Presentado por Yilmer Palacios, Betina Cortés, Lida Jimena Rivera, Nelson Fabián López

Problem Set 1: Predicting Income

I. Introducción

La sub-declaración de los ingresos por parte de los ciudadanos representa uno de los mayores retos del sistema fiscal, pues conocer el valor real de la renta individual es esencial para el cálculo los impuestos. De acuerdo con estudios realizados para Estados Unidos por el Servicio de Impuestos Internos (IRS), alrededor del 83,6 % de los impuestos se pagan de manera voluntaria y oportuna, con el reporte de ingresos más bajos que los reales como la principal causa de esta brecha. Ahora bien, resulta fundamental conocer el contexto colombiano, en donde el mercado laboral y el ingreso están permeados por variables sociales complejas. Así, a lo largo de este *Problem Set* se desarrolla un modelo de predicción de ingresos basado en características individuales usando datos de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 del Departamento Administrativo Nacional de Estadística (DANE) para conocer información sociodemográfica de la población y señalar casos de fraude que podrían conducir a la reducción de la brecha en el país e identificar a las familias vulnerables que podrían necesitar mayor asistencia por parte del gobierno.

Las principales conclusiones indican que las variables de sexo y edad, tipo de ocupación, nivel de educación, experiencia y estrato socioeconómico, tienen impacto significativo en el salario de los incluidos estudiados y son clave para predecir sus ingresos por hora. Considerar estos resultados es crucial para la implementación de políticas fiscales efectivas y para mejorar la eficiencia del sistema de recaudo de impuestos.

Nota:
La base de datos usada, al igual que el script de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace: https://github.com/Yilap/Repositorio_Taller1

Contexto

En el mercado colombiano se presentan dos grandes fenómenos que afectan el recaudo de impuestos por parte de las entidades del Estado: i) la **evasión de impuestos** *(tax evasion)* que implica actos ilícitos por parte de los contribuyentes los cuales violan los deberes derivados de la relación jurídica tributaria - tales como presentar declaraciones verdaderas o mantener los libros comerciales regulares – y ii) la **elusión de impuestos** *(tax avoidance),* conocida como la práctica de actos mediante los cuales se influencian los canales de conexión para evitar la aplicación de ciertos gravámenes tributarios (Sentencia C-360 de 2016 Corte Constitucional de Colombia, 2016). La evasión de impuestos se genera en múltiples formas, entre las que se desatacan la omisión de ingresos, declaración de costos, deducciones y descuentos inexistentes, subvaloración de activos, mimetización ilegal de ingresos, entre otros. Estas acciones dificultan el ejercicio de las autoridades estatales para recaudar y hacer cumplir las obligaciones tributarias

de los usuarios. Adicionalmente, se considera el fraude fiscal como delito, toda vez que el contribuyente tiene la intención de evadir el pago de sus obligaciones fiscales, como consecuencia de su comportamiento premeditado (Sentencia C-360 de 2016 Corte Constitucional de Colombia, 2016).

La evasión y elusión de impuestos ocasionan una pérdida para el gobierno de Colombia de entre COP $50 y $80 billones anuales. En 2021, según el Banco Interamericano de Desarrollo (BID), estos dos fenómenos representaron aproximadamente US$17 mil millones al año (o cerca de COP $68 mil millones) y, de acuerdo con Fedesarrollo, se pierden alrededor de 5,4 puntos porcentuales del PIB anual debido a estos factores. Sobre esto, es relevante mencionar que la mayor pérdida de recaudo se da por la evasión del impuesto de renta de las empresas, con un 3,4% del PIB; seguida de la evasión del IVA, que representa el 1,3% del PIB, y finalmente la evasión por concepto de impuesto de renta a personas, con cerca de un 0,7% del PIB (La República, 2022).

II. Datos

a. Descripción de las fuentes de datos

Para el desarrollo de este Problem Set se utilizarán los datos de la Gran Encuesta Integrada de Hogares (GEIH) del Departamento Administrativo Nacional de Estadística (DANE). Esta encuesta contiene información sobre las condiciones de empleo de las personas (si trabajan, en qué trabajan, cuánto ganan, fuentes de ingresos, si tienen seguridad social en salud o si están buscando empleo), adicional a las características generales de la población como sexo, edad, estado civil y nivel educativo (DANE, 2018). La GEIH consolida información no solo a nivel nacional sino a nivel regional, departamental, cabecera y ciudades capitales.

La GEIH es una fuente de información pertinente para el análisis de la evasión y elusión de impuestos en Colombia, ya que, al condensar información sociodemográfica sobre los individuos, los niveles de ingresos laborales y no laborales y la estructura de la fuerza de trabajo en el territorio nacional (tasas de ocupación, desempleo e informalidad laboral), permite identificar las características más relevantes que influyen sobre la renta personal. Usando estos datos como insumo, es posible predecir diferencias significativas entre los ingresos declarados y los reales.

b. Adquisición de los datos

Para la obtención de los datos de la GEIH se utilizaron técnicas de web scraping. El conjunto de datos contiene todos los individuos muestreados en Bogotá y está disponible en el siguiente sitio web https://ignaciomsarmiento.github.io/GEIH2018_sample/. En este caso, la página web fuente de los datos divide la información en 10 "*chunks*" de datos. Para conocer la estructura de cada enlace fue necesario analizar el código HTML de la página web, logrando identificar su

naturaleza dinámica. El hecho de ser dinámica plantea un reto para la extracción, pues el enlace original no es donde realmente reposa la información. Es necesario esperar a que cargue por completo para inspeccionarlo y encontrar enlace particular de la tabla a extraer, de lo contrario, no es posible detectarla mediante el código de R.

El web scraping se realizó usando el paquete *rvest*. En este caso, se identificó un patrón en los enlaces de cada data chunk, permitiendo importar la información por medio del loop a continuación y uniendo las 10 porciones de los datos en un solo data frame:

```
df_list <- list()

for (i in 1:10) {
  html_i <-
read_html(paste0("https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_", i,
".html")) %>%
    html_table()
  df_i <- as.data.frame(html_i)
  df_list[[i]] <- df_i
}

GEIH <- do.call(rbind, df_list)
```

c. Descripción del proceso de limpieza de datos

La base de datos extraída contiene un total de 32.177 observaciones de todos los individuos muestreados en la GEIH en Bogotá y 178 variables. Para la limpieza de datos, se eliminan las variables que no resultan útiles para el modelo de predicción. Además, el análisis se centra únicamente en las personas empleadas mayores de 18 años, por lo que la muestra se limita a personas que cumplen estas características.

De acuerdo con la literatura económica, algunas de las variables más relevantes para la predicción de la renta individual son: edad, género, educación, experiencia, tipo de ocupación (relab) y estrato socioeconómico (estrato). A continuación, se justifica de manera detallada la inclusión de cada una de las variables en el modelo:

Cabe mencionar que previo a elegir las variables objeto de análisis se tuvo en cuenta la población en edad de trabajar (pet) toda vez que es fundamental contar con una segmentación por edades, ya que eso permite tener un panorama más claro para proceder con el análisis. La población en edad de trabajar representa aquellos individuos que pueden generar ingresos por concepto de trabajo y ser jefes de los hogares, haciendo que, esta variable sea necesaria para contar con un modelo objetivo y claro sobre cuál será la población para describir. Este segmento está constituido por las personas de 12 años y más en las zonas urbanas y 10 años y más en las zonas

rurales. Además, se divide en población económicamente activa y población económicamente inactiva (DANE, s.f.).

- **Edad (Edad):** La edad de un individuo tiende a representar sus necesidades, oficios, intereses y preferencias. Por lo tanto, conocer la edad de los individuos nos permite generar un filtro para observar cuál es la población objetivo para cada investigación y planteamiento que se desee presentar. Por ejemplo, en este modelo de ingresos, los menores de edad no aportan información representativa, ya que cuentan usualmente con un jefe de hogar, quien percibe sus ingresos para manutención y demás necesidades. Por consiguiente, sus preferencias, oficios e intereses no serán analizadas en este espacio.

- **Género (Sex):** El género es fundamental en el análisis de los ingresos de los individuos, toda vez que, en el contexto colombiano, por ejemplo, existe una brecha entre hombres y mujeres en el momento de obtener trabajo y ganar un salario determinado, por lo tanto, ser hombre o mujer sí tiene influencia en la cantidad de ingresos que se perciben. Es por eso por lo que en el modelo tiene que estar presente esta variable, ya que ayudará a conocer el impacto en el salario dependiendo del género que tenga dicho individuo.

  Adicionalmente, el enfoque de género tiene como objetivo identificar y caracterizar las particularidades contextuales y situaciones vividas por cualquier persona de acuerdo con su sexo, lo cual implica constructos sociales asociados, implicaciones y diferencias económicas, políticas, psicológicas, culturales y jurídicas, que pueden incidir en brechas sociales y eventuales situaciones de discriminación (DANE, 2022).

