1. INTRODUCCCIÓN

En Colombia, de acuerdo con la Dirección de Impuestos y Aduanas Nacionales (DIAN) la evasión de impuestos es cercana a los $65 billones de pesos, estimaciones indican que la evasión esta cercana a 5,4 puntos del PIB, de los cuales 0,7 puntos corresponden evasión de impuestos de personas naturales. [[1]](#footnote-1) Algunos trabajos han estimado el subreporte de ingresos en 1.66 veces mas que el reporte inicial. (Rocha (2014)). Por lo anterior, usando metodologías aprendizaje estadístico, aun incipientes en economía, estimamos un modelo de predicción de salarios basado en información de la GEIH para el año 2018 en Bogota. Este modelo predictivo es potencialmente útil para encontrar casos de fraude fiscal y adicionalemente, apoyar la toma de decisiones de política para intervenir familias vulnerables.

1. DATOS

Usamos datos de la Gran Encuesta Integrada de Hogares GEIH en Bogota para el año 2018, la cual contiene información sobre las condiciones de empleo de las personas, características generales de la población como sexo, edad, estado civil y nivel educativo. Contiene también información sobre fuentes de ingreso.

El proceso de adquisición de la información se realizo mediante el scraping la información contenida en el sitio web <https://ignaciomsarmiento.github.io/GEIH2018_sample/>, sobre esta información no existe ningún tipo de restricción. El sitio web contiene información extraída y clasificada de la GEIH con variables originales y construidas que proporcionan datos relevantes para el análisis. Se realizo un web scrapping de los 10 bloques de información los cuales estaban particionados en diferentes URL´s. Estos bloques de información se consolidaron para terminar con una sola base de datos integrada de la información relevante.

* 1. *Limpieza de datos*

Considerando que la GEIH en lo concerniente a los procedimientos de conformación, depuración, imputación y empalme del ingreso contiene 137 variables con las características de los individuos pero sobre todo aproximadamente 40 variables que contienen información relacionada a ingreso, es necesario realizar un procedimiento de limpieza de la base en el cual se restringe el universo de datos en aquellos relevantes para los objetivos del trabajo, en particular usamos procedimientos para limpiar la base y dejar únicamente las siguientes variables: *i) Sexo*; Variable dicótoma que toma valor 1 si es hombre y 0 si es mujer. Usamos esta variable con la finalidad de identificar brechas salariales entre hombres y mujeres. *ii) Edad;* contiene la edad en años del individuo, la muestra se analiza para aquellos mayores de 18 años. Esta variable se usa para identificar patrones predictivos entre la edad y el salario de las personas *iii) Educación*; contiene e máximo grado escolar alcanzado iv) *Tamaño empresa;* que muestra el numero de empleados de la empresa en que el individuo labora v) *Salario;* la variable de salario usada es el salario mensual por hora.

* 1. *Estadísticas descriptivas*

Como análisis base del objetivo del trabajo podemos analizar las estadísticas descriptivas de las variables de interés. La Tabla 1. Indica, respecto al sexo, que la muestra esta balanceada en la proporción de hombres y mujeres encuestados, la edad promedio esta en aproximadamente 36 años con individuos de hasta 86, recordemos que limitamos el estudio a los mayores de 18 años, que es la mayoría de edad en Colombia. En cuanto a la educación, la población Bogotana en promedio tiene un alto nivel educativo, pues los individuos promedian 6.08 de Educación en una escala que tiene su máximo en 7, por tanto la mayoría de la población encuestada tiene niveles de educación altos respecto a la escala determinada.

Tabla 1.

