Problem Set 1 RMarkdown

Irina Vélez - Lucía Fillippo - Daniel Casas - Miguel Victoria

2023-06-21

# Taller 1

El presente informe presenta la solución al Problem Set 1 de la clase Big Data & Machine Learning, con el objetivo de aplicar diversos conceptos y herramientas para la predicción de modelos, el manejo de bases de datos grandes, entre otros. Para el desarrollo del trabajo se utilizó el repositorio GitHub el cual contiene información de la Gran Encuesta Integrada de Hogares - GEIH para el año 2018 (luego en word ponemos este hiperbínculo <https://ignaciomsarmiento.github.io/GEIH2018_sample/>); además, se empleó el sotware Rstudio para el manejo de los datos, generación de resultados y desarrollo del taller, cuyo código se encuentra en el siguiente link: (link del Github con el código)

## 1. Introducción

El valor de ingresos de las personas es un insumo esencial para el desarrollo de políticas públicas, ya sea para identificar a los hogares que tienen la posibilidad de pagar más impuestos, así como para lograr una mejor focalización en aquellos hogares que requieren apoyos sociales; no obstante, en algunas ocasiones los ingresos de las personas no son reportados, de manera que esto se convierte en una barrera para el desarrollo de políticas públicas eficientes. En virtud de lo anterior, poder determinar el valor de los ingresos de las personas se convierte en un gran insumo para el desarrollo de políticas tributarias y sociales, razón por la cual el objetivo principal de este documento es construir un modelo predictivo de los salarios por hora de los individuos, a partir del siguiente modelo:

donde “w” representa el salario por hora y “X” es una matriz de potenciales variables que explican el salario. Como se mencionó previamente, para la creación de este modelo se utilizarán datos de la Gran Encuesta Integrada de Hogares – GEIH del año 2018.

Para importar los datos, es importante conocer qué tipo de página web contiene la información, en este caso, la página web que contiene las bases de datos es dinámica, razón por la cual es pertinente identificar el link principal a partir del cual se realizará la extracción de la información. Para esto se aplicó un código en bucle para que la extracción de la informacíon de las distintas ventanas de la página web fuese más eficiente; además, se realizaron una serie de filtros a los datos, siguiendo las instrucciones dadas, con la finalidad de eliminar las variables “N/A” y considerar únicamente a los individuos mayores de 18 años.

# Descripción de los datos

## Descripción general

## y\_salary\_m\_hu pet mes age   
## Min. : 151.9 Min. :1 Min. : 1.00 Min. :19.00   
## 1st Qu.: 3797.7 1st Qu.:1 1st Qu.: 4.00 1st Qu.:27.00   
## Median : 4522.3 Median :1 Median : 6.00 Median :34.00   
## Mean : 7984.7 Mean :1 Mean : 6.44 Mean :36.44   
## 3rd Qu.: 7291.7 3rd Qu.:1 3rd Qu.: 9.00 3rd Qu.:45.00   
## Max. :291666.7 Max. :1 Max. :12.00 Max. :86.00   
## sex ocu oficio maxEducLevel totalHoursWorked  
## Min. :0.0000 Min. :1 Min. : 1.00 Min. :1.000 Min. : 1.0   
## 1st Qu.:0.0000 1st Qu.:1 1st Qu.:33.00 1st Qu.:6.000 1st Qu.: 48.0   
## Median :1.0000 Median :1 Median :45.00 Median :6.000 Median : 48.0   
## Mean :0.5017 Mean :1 Mean :48.27 Mean :6.098 Mean : 48.4   
## 3rd Qu.:1.0000 3rd Qu.:1 3rd Qu.:58.00 3rd Qu.:7.000 3rd Qu.: 48.0   
## Max. :1.0000 Max. :1 Max. :99.00 Max. :7.000 Max. :130.0   
## exp   
## Min. : 0.000   
## 1st Qu.: 0.000   
## Median : 2.000   
## Mean : 3.989   
## 3rd Qu.: 5.000   
## Max. :58.000