- **Educación (Educ):** La educación representa el nivel de cualificación de un individuo y, a su vez, muestra el nivel salarial que obtiene dado sus condiciones académicas. Por lo tanto, se asume que, entre más educación posea un individuo, es probable que sea es más competente y con ello, tiende a ser más productivo. Por lo tanto, contar con esta variable dentro del modelo permite analizar cuán importante es para saber el nivel de ingresos que puede obtener un individuo si aumenta uno o más años de educación.

- **Experiencia (Exp):** La experiencia permite conocer cuánto tiempo ha durado trabajando una persona, en este caso, la base de datos nos presenta los datos del tiempo que lleva trabajando la persona en la empresa actual. Esta variable es relevante para el análisis dado que aporta evidencia de la influencia que tiene la experiencia sobre el ingreso de una persona, es decir, entre más tiempo laboral posea, probablemente puede ser más productivo porque cuenta con más conocimientos, capacidades y habilidades para desempeñar sus actividades. Así pues, el modelo propuesto por Mincer (1974) sugiere que el salario de un individuo depende de su nivel educativo y su experiencia laboral, teniendo en cuenta otras variables relevantes que lo describan. Con esta ecuación, se pueden hacer ajustes para considerar las variaciones individuales. Por tanto,

se puede afirmar que, los individuos con una mayor educación y experiencia tenderán a tener un mejor ingreso, lo que implica causalidad entre estas variables y el ingreso, lo que significa que entre mayores sean estas variables, mayores serán los ingresos.

- **Tipo de ocupación (Relab):** El tipo de ocupación se refiere a las categorías homogéneas de tareas que constituyen un conjunto de empleos, desempeñados por una persona, dadas sus capacidades y habilidades adquiridas por los años de educación y/o de experiencia adquirida, y por lo cual recibe un ingreso (DANE, 2005). Se puede inferir que, dependiendo del oficio, las personas obtendrán más o menos ingresos. Esta variable resume las principales tareas y deberes desempeñados en las ocupaciones y proporciona las categorías ocupacionales (DANE, 2015).

- **Estrato socioeconómico (Estrato):** El estrato socioeconómico es una clasificación en estratos de los inmuebles residenciales que deben recibir servicios públicos. Particularmente, es necesaria para cobrar de manera diferencial los servicios públicos domiciliarios, con el propósito de asignar subsidios y cobrar contribuciones. En este sentido, quienes tienen más capacidad económica pagan más por los servicios públicos y contribuyen para que los estratos bajos puedan pagar sus tarifas (DANE, s.f.).

  Por otro lado, sirve para identificar geográficamente sectores con distintas características socioeconómicas para orientar la inversión pública, la asignación de programas sociales como mejoramiento de infraestructura de servicios públicos, vías, salud, saneamiento básico y servicios educativos y recreativos. Permite también el cobro de tarifas de impuesto predial diferenciales por estrato (DANE, s.f.). Lo anterior sirve como proxy para identificar en qué lugares se ubican los individuos con mayores o menores ingresos.

d. Análisis descriptivo de los datos (estadísticas descriptivas)

Para iniciar con el análisis descriptivo de los datos, se procede a eliminar las observaciones de las personas menores de 18 años y las personas que se encuentran desocupadas (esto teniendo en cuenta el enunciado del Problem Set y que ya a los 18 años se puede laborar formalmente en Colombia). Adicionalmente, dado que en los puntos siguientes se utilizará la variable ingreso como logaritmo natural, se decidió eliminar las observaciones con valor de cero para que no existiesen datos incongruentes en la base. Se utilizaron los siguientes comandos:

```
GEIH <-GEIH[GEIH$age>=18,] y GEIH <- GEIH[GEIH$ocu == 1, ].
```

Posteriormente, se renombra la variable de máximo nivel de educación para tener mayor claridad. Este proceso se hace con el comando GEIH <- rename(GEIH, educ = p6210). En seguida, las variables, educación y tipo de ocupación, se establecen como categóricas de la siguiente manera:

```
GEIH$educ <- factor(GEIH$educ)
```

```
class(GEIH$educ)
GEIH$relab <- factor(GEIH$relab)
class(GEIH$relab)# Cálculo de la experiencia potencial
```

Por otro lado, en primer lugar, se estiman los años de educación dependiendo del máximo nivel alcanzado como se observa a continuación:

```
GEIH$añoseduc <- ifelse(GEIH$educ == 3, 5,
                  ifelse(GEIH$educ == 4, 9,
                        ifelse(GEIH$educ == 5, 11,
                              ifelse(GEIH$educ == 6, 16,
                                    ifelse(GEIH$educ == 9, 0, 0)))))
```

Se aplica la fórmula de experiencia potencial, en la cual los valores negativos se aproximan a 0 experiencia y se eliminan las personas que tengan una experiencia de 0 años.

```
GEIH$experp <- GEIH$age - 5 - GEIH$añoseduc
GEIH$experp <- ifelse(GEIH$experp < 0, 0, GEIH$experp)
GEIH<- GEIH[GEIH$experp>0,]
```

Ahora bien, para el cálculo de las horas totales trabajadas, se suman las horas trabadas en el empleo principal y secundario en una sola variable denominada **horast,** aplicando los siguientes comandos:

```
GEIH$hoursWorkActualSecondJob <- ifelse(is.na(GEIH$hoursWorkActualSecondJob), 0,
GEIH$hoursWorkActualSecondJob)
GEIH$horast <- GEIH$hoursWorkUsual + GEIH$hoursWorkActualSecondJob
```

Seguido a esto, se eliminan las personas que tengan un ingreso total de 0.

```
GEIH<- GEIH[GEIH$ingtot>0,]
```

Y para obtener el salario por hora se realiza la siguiente operación: (Ingresos mensuales, por 12 meses, dividido en las horas semanales trabajadas por 52 semanas del año).

```
GEIH$inghora <- (GEIH$ingtot*12)/(GEIH$horast*52)
```

Finalmente, se hace un subset de las variables de interés objeto de análisis para obtener subconjuntos, como se evidencia a continuación:

```
GEIHf <-subset(GEIH, select = c("inghora","age","sex","educ","experp","horast","estrato1","relab"))
GEIHf <- na.omit(GEIHf)
```

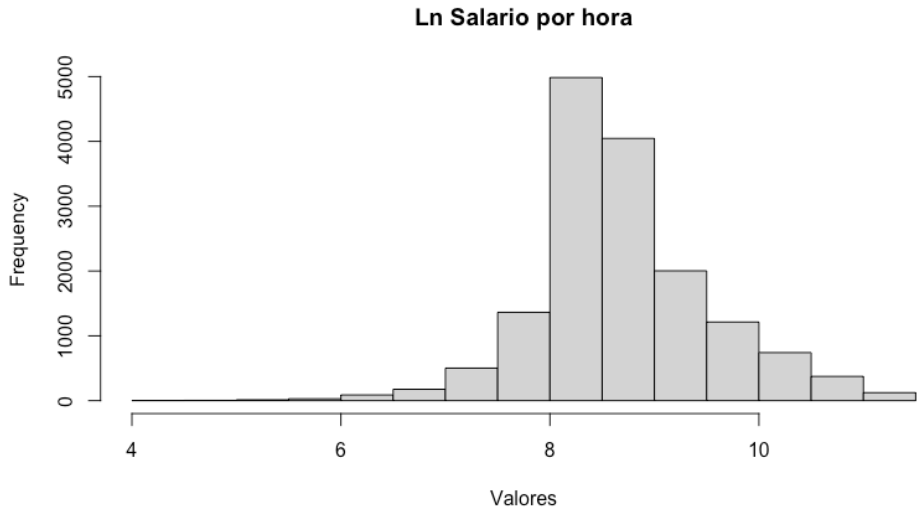
Y se identificaron y eliminaron los outliers, estos serán los que estén 3 desviaciones estándar alejados de la media, con lo que se creó la nueva base de datos denominada GEIHSO.

Con lo anterior, se procede a realizar las respectivas estadísticas descriptivas de nuestras variables de interés ("inghora","age","sex","educ","experp","horast","estrato1","relab")), con lo que se puede inferir que se tienen 15.661 observaciones con 8 variables.

