuyo estadístico para el contraste de normalidad obtenido es de Chi-Cuadrado(2) = 4,867[0,0877] apoyando nuestra hipótesis conjunta de significación del modelo y

aceptando la normalidad del modelo. Por lo tanto en este modelo, se podría aceptar la normalidad, sin la necesidad de convertir el modelo.

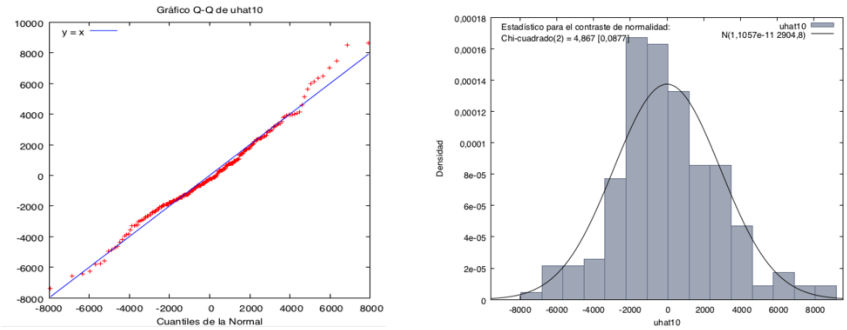


figura 27: Contraste de la normalidad sobre el modelo final MCO A, fuente: elaboración propia

Para estudiar la heterocedasticidad del modelo en primer lugar se realizara un contraste de Breusch-Pagan donde estudiaremos la varianza estimada de los residuos en la regresión

```
Suma de cuadrados explicada = 112,947

Estadístico de contraste: LM = 56,473377,
con valor p  = P(Chi-cuadrado(12) > 56,473377) = 0,000000
```

figura 28: Contraste de Breusch-Pagan sobre el modelo final MCO A, fuente propia

Analizando el contraste, vemos que nuestro p-valor devuelve un resultado de 0,00000 rechazaremos el contrate de la homocedasticiad de los residuos.

Fijándonos en el gráfico de los residuos, podemos observar que la cantidad de datos atípicos es mayor al resto de los modelos y posee una significación mayor que el resto para la elaboración del modelo. Por otro lado de dispersión de los residuos parece seguir una tendencia, la cual se corregirá posteriormente mediante la inclusión de nuevas variables.

Para finalizar se han estudiado la estabilidad estructural del modelo mediante un contraste de Chow en el individuo que se encuentra en el punto medio de nuestra muestra, así como un estudio de la especificación del modelo mediante un contraste de Ramsey (RESET) sobre todas las variantes. Ámbos contrastes devolvieron un p-valor de prácticamente de 0, por lo que no apoyan nuestra hipótesis nula de significación.

B) Ganancia media respecto a experiencia, educación, sexo y comunidades autónomas

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	816,558	845,207	0,9661	0,3352	
experiencia	4210,95	146,643	28,72	1,06e-72	***
educacion	4112,52	142,912	28,78	7,57e-73	***
Sexo	-7703,78	487,103	-15,82	5,04e-37	***
Comunidad	666,788	296,308	2,250	0,0255	**

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	2,40e+09	D.T. de la regresión	3473,311
R-cuadrado	0,905841	R-cuadrado corregido	0,903949
F(4, 199)	478,6136	Valor p (de F)	7,2e-101
Log-verosimilitud	-1950,116	Criterio de Akaike	3910,233
Criterio de Schwarz	3926,824	Crit. de Hannan-Quinn	3916,944

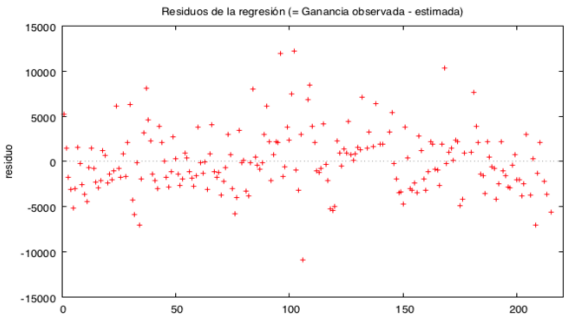


figura 29: Modelo MCO ganancia media con respecto a la experiencia, educación, sexo y comunidad autónoma, fuente: elaboración propia

Observando este modelo donde:

Ganancia media = $\beta o + \beta_1$Experiencia + β_2Educación + β_3Sexo + β_4Comunidad + ε

Finalmente, se ha elaborado un último modelo aplicando las variables codificadas nuevamente, en lugar de sus variables ficticias correspondientes, con la finalidad de comparar ambos modelos.