<<<<<<< HEAD

# ## Análisis descriptivo

De manera general, se identifica que nuestra base de datos está compuesta por 9.784 filas y por 151 columnas. Las variables que hacen parte de la base son:

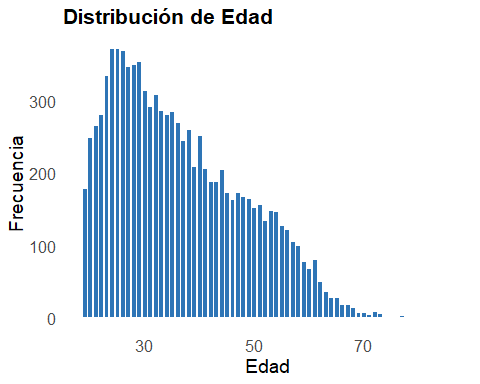
* (y\_salary\_m\_hu): Indica el salario mensual por hora de la persona
* (pet): Indica si la persona hace parte de la Población en Edad de Trabajar - PET
* (mes): Contiene el mes de referencia
* (age): Contiene la edad de la persona
* (sex): Contiene el sexo de la persona
* (ocu): Señala si la persona es ocupada o no ocupada
* (oficio): Indica el oficio de la persona
* (maxEducLevel): Indicar el máximo nivel educativo alcanzado
* (totalHoursWorked): Indica el total de horas trabajadas en el último mes
* (exp): Hace referencia a la experiencia en años que tiene la persona

Dicho lo anterior, a continuación se procede a realizar una descripción más amplia y gráfica de las variables que harán parte del modelo.

======= ## Descripción edad ======= La variable de edad es una variable con números enteros donde se observa un mínimo de 19 años, lo cual guarda sentido con la filtración inicial de los datos, donde se tuvo en cuenta únicamente a las personas mayores de 18 años, y en contraste se identifica un máximo de 86 años. En el primer cuartil de la base se observa una edad de 27 años, en el tercer cuartil una de 45 años. La mediana de los datos es de 34 años, la media es de 36 años y la moda es de 24 años. A continuación se presenta una gráfica de barras que permite observar la distribución de la edad a lo largo de la muestra, donde se identifica que, en general, se cuenta con una población relativamente joven.

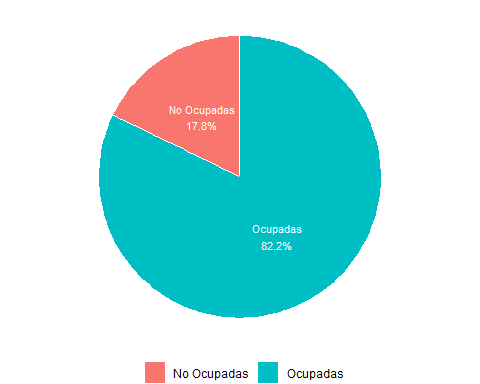
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.00 27.00 34.00 36.44 45.00 86.00

## [1] 24



======= ## Descripción ocupación =======

La variable de ocupación es una dummy que toma el valor de “1” si la persona está ocupada y “0” si no lo está. Inicialmente en la base sin filtrar se tenía una distribución de ocupación con 16.277 personas ocupadas y 3.524 no ocupadas, como la que se observa a continuación:



Ahora bien, como resultado de la filtración inicial llevada a cabo, y siguiendo la instrucción impartida, la base de datos final únicamente se cuenta con personas mayores de 18 años ocupadas lo que da como resultado un total de 9.784 personas ocupadas.