Statistic	N	Mean	St. Dev.	Min	Max
inghora	15,661	8,757.570	10,148.970	76.923	84,935.900
age	15,661	39.678	12.977	18	79
sex	15,661	0.534	0.499	0	1
educ	15,661	3.986	1.068	1	6
experp	15,661	22.809	14.588	1	66
horast	15,661	47.945	15.215	1	130
estrato1	15,661	2.535	0.988	1	6
relab	15,661	2.222	1.492	1	9

- **Ingreso por hora**

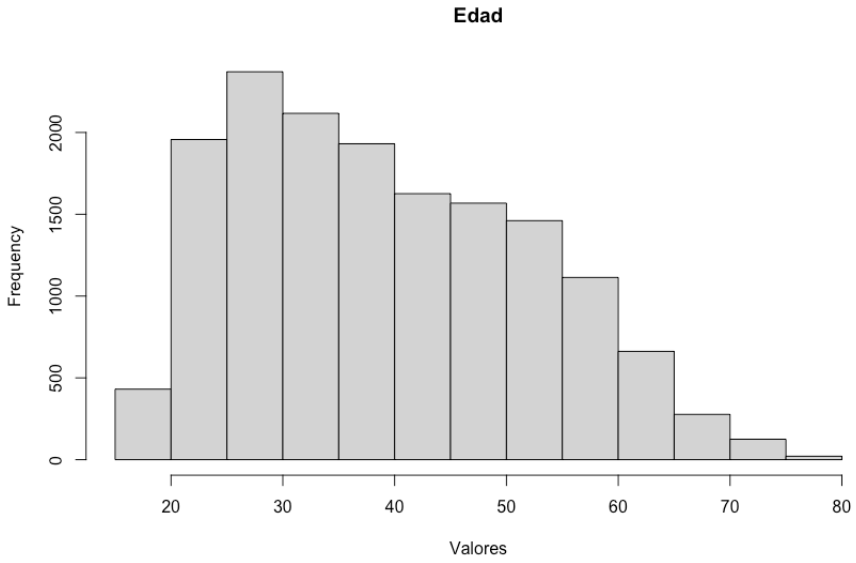
La Ingreso por hora se refiere al salario devengado en promedio por las personas. Se puede observar una asimetría hacia la derecha. El salario mínimo es 76 pesos y el máximo es de 84.935 pesos por hora, con un promedio de 8.757 pesos.



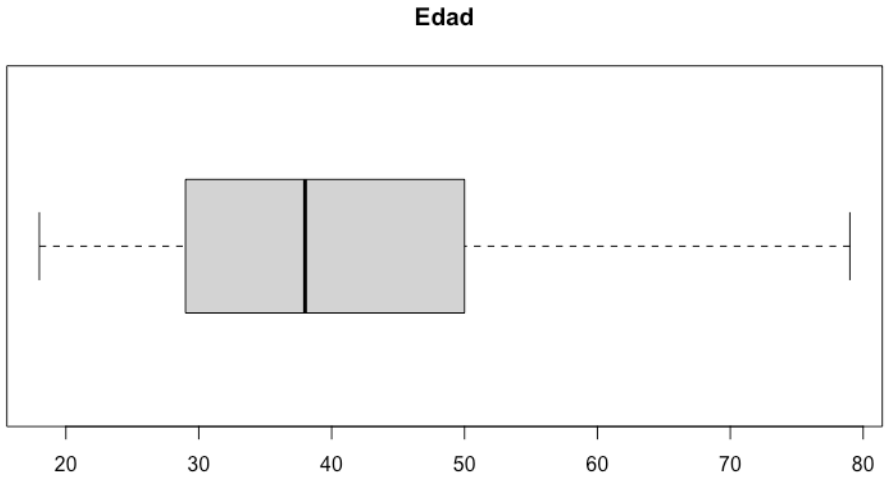
- **Edad**

La Edad hace referencia a que la variable maneja números enteros. Se puede observar una asimetría hacia la izquierda y gran parte de los individuos están entre los 20 y 40 años. Adicionalmente, se puede inferir que, de acuerdo con la limpieza de la base de datos y los filtros

realizados, la edad mínima es de 18 años, la edad máxima es de 79 años y el promedio de edad 39 años.



En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre 30 y 50 años con una media situada en 39 años.



- **Sexo**

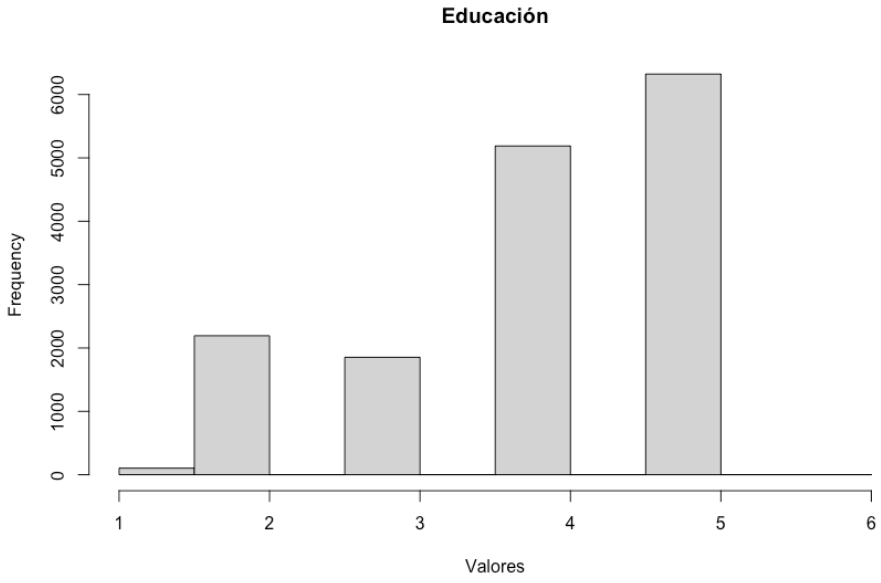
La variable es categórica, toda vez que hace referencia al género de los individuos: Hombre o Mujer. Por lo tanto, cuenta con dos niveles: 1= Hombre y 0 = Mujer, evidenciando que, del total de la muestra (15.661), el 53,4% es equivalente a los hombres, es decir, 8.363; el restante 46,6% equivale a las mujeres, es decir, 7.298. Lo anterior, evidencia que los hombres tuvieron mayor participación en comparación con las mujeres.

- **Educación**

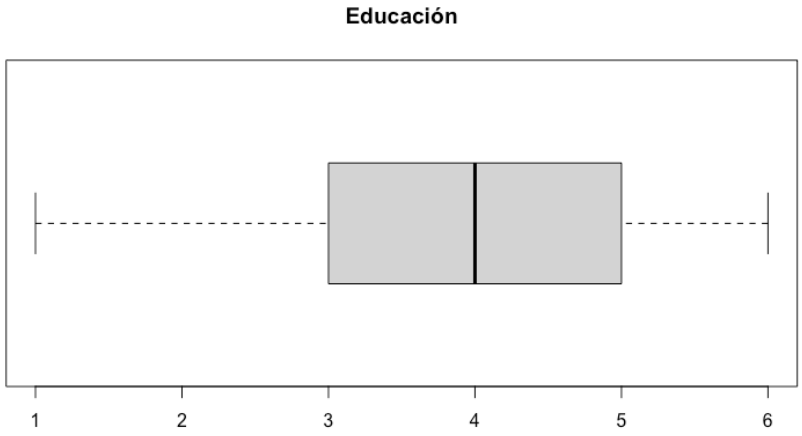
Esta variable responde a la pregunta ¿Cuál es el nivel educativo más alto alcanzado por .... y el último año o grado aprobado en este nivel? Con esta pregunta se trata de obtener el nivel educativo más alto alcanzado y el último grado de ese nivel. Si bien nos indica que la variable cuenta con número enteros, se pude conocer que dichos números representan diferentes variables, tales como:

1. Ninguno
2. Preescolar
3. Básica primaria (1º a 5º)
4. Básica secundaria (6º a 9º)
5. Media (10º a 13º)
6. Superior o universitaria
9. No sabe, no informa

Por lo tanto, podemos deducir que el promedio de los encuestados cuenta con educación media. Sin embargo, la categoría con mayor frecuencia es la 6, es decir, superior o universitaria, tal como se demuestra en el histograma. Podemos inferir que gran parte de la muestra cuenta con una educación superior o universitaria.

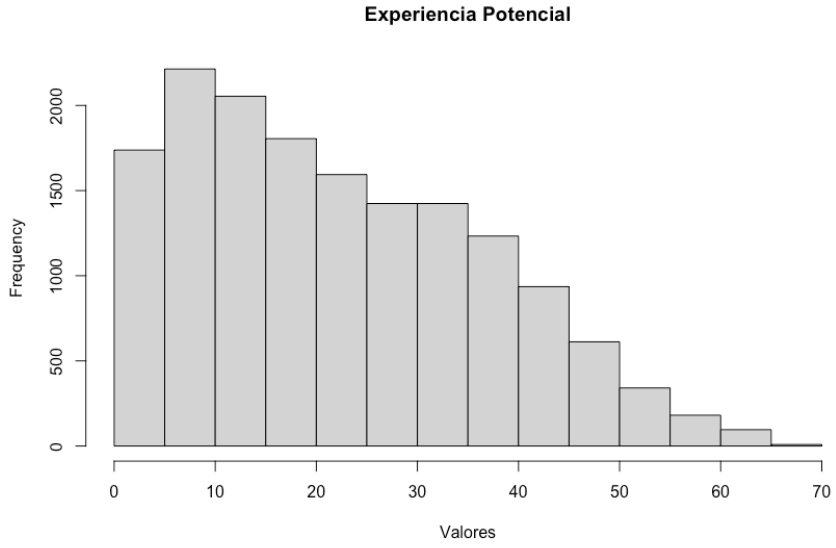


En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre 4 y 6 con una media situada en 5.