La ganancia media de los individuos siendo la ganancia base **βo**: 816,558€, la cual aumentará positivamente **β_1:** 4210,95€ por de cada uno de los distintos niveles de experiencia en los que se encuentre el individuo, así como esta aumentara positivamente **β_2:** 4112,56€ en función a que nivel de educación el individuo llegue. Por otro lado la variable independiente del sexo tiene un efecto negativo **β_3:** -7703,78€ sobre la ganancia media siendo el individuo una mujer. Finalmente el echo de que nuestro individuo pertenezca a la comunidad autónoma de Madrid o a Cataluña aumentara su ganancia media en **β_4:** 666,788€ **β_4**

Atendiendo a la desviación típica de las **β_1** , **β_2** , **β_3** y **β_4** de este modelo, son significativamente menor que su respectivos coeficientes por lo que los resultados de este modelo parece ser, que tienen cierto grado de precisión, Además, si observamos la $\overline{R}^2$, observamos que el conjunto de nuestras variables independientes, explican el 90% de la variabilidad de la ganancia media del modelo, un 3% inferior al modelo final A. Si atendemos a los criterios de Akaike (AIC) y de Schwarz (BIC) los cuales ofrecen una estimación relativa de la información perdida en la elaboración del modelo de 3916,944€

y 3926,824€ respectivamente, por lo que mediante la inclusión de la variable independiente **X4:** Comunidad autónoma**,** no perderemos ninguna información adicional en comparación a nuestro modelos en los apartados anteriores.

Contrastando la hipótesis $\mathbf{H_o : \beta_1 = \beta_2 = \ldots. = \beta_k = 0}$ frente a $\mathbf{H_1 :}$ **al menos un** $\beta_j \neq 0$, atenderemos al p-valor del estadístico F del modelo así como a su p-valor, los cuales nos llevan a rechazar la hipótesis de nula de que las variable independientes no sean significativas para nuestro modelo.

Para estudiar la normalidad de los residuos, hemos realizado un grafico Q-Q. En el cual se aprecia, la mayoría de los residuos caen bastante alineados en la media a excepción de los puntos extremos tanto inferiores como superiores lo cual supone un problema de normalidad. Además, mediante un test de normalidad de los residuos cuyo estadístico para el contraste de normalidad obtenido es de Chi-Cuadrado(2) = 14,020[0,0009] lo cual nos indicara nuevamente un problema de normalidad del modelo.

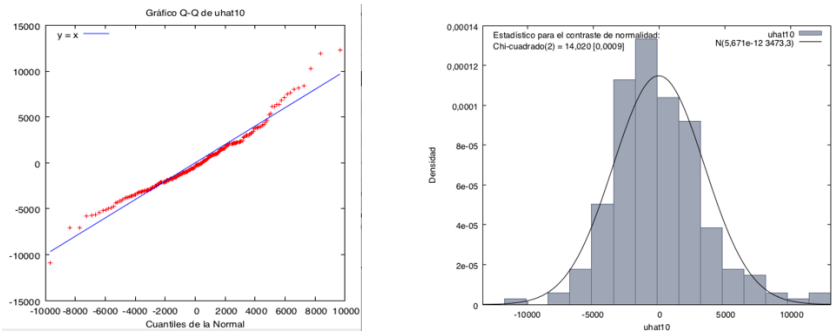


figura 30: Contraste de la normalidad sobre el modelo final MCO B, fuente: elaboración propia

Para estudiar la heterocedasticidad del modelo en primer lugar se realizara un contraste de Breusch-Pagan donde estudiaremos la varianza estimada de los residuos en la regresión

```
Suma de cuadrados explicada = 45,3579

Estadístico de contraste: LM = 22,678928,
con valor p  = P(Chi-cuadrado(4) > 22,678928) = 0,000147
```

figura 31: Contraste de Breusch-Pagan sobre el modelo final MCO B, fuente: elaboración propia

Analizando el contraste, vemos que nuestro p-valor devuelve un resultado de 0,000147 rechazaremos el contrate de la homocedasticiad de los residuos.