## ocu  
## 1   
## 9784

## Descripción educación

La variable de educación es categórica y tiene la siguiente clasificación por categorías:

1. Ninguno: Que corresponde a aquellas personas sin educación 2 Preescolar: Que corresponde a aquellas personas que solamente terminaron preescolar 3 Basica primaria: Que corresponde a aquellas personas que solo terminaron básica primaria, esto es, los grados de primero a quinto 4 Basica secundaria: Que corresponde a aquellas personas que solo terminaron básica secundaria, esto es, los grados de secto a noveno 5 Media: Que corresponde a aquellas personas que solo terminaron la educación media, esto es, los grados de noveno a once 6 Superior - Universitaria: Que corresponde a aquellas personas que terminaron educación superior y/o universitaria 7 No sabe, No informa

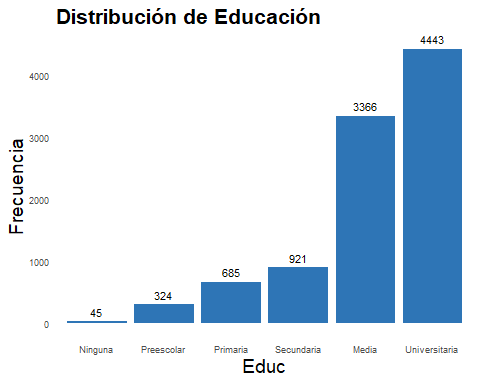
Dicho lo anterior, el análisis de esta variable muestra que la media y la mediana son personas que tienen educación universitaria, lo cual guarda sentido con que sean aquellas que han podido acceder al mercado laboral y encontrarse ocupadas.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 6.000 6.000 6.098 7.000 7.000

## La moda de la variable ocupación corresponde a la clasificación 7

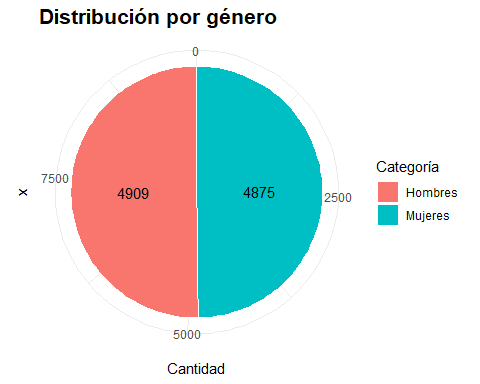
Lo anterior se puede constatar de manera visual en la siguiente gráfica de barras, donde se muestra la manera en que se encuentra distribuida la variable de educación, de acuerdo con la base de datos obtenida:

## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(count)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



## Descripción sexo

La variable de género es dummy y toma el valor de 0 si la persona es mujer y toma el valor de 1 si es hombre. Los datos reflejan un total de 4.909 hombres y 4.875 mujeres, como se muestra en la siguiente gráfica de pie.



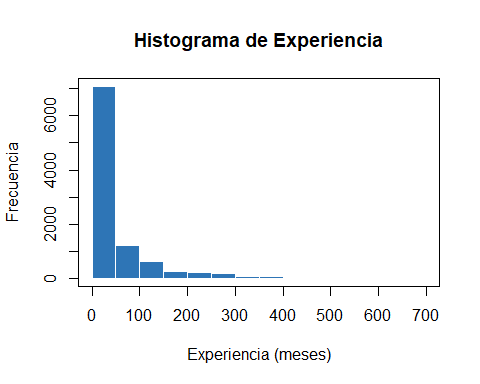
## Descripción de la experiencia

La variable de experiencia es coninua e indica los meses que lleva trabajando la persona en su trabajo actual. La mediana y la moda de esta variable es de 24 meses, la media de 50 meses, y el valor máximo es de 696 meses.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 6.0 24.0 50.2 60.0 696.0

## [1] 24

En virtud de lo anterior, el siguiente histograma refleja la distribución de la experiencia, donde se observa que la mayoría de observaciones se ubican entre 0 y 50 meses, lo cual guarda sentido con que los datos de media señalados previamente.



## Perfil de Salario y Edad

De acuerdo con la forma específicación funcional establecida para el perfil de edad-salario, presentada a continuación:

Se detalla que se escogió la variable edad y edad elevado al cuadrado, precisamente porque de acuerdo a la teoría de la economía laboral, se ha observado empíricamente que la relación entre el salario y la edad de los trabajadores sigue una tendencia en forma de “pico de edad”, perseguido por un decrecimiento en el mediano/largo plazo.