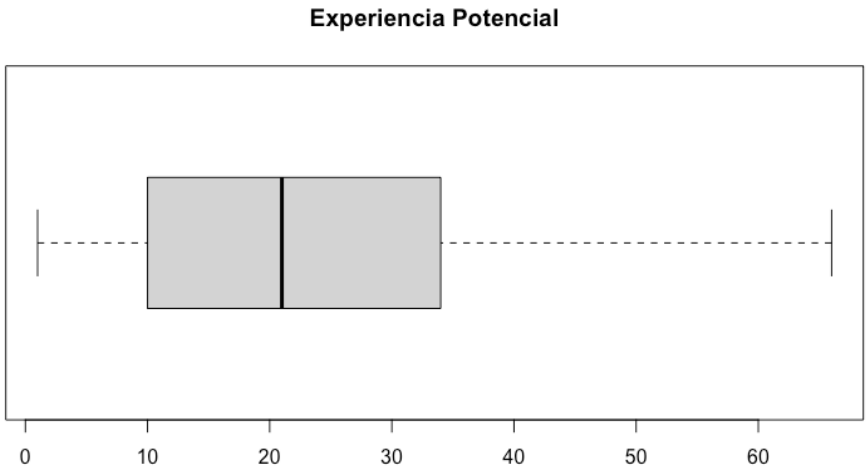


- **Experiencia**

Esta variable es numérica, en donde se mide el número de años que el individuo lleva en su último trabajo. Por lo tanto, se puede observar que la frecuencia de los datos es asimétrica hacia la izquierda, el dato mínimo es 1, es decir, un año, el máximo son 66 años, por consiguiente, el dato más frecuente es que los individuos lleven un año o menos en su oficio actual.



En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre 10 y 35 años con una media situada en 22 años.

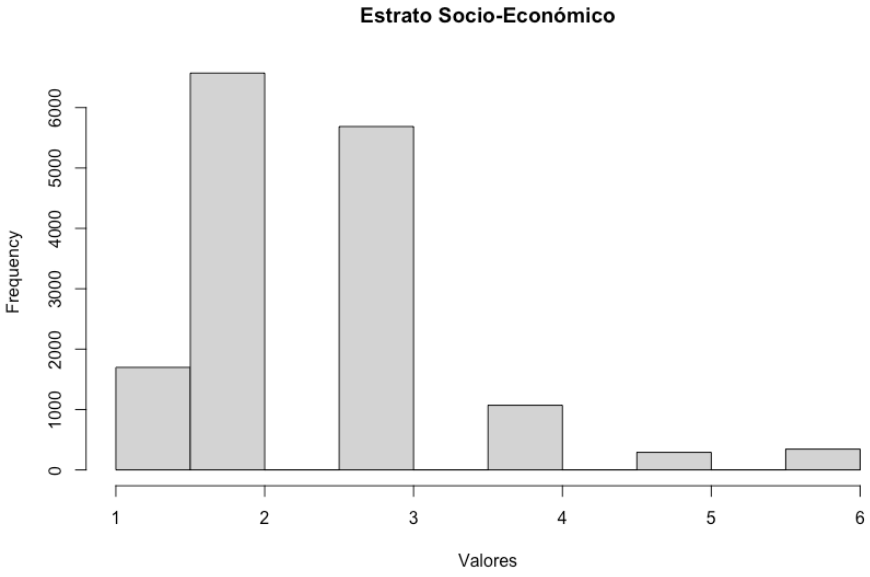


- **Horas totales**

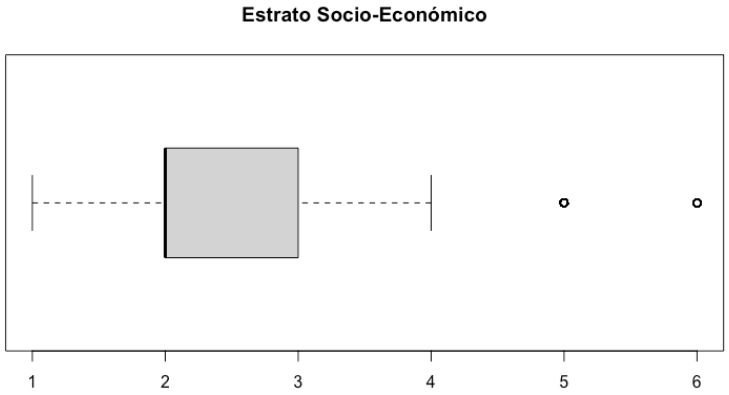
La variable horas totales es cuantitativa que mide el total de las horas trabajadas en promedio por una persona. En razón a esto, se puede observar que las horas totales mínimas laboradas es de 1 y las máximas son de 130 horas. El promedio es de 47 horas.

- **Estrato socioeconómico**

El estrato se refiere a una variable que toma valores entre 1 y 6, donde 1 es el estrato más bajo y 6 el estrato más alto. El promedio de los datos se encuentra en el estrato 2, seguido del estrato 3, luego estrato 1, y finalmente, los estratos 4, 5 y 6.



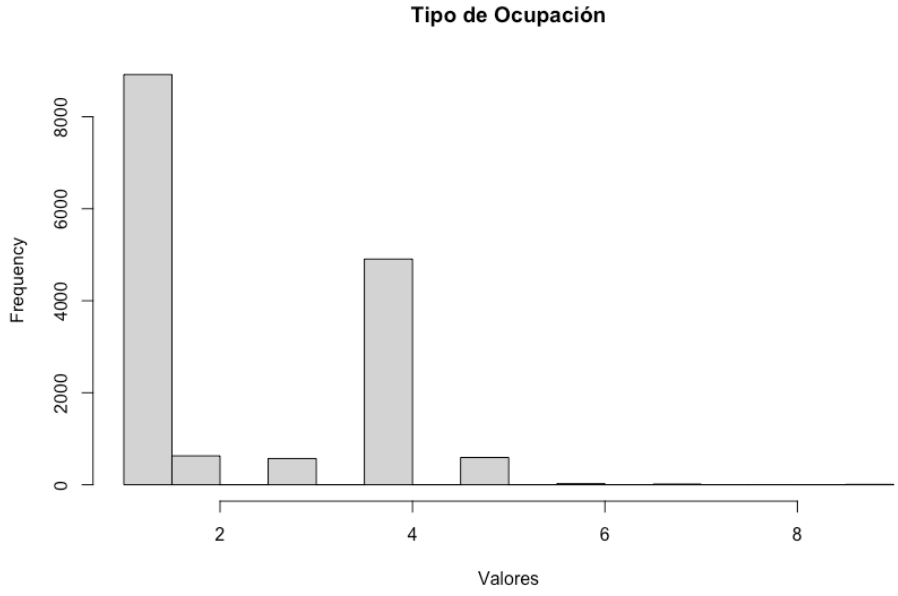
En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre los estratos 2 y 3 con una media situada en el estrato 2. Es importante mencionar que, se presentan outliers, representados en los estratos 5 y 6, por ser los estratos donde se concentran menos observaciones.



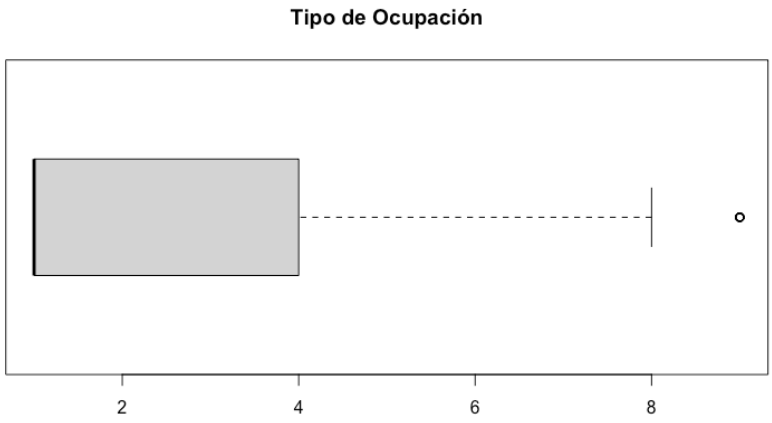
- **Tipo de ocupación**

Como se puede observar, el Tipo de Ocupación hace referencia a una variable categórica, en donde 1 es Obrero o empleado de empresa particular; 2 Obrero o empleado del gobierno; 3 Empleado doméstico; 4 Trabajador por cuenta propia; 5 Patrón o empleador; 6 Trabajador familiar sin remuneración; 7 Trabajador sin remuneración en empresas o negocios de otros hogares; 8 Jornalero o peón; y 9 Otro.

En este sentido, se observa que, la mayor parte de la población trabaja como obrero o empleado de empresa particular o como trabajadores por cuenta propia.



En el siguiente gráfico boxplot se presenta el 50% de las observaciones centrales entre los tipos de ocupación 1 y 4, con una media situada en el tipo de ocupación 1. Es importante mencionar que, se presentan outliers, representados en el tipo de ocupación 9, es decir, en otros.