Fijándonos en el gráfico de los residuos, a simple vista, se puede observar como estos empeoran considerablemente en comparación con nuestro modelo final anterior. Sin embargo mejora la estimación de los coeficientes. También podemos observar que la cantidad de datos atípicos es mínima y poco significativos para la elaboración del modelo. Por otro lado de dispersión de los residuos parece seguir una tendencia, la cual se corregirá posteriormente mediante la inclusión de nuevas variables.

Para finalizar se han estudiado la estabilidad estructural del modelo mediante un contraste de Chow en el individuo que se encuentra en el punto medio de nuestra muestra, así como un estudio de la especificación del modelo mediante un contraste de Ramsey (RESET) sobre todas las variantes. Ambos contrastes devolvieron un p-valor de prácticamente de 0, por lo que no apoyan nuestra hipótesis nula de significación.

Teniendo en cuenta el modelo de mínimos cuadrados estimado con anterioridad, se ha determinado que la **X4:**comunidad autónoma de nuestro individuo será una variable significativa. Para analizar mas a fondo la evolución de la ganancia media en función de la comunidad autónoma del individuo, se han realizado dos modelos adicionales:

a) **Ganancia media** = $\beta o + \beta_1$Experiencia + β_2Educación + β_3Sexo + β_4Madrid + ε

b) **Ganancia media** = $\beta o + \beta_1$Experiencia + β_2Educación + β_3Sexo + β_4Cataluña + ε

tras analizar los resultados principales de cada modelo, se ha podido concluir que para la Comunidad autónoma de Madrid es significativo construir un modelo propio, mientras que para Cataluña los datos indican que no es significativo pues únicamente aportará 143€ a la ganancia media individual de un trabajador. El modelo propio de la comunidad autónoma de Madrid, será el siguiente:

Variable dependiente: Ganancia

	Coeficiente	Desv. Típica	Estadístico t	valor p	
const	970,648	788,942	1,230	0,2200	
experiencia	4198,43	144,204	29,11	1,16e-73	***
educacion	4103,24	140,757	29,15	9,43e-74	***
Sexo	-7701,44	479,643	-16,06	9,25e-38	***
Mad	1725,23	510,050	3,382	0,0009	***

Media de la vble. dep.	26320,10	D.T. de la vble. dep.	11207,07
Suma de cuad. residuos	2,33e+09	D.T. de la regresión	3420,279
R-cuadrado	0,908695	R-cuadrado corregido	0,906859
F(4, 199)	495,1253	Valor p (de F)	3,4e-102
Log-verosimilitud	-1946,978	Criterio de Akaike	3903,955
Criterio de Schwarz	3920,546	Crit. de Hannan-Quinn	3910,667

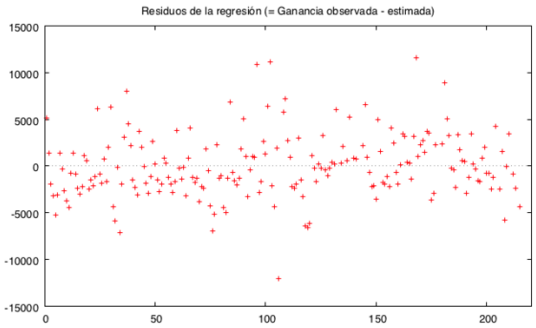


figura 32:Modelo MCO ganancia media con respecto a la experiencia educación, sexo y la comunidad autónoma de Madrid, fuente: elaboración propia

Donde se observa la nueva β_4 , que el trabajador pertenezca a la comunidad autónoma de Madrid con un aumento en la ganancia media de 1725,23€, sin embargo, realizando el resto de criterios, esta nueva variable no parece aportar mayor cantidad de información al modelo.

Imputación de Valores ausentes.

El objetivo de este apartado, será imputar los modelos que faltan por lo que se ha empleado nuestro modelo final B. Esto se debe a que se intentara reducir las dispersiones de los residuos de la ganancia media, pues pese a que el modelo más correcto es nuestro modelo final A, el cual tiene en cuenta cada uno de los distintos niveles de **X1:** Experiencia y **X2:** Educación, las dispersiones son hasta 7 veces superiores que las del modelo sin la inclusión de variables ficticias. Por esto, emplearemos nuestro modelo final B, pues se espera obtener un valor mas ajustado.

El principal motivo de obtener dispersiones tan elevadas en el modelo final A, se encuentra en que como el numero de grados de libertad es mucho menor con las ficticias y la n no varia con respecto a ambos modelos, los errores estándar serán menores al utilizar menos variables.

Tras analizar los distintos modelos, seleccionaremos aquellos modelos cuyas variables independientes permiten una mejor adecuación del modelo. Se emplearan dichos modelos con la finalidad de imputar los valores de la posible ganancia media de nuestros 12 valores ausentes. Los modelos que emplearemos serán:

A) **Ganancia media =** $\beta o + \beta_1 Experiencia + \beta_2 Educación + \beta_3 Sexo + \beta_4 Comunidad + \varepsilon$

B) **Ganancia media =** $\beta o + \beta_1 Experiencia + \beta_2 Educación + \beta_3 Sexo + \beta_4 Madrid + \varepsilon$

Pese a existir distintas técnicas de imputación en este estudio se emplearan los modelos de mínimos cuadrados obtenidos como modelos óptimo, pues pese a necesitar la inclusión de variables independientes adicionales para el total de la explicación de la variabilidad de la ganancia media, de nuestros modelos explican un 90%, por lo que los resultados podrían ser bastante adecuados. Los valores ausentes a imputar serán los siguientes:

Cataluña, donde emplearemos nuestro modelo A , puesto que un trabajador pertenezca a Cataluña no resulta significativo, por lo que no contara con modelo propio:

i) Mujer con un nivel de educación primaria y entre 21 y 29 años de experiencia: 18946,836€

ii) Mujer con un nivel de educación primaria con 30 o mas años de experiencia: 23167,786€

iii) Mujer con un nivel de educación de enseñanza de formación profesional de grado superior y similar con 30 o mas años de experiencia: 35495,346€

iv) Mujer con un nivel de educación de licenciados y similares, y doctores universitarios con 30 o mas años de experiencia: 43720,386€

v) Hombre con un nivel de educación de enseñanza de formación profesional de grado superior y similar con 30 o mas años de experiencia: 43199,156€

vi) Hombre con un nivel de educación de licenciados y similares, y doctores universitarios con 30 o mas años de experiencia: 51424,166€

Comunidad autónoma de Madrid, donde emplearemos nuestro modelo B, puesto que un trabajador pertenezca a la comunidad autónoma de Madrid tiene una incidencia significativa en la ganancia media de un trabajador:

i) Mujer con un nivel de educación primaria y entre 21 y 29 años de experiencia: 20089,828€

ii) Mujer con un nivel de educación primaria con 30 o mas años de experiencia: 24288,258€

iii) Mujer con un nivel de educación de enseñanza de formación profesional de grado superior y similar con entre 21 y 29 años de experiencia: 32399,548€

iv) Mujer con un nivel de educación de enseñanza de formación profesional de grado superior y similar con 30 o mas años de experiencia: 36597,978€

v) Mujer con un nivel de educación de diplomados universitarios y similares con 30 o mas años de experiencia: 40701,218€

vi) Hombre con un nivel de educación de diplomados universitarios y similares con 30 o mas años de experiencia: 52505,89€

5. Resultados y conclusiones

Tras realizar un análisis exhaustivo de los factores considerados en el estudio, en primer lugar, se ha llegado a la conclusión de que todas y cada una de las variables independientes, son significativas a la hora de construir la ganancia media de un individuo.

En cuanto a la variable independiente **X1:** Experiencia, se podría concluir que los años de experiencia que un trabajador posea, afectaran positivamente a la ganancia media este de manera progresiva lo cual se puede observar en el modelo final A . Tras el estudio de la variable independiente **X2:** Educación, podemos decir que a día de hoy podría seguir siendo un factor determinante pues desde la educación primaria hasta llegar al nivel de licenciados y similares y doctores universitarios, la ganancia media de un individuo aumentará hasta un total de un 45%. Tras estudiar la variable independiente **X3:** Sexo, se ha concluido a través del estudio que efectivamente existe una brecha salarial en función del género de la persona, la cual supone una diferencia negativa de un 27% para las mujeres con respecto a los hombres, sin embargo en los hombres existe una desviación inexplicada de la ganancia media un 18% superior a la de las mujeres. Finalmente, con respecto a la variable independiente **X4:** Comunidad autónoma, se podría argumentar, que todos aquellos trabajadores pertenecientes a la Comunidad de Madrid o a Cataluña obtendrán una ganancia media superior al resto de los trabajadores de la península, sin embargo esta ganancia media solamente será significativamente superior en la Comunidad autónoma de Madrid, donde la diferencia es de 1725€ mientras que en Cataluña supone una diferencia de 143€ con respecto al resto.