Variables incluidas en la base de Datos

==============================================================================

Statistic N Mean St. Dev. Min Max

------------------------------------------------------------------------------

sex 9,967 0.506 0.500 0 1

edad 9,967 36.078 11.942 18 86

maxEducLevel 9,967 6.084 1.112 1 7

hoursWorkUsual 9,967 48.027 12.336 1 130

salario\_mensual 9,967 1,524,879.000 2,040,110.000 10,000 30,200,000

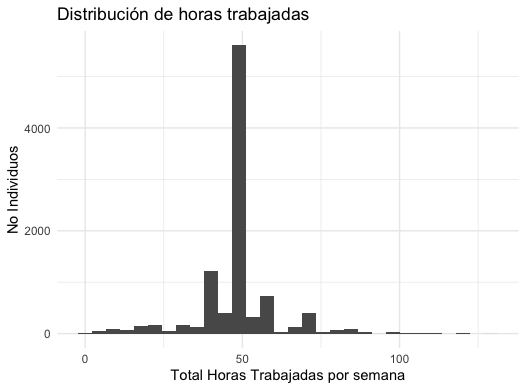
salario\_mensual\_hora 9,967 8,314.517 11,832.650 208.333 312,500.000

log\_salario\_mensual\_hora 9,967 8.681 0.705 5.339 12.652

------------------------------------------------------------------------------

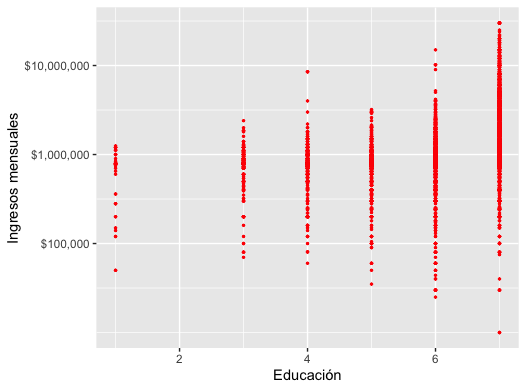
Respecto a las horas promedio trabajadas por semana coinciden con la jornada laboral legal, ver Grafico 1. Con algunos individuos (outliers) que trabajan mucho mas ( 130 horas) y otros mucho menos (1 hora).

Grafico 1.



Respecto a el salario mensual promedio de los individuos, este se ubico en $ 1.524.879, esto quiere decir, que el salario promedio de los bogotanos para el año 2018 fue aproximadamente 1.75 veces mayor al salario mínimo mensual vigente para ese año, el cual se ubicaba en $869.453 incluyendo auxilio de transporte. No obstante, existe una variabilidad importante en los datos, pues la desviación estándar es 1.33 veces mayor que el salario promedio del grupo, esta variabilidad de los ingresos es bien ilustrada por el grafico 2.

Grafico 2. Relación de Ingresos mensuales y Educación



Nuestro principal objetivo es encontrar un modelo predictivo con el potencial de darnos valores predichos de los salarios de forma precisa y de esta manera usar el modelo para diferentes fines importantes, como encontrar casos de fraude o subreporte de ingresos, o a su vez, determinar aquellos individuos vulnerables donde se podría eventualmente focalizar la asistencia. Dentro de nuestras variables, características como la edad, el sexo y la educación han sido largamente estudiadas en economía laboral como características claves que explican el salario de los individuos y que además tienen patrones predictibles, como es el caso de la edad.

1. AGE-WAGE PROFILE

El salario es el pago regular de dinero de acuerdo a las horas trabajadas pactado mediante un contrato. La evidencia empírica sobre los efectos de la edad en el salario encuentra que, a medida que aumenta la edad, el salario también crece hasta cierta edad, a partir del cual comienza su descenso, cuya especificación corresponde al siguiente modelo:

Donde: es el logaritmo natural de salario por hora y es la edad.

En esta sección se estima esta especificación del salario. La variable dependiente corresponde al logaritmo del salario por hora, dicha transformación busca reducir el rango de la variable a una unidad más pequeña que el salario en niveles, y facilita su interpretación. La variable pre determinada es la edad. Los datos son obtenidos de la Gran Encuesta Integrada de Hogares (GEIH) correspondiente a Bogotá.

Tabla 2.

Estadisticas Descriptivas

Statistic N Min Pctl(25) Median Mean Pctl(75) St. Dev. Max

Salario 9,967 208.3 4,069.0 4,846.9 8,314.5 7,552.1 11,832.6 312,500.0

Edad 9,967 18 26 34 36.1 44 11.9 86

El salario promedio en la ciudad de Bogota ascienda a $ 8,314.1 por hora, cuya variabilidad es elevada, revelada por la desviación estándar de $ 11,832.1 y los cuartiles 1 y 3, con un rango de $ 3,483.115. La edad promedio de las personas mayores a 18 años, que trabajan está en 36 años. En el Grafico 3 se observa que la amplitud de variación del salario para cada edad es importante, que está asociada con las características individuales de cada persona. Tambien se puede notar la presencia de outliers.