III. Perfil Edad – Salarios

a. Regresión lineal

$$log(w) = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u$$

	Dependent variable:
	lningresoh
age	0.043***
	(0.003)
age2	-0.0005***
	(0.00004)
Constant	7.806***
	(0.064)
Observations	15,661
R2	0.013
Adjusted R2	0.013
Residual Std. Error	0.820 (df = 15658)
F Statistic	103.486*** (df = 2; 15658)
Note:	*p<0.1; **p<0.05; ***p<0.01

b. Interpretación de los coeficientes y su significancia

La regresión cuenta con la variable Age2 que representa las edades al cuadrado, ya que se está considerando que después del crecimiento del individuo, llega un punto en el que esa edad genera una relación negativa frente al ingreso, toda vez que alcanzó un punto máximo. Al correr la regresión se presentan estos datos. Los coeficientes representan el impacto que tiene dicha variable en la variable independiente, es decir, ingreso por hora.

La variable ingresoh (ingreso por hora) se transformó para poder analizar correctamente el efecto de los coeficientes de age y age2 eliminando efecto de dichas unidades y así, lograr la interpretación del modelo porcentual y que los datos sean tratados de manera efectiva y sin inconvenientes en la ejecución de la regresión por MCO.

La constante no suele generar ningún impacto en el modelo, ya que esta es representativa cuando Xi puede tomar el valor 0. Sin embargo, en este modelo no es posible que las variables tomen ese valor, ya que, si fuese el caso, no estarían dentro del modelo. De acuerdo con lo anterior, la constante no genera ningún análisis más allá de ser la intersección que define la relación entre dos variables. (ingresohy age o ingresoh y Age2).

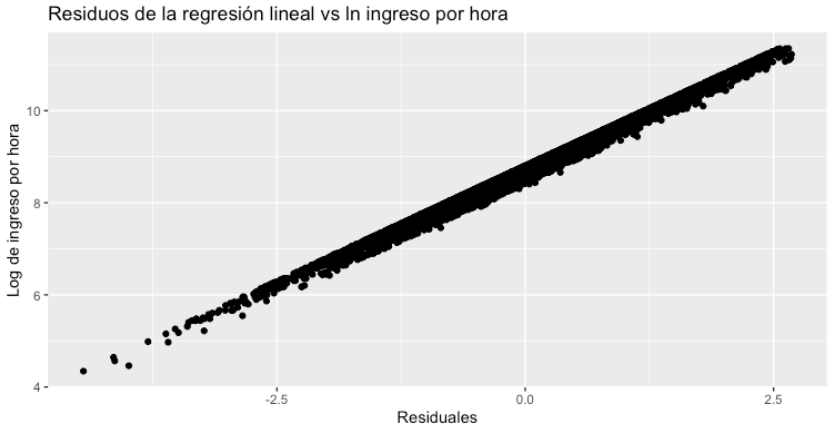
El coeficiente de Age hace referencia a que cuando un individuo aumenta un año de vida, el ingreso de este aumenta en 4.3% su ingreso, contando con un error estándar de la variable de 0.03, siendo este el que mide la precisión con la que cuenta la variable respecto a los valores estimados. A su vez, el coeficiente de Age2 nos indica una relación negativa entre la variable dependiente y la independiente, es decir, por cada año que envejezca el individuo al cuadrado el ingreso disminuye 0.05%, junto con su error estándar de 0.00004, siendo este valor muy pequeño.

Debido a la limpieza de datos generada, se contó con 15.661 observaciones y un mismo valor de R y R ajustado de 0.013, representando el poco ajuste que tienen las variables del modelo a la variable independiente, ingresos por hora, se puede identificar que es necesario contar con más variables explicativas para poder identificar qué genera el ingreso por hora en la población. Por otra parte, se cuenta con el estadístico F con 2 grados de libertad, no rechazando la hipótesis nula de falta de capacidad explicativa de las variables.

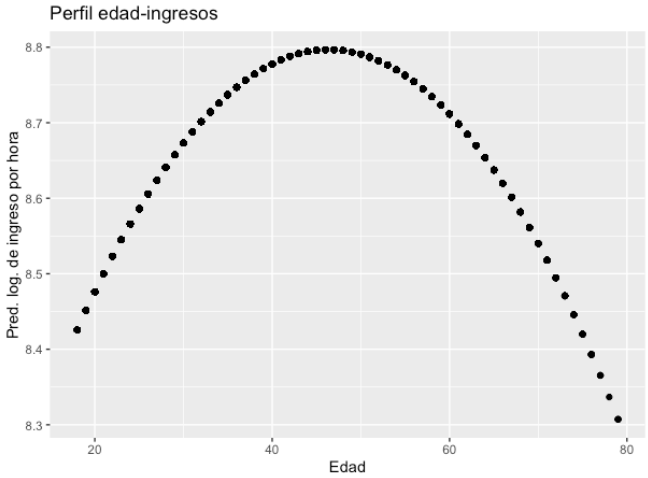
Finalmente, cada variable cuenta con una significancia aceptada por el 5%. Esto está representado que se puede tratar a los estimados diferentes de 0.

c. Discusión sobre el ajuste del modelo

El error estándar residual es de 0.820 siendo este el valor que nos indica qué tan bien se están ajustando los datos a la recta de la regresión, aunque este no sea un número muy pequeño, se puede observar que no ajusta todos los datos, pero gran parte de los datos si están cerca de la recta.



d. Gráfica de perfil estimado de ingresos por edad e intervalos de confianza



La gráfica señala el ingreso marginal decreciente, en donde a medida que la edad de un individuo aumente, la tendencia del aumento del ingreso va disminuyendo, haciendo que el individuo de una edad en adelante, no aumente significativamente sus ingresos. Por ejemplo, los ingresos de una persona de 30 años con el paso del tiempo pueden aumentar considerablemente en comparación con una persona de 50 años, en donde a medida que aumente su edad, como se observa en la gráfica, este no tendrá un crecimiento en el porcentaje de ingresos.

Ahora bien, se construyen los intervalos de confianza usando Bootstrap la cual ayuda a caracteriza la viabilidad de cada una de las variables, los cuales arrojan la siguiente información:

(Intercept)	7.8060988293
age	0.0427107403
age2	-0.0004603549

Contando con los coeficientes evidenciados en la anterior tabla que nos indican el impacto en el ingreso respecto a las variables independientes podemos continuar con la herramienta Bootstrap la cual ayuda a caracterizar la variabilidad. Por lo tanto, se realizó el ejercicio aplicando los conceptos de Bootstrap, utilizando la semilla 10101 y R=1000, se evidencian la siguiente distribución:

```
Bootstrap Statistics :
         original        bias     std. error
t1*  7.8060988293 -8.836550e-04 6.461223e-02
t2*  0.0427107403  2.707165e-05 3.427084e-03
t3* -0.0004603549 -1.918641e-07 4.171466e-05
```

Finalmente, el objetivo era maximizar la función, se decidió que dentro de todos los puntos de la distribución, se escogió la media de cada variable para que dentro del bootstrap se pudiera generar el PeakAge. Finalmente, se realiza el peak age con la siguiente fórmula PeakAge <- b1/(-2*b2) que se obtiene de realizar el siguiente procedimiento:

$$\log(salario\ por\ hora) = \beta_0 + \beta_1 age + \beta_2 age^2 + u$$

$$\frac{d\ (log(salario\ por\ hora)}{d\ age} = \beta_1 + 2\beta_2 age = 0\ (Para\ el\ máximo)$$

$$Age_{peak} = -\frac{\beta_1}{2\beta_2}$$

El resultado de esta "edad pico" es una edad aproximada de 46 años:

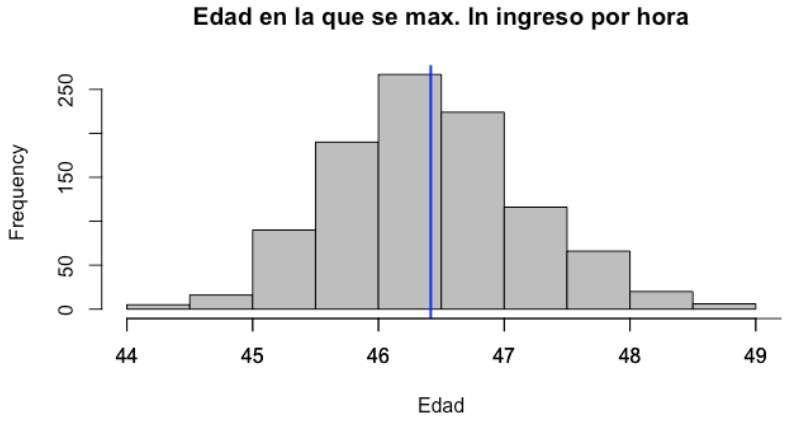
```
> summary(est_reg1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  44.52   45.93   46.39   46.45   46.93   50.01
```

$$\log(income) = \beta_0 + \beta_1 age + \beta_2 age^2 + u$$

$$\frac{d\ (log(income)}{d\ age} = \beta_1 + 2\beta_2 age = 0\ (Para\ el\ máximo)$$

$$Age_{peak} = -\frac{\beta_1}{2\beta_2}$$

El resultado de esta "edad pico" es una edad aproximada de 46 años:




```
> edadpico
     age
46.08629
```

IV. Brecha salarial de género

La brecha salarial de género se ha convertido en un tema de vigilancia en las últimas décadas para muchos gobiernos, es por esto que identificarla dentro de las poblaciones de estudio es de relevancia para quienes desean proponer y ejecutar políticas públicas que busquen igualdad de condiciones para hombres y mujeres, este será el foco de los próximos modelos.

a) Estimación y discusión de la brecha salarial incondicional de género.

Partimos del modelo propuesto siguiente:

$$\log(w) = \beta_0 + \beta_1 Sex + u$$

Donde w es es salario, Sex es una dicótoma que toma el valor de 0 si el individuo en mujer y 1 si el individuo es hombre, y u son los errores.

Los resultados del modelo son:

```
================================================
                         Dependent variable:
                     ---------------------------
                              lningresoh
------------------------------------------------
sex                            0.036***
                               (0.013)

Constant                       8.679***
                               (0.010)

------------------------------------------------
Observations                    15,661
R2                              0.0005
```

```
Adjusted R2                    0.0004
Residual Std. Error     0.825 (df = 15659)
F Statistic          7.566*** (df = 1; 15659)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Como podemos apreciar en la tabla, el estimador de la variable sexo es de 0.036, esto es, dado que es un modelo log-lin, un aumento del 3.6% en el salario cuando el individuo es hombre, mostrando que, de acuerdo con este modelo y nuestra base de datos, las mujeres ganaron menores salarios que los hombres en la ciudad de Bogotá para el 2018.

Sin embargo, se deben considerar otros factores para apoyar la afirmación, pues no todos los individuos tienen el mismo tipo de empleos, grados educativos, estrato socioeconómicos, edad, entre otros. Por esto, se hace necesarios controlar por otras variables.

b) Como se mencionó anteriormente, se hace necesario controlar por características propias de los individuos, en nuestro caso, usaremos las variables de control mencionadas en la sección "Datos" del presente documento, estas son:

- Sexo (dicótoma)
- Edad
- Nivel Educativo (categórica)
- Experiencia Potencial
- Tipo de trabajo (categórica)
- Estrato Socioeconómico (categórica)

El modelo planteado es el siguiente:

$$\log(w) = \beta_0 + \beta_1 Sex + \beta_2 Edad + \beta_3 Edad^2 + \beta_4 Experiencia\ Potencial + \sum \alpha_i Nivel\ Educativo_i + \sum \alpha_j Tipo\ de\ Ocupación_j + \sum \alpha_k Estrato\ Socioeconómico_k + u$$

Donde i, j y k, son el número de variables dicótomas creadas a partir de las variables categóricas "nivel educativo", "Tipo de Ocupación" y "Estrato Socioeconómico".

A continuación, se muestran los resultados de la regresión lineal del modelo anteriormente mostrado.

```
================================================
                        Dependent variable:
                    ---------------------------
                              lningresoh
------------------------------------------------
sex                            0.114***
                               (0.011)

Controls                          -
```

```
Constant                          7.297***
                                  (0.083)

-----------------------------------------------
Observations                       15,661
R2                                  0.386
Adjusted R2                         0.386
Residual Std. Error     0.647 (df = 15639)
F Statistic         469.045*** (df = 21; 15639)
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Cómo podemos apreciar, el estimador de la variable sexo es 0.114, esto es que, *ceteris paribus*, en promedio, un hombre gana 11,4% salario que las mujeres, esto con una significancia mayor o igual al 99%. Cabe resaltar que, en nuestra regresión lineal, y según el modelo planteado, la experiencia potencial da alta colinealidad.

Los anteriores resultados nos muestran, que, controlando por los efectos de la edad, el nivel educativo, el tipo de ocupación y el estrato socioeconómico; los hombres ganan más que las mujeres y esta brecha es mayor en comparación al primer modelo expuesto; lo anterior contrario a la afirmación de que hombres y mujeres con características similares tienes salarios similares.

Es importante mencionar, que pueden haber efectos observables no incluidos como por ejemplo el número de hijos del hogar o si son jefes o no del hogar, o también variables no observables como el interés y el gusto por trabajar, las preferencias personales, entre otras.

El anterior modelos también tiene la debilidad de no mostrar si hay efectos correlacionados entre sus variables, es por esto, que aplicar el teorema de FWL al modelo es de utilidad.

A continuación, mostramos la regresión del mismo modelo aplicando FWL, usando la muestra trabajada o simulando muestras mediante *bootstrapping*, los resultados se muestran a continuación:

i) Modelo FWL sin Bootstrap.

```
================================================================================
                                    Dependent variable:
                        ------------------------------------------------------
                                lningresoh                    lnwageResidF
                                    (1)                           (2)
--------------------------------------------------------------------------------
sex                           0.1138400***
                               (0.0106631)

Controls

SexResidF                                                     0.1138400***
                                                              (0.0106563)

Constant                      7.2966420***                     -0.0000000
                               (0.0829754)                    (0.0051642)

--------------------------------------------------------------------------------
Observations                     15,661                         15,661
R2                              0.3864397                      0.0072353
Adjusted R2                     0.3856158                      0.0071719
Residual Std. Error     0.6466788 (df = 15639)         0.6462657 (df = 15659)
```

```
F Statistic         469.0446000*** (df = 21; 15639) 114.1237000*** (df = 1; 15659)
================================================================================
Note: *p<0.1; **p<0.05; ***p<0.01
```

La tabla anterior muestra los resultados de dos modelos, el primero es la regresión de interés realizada convencionalmente, la segunda muestra el teorema FWL aplicado, en el modelo 2 se puede apreciar que el estimador de residuales de la variable Sexo "SexResidF" coincide con el estimador de interés de nuestro primer modelo, sin embargo, el estimador obtenido de FWL nos garantiza que este es el efecto directo sobre la variable ln(wage) pues retira los efectos de correlación de las variables de control. Por último, como se puede apreciar, los errores estándar no son los mimos, esto es porque el software R realiza el SE para el segundo modelo con distintos grados de liberar (tiene menos variables en su modelo regresivo). Realizando el cálculo manual obtenemos un SE de 0.01066313 siendo el mismo que del modelo 1.

ii) Modelo FWL con Bootstrap.

Corremos el mismo modelo propuesto anteriormente, sin embargo, en este caso aplicamos *Bootstrapping* para hallar el estimador y su error estándar, haciendo *sampling* con reemplazo y 1000 muestras, los resultados obtenidos son:

Estimador	Sesgo	SE
0.11384	-0.0001531979	0.01051701

Como podemos apreciar el estimador es el mismo que en anteriores modelos, sin embargo su error estándar disminuyó ligeramente, esto confirma que con varias muestras simuladas, la estimación es la misma.

c) Como parte del estudio realizado, vale la pena realizar un modelo que discrimine los efectos de la edad condicionado al sexo del individuo, usaremos el modelo propuesto en el punto 3, pero discriminando la muestra en hombres y mujeres. A continuación, mostramos nuevamente el modelo a regresar:

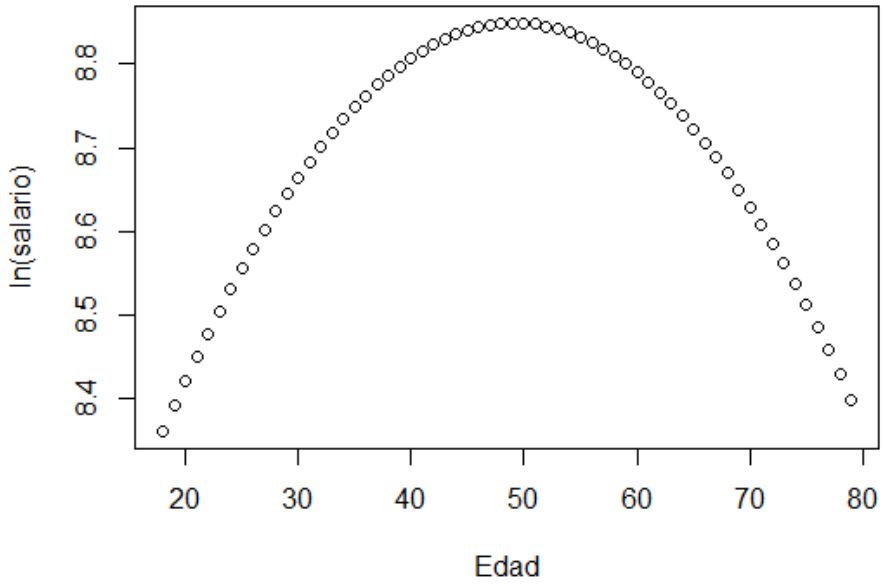
$$\log(w) = \beta_0 + \beta_1 age + \beta_1 age^2 + u$$

Para hombres, los resultados de la regresión son:

```
================================================
                         Dependent variable:
                     ---------------------------
                              lningresoh
------------------------------------------------
age                            0.050***
                               (0.004)

age2                           -0.001***
                               (0.00005)

Constant                       7.633***
                               (0.082)
```

```
------------------------------------------------
Observations                     8,363
R2                               0.025
Adjusted R2                      0.025
Residual Std. Error        0.787 (df = 8360)
F Statistic           109.059*** (df = 2; 8360)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Como podemos apreciar, efectivamente hay un efecto significativo de la edad sobre el salario en el caso de los hombres, y este efecto es cuadrático, a continuación, modelamos la curva ln(wage) vs age usando los estimadores de la regresión anterior.



Para hallar la edad en la que se encuentra el máximo salario, procedemos a maximizar la función de la curva anteriormente expuesta, sin embargo, usaremos *Bootstrapping* para el cálculo. Los resultados del máximo, error estándar e intervalos de confianza se muestran como sigue:

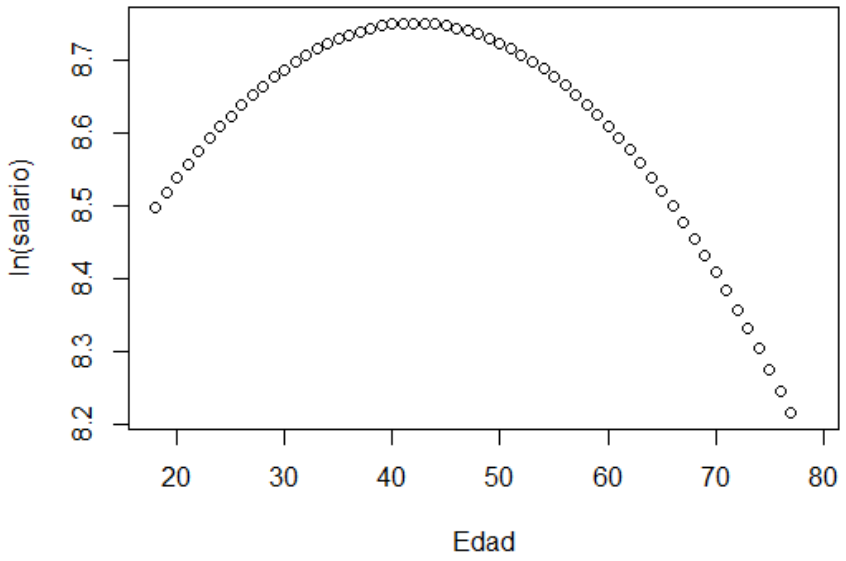
Edad (salario-max)	S.E	IC inferior	IC superior
49.13	1.14	46.93	51.33

Procedemos a hacer los mismos procesos para la muestra de mujeres, los resultados son:

```
================================================
                         Dependent variable:
                     ---------------------------
                              lningresoh
------------------------------------------------
age                            0.037***
                               (0.005)

age2                          -0.0004***
                               (0.0001)

Constant                       7.977***
```

```
                                        (0.101)

-----------------------------------------------
Observations                             7,298
R2                                       0.007
Adjusted R2                              0.007
Residual Std. Error          0.852 (df = 7295)
F Statistic            27.442*** (df = 2; 7295)
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Como podemos apreciar, también hay un efecto significativo de la edad sobre el salario en el caso de las mujeres, y este efecto es cuadrático, a continuación, modelamos la curva ln(wage) vs age usando los estimadores de la regresión anterior.



De igual manera, exponemos los resultados del estimador, SE e IC obtenidos por Bootstrapping.

Edad (salario-max)	S.E	IC inferior	IC superior
42.03	0.99	40.09	43.98

Como vemos en los resultados anteriores, las mujeres adquieren sus salarios máximos antes que los hombres (mujeres = 42 años, hombres = 49 años), por otro lado, los intervalos de confianza son similares (+/- 2 años), otro punto a destacar es que, según las curvas graficadas, a pesar de que las mujeres obtienen su salario máximo antes que los hombres, este es menor que al de los hombres reforzando a la afirmación de que hay una brecha salarial de género.

Una vez más es importante señalar que este modelo incondicional puede omitir variables observables y no observables, y por consiguiente puede tener sesgos por variables omitidas.

V. Predicción de ingresos

En esta sección del Problem Set se hace una evaluación del desempeño predictivo de las especificaciones planteadas en los literales anteriores y se proponen modelos adicionales que buscan mejorar el poder de predicción.

Para asegurar la reproducibilidad del ejercicio se establece una semilla, en este caso set.seed(1111). Para empezar, se divide la base de datos en dos: una muestra de entrenamiento (que representa el 70% de los datos) y otra de prueba (30%). Como resultado, la muestra de entrenamiento (*train_data)* se compone de 10.963 observaciones y la de prueba (test_data) de 4.698 observaciones.

Habiendo separado los datos, los modelos estimados en los puntos anteriores se replican en esta sección y, sumado a esto, se incluye un modelo simple sin covariables y con únicamente una constante para usar como caso base. Dado que el principal interés es predecir bien fuera de la muestra, es necesario evaluar los modelos en los datos de prueba. Así, se utiliza el coeficiente estimado en los datos de entrenamiento y luego se usa como predictor en los datos de prueba. Los modelos usados previamente se resumen a continuación:

Modelo base - 0	$\log(ingresoh) = \beta_0 + u$
Modelo previo - 1	$\log(ingresoh) = \beta_0 + \beta_1 age + \beta_2 age^2 + u$
Modelo previo - 2	$\log(ingresoh) = \beta_0 + \beta_1 sex + u$
Modelo previo - 3	$\log(ingresoh) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 educ + \beta_5 experp + \beta_6 relab + \beta_7 estrato1 + u$

Posteriormente, se plantean 5 modelos nuevos con especificaciones adicionales que incluyan no-linealidades y complejidades respecto a los anteriores. Estos son:

Modelo nuevo – 1	$\log(ingresoh) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 educ + \beta_5 \log(experp) + \beta_6 relab + \beta_7 estrato1 + u$
Modelo nuevo – 2	$\log(ingresoh) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 educ + \beta_5 experp + \beta_6 relab + \beta_7 estrato1 + \beta_8 sex * estrato1 + u$
Modelo nuevo – 3	$\log(ingresoh) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 educ + \beta_4 experp + \beta_5 experp^2 + \beta_6 relab + \beta_7 estrato1 + u$
Modelo nuevo – 4	$\log(ingresoh) = \beta_0 + \beta_1 sex + \beta_2 age * sex + \beta_3 age + \beta_4 age^2 + \beta_5 educ + \beta_6 experp + \beta_7 relab + \beta_8 estrato1 + u$
Modelo nuevo – 5	$\log(ingresoh) = \beta_0 + \beta_1 sex + \beta_2 age * sex + \beta_3 age + \beta_4 educ + \beta_5 experp + \beta_6 relab + \beta_7 estrato1 + u$

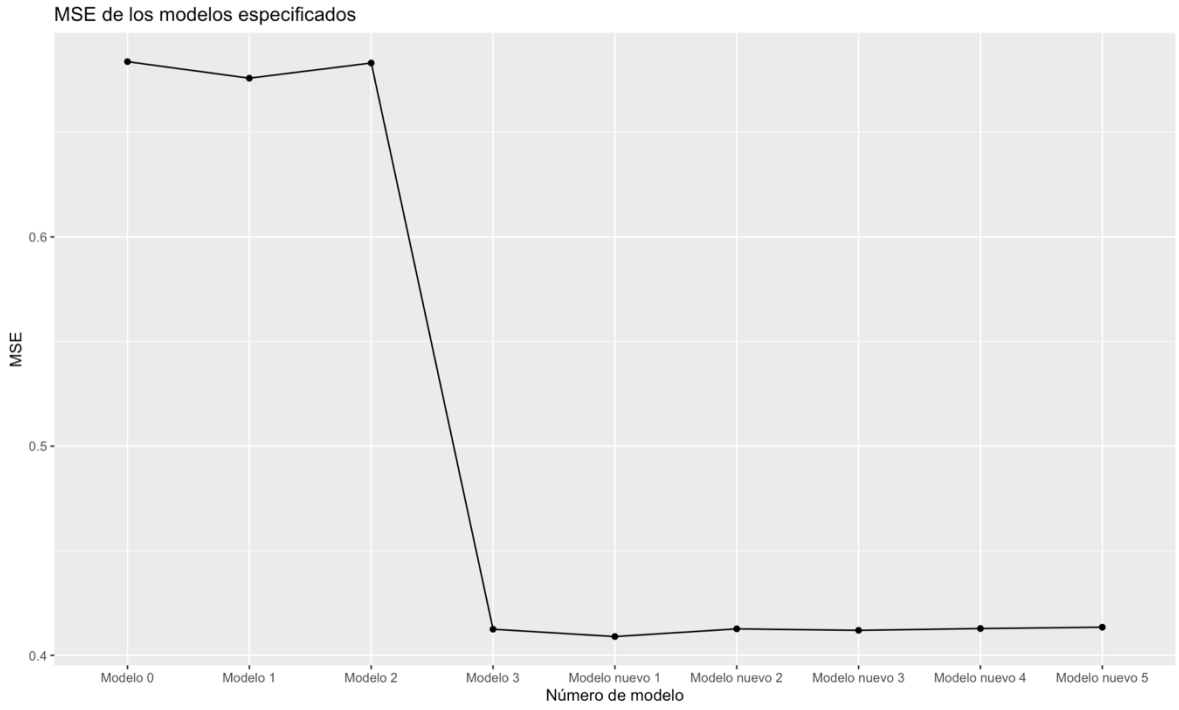
La métrica de performance elegida para este punto es el Error Cuadrático Medio (MSE, por sus siglas en inglés). Se considera adecuada, pues que mide la diferencia entre los valores reales y los