Esto, considerando que los trabajadores adquieren experiencia y habilidades a lo largo de su vida laboral, su productividad tiende a aumentar, lo que se refleja a través de mejores y más altos salarios. Esto, nos introduce al concepto de “pico de edad”, en el que los trabajadores alcanzan su punto máximo de productividad y, por tanto obtienen salarios más altos.

Sin embargo, a medida que los trabajadores envejecen, de acuerdo al ciclo de la vida, empiezan a enfrentar desventajas relacionadas con la obsolescencia en habilidades, menor capacidad física y adaptabilidad, entre otras, que en últimas se traduce en niveles decrecientes de productividad.

En ese sentido, para capturar esta relación no lineal entre la edad y el salario, naturalmente en los modelos de regresión se utiliza una forma funcional en la que la edad se eleva al cuadrado, permitiendo capturar el crecimiento inicial y su posterior decrecimiento.

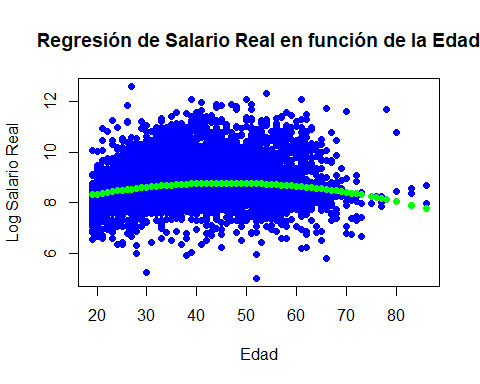
A continuación se presentan los coeficientes de la salida de regresión y su interpretación:

##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## log\_salarioreal   
## -----------------------------------------------  
## age 0.058\*\*\*   
## (0.004)   
##   
## age\_2 -0.001\*\*\*   
## (0.00005)   
##   
## Constant 7.429\*\*\*   
## (0.070)   
##   
## -----------------------------------------------  
## Observations 9,784   
## R2 0.035   
## Adjusted R2 0.035   
## Residual Std. Error 0.708 (df = 9781)   
## F Statistic 176.374\*\*\* (df = 2; 9781)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Si se obtienen los valores y estimados a través de la siguiente derivada:

Así por ejemplo, de manera matemática y manual, encontramos la edad pico, o el punto máximo de nuestro modelo de regresión considerando los estimadores obtenidos.

Graficando los valores estimados sobre la nube de puntos, obtenemos en efecto que la edad pico se podría determinar en los 46 años tal como se observa en la imagen presentada a continuación y luego, comienza a decrecer.



Ahora, bien los intervalos de confianza obtenidos por el método de Bootstrap representan un margen de error del 5% con el 95% de confianza los siguientes valores

## Variables Intervalos.de.confianza  
## 1 edad (0.0503, 0.0663)  
## 2 edad^2 (-0.0007, -0.0005)

Esto quiere decir, que por ejemplo, para la variable edad o en inglés “age”, el intervalo de confianza calculado es (0.0498,0.0669). Esto quiere decir que con un nivel de confianza del 95%, es posible afirmar que el coeficiente poblacional de la variable “edad”, se encuentra en ese intervalo. Es decir, esperamos que por cada unidad adicional de edad, el logaritmo del salario real aumente en un valor comprendido entre 0.0498 y 0.669, ceteris paribus.

## GAP en salario por género

La brecha de género en el salario es una de las manifestaciones más evidentes de la desigualdad entre hombres y mujeres en el ámbito laboral. Para analizar este fenómeno, se realizarán dos modelos, el primero será la variable de salario contra la variable Female, en el segundo se añadrirán variables de control, tales como la edad, la educación y el total de horas trabajadas.

En donde Female es una variable dummy con valor de 1 en caso de mujer y 0 en otro caso.