Grafico 3.



* 1. *Regresión*

Resultados de la Regresion

=================================================

Dependent variable:

-----------------------------

Logaritmo del Salario

-------------------------------------------------

edad 0.0595914\*\*\*

(0.0035464)

edad al cuadrado -0.0006548\*\*\*

(0.0000441)

Constant 7.4769350\*\*\*

(0.0663397)

-------------------------------------------------

Observations 9,967

R2 0.0379413

Adjusted R2 0.0377482

Residual Std. Error 0.6920231 (df = 9964)

F Statistic 196.4784000\*\*\* (df = 2; 9964)

=================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

* 1. *Significancia e interpretación*

*Significancia:* El modelo estimado muestra que la variable edad es estadísticamente significativa a un nivel de significancia del 1% (p-value<0.01) tanto la parte lineal como no lineal (cuadrático). Es decir, la estimación confirma una relación no lineal entre la edad y el salario. También, el estadístico F muestra que la especificación no lineal de la relación salario-edad es estadísticamente significativa de manera conjunta al 1% de nivel de significancia (p-value<0.01).

*Interpretación:* Al mostrar el modelo una relación no lineal entre la edad y el salario, para ver el efecto que tiene la edad en el salario es necesario derivar el logaritmo del salario respeto a la edad, lo cual resulta en una semi elasticidad.

El cálculo de la semi elasticidad requiere utilizar el promedio de la edad de los bogotanos, que es años; también puede ser evaluada para diferentes edades.

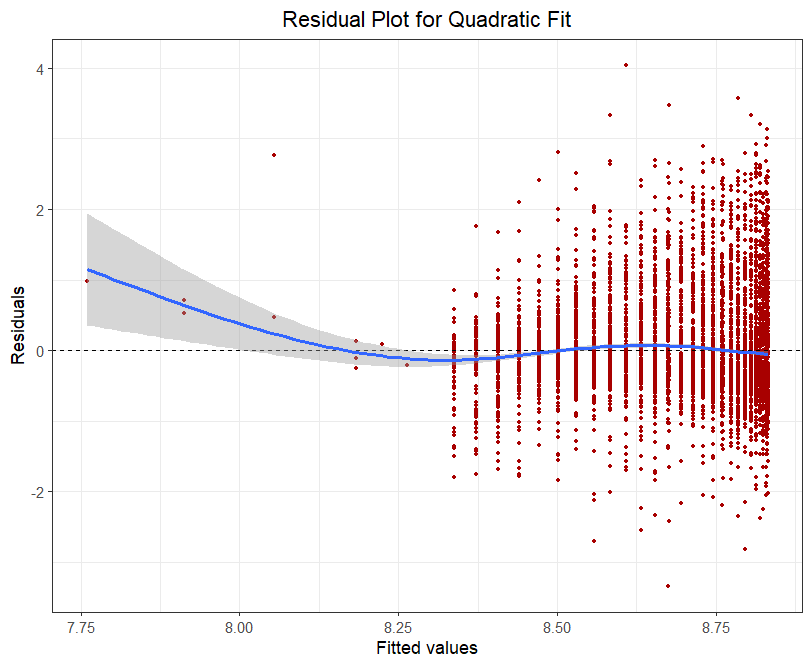
Esta semi elasticidad nos indica que una persona que tiene una edad promedio de 36.1 años, ante el incremento de un año más en su edad, su salario aumenta 1.23%. Pero esta semi elasticidad cambia según la edad. Es decir siendo joven, cuanta más edad tenga, su salario aumentará cada menos hasta llegar a cierta edad (rendimientos marginales decrecientes) y comenzará a disminuir conforme se haga más adulto. Por ejemplo, para alguien que tiene 18 años, tener un año más implica un incremento en su salario de 3.6% y para otro que tiene una edad de 60 años, tener un año más se traduce en una disminución del 1.9% de su salario.