valores predichos por los modelos. Adicionalmente, tiene múltiples bondades frente a otras métricas de rendimiento, por ejemplo, es de fácil interpretación y comparabilidad con otros modelos, además de ser insensible la escala de los datos, ya que es una medida de la magnitud del error en términos cuadráticos.

Dicho esto, la tabla siguiente presenta el desempeño predictivo en términos de MSE de los 9 modelos estimados:

Modelo estimado	**MSE**
Modelo 0	0.6837741
Modelo 1	0.6758730
Modelo 2	0.6831309
Modelo 3	0.4124815
Modelo nuevo 1	**0.4089766**
Modelo nuevo 2	0.4126468
Modelo nuevo 3	**0.4119512**
Modelo nuevo 4	0.4127990
Modelo nuevo 5	0.4134420

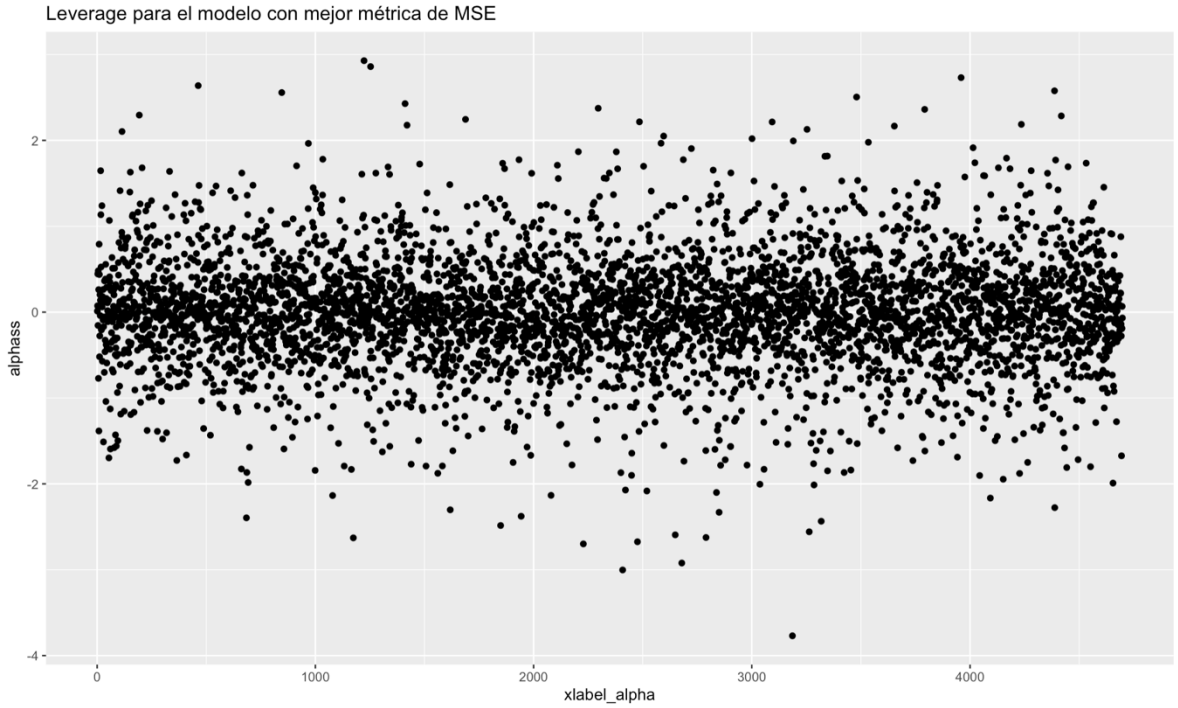
Se observa que los modelos con menores MSE son el modelo nuevo 1 y el modelo nuevo 3, señalados en la tabla anterior. Sin embargo, los valores son muy cercanos al modelo previo 3 y los modelos nuevos 2, 4 y 5. Estos resultados se muestran también en la gráfica a continuación:



Ahora bien, para el modelo con menor MSE, es decir, con mejor capacidad de predicción del logaritmo del ingreso por hora de los trabajadores mayores de 18 años en Bogotá, se estudia la existencia de outliers que puedan levantar alarmas sobre individuos que no reporten de forma transparente su salario. Para recapitular, el modelo con mejor métrica fue el siguiente:

$$\log(ingresoh) = \beta_0 + \beta_1 sex + \beta_2 age + \beta_3 age^2 + \beta_4 educ + \beta_5 \log(experp) + \beta_6 relab + \beta_7 estrato1 + u$$

La identificación de outliers se realiza por medio del apalancamiento o leverage, ya que mide la influencia de un punto específico en un modelo estadístico. Por lo tanto, es una herramienta que mide la capacidad de un punto en particular para alterar los resultados del modelo. De tal forma, entre más alto sea el leverage de un punto, más probable es que ese punto sea un outlier. Para efectos de este Problem Set, se tienen en cuenta valores que salgan del rango de 1 a -1. Mediante el código de R, se determinó que cerca del 11% de la muestra está por fuera de este rango. Este ejercicio se evidencia en la gráfica siguiente:



Los valores máximos y mínimos obtenidos fueron de 2.93 y -3.76 aproximadamente. Con esta información, podría decirse que no existen datos que represente un outlier significativo frente a la media de los datos y que levanten alarmas para la DIAN.

Finalmente, para los dos modelos con mejor capacidad de predicción de acuerdo con sus MSE, se calcula la performance utilizando LOOCV. Los resultados obtenidos para el modelo nuevo 1 fueron de un RMSE de 0.6448116, es decir, mayor al del enfoque de validación cruzada, el cual

era de 0.4089766. De igual forma, esto ocurrió para el modelo nuevo 1, para el que se obtuvo un RMSE de 0.6469727, mientras que con validación cruzada fue de 0.4119512.

VI. Bibliografía

Alcaldía de Bogotá., (2016). Documentos para DELITOS CONTRA EL ORDEN ECONÓMICO Y SOCIAL: Evasión Fiscal. Recuperado de https://www.alcaldiabogota.gov.co/sisjur/listados/tematica2.jsp?subtema=32520&cadena=

Corte Constitucional de Colombia., (2016). Sentencia C-360/16. https://www.corteconstitucional.gov.co/relatoria/2016/C-360-16.htm

DANE., (2005). Clasificación Internacional Uniforme de Ocupaciones Adaptadas para Colombia. Recuperado de https://www.dane.gov.co/files/sen/nomenclatura/ciuo/CIUO_88A_C_2006.pdf

DANE., (2015). Clasificación Internacional Uniforme de Ocupaciones CIU 08 A.C. Adaptada para Colombia. Recuperado de https://www.dane.gov.co/files/sen/nomenclatura/ciuo/CIUO_08_AC_2015_07_21.pdf

DANE., (2018). Mercado laboral (Empleo y desempleo) Históricos. Recuperado de: https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo/geih-historicos

DANE., (2022). Enfoque diferencial e interseccional. Enfoques de género. Recuperado de https://www.dane.gov.co/index.php/estadisticas-por-tema/enfoque-diferencial-e-interseccional/enfoque-de-genero#:~:text=El%20enfoque%20de%20g%C3%A9nero%20tiene,%2C%20psicol%C3%B3gicas%2C%20culturales%20y%20jur%C3%ADdicas%2C

DANE., (s.f.) EMPLEO – Preguntas frecuentes. Recuperado de https://www.dane.gov.co/files/faqs/faq_ech.pdf

DANE., (s.f.). Estratificación socioeconómica. Recuperado de https://www.dane.gov.co/index.php/69-espanol/geoestadistica/estratificacion/468-estratificacion-socioeconomica

La República., (2022). La evasión de impuestos le estaría quitando a Colombia cerca de $80 billones al año. Recuperado de: https://www.larepublica.co/economia/la-evasion-de-impuestos-le-estaria-quitando-a-colombia-cerca-de-80-billones-al-ano-3418446