A continuación podemos observar los resultados de las dos regresiones:

##   
## ======================================================================  
## Dependent variable:   
## --------------------------------------------------  
## log\_salarioreal   
## (1) (2)   
## ----------------------------------------------------------------------  
## female -0.047\*\*\* -0.181\*\*\*   
## (0.015) (0.012)   
##   
## age 0.061\*\*\*   
## (0.003)   
##   
## age\_2 -0.001\*\*\*   
## (0.00004)   
##   
## totalHoursWorked -0.010\*\*\*   
## (0.0005)   
##   
## educ3 0.212\*\*   
## (0.092)   
##   
## educ4 0.279\*\*\*   
## (0.089)   
##   
## educ5 0.317\*\*\*   
## (0.089)   
##   
## educ6 0.514\*\*\*   
## (0.087)   
##   
## educ7 1.189\*\*\*   
## (0.087)   
##   
## Constant 8.648\*\*\* 7.038\*\*\*   
## (0.010) (0.105)   
##   
## ----------------------------------------------------------------------  
## Observations 9,784 9,784   
## R2 0.001 0.355   
## Adjusted R2 0.001 0.354   
## Residual Std. Error 0.721 (df = 9782) 0.579 (df = 9774)   
## F Statistic 10.503\*\*\* (df = 1; 9782) 597.355\*\*\* (df = 9; 9774)  
## ======================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Podemos observar que al realizar el modelo solo con la variable Female, la diferencia salarial entre los hombres y mujeres es de 4.7%; sin embargo, luego de añadir las variables de control, se evidencia que la diferencia salarial es aún más significativa, las mujeres ganan en promedio 18.1% menos que los hombres, manteniendo las demás variables constantes, en ambos casos, el coeficiente es significativo al 99% de confianza.

Tambien podemos evidenciar que las variables de control edad, edad^2 y Horas trabajadas son significativas al 99% de confianza, con valores respectivos de 6.1%, -0.1% y -1%. Los dos primeros coeficientes muestran que la edad es un factor fundamental en el momento de definir el salario y, como en el punto anterior, llega un punto en el que más edad impacta de manera negativa el salario. El coeficiente relacionado con horas trabajadas, muestra que el incremento en horas trabajadas no necesariamente impacta de forma positiva el salario, lo cual puede estar ocasionado porque los individuos con menores salarios deben trabajar más tiempo para sostener sus gastos básicos.

Ahora, se procederá a realizar el mismo modelo con el método FWL, a continuación se muestra una tabla en donde la primera columna corresponde al resultado de FWL y la segunda columna es el resultado de la regresión anterior.

##   
## =======================================================================  
## Dependent variable:   
## ---------------------------------------------------  
## lnw\_resid log\_salarioreal   
## (1) (2)   
## -----------------------------------------------------------------------  
## female\_resid -0.181\*\*\*   
## (0.012)   
##   
## female -0.181\*\*\*   
## (0.012)   
##   
## age 0.061\*\*\*   
## (0.003)   
##   
## age\_2 -0.001\*\*\*   
## (0.00004)   
##   
## totalHoursWorked -0.010\*\*\*   
## (0.0005)   
##   
## educ3 0.212\*\*   
## (0.092)   
##   
## educ4 0.279\*\*\*   
## (0.089)   
##   
## educ5 0.317\*\*\*   
## (0.089)   
##   
## educ6 0.514\*\*\*   
## (0.087)   
##   
## educ7 1.189\*\*\*   
## (0.087)   
##   
## Constant 0.000 7.038\*\*\*   
## (0.006) (0.105)   
##   
## -----------------------------------------------------------------------  
## Observations 9,784 9,784   
## R2 0.023 0.355   
## Adjusted R2 0.023 0.354   
## Residual Std. Error 0.579 (df = 9782) 0.579 (df = 9774)   
## F Statistic 229.296\*\*\* (df = 1; 9782) 597.355\*\*\* (df = 9; 9774)  
## =======================================================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Se identifica que luego de realizar el método, el coeficiente de la regresión salida de los resuduales con el método FWL es el mismo que el modelo realizado anteriormente.