Con respecto a la elaboración y el estudio de los distintos modelos elaborados, el modelo final A:

$$\textbf{Ganancia media} = \beta_o + \beta_1 de1a3 + \beta_2 de4a10 + \beta_3 de11a20 + \beta_4 de21a29 + \beta_5 masde30 + \beta_6 Ed2 + \beta_7 Ed3 + \beta_8 Ed4 + \beta_8 Ed5 + \beta_{10} Ed6\beta_7 + \beta_{11} Sexo + \beta_{12} Comunidad\ autónoma + \varepsilon$$

Es el modelo, que mejor ha resultado adaptarse a nuestro objetivo, pues pese a no ser un modelo perfecto, no solo es el modelo que mayor variabilidad de la ganancia media explica, sino que también es aquel que mas detalladamente explica la ganancia media en

función de cada uno de los distintos valores posibles para nuestras distintas variables independientes **X1:** Experiencia y **X2:** Educación. Sin embargo, este modelo, debido al uso de un número tan elevado de variables independientes, causa que la desviación típica de los coeficientes estimados individuales sea mucho mayor en relación al modelo final obtenido en el cual se mantiene la codificación de las variables independientes como esta ha sido extraída del INE. Como ya se ha comentado con anterioridad, esta alta desviación típica se debe al gran aumento del número de variables explicativas, unido a que el número de individuos se mantiene estable en todos los modelos, por lo que posiblemente será menos exacto el calculo de los trabajadores.

Otra de las ventajas de este modelo frente al resto de modelos estudiados, reside en el hecho de que este, es el único que acepta la normalidad de los errores sin la necesidad de convertir el modelo y que además, tiene una perdida de información inferior durante la elaboración del modelo. Sin embargo, no ha sido posible encontrar un modelo en el cual exista una homocedasticidad de los residuos o que cumpla con los contrastes de estabilidad estructural así como de especificación del modelo, con los datos extraídos.

Durante la Imputación de los valores ausentes, tras emplear el modelo final B, hemos obtenido unos valores, que aparentemente se ajustarían correctamente a los valores, mientras que el modelo final A tendía a devolver unos valores mas extremos.

Otros autores, además de la elaboración de los modelos, probaron transformaciones complejas de los mismos con la finalidad de obtener un mejor ajuste. En este estudio, al tratarse de un TFG y por cuestiones de tiempo, no se ha tratado la conversión del modelo final B, sin embargo, una manera de proceder con el estudio, podría basarse en la transformación del modelo.

6. Bibliografía

1. Información Teórica

- Newbold, P., W. Carlson and B. Thorne. *Statistics for Business and Economics*. Pearson-Prentice Hall.
- Novales Cinca, A.

  Análisis de Regresión. Apuntes de Econometría Superior.

  https://www.ucm.es/data/cont/docs/518-2013-11-13-Analisis%20de%20Regresion.pdf

  (Visitado: Junio 2017)
- Montgomery, D.C., E.A. Peck and G.G. Vining (2001, 3rd ed). Introduction to Linear Regression
- Verbeek, M. A guide to modern Econometrics, John Wiley and Sons.

2. Guias de Gretl

- J. Pérez, C

  Guía rápida Gretl

  http://www.eco.uc3m.es/~cavelas/Econometrial/Guia%20rapida%20de%20gretl.pdf

  (Visitado: Marzo-Mayo 2017)

- Cottrell, A

  Guía del usuario de Gretl

  http://ocw.uniovi.es/pluginfile.php/2990/mod_resource/content/1/T_1C%2CA_668/Gretl/Guia_Gretl.pdf

  (Visitado: Marzo-Mayo 2017)

3. Base de datos

- Instituto Nacional de estadística

  Encuesta cuatrienal de estructura salarial.

  http://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177025&menu=resultados&idp=1254735976596

  (visitado: Marzo 2017)