* 1. *Ajuste del modelo*

El ajuste del modelo medido a partir del R-cuadrado nos dice cuánto de la varianza del logaritmo del salario es explicado por el modelo. En este caso un R-cuadrado de 0.0379413 significa que solo el 3.79% de la varianza del salario es explicado por el modelo. Este es bastante bajo y sugiere que el modelo no tiene buen ajuste. Sin embargo, esto no necesariamente es malo, pues los predictores son estadísticamente significativos y los coeficientes del modelo aun representan el cambio en el salario frente a cambios en la edad. Por lo tanto, se justifica en parte el ajuste del modelo.

Otra manera de ver es a partir del gráfico de los errores del modelo. Este grafico muestra que los errores no se alejan sistemáticamente de cero y no tienen un patrón (línea azul). En consecuencia, sugiere que la edad al cuadrado mejora el ajuste a los datos.

Grafico 3.

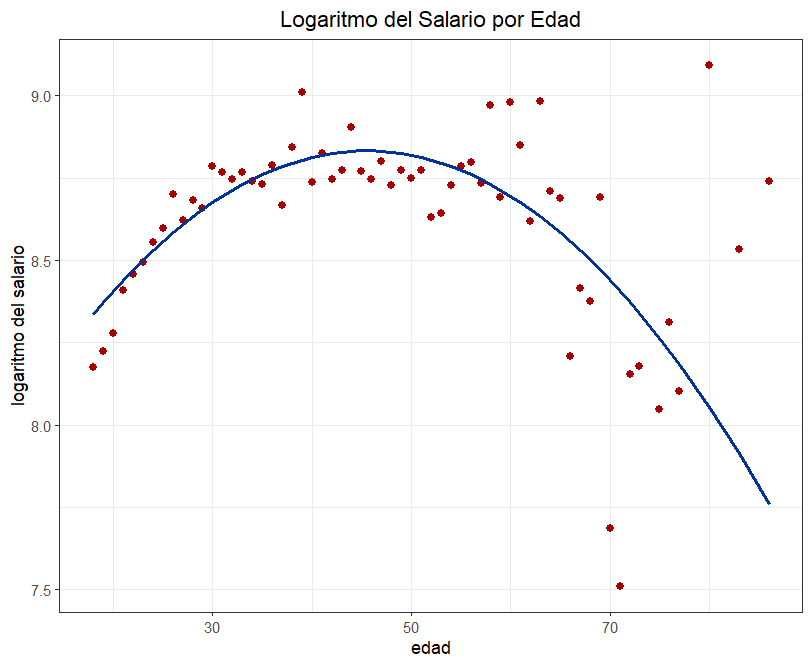


* 1. *Edad-salario estimado, “peak age” e intervalos de confianza*

La relación entre la edad y el salario es no lineal como se ve en el gráfico, en una primera etapa el salario crece conforme aumenta la edad del trabajador, hasta llegar a los 46 años (45.50549), edad a partir del cual entra a una segunda etapa, cuando su salario comienza a descender.

El dato 45.50549de edad (*peak age*) se ha calculado usando Bootstrap para obtener 10000 estimaciones de la edad, basado en 10000 muestras con reemplazo sobre la muestra original. El Bootstrap genera una distribución cercana a la normal de la edad, tal como se observa en el histograma y el gráfico Cuantil-Cuantil.

Grafico 4.



Por último, el intervalo de confianza de 95% para la edad es IC:[44.04, 46.91], en otras palabras, la edad a partir del cual el salario comienza a caer estará entre 44 y 47 años para los bogotanos.

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:

boot(data = base\_nueva, statistic = eta\_reglsalario\_mes\_hora\_fn,

R = 10000)

Bootstrap Statistics :

original bias std. error

t1\* 45.50549 0.03994752 0.7383577

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 10000 bootstrap replicates

CALL :

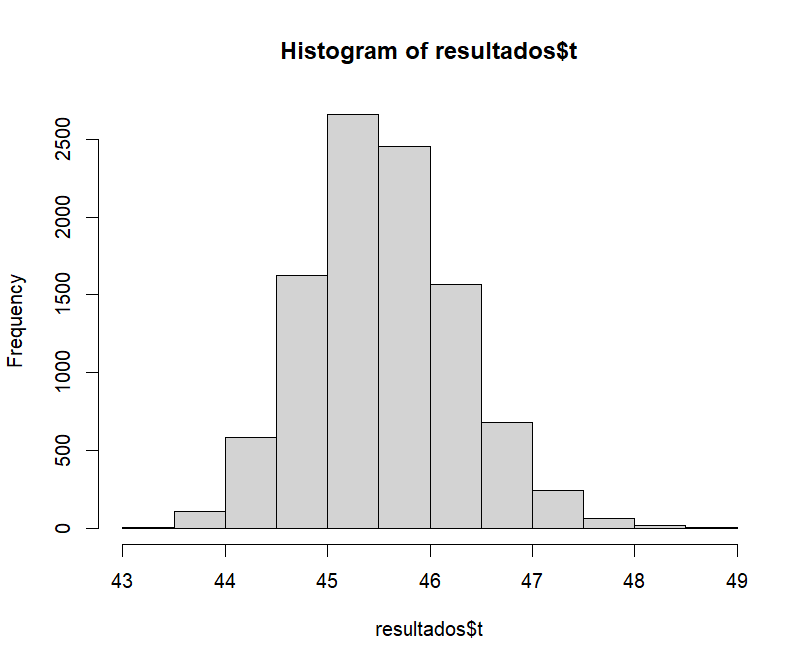
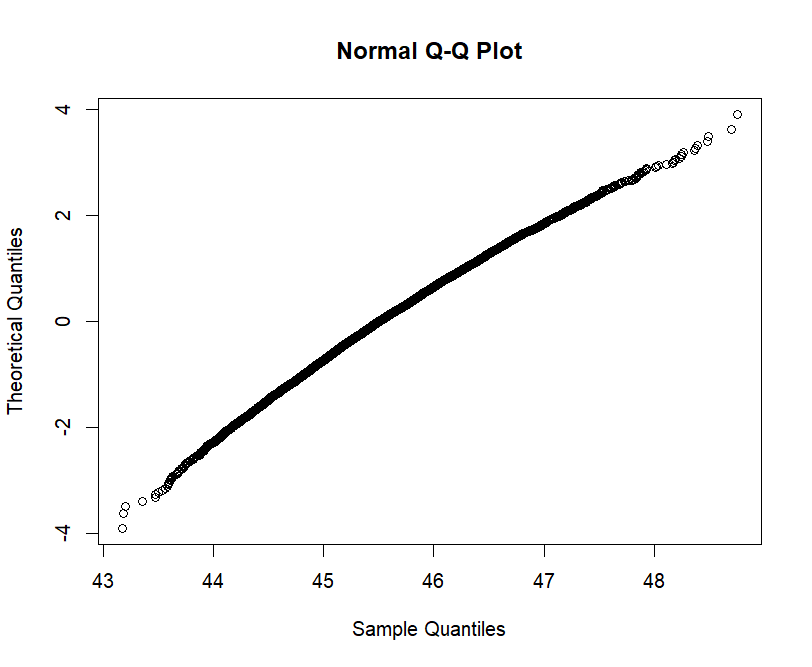
boot.ci(boot.out = resultados, type = c("norm", "basic"))

Intervals :

Level Normal Basic

95% (44.02, 46.91 ) (43.90, 46.82 )

Calculations and Intervals on Original Scale

1. THE GENDER EARNINGS GAP
2. PREDICTING EARNINGS
   1. Para este punto se dividide la muestra en dos, 70% para entrenamiento y 30% para prueba. Con el primero obtenemos los estimadores que mejor se ajustan, con el modelo estimado predecimos y evaluamos el mejor modelo con el más bajo error de predicción fuera de la muestra.
   2. Se contaba con tres modelos previamente estimados de edad - salario, sexo - salario, y sexo – salario con controles. Adicional a estos se estiman 7 modelos más con relaciones no lineales de tipo polinómicas, de raíz cuadrada y de interacción entre los predictores, de los mimos se calculan los MSE para evaluar su desempeño predictivo mediante el error de predicción (MSE) fuera de la muestra (prueba). Los resultados se muestran en la siguiente tabla.

model MSE

1 model1 0.4852

2 model2 0.5015

3 model3 0.3400

4 model4 0.4167

5 model5 0.3390

6 model6 0.3375

7 model7 0.3365

8 model8 0.3195

9 model9 0.3369

10 model10 0.3361

* 1. Evaluación de resultados

1. De los tres primeros modelos en las dos secciones anteriores, aquel que tiene un mejor desempeño fuera de la muestra es el salario explicado por la no linealidad de la edad, sexo, relación de trabajo, máximo nivel educativo y tamaño de la empresa, cuyo MSE es 0.34.

Al compararlos con los modelos adicionales, el **modelo 8** que establece relaciones no lineales de tipo raíz cuadrada y su interacción entre y entre los otros predictores, es el que muestra el menor error predictivo fuera de la muestra con un **MSE de 0.3195**. Así, este modelo es superior a los otros nueve, pues cuando complejizamos más los modelos con relaciones polinómicas de grado cinco y ocho, como en los modelos 9 y 10, el error predictivo comienza a aumentar.

Las estimaciones muestran que mientras más complejidad haya en el modelo el performance mejor y a partir del modelo 9 el performance es menor, medido con el MSE.

1. El modelo con el menor error predictivo es el modelo 8, cuya especificación es siguiente:

Regresion con mejor error predictivo (model8)

====================================================

Dependent variable:

---------------------------

Logaritmo del Salario

----------------------------------------------------

edad 0.132\*\*\*

(0.041)

sqrt(edad) -1.838\*\*\*

(0.526)

female -0.512

(0.524)

relab -0.995\*\*

(0.503)

tam\_empresa -0.324\*\*\*

(0.095)

maxEducLevel -1.113\*\*\*

(0.244)

edad:female -0.003

(0.014)

sqrt(edad):female 0.034

(0.174)

edad:maxEducLevel -0.025\*\*\*

(0.006)

sqrt(edad):maxEducLevel 0.340\*\*\*

(0.078)

edad:relab -0.025\*

(0.013)

sqrt(edad):relab 0.339\*\*

(0.165)

edad:tam\_empresa 0.002

(0.003)

sqrt(edad):tam\_empresa 0.006

(0.031)

female:relab 0.040

(0.046)

female:maxEducLevel 0.047\*\*\*

(0.015)

female:tam\_empresa 0.004

(0.006)

maxEducLevel:tam\_empresa 0.042\*\*\*

(0.003)

maxEducLevel:relab -0.028\*\*

(0.013)

relab:tam\_empresa 0.033\*\*\*

(0.006)

Constant 14.473\*\*\*

(1.643)

----------------------------------------------------

Observations 6,977

R2 0.356

Adjusted R2 0.354

Residual Std. Error 0.566 (df = 6956)

F Statistic 192.082\*\*\* (df = 20; 6956)

====================================================

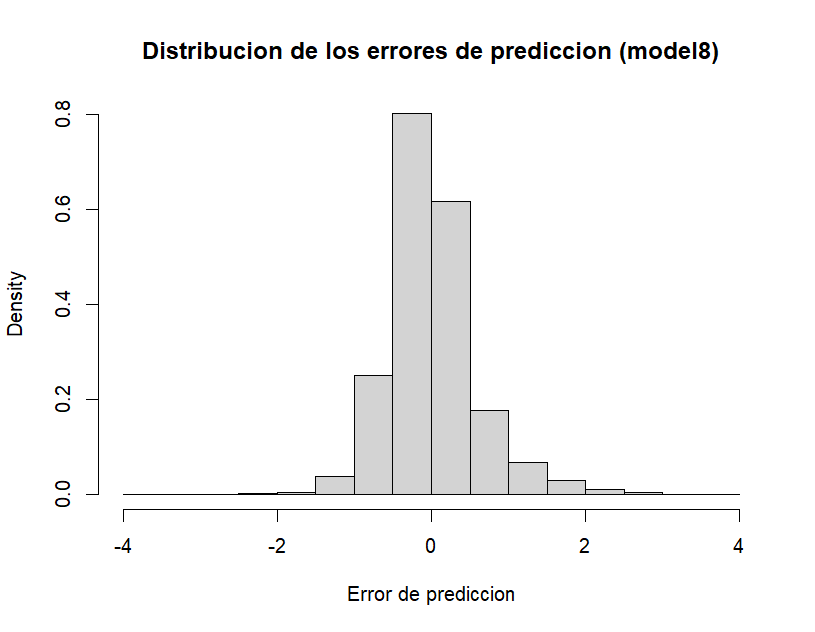
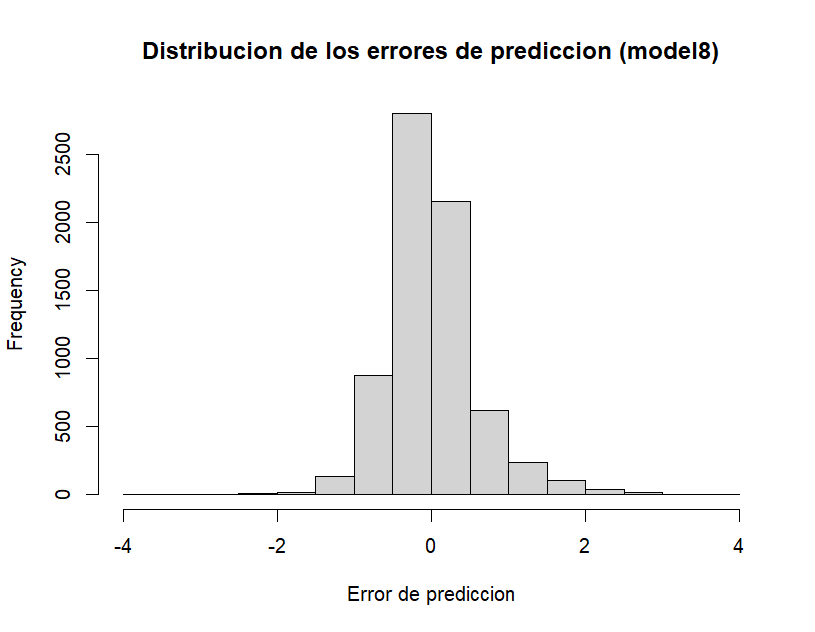
Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

La mayoría de predictores considerados es estadísticamente significativa, con excepción de si es mujer y su interacción con la edad que no son estadísticamente significativas, al igual que la edad con el tamaño de la empresa. Por su parte, el ajuste del modelo, medido por el R-cuadrado mejora a 0.356, lo que significa que las variables independientes, expresadas en no linealidades e interacciones, explican en un 35.6% el comportamiento del logaritmo del salario. Además, predice mejor fuera de la muestra por presentar el MSE más bajo frente a los otros nueve modelos.

1. Distribución de los errores de predicción del mejor modelo

El histograma muestra que el modelo para la mayor parte de la muestra de prueba predice muy buen, pues la mayoría de los datos de concentran alrededor de cero, es decir considerando determinadas variables que caracterizan a cada persona, el ingreso estimado por el modelo es similar al salario declarado. Sin embargo, la distribución presenta colas, lo que significa en este caso, es que el salario estimado es muy diferente del salario observado (dato real).

La cola izquierda de la distribución refleja a aquellas personas que manifiestan percibir salarios por debajo de lo que en realidad podrían tener dadas sus características (salario estimado por el modelo), sugiriendo un comportamiento cuyo incentivo, entre otros, es evadir el pago de impuestos, lo que a la autoridad tributaria debería llamarle la atención y observar con cuidado. En cuanto a la cola derecha de la distribución, están en ella declaraciones de salarios por encima de lo que sus características individuales sugieren de su salario.



* 1. LOOCV

1. Ver:<https://www.larepublica.co/especiales/reforma-tributaria-2022/segun-la-dian-y-el-minhacienda-la-evasion-de-impuestos-es-cercana-a-80-billones-3422523> [↑](#footnote-ref-1)