Luego de esto, se procede a realizar el modelo FWL con boostrap:

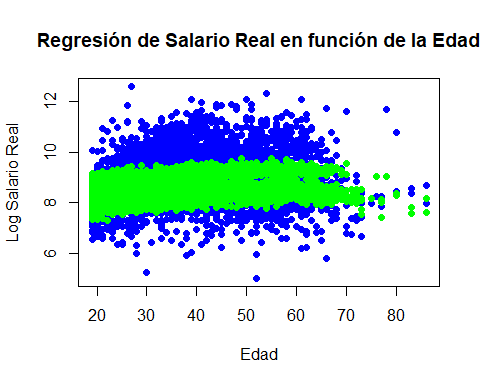
##   
## ORDINARY NONPARAMETRIC BOOTSTRAP  
##   
##   
## Call:  
## boot(data = geih\_filtered, statistic = eta\_fn3, R = 1000)  
##   
##   
## Bootstrap Statistics :  
## original bias std. error  
## t1\* 3.230525e-17 -3.257348e-17 5.608839e-17  
## t2\* -1.811997e-01 4.148179e-04 1.202015e-02

En este caso, también podemos evidenciar que las mujeres ganan 18.1% menos que los hombres, manteniendo las demás variables constantes; además, estos resultados son robustos a la heterocedasticidad. Para completar el análisis, a continuación se presente el Mean Squared Error (MSE) de los tres modelos realizados:

## Modelo MSE  
## 1 Modelo long 0.33537747  
## 2 Modelo FWL 0.33537747  
## 3 Modelo FWL con Boostrap 0.01641376

La tabla anterior muestra que el modelo FWL con Boostrap presenta un mejor ajuste en los datos que en los casos del modelo principal (Modelo long) y el Modelo FWL, lo cual es consistente con la teoría.

Por último, a continuación se muestra la tabla de los valores estimados del modelo:



En este caso, se evidencia que los datos estimados parecen seguir el comportamiento de los valores reales y, similar a la gráfica en donde solo se tiene en cuenta la edad, se observa un crecimiento de los datos hasta mediados de los años 50, luego esto, se nota un leve decrecimiento en el salario.

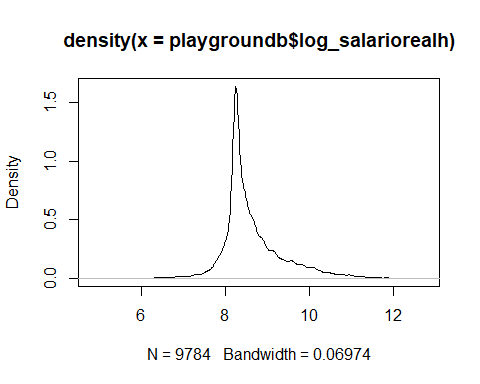
# Notas

Sustentar todas las medidas que salgan en las regresiones Exportar salidas de regresión con stargazer

# Punto 5

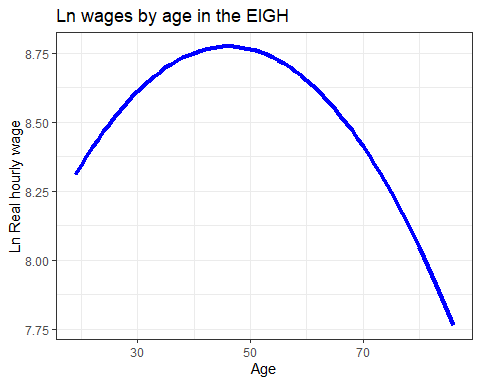
Dividiendo la muestra en dos submuestras, la primera de ellas (70%) para entrenamiento y la segunda (30%) como muestra de prueba, se procedió a realizar una predicción del ingreso. Ahora bien, para lograr esto se incluyó además una semilla de 10101. Se eligirán a partir de dicha base de datos una serie de posibles predictores:

##   
## Variables included in the Selected Data Set  
## ============================================================================  
## Statistic N Mean St. Dev. Min Max   
## ----------------------------------------------------------------------------  
## y\_salary\_m\_hu 9,784 7,984.690 11,629.940 151.910 291,666.700  
## maxEducLevel 9,784 6.098 1.110 1 7   
## exp 9,784 50.197 73.464 0 696   
## age 9,784 36.438 11.937 19 86   
## sex 9,784 0.502 0.500 0 1   
## hoursWorkUsual 9,784 48.081 12.062 1 130   
## totalHoursWorked 9,784 48.404 12.166 1 130   
## hoursWorkActualSecondJob 283 11.155 8.258 1 50   
## p6870 9,784 6.440 2.872 1 9   
## p6610 9,784 1.973 0.194 1 9   
## p7500s1 697 1.369 0.483 1 2   
## p7500s1a1 9,784 37,003.170 271,329.600 0 12,000,000   
## p7510s5 5,147 2.059 0.836 1 9   
## p7510s5a1 9,784 46,675.030 1,491,437.000 0 80,000,000   
## p7510s6 5,147 1.307 1.087 1 9   
## p7510s6a1 9,784 135,983.500 804,533.800 0 30,000,000   
## p7510s7 5,147 1.773 0.600 1 9   
## p7510s7a1 9,784 251,833.600 2,043,606.000 0 80,000,000   
## log\_salariorealh 9,784 8.624 0.721 5.023 12.583   
## exp2 9,784 7,916.178 23,857.650 0 484,416   
## age2 9,784 1,470.244 969.984 361 7,396   
## ----------------------------------------------------------------------------



##   
## ===============================================  
## Dependent variable:   
## ---------------------------  
## log\_salariorealh   
## -----------------------------------------------  
## age 0.058\*\*\*   
## (0.004)   
##   
## age2 -0.001\*\*\*   
## (0.00005)   
##   
## Constant 7.429\*\*\*   
## (0.070)   
##   
## -----------------------------------------------  
## Observations 9,784   
## R2 0.035   
## Adjusted R2 0.035   
## Residual Std. Error 0.708 (df = 9781)   
## F Statistic 176.374\*\*\* (df = 2; 9781)   
## ===============================================  
## Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



## (Intercept) age age2   
## 7.4290366056 0.0583634003 -0.0006334194

## age   
## 46.07011

En virtud de lo anterior se crearon una serie de modelos, con el objetivo de calcular el menor MSE, esto es, identificar el modelo que mejor prediga el salario.

## Modelo 1: Edad de la persona

## Modelo 2: Al modelo anterior añade la edad al cuadrado

## Modelo 3: Al modelo anterior añade el nivel de educación

## Modelo 4: Al modelo anterior añade la experiencia de la persona

## Modelo 5: Al modelo anterior añade el cuadrado de la experiencia

## Modelo 6: Al modelo anterior añade las horas trabajadas

## Modelo 7: Al modelo anterior añade el género si es mujer

## Modelo 8: Al modelo anterior añade si se recibe un ingreso adicional

## Modelo 9: Al modelo anterior añade si cotiza a pensión

## Modelo 10: Al modelo anterior añade si la persona es informal

## Modelo 11: Al modelo anterior añade el tamaño de la firma

## Modelo 12: Al modelo anterior añade el estrato de la persona

## Modelo 13: Al modelo anterior añade si tiene un trabajo adicional

Tras realizar el ejercicio de iteración, se observa que al pasar del Modelo 12 al Modelo 13 el MSE se incrementa; es decir, el Modelo 13 es el que mejor predice el ingreso. A continuación se resumen los MSE obtenidos para cada modelo probado, donde se confirma lo anteriormente dicho:

## model MSE  
## 1 Model1 0.5342218  
## 2 Model2 0.5212888  
## 3 Model3 0.3640148  
## 4 Model4 0.3528889  
## 5 Model5 0.3523483  
## 6 Model6 0.3394469  
## 7 Model7 0.3299643  
## 8 Model8 0.3251869  
## 9 Model9 0.3081166  
## 10 Model10 0.3080457  
## 11 Model11 0.2999375  
## 12 Model12 0.2343123  
## 13 Model13 0.2344121

Producto de lo anterior, se considera que el modelo que mejor predice el ingreso es: