

Universidad de Los Andes

Big Data and Machine Learning for Applied Economics

MECA 4107

Junio 26,2022

Bogotá D.C, Colombia

## **PROBLEM SET 1: PREDICTING INCOME**

Maria Valeria Gaona Guevara

Código: 202214418

Correo: [mv.gaona@uniandes.edu.co](mailto:mv.gaona@uniandes.edu.co)

Andrea Margarita Beleño Hernández

Código: 200620739

Correo: [a.beleno@uniandes.edu.co](mailto:a.beleno@uniandes.edu.co)

**Resumen:** El siguiente documento presenta la solución al Problem Set 1 de la clase Big Data, en donde se aplicaron diversos conceptos y herramientas para la predicción de modelos, el manejo de bases de datos grandes, entre otros. El repositorio GitHub se encuentra este documento y el código R, en donde se generaron todos los resultados. Este repositorio se encuentra en el siguiente link: <https://github.com/mvgaona/Taller-1>

### **1. Información General**

#### **1.1. *Adquisición de datos.***

**El objetivo principal es construir un modelo predictivo de la renta individual:**

$$\text{Ingreso} = f(\mathbf{X}) + u$$

**1.1.1. Extraer los datos que se encuentran disponibles en el siguiente sitio web**

**[https://ignaciomsarmiento.github.io/GEIH2018\\_sample/](https://ignaciomsarmiento.github.io/GEIH2018_sample/).**

Para realizar la importación de datos es importante conocer qué tipo de página web es. La página web que contiene la base de datos es dinámica, haciendo la extracción de datos sea más compleja.

Los datos fueron extraídos de cada uno de los 10 enlaces presentados en el sitio web. El código utilizado se encuentra en el archivo: “GEIH-Archivo de trabajo oficial.r” y la base de datos en la carpeta: “Elementos guardados” con el nombre: “Datos\_GEIH.rds”, luego de hacer una inspección previa del código html de la página para determinar la fuente de los datos. En las siguientes, se muestra que al realizar la inspección, se encuentra que la página no cuenta con la tabla de datos directa, por lo cual se procedió a revisar el enlace de la página donde residen los datos, los cuales se encuentran contenidos en el Script.

```
PS11.html
</button>

</div>
<!-- Collect the nav links, forms, and other content for toggling -->
<div class="collapse navbar-collapse" id="bs-example-navbar-collapse-1">
  <ul class="nav navbar-nav navbar-right navbar-li-color">
<li id="active"><a href="index.html">Home</a></li>
<li id="active"><a href="dictionary.html">Dictionary</a></li>
  </ul>
</div>
<!-- /.navbar-collapse -->
</div>
<!-- /.container -->
</nav><div class="container-fluid">
  <div class="row-fluid">
    <div class="col-md-3">
      <br><br><h2 class="name">Problem Set 1</h2>
      <h3 class="tagline">BDML MECA 4107</h3>
    </div>
    <div class="col-md-9">
      <div w3-include-html="pages/geih_page_1.html"></div>
    </div>
  </div>
</div>
```

```
35      <br><br><h2 class="name">Problem Set 1</h2>
36      <h3 class="tagline">BDML MECA 4107</h3>
37    </div>
38
39    <div class="col-md-9">
40      <div w3-include-html="pages/geih_page_1.html"></div>
41    </div>
42  </div>
43
44  </div>
45</div>
46
47<script>
48function includeHTML() {
49  var z, i, elmnt, file, xhttp;
50  /* Loop through a collection of all HTML elements: */
51  z = document.getElementsByTagName("*");
52  for (i = 0; i < z.length; i++) {
53    elmnt = z[i];
54    /*search for elements with a certain attribute:*/
55    file = elmnt.getAttribute("w3-include-html");
56    if (file) {
57      /* Make an HTTP request using the attribute value as the file name: */
58      xhttp = new XMLHttpRequest();
59      xhttp.onreadystatechange = function() {
60        if (this.readyState == 4) {
61          if (this.status == 200) {elmnt.innerHTML = this.responseText;}
62          if (this.status == 404) {elmnt.innerHTML = "Page not found.";}
63        }
64      };
65      xhttp.open("GET", file, true);
66      xhttp.send();
67      /* The HTML output will be injected here: */
68    }
69  }
70}
```

Dentro de la realización del script en R, es necesario descargar un paquete llamado *rvest*, el cual se utiliza como herramienta para hacer el Web Scraping, es decir, importar los datos desde una página web.

1. Dado el tamaño de cada base datos, se procedió con la descarga desde la página web específica y cada base de datos se convirtió en matriz:

```
Base1<read_html("https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_1.html")%>% html_table()
```

```
Base1 <- data.frame(Base1)
```

```
Base1<read_html("https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_2.html")%>% html_table()
```

```
Base2 <- data.frame(Base2), hasta terminar de descargar las 10 bases de datos.
```

2. Posteriormente, se genera una fusión entre las 10 bases datos para que se conforme la GEIH completa

```
DatosGEIH<- rbind(Base1, Base2, Base3, Base4, Base5, Base6, Base7, Base8, Base9, Base10).
```

Con este código se obtuvo la base de datos completa con 178 variables y 32177 observaciones.

### **1.1.2. ¿Existen restricciones para acceder o extraer estos datos?**

Para acceder a los datos es necesario esperar un momento, ya que los datos no se cargan automáticamente, debido a que los datos no se encuentran en ese mismo enlace web, sino que los extrae de otra página web donde realmente reposan los datos. De acuerdo con lo anterior, sí existen restricciones para extraer directamente los datos debido al inconveniente mencionado anteriormente; cuando se corre el código de extracción de datos en R, no detecta la tabla en html que se visualiza en la página web, porque esta se importa a través de la función “*includehtml()*”, que se encuentra en el script. Sin embargo, al obtener el enlace de página web en donde reposan los datos en html, no existen restricciones para descargar o realizar “scraping” a los datos.

### **1.1.3. Usando pseudocódigo, describa su proceso de adquisición de datos**

01. Ingresar al enlace: [https://ignaciomsarmiento.github.io/GEIH2018 muestra/](https://ignaciomsarmiento.github.io/GEIH2018_muestra/).

02. Ingresar al enlace: Data chunk 1

([https://ignaciomsarmiento.github.io/GEIH2018\\_sample/page1.html](https://ignaciomsarmiento.github.io/GEIH2018_sample/page1.html))

03. Realizar la inspección del código fuente de la página para determinar si la tabla existe.

04. Al verificar que la tabla no se encuentra, verificar la fuente de la tabla de datos en html.

05. Obtener el enlace a la tabla donde se encuentra la tabla de datos en html.

06. Validar que para los demás datasets el enlace a la página web.

07. Elaborar el código en R para realizar la extracción precisa de los datos

- a. Instalar los paquetes necesarios
- b. Realizar el vector de los enlaces de las páginas web donde se extraerán los datos
- c. Realizar el scrape de los datos usando la función de “rvest” [readhtml () y  
html\_table()]
- d. Unir los datos en una base de datos de extensión “.rds”

## 1.2. Limpieza de datos

**1.2.1 El set de datos incluye múltiples variables que pueden ayudar a explicar el ingreso individual. Guiado por su intuición y conocimiento en economía, escoja las variables más relevantes y realice un análisis descriptivo de estas variables. Por ejemplo, puede incluir variables que midan la educación y la experiencia, dadas las implicaciones del modelo de acumulación de capital humano (Becker, 1962, 1964; y Mincer (1962, 1975)**

Para conocer el ingreso total es fundamental contar con las siguientes variables:

1. Pet = Población en edad de trabajar
2. Edad
3. Educ = Hace referencia a la educación con la que se cuenta. En la base de datos se trabajará la variable p6210, en donde se evidencia el nivel de educativo más alto alcanzado
4. Ocu= Ocupación
5. Sex = Género.
6. Exp= Experiencia. En la base de datos se trabajará la variable p6426, en donde se dividirá este dato entre 12 para tener los datos en formato años.

A continuación, se describen las variables para tener en cuenta de la base de datos:

### Edad y población en edad de trabajar. (Edad y Pet)

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. En este modelo de ingresos, los menores de edad, por ejemplo, no representan información representativa, ya que cuentan usualmente con un jefe de hogar quien es el que percibe sus ingresos para manutención y demás necesidades. Por consiguiente, sus preferencias, oficios e intereses no serán analizadas en este espacio.

De acuerdo con lo anterior, es fundamental contar con una segmentación por edades, ya que eso permite contar con 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 a describir

### Educación (Educ)

La educación representa cuán capacitado y certificado está el individuo. La educación le permite al individuo poder contar con mejor salario, ya que se asume que el individuo entre más educación posea, es más competente y con ello, tiende a ser más productivo. Por lo tanto, contar con la educación en el modelo, permite analizar cuán importante es la educación para saber cuántos ingresos puede a llegar a obtener si aumenta uno o más años de estudio.

### Ocupación (Ocu)

La ocupación permite filtrar a aquellos individuos que se encuentran en edad de trabajar y en este momento se encuentran con trabajo, u ocupados. Esto permite contar con un espectro más claro en el modelo, ya que permite analizar cómo sus ingresos dependen si están ocupados o no.

### Género (Sex)

El género es fundamental en el análisis de los ingresos de los individuos, ya 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 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.

### Experiencia (Exp)

La experiencia permite conocer cuánto tiempo ha durado una persona trabajando, 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 muy importante, ya que aporta al conocimiento la influencia que tiene la experiencia al ingreso de una persona, ya que entre más tiempo lleve trabajando, puede ser más productivo porque cuenta con más conocimiento y habilidades para realizar sus actividades.

### Oficio (oficio)

El oficio nos permitirá saber si, dependiendo de dicha categoría, las personas obtienen ingresos similares o no, independiente del género. Si bien esta variable es categórica con 99 opciones, se realizará una abreviación de dicha categoría para determinar si es relevante o no el oficio para estimar los ingresos. Se puede inferir que dependiendo del oficio, las personas obtendrán más ingresos.

**1.2.2 Observe que hay muchas observaciones con datos faltantes. Dejo en sus manos el encontrar la manera de tratar con estos datos. En su explicación, describa los pasos que usted realizó para limpiar los datos y justifique su decisión.**

Antes de realizar la limpieza formal de la base de datos, realizamos un análisis en el cual concluimos lo siguiente:

- Antes de construir la base de datos final, se debe hacer el filtro de los datos para retirar los datos de las personas menores a 18 años (esto teniendo en cuenta el enunciado del Problem Set y que ya a los 18 años se puede laborar formalmente en Colombia).
- Como 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.

Por lo anterior, se procedió a hacer el filtrado de la base con las condiciones mencionadas anteriormente. Luego de realizar este filtro, se construyó la base final de trabajo con las variables seleccionadas en el punto 1.2.1.

Para las variables mencionadas anteriormente, existen datos faltantes para la variable “exp”, y para la variable “oficio” como se presenta en la tabla a continuación:

ingtot	pet	mes	age	sex	ocu	oficio	exp	educ
Min. : 4167	Min. :1	Min. : 1.000	Min. : 18.00	Min. :0.0000	Min. :0.000	Min. : 1.00	Min. : 0.00	Min. :1.000
1st Qu.: 780000	1st Qu.:1	1st Qu.: 4.000	1st Qu.: 29.00	1st Qu.:0.0000	1st Qu.:1.000	1st Qu.:33.00	1st Qu.: 0.00	1st Qu.:4.000
Median : 1000000	Median :1	Median : 6.000	Median : 40.00	Median :1.0000	Median :1.000	Median :45.00	Median : 2.00	Median :5.000
Mean : 1698982	Mean :1	Mean : 6.434	Mean : 42.95	Mean :0.5074	Mean :0.822	Mean :49.72	Mean : 5.15	Mean :4.899
3rd Qu.: 1695333	3rd Qu.:1	3rd Qu.: 9.000	3rd Qu.: 55.00	3rd Qu.:1.0000	3rd Qu.:1.000	3rd Qu.:70.00	3rd Qu.: 7.00	3rd Qu.:6.000
Max. :8583333	Max. :1	Max. :12.000	Max. :106.00	Max. :1.0000	Max. :1.000	Max. :99.00	Max. :60.00	Max. :9.000
						NA's :3524	NA's :3524	

Así mismo, se analiza el porcentaje de NA incluido en las variables “exp” y “oficio”, en donde se observa que para estas variables el 17,8% de la base de datos contiene NA.

	cantidad_na
ingtot	0.00000
pet	0.00000
mes	0.00000
age	0.00000
sex	0.00000
ocu	0.00000
oficio	17.79708
exp	17.79708
educ	0.00000

Para las variables “exp” y “oficio”, se tiene una gran cantidad de NA, y si se fuese a hacer una regresión lineal, lo ideal en ese caso sería eliminar dichas observaciones. Sin embargo, como se está realizando una aplicación de Machine Learning y nuestro principal objetivo es predecir valores, tomaremos la decisión de imputar valor cero a estos NA, teniendo en cuenta que si no reportan experiencia laboral es como si no tuviesen, en nuestro análisis. En el script en R se encuentra el procedimiento.

En el caso de la variable oficio, se creará una nueva categoría, la número 100, la cual indica: No reporta; se imputará este número 100 a dichos datos faltantes de “oficio”, esto con el fin de saber si estas personas perciben mayores ingresos que las que tienen un oficio definido y pueden ser también objeto de revisión por parte de la dirección de impuestos, en dado caso.

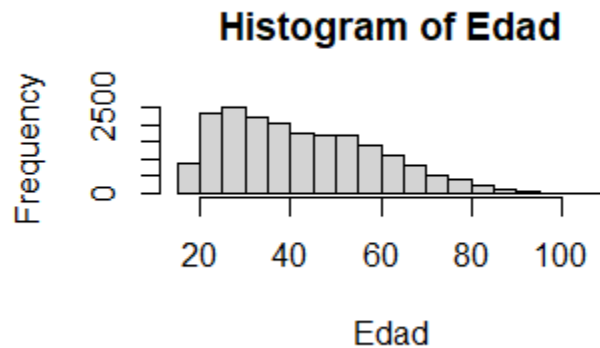
### 1.2.3 Descripción estadística de las variables.

#### Edad

Para iniciar el análisis descriptivo de la variable es necesario conocer la clase de esta:

```
> class(Edad)
[1] "integer"
```

Hace referencia a que la variable maneja números enteros. Por otra parte, observaremos la frecuencia y distribución de los datos dentro de la variable.



Se puede observar una asimetría hacia la izquierda y gran parte de los individuos están entre los 20-40 años. Sin embargo, analizaremos la edad mínima, máxima, el promedio de edad y finalmente, la moda, la edad más común entre los individuos.

```
> min(Edad)
[1] 18
> max(Edad)
[1] 106
> mean(Edad)
[1] 42.95419
> modeEdad(Edad)
[1] 25
```

La edad mínima de los individuos en la muestra es 18 años, debido al filtro realizado para la realización del modelo, la edad máxima es 106 años, la media se encuentra en 43 años, siendo el promedio de edad de los individuos y finalmente, la edad que más se repite es de 25 años.

### PET

La variable Población en Edad de Trabajar (PET) señala:



```

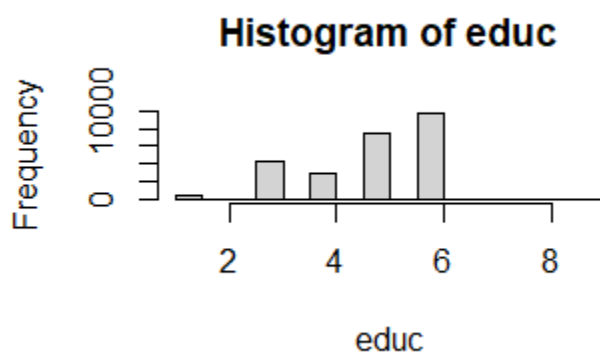
> class(PET)
[1] "factor"
> levels(PET)
[1] "1"
> summary(PET)
      1
19801

```

La clase indica que la variable está dada por “factor”, es decir, representa valores categóricos, ya que cada nivel tiene una etiqueta asociada. En este caso, es 1 si el individuo está en edad de trabajar y 0 si no lo está. Sin embargo, se puede observar que solo existe un nivel (1) ya que dentro de la organización y limpieza de datos se realizó un filtro en donde solo las personas en edad de trabajar hicieran parte de la muestra, es decir 19801 personas.

### Edad

Para la variable edad, el análisis descriptivo señala:



```

> class(educ)
[1] "integer"
> mean(educ)
[1] 4.873413
> modeEduc(educ)
[1] 6

```

Si bien nos indica que la variable cuenta con número enteros, es conocido que dichos números representan diferentes variables.

1. Ninguno

2. Preescolar
3. Básica primaria
4. Básica secundaria
5. Media
6. Superior o universitaria

Por lo tanto, podemos deducir que el promedio de los encuestados cuentan 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.

### Ocupación (Ocu)

La variable ocupación contiene la siguiente descripción

```
> class(ocu)
[1] "factor"
> levels(ocu)
[1] "0" "1"
> summary(ocu)
      0      1
3524 16277
```

Como se puede observar, la variable tiene clase *factor* haciendo referencia a que es una variable categórica, en donde existen 2 niveles: 1 Ocupado, 0 desocupado. Por lo tanto, al realizar una inspección de los datos, se evidencia que de 19801 observaciones, 16277 se encuentran ocupados y 3524 no. Esta proporción puede evidenciarse en el siguiente gráfico de torta:



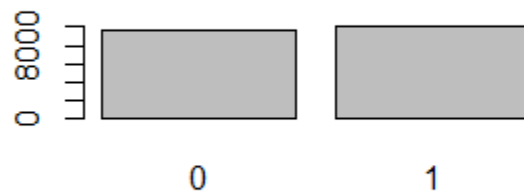
### Género (Sex)

```

> class(sex)
[1] "factor"
> levels(sex)
[1] "0" "1"
> summary(sex)
      0      1
9754 10047
> table(sex)
sex
  0    1
9754 10047

```

La variable es categórica, ya 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 dentro de la muestra, los hombres tuvieron mayor participación en comparación con las mujeres: 10047 fueron encuestados. Sin embargo, en el siguiente gráfico de barras se puede identificar la proporción de hombres y mujeres encuestados.



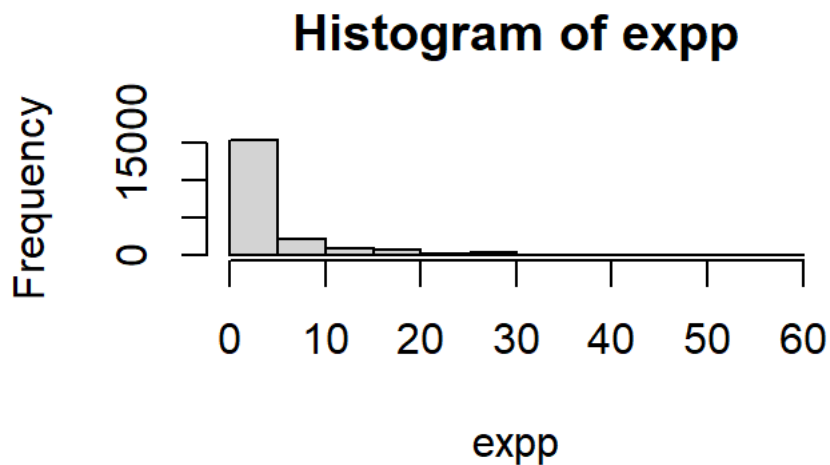
### Experiencia (Exp)

La variable experiencia cuenta con la siguiente descripción

```

> class(expp)
[1] "numeric"
> plot(hist(expp))
> mean(expp)
[1] 4.233726
> min(expp)
[1] 0
> max(expp)
[1] 60
> modeExp(expp)
[1] 0

```



Esta variable es numérica, en donde se miden el números 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 0, es decir, menos de un año, el máximo son 60 y finalmente, la moda está dada por 0, por consiguiente, el dato más frecuente es que los individuos lleven menos de un año en su oficio actual.

### Oficio

La variable oficio, como se ha mencionado anteriormente permite establecer si las personas tienen una ocupación. Las personas que no tenían una ocupación clara, aparecerán en la base con el valor 100.

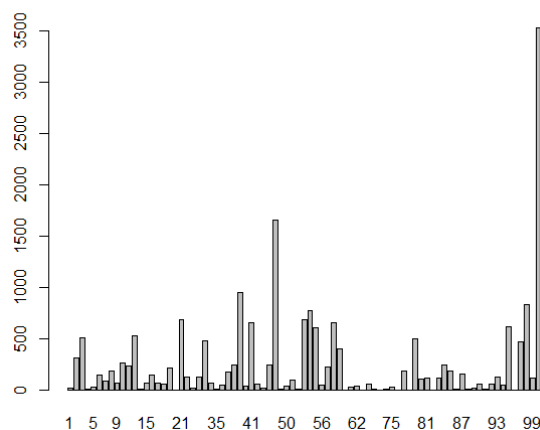
Se observa que esta variable es de clase “factor”, es decir, es categórica.

```
> class(oficio_)
[1] "factor"
```

También tiene 81 niveles, es decir, no se relacionan las 100 categorías como tal.

```
> levels(oficio_)
[1] "1" "2" "3" "4" "5" "6" "7" "8" "9" "11" "12" "13" "14" "15" "16" "17" "18" "19" "20" "21" "30" "31" "32" "33"
[25] "34" "35" "36" "37" "38" "39" "40" "41" "42" "43" "44" "45" "49" "50" "51" "52" "53" "54" "55" "56" "57" "58" "59" "60"
[49] "61" "62" "63" "70" "72" "73" "74" "75" "76" "77" "78" "79" "80" "81" "82" "83" "84" "85" "86" "87" "88" "89" "90" "91"
[73] "92" "93" "94" "95" "96" "97" "98" "99" "100"
```

En el diagrama de barras, se observa que la categoría no tiene una función uniforme.



La moda para esta variable es la categoría 100, es decir, personas que no reportaron ocupación alguna.

### 1.3. Perfil edad-ingresos

**1.3.1. 1.3.1 En el set de datos, múltiples variables describen el ingreso. Escoja una que Ud. considere la más representativa para el ingreso total de un trabajador, justificando su selección.**

La variable más representativa de los ingresos totales es la denominada “*ingtot*” (Ingresos totales) ya que esta agrupa las diversas variables acerca de los ingresos. Al comparar diferentes variables de los ingresos, se puede observar que los ingresos totales son la suma de cada una del resto de las variables. Por ejemplo:

Se comparó que la Variable *ingtot* es la suma entre *ingtotes* e *ingtotob*

	↑ ingtot	ingtotes	ingtotob
<b>14</b>	883357.0	NA	883357.0
<b>15</b>	0.0	NA	0.0
<b>16</b>	1200000.0	500000	700000.0
<b>17</b>	0.0	NA	0.0
<b>18</b>	1000000.0	NA	1000000.0
<b>19</b>	981000.0	NA	981000.0

A pesar de tener evidencia de que *ingtot* es la variable que representa estas otras dos variables, es pertinente analizar las demás variables que pueden componer el ingreso, siendo estos, los ingresos por intereses, por ayudas, monetarios, arriendos, especie y monetarios para tener certeza acerca de la variable representativa *ingtot*. En la siguiente tabla, se encuentran las variables y su significado

iof1	Ingreso por intereses y dividendos antes de imputación
iof1es	Ingreso por intereses y dividendos imputado (sólo para faltantes o extremos)
iof2	Ingreso por jubilaciones y pensiones antes de imputación
iof2es	Ingreso por jubilaciones y pensiones imputado (sólo para faltantes o extremos)
iof3h	Ingreso por ayudas de hogares, antes de imputación
iof3hes	Ingreso por ayudas de hogares, imputado (sólo para faltantes o extremos)
iof3i	Ingreso por ayudas de instituciones, antes de imputación
iof3ies	Ingreso por ayudas de instituciones, imputado (sólo para faltantes o extremos)
iof6	Ingreso por arriendos antes de imputación
iof6es	Ingreso por arriendos imputado (sólo para faltantes o extremos)
isa	Ingreso monetario de la segunda actividad antes de imputación
isaes	Ingreso monetario de la segunda actividad imputado (sólo para faltantes o extremos)

Impa	Ingreso monetario de la primera actividad antes de imputación
impaes	Ingreso monetario de la primera actividad imputado (sólo para faltantes, extremo)

	ingtot	iof1	iof1es	iof2	iof2es	iof3h	iof3hes	iof6	iof6es	isa	isaes	ie	iees	imdies	impa
6	737717.0	0	NA	0	NA	0.0	NA	0	NA	NA	NA	NA	NA	NA	NA
7	0.0	0	NA	0	NA	0.0	NA	0	NA	NA	NA	NA	NA	NA	NA
8	1500000.0	0	NA	0	NA	0.0	NA	1500000	NA	NA	NA	NA	NA	NA	NA
9	1878973.3	0	NA	0	NA	0.0	NA	500000	NA	0	NA	0	NA	NA	1378973.3
10	0.0	0	NA	0	NA	0.0	NA	0	NA	NA	NA	NA	NA	NA	NA
11	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Al observar que efectivamente algunas de estas variables de ingreso suman en algunas observaciones el ingreso total se asume que efectivamente la variable que describe todo el ingreso. Sin embargo, para realizar una última comparación, la variable *ingtot* se comparará con las diferentes primas y subsidios:

_y_primaServicios_m	Ingreso prima servicios monetario ultimos 12 meses
y_primaVacaciones_m	Ingreso prima vacaciones monetario ultimos 12 meses
y_primas_m	Ingreso primas monetario en el mes
y_salarySec_m	salary nominal mensual occ. secundario
y_subEducativo_m	Ingreso subsidio educativo monetario en el mes
y_subFamiliar_m	Ingreso subsidio familiar monetario en el mes

	ingtot	y_primas_m	y_primaVacaciones_m	y_primaServicios_m	y_primaNavidad_m
170	11177083.3	NA	3800000	7500000	3800000
171	6018333.3	NA	NA	5200000	2800000
172	1461473.3	NA	NA	700000	NA
173	0.0	NA	NA	NA	NA
174	0.0	NA	NA	NA	NA
175	0.0	NA	NA	NA	NA

Como se puede observar, algunas de estas variables suman en la contabilidad de algunas observaciones del ingreso total. Por lo tanto, tomando como evidencia todas las verificaciones anteriores, se comprueba que la variable que describe el ingreso es *ingtot* (Ingreso total), ya que esta contiene todas las demás variables acerca del ingreso que se encuentran en la base de datos de la GEIH

**Nota:** Si desea ver cada una de las tablas sustentadas anteriormente, diríjase al script en R que contiene toda la programación.

### 1.3.2. Con base en esta estimación utilizando OLS, la ecuación del perfil de edad-ingresos:

$$\text{Ingreso} = \beta_1 + \beta_2 \text{Edad} + \beta_3 \text{Edad}^2 + u \quad (2)$$

### 1.3.3. ¿Qué tan bueno es este modelo en el ajuste de la muestra?

Dependent variable:	
----- lningtot -----	
age	0.048*** (0.002)
Age2	-0.0005*** (0.00002)
Constant	12.881*** (0.047)
-----	
Observations	19,801
R2	0.025
Adjusted R2	0.025
Residual Std. Error	0.964 (df = 19798)
F Statistic	256.001*** (df = 2; 19798)
=====	
Note:	*p<0.1; **p<0.05; ***p<0.01

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 con el ingreso ya que alcanzó un 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 total.

La variable *ingtot* (Ingreso Total) 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 se los datos sean tratados de manera efectiva y sin inconvenientes en la ejecución de la regresión por MCO.

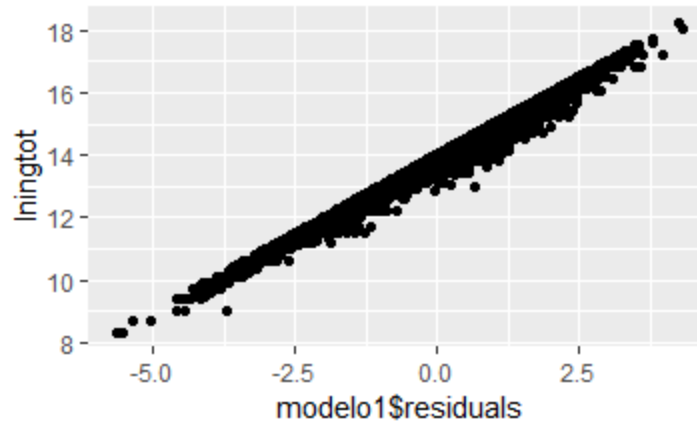


La constante no suele generar ningún impacto en el modelo, ya que esta es representativa cuando  $X_i$  puede tomar el valor 0. Sin embargo, en este modelo no es posible que las variables tomen ese valor, ya que si fuese el caso, no estarían dentro del modelo. De acuerdo con lo anterior, la constante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

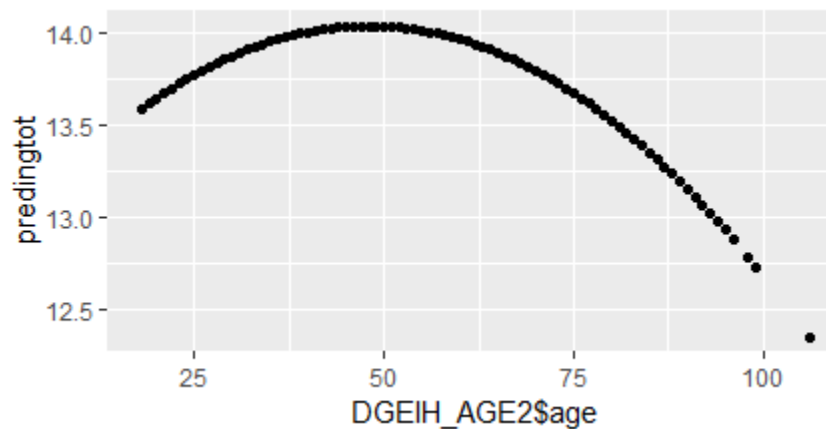
El coeficiente de *Age* hace referencia a que cuando un individuo aumenta un año de vida, el ingreso de este aumenta en 4.8% su ingreso, contando con un error estándar de la variable de 0.02, siendo este el que mide la precisión con la que cuenta la variable respecto a los valores estimados. A su vez, 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 0.00002, siendo este valor muy pequeño.

Debido a la limpieza de datos generada, se contó con 19801 observaciones y un mismo valor de  $R^2$  y  $R^2$  ajustado de 0.025, representando el poco ajuste que tienen las variables del modelo a la variable independiente, ingresos totales, se puede identificar que es necesario contar con más variables explicativas para poder identificar qué genera el ingreso total 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. El error estándar residual es de 0.964 siendo este el valor que nos indica que 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.



**1.3.4. Grafique el perfil proyectado de ingresos por edad implícito en la ecuación anterior.**

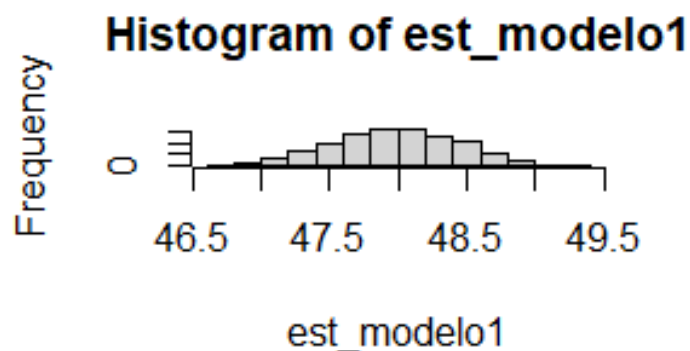


La gráfica señala el ingreso marginal decreciente, en donde a medida que la edad de un individuo aumenta, la tendencia del aumento del ingreso va disminuyendo, haciendo que el individuo de una edad en adelante, no aumente significativamente sus ingresos. Por ejemplo, los ingresos de una persona de 22 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.

**1.3.5. ¿Cuál es la “edad pico” sugerida por la ecuación anterior? Usar bootstrap para calcular los errores estándar y construir los intervalos de confianza.**

Dependent variable:	
lningtot	
age	0.048*** (0.002)
Age2	-0.0005*** (0.00002)
Constant	12.881*** (0.047)

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:



```
> mean(est_modelo1)
[1] 47.98265
> sqrt(var(est_modelo1))
[1] 0.4503915
> quantile(est_modelo1, c(0.025, 0.975))
      2.5%      97.5%
47.11488 48.83212
```

Aquel vector está centrado en 48, contando con una distribución aparentemente normal. Y la variabilidad corresponde a 0.45 y junto con un intervalo de confianza de 95%, existe el rango de que el dato verdadero esté entre 47.11 - 48. Por otra parte, se encuentran los errores estándar por medio de la función boot, demostrando que el modelo no es el más eficiente para explicar todo el

ingreso total, es decir, hay más variabilidad que el modelo no puede capturar, aún teniendo el  $Age^2$ . Además, señala el sesgo que cada una de las variables presenta.

```

Bootstrap Statistics :
      original      bias      std. error
t1* 12.8807652420  2.746925e-04  4.909573e-02
t2*  0.0475273221 -2.350207e-05  2.312220e-03
t3* -0.0004950201  1.791477e-07  2.461502e-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  $PeakAge <-((-Beta1/-2*Beta2)$  que se obtiene de realizar el siguiente procedimiento:

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

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

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

El resultado de esta “edad pico” es una edad aproximada de 48 años:

```

> PeakAge
      age
47.56142

```

#### 1.4.La brecha de ganancias

Estimar la brecha de ingresos incondicional:

$$\log(ingreso) = \beta_1 + \beta_2 Femenino + u$$

**1.4.1. ¿Cómo debemos interpretar el coeficiente  $\beta_2$ ? ¿Qué tan bueno es este modelo en samplefit?**

Este  $\beta_2$  es la diferencia porcentual ( $100 * \beta_2$ ) en el ingreso promedio entre hombres y mujeres. Si dejamos que esta variable sea igual a 1 para los hombres como está en la base de datos, significa que, puede ser la diferencia positiva o negativa con respecto a lo que ganan las mujeres.

Para realizar la regresión, se creó la variable “sex\_female”, que es el opuesto de la variable “sex”, ya que en “sex” el 1 representaba a los hombres y cero a las mujeres.

Para el Samplefit, se presenta el resultado de la regresión:

Dependent variable:	
lningt	
sex_female	-0.242*** (0.014)
Constant	13.990*** (0.010)
Observations	19,801
R2	0.015
Adjusted R2	0.015
Residual Std. Error	0.969 (df = 19799)
F Statistic	309.299*** (df = 1; 19799)
Note: *p<0.1; **p<0.05; ***p<0.01	

Por los resultados obtenidos, se observa que el R2 es bajo, por lo cual se puede inferir que el modelo no se ajusta a la muestra. Sin embargo, al revisar la significancia de los coeficientes, podemos observar que estos son significativos al 1%, por lo cual, hay suficiente evidencia estadística para afirmar que en promedio, las mujeres tienen un ingreso 24,2% inferior que los hombres. Se observa también que el error estándar residual es alto, con ello se refuerza que el modelo no se ajusta completamente a los datos.

#### **1.4.2. Estimar y trazar el perfil de edad-ingresos pronosticado por género. ¿Los hombres y las mujeres en Bogotá tienen la misma intersección y pendientes?**

Se realizó la regresión utilizando el modelo del punto 1.3, sin embargo, se calculó para la base de datos cuando sex\_female=1 (mujeres) y sex\_female=0 (hombres).

Se utilizó R para realizar la regresión incorporando la edad para presentar lo solicitado, utilizando la transformación logarítmica para el ingreso, con lo cual, el resultado de la regresión se muestra en la siguiente figura:

Dependent variable:		
	lningtot	
	(1)	(2)
age	0.031*** (0.003)	0.063*** (0.003)
Age2	-0.0004*** (0.00003)	-0.001*** (0.00003)
Constant	13.159*** (0.071)	12.619*** (0.062)
Observations	9,754	10,047
R2	0.015	0.047
Adjusted R2	0.015	0.047
Residual Std. Error	1.017 (df = 9751)	0.889 (df = 10044)
F Statistic	73.557*** (df = 2; 9751)	249.737*** (df = 2; 10044)

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

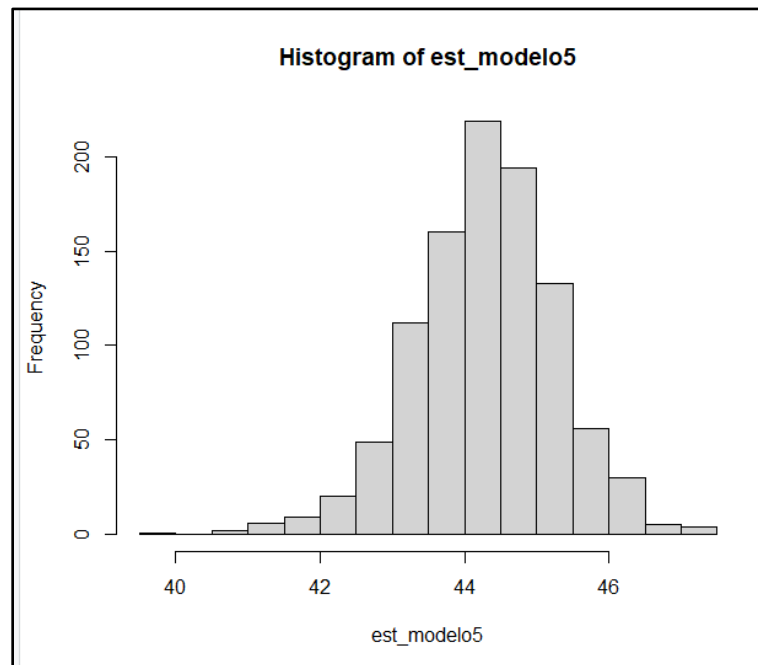
Se observa de la tabla de la regresión que las pendientes de ambas regresiones son diferentes (mirando el coeficiente  $\beta_2$  de age, se presenta que para mujeres y hombres es diferente, es menor para la regresión realizada con el subconjunto de mujeres), y también el intercepto es diferente para cada regresión. En la siguiente gráfica, donde la línea color azul representa a los hombres y la roja a las mujeres, se puede evidenciar lo anteriormente mencionado. De la regresión anterior, se evidencia que las variables son significativas para los modelos, sin embargo, el  $R^2$  presenta un bajo valor, que quiere decir que el modelo solo explica el 1,5% de los datos.



**1.4.3. ¿Cuál es la “edad pico” implícita por género?. Utilice bootstrap para calcular los errores estándar y construir los intervalos de confianza. ¿Se superponen estos intervalos de confianza?**

Así como en el punto anterior, se utilizará la herramienta “Bootstrap” para caracterizar la variabilidad del modelo para hombres y mujeres. Se empezará con el modelo para las mujeres:

Se utilizó la semilla= 10101 y R=1000, con lo cual, se presenta la siguiente distribución:



```
> mean(est_modelo5)
[1] 44.28249
> sqrt(var(est_modelo5))
[1] 0.9901475
> quantile(est_modelo5, c(0.025, 0.975)) #Los intervalos de confianza
      2.5%      97.5%
42.23272 46.22695
```

El vector está centrado en 44.28, contando con una distribución aparentemente normal, con un ligero sesgo a la derecha (sesgo negativo). Y la variabilidad, muy pequeña es 0.99 y junto con un intervalo de confianza de 95%, se tiene que el valor verdadero para esta variable se encuentra entre 42.23 y 46.23. Por otra parte, se encuentran los errores estándar por medio de la función boot, demostrando que el modelo, al igual que en el punto 1.3, no es el más eficiente para explicar todo el ingreso total, es decir, hay más variabilidad que el modelo no puede capturar, aún teniendo el  $Age^2$

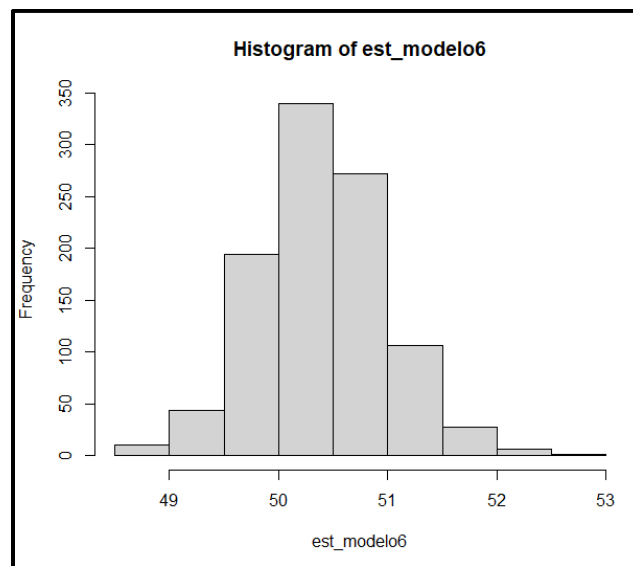
```
Call:
boot(data = DGEIH_AGE2_female, statistic = estimboot5, R = 1000)

Bootstrap Statistics :
      original      bias      std. error
t1* 13.1593348282  1.114962e-03  7.099219e-02
t2*  0.0311523085 -6.545399e-05  3.237560e-03
t3* -0.0003513574  6.565673e-07  3.343081e-05
```

Finalmente, se realiza el *peak age* con la siguiente fórmula  $\text{Peak\_Age5} <- ((-\text{Beta1})/\text{Beta2}) * (1/2)$ , aproximado para las mujeres es de 45 años:

```
> PeakAge5
age
44.61974
```

A continuación, se realizará el análisis para los hombres. Se utilizó la semilla= 10101 y R=1000, con lo cual, se presenta la siguiente distribución:



```
> mean(est_modelo6)
[1] 50.4014
> sqrt(var(est_modelo6))
[1] 0.5814119
> quantile(est_modelo6, c(0.025, 0.975)) #Los intervalos de confianza
      2.5%      97.5%
49.29303 51.59322
```

El vector está centrado en 50.4, contando con una distribución aparentemente normal, con un sesgo positivo (datos a la izquierda). La variabilidad fue de 0.58 y junto con un intervalo de confianza



de 95%, se tiene que el valor verdadero para esta variable se encuentra entre 49.29 y 51.59. Por otra parte, se encuentran los errores estándar por medio de la función boot.

```
Call:
boot(data = DGEIH_AGE2_male, statistic = estimboot6, R = 1000)

Bootstrap Statistics :
      original      bias      std. error
t1* 12.6192974712 -8.086208e-04 6.599666e-02
t2*  0.0626568121  2.357912e-05 3.203377e-03
t3* -0.0006216833 -2.729764e-07 3.533921e-05
```

Finalmente, se realiza el *peak age* con la siguiente fórmula  $\text{PeakAge6} <- ((-\text{Beta1})/\text{Beta2})^{(1/2)}$ , aproximado para los hombres es de 51 años:

```
> PeakAge6
age
50.91967
```

Con respecto a lo encontrado en la función de ingresos considerando la edad y la edad al cuadrado, el resultado del punto 1.3 y los hallados en este punto son diferentes. Sin tener en cuenta el sexo, la edad máxima para recibir el ingreso mayor se encontró en 48 años; sin embargo, en este ejercicio realizado por los géneros aparte, da como resultado que las mujeres logran un ingreso mayor a menos edad que los hombres (45 vs 51 años), lo cual quiere decir que, en ese sentido, también existe una brecha entre hombres y mujeres, ya que las mujeres tienen menor tiempo para lograr el máximo de ingresos, según lo obtenido utilizando el Bootstrap. Así mismo, se observa que los intervalos de confianza para el análisis entre hombres y mujeres no se traslapan, siendo los menores valores obtenidos para el coeficiente  $\beta_2$  los obtenidos para las mujeres.

#### **1.4.4. Estime una brecha de ingresos condicional que incorpore variables de control tales como características similares del trabajador y del puesto (X).**

En la base de datos, se encontró la variable: “Oficio”, la cual contiene 100 categorías describiendo los oficios de las personas encuestadas de la base de datos. Con el fin de realizar un análisis de la variable “características similares del trabajador y del puesto”, asumiremos que el oficio es la base para determinar la similitud en el trabajo. Ahora bien, como tener 100 categorías podría ser un número considerable para analizar, se decidió crear una nueva variable a partir del ingreso total y

del oficio para reducir estos oficios en 5 categorías. ¿Cómo se hizo dicha variable? Se describirá el proceso por el cual se creó la variable: “Clasificación\_oficio\_similar”:

- i. Se tomó el promedio de ingresos para cada categoría de oficio.
- ii. Se obtuvo el cuartil 20%, 40%, 60% y 80% para este promedio de ingresos por oficio.
- iii. Con base en estos cuartiles para dividir los datos de manera igual, se les asignó una categoría a cada cuartil, y se obtuvieron los oficios para los cuales:
  1. Para las personas con determinado oficio e ingreso promedio menor al 20%, se les asignó el número 1.
  2. Para las personas con determinado oficio e ingreso promedio entre el 20% al 40%, se les asignó el número 2.
  3. Para las personas con determinado oficio e ingreso promedio entre el 40% al 60%, se les asignó el número 3.
  4. Para las personas con determinado oficio e ingreso promedio entre el 60% al 80%, se les asignó el número 4.
  5. Para las personas con determinado oficio e ingreso promedio superior al 80%, se les asignó el número 5.

Los cuantiles obtenidos para este caso, se presentan a continuación:

```
> q
      20%      40%      60%      80%
1010547 1185972 1686921 2911667
```

Los resultados de la asignación de la categoría de dicha variable, se encuentra en la figura presentada a continuación, donde la columna “Group.1” hace referencia a las categorías de oficio encontrada en la base de datos, “x” el promedio de salarios para ese oficio y “Clasificación\_oficio\_similar” es la nueva variable donde se obtiene el oficio con característica similar para los ingresos.

Group.1	x	Clasificacion_oficio_similar
59	78	700000.0
61	80	775951.7
56	75	778293.6
44	56	784910.0
80	99	804236.9
42	54	825460.4
27	36	837644.2
43	55	854662.0
75	94	867121.5
60	79	872151.3
63	82	875000.0
74	93	935750.5
45	57	953067.6
78	97	969445.8
62	81	979456.4
41	53	997421.5
72	91	1010546.7
53	72	1012000.3
28	37	1022911.8
51	63	1025464.7
55	74	1042869.1
57	76	1054248.2
68	87	1054935.6
70	89	1056019.3
58	77	1071778.5
71	90	1077019.5
36	45	1102971.7
40	52	1120889.6
76	95	1122555.1
64	83	1155455.8
50	62	1165297.9
29	38	1171568.7
13	14	1185972.2
47	59	1203849.0
73	92	1206646.0
65	84	1229157.1
81	100	1240770.0
79	98	1246160.4
66	85	1312864.2
25	34	1359352.4
17	18	1420832.8
32	41	1426648.2
24	33	1510671.1
46	58	1530167.5
67	86	1530950.4
39	51	1572365.5

#### 1.4.4.1 (a) Estime la brecha de ingresos condicional

$$\log(\text{ingreso}) = \beta_1 + \beta_2 \text{Femenino} + \theta X + u$$

Se realizó la regresión presentada, con las siguientes variables que se encuentran en la base de datos utilizada:

$$\log(\text{ingreso}) = \beta_1 + \beta_2 \text{Sex}_{\text{female}} + \theta(\text{Clasificacion}_{\text{oficio}_{\text{similar}}}) + u$$

Los resultados de la regresión se muestran en la siguiente tabla:

Dependent variable:	
lningtot	
sex_female	-0.236*** (0.013)
categor_oficio2	0.082*** (0.021)
categor_oficio3	0.167*** (0.017)
categor_oficio4	0.780*** (0.029)
categor_oficio5	1.340*** (0.022)
Constant	13.669*** (0.016)
-----	
Observations	19,801
R2	0.221
Adjusted R2	0.220
Residual Std. Error	0.862 (df = 19795)
F Statistic	1,120.283*** (df = 5; 19795)
=====	
Note: *p<0.1; **p<0.05; ***p<0.01	

Se observa en la tabla que la variable categoría del oficio es relevante solo para los rangos de 2 a 5, es decir, cuando se tiene un ingreso mayor. Ahora bien, se observa que la variable *sex\_female*, si bien se disminuye el porcentaje con respecto al caso de la regresión en el punto 1.4.1, se observa que persiste la brecha de ingresos, ya que en promedio las mujeres ganan -23,6% que los hombres, por lo cual, independiente de si laboran en las ocupaciones que pertenecen a los rangos 2 a 5 de la categoría de oficios, sigue existiendo brechas para el ingreso de las mujeres.

El  $R^2$  es un poco mayor que en las regresiones (se obtuvo 0.221), por lo cual, se tiene una baja bondad de ajuste para el modelo.

#### **1.4.4.2 Use FWL para repetir la estimación anterior, donde el interés radica en $\beta_2$ . ¿Obtiene las mismas estimaciones?**

Se utilizó el teorema FWL para realizar la estimación anterior. Los modelos a utilizar para realizar esta estimación se muestran a continuación:

$$\log(ingtot) = \theta(\text{Clasificacion}_{oficio_{similar}})(1)$$

$$sex\_female = \theta(\text{Clasificacion}_{oficio_{similar}})(2)$$

$$e_1 = e_2(3)$$

Donde los  $e$  son los residuos de las regresiones (1) y (2). Luego de correr el script en R, se obtienen los siguientes resultados de las regresiones:

Dependent variable:			
	lningtot (1)	sex_female (2)	residuals (3)
Categor_oficio2	0.155*** (0.021)	-0.310*** (0.012)	
Categor_oficio3	0.217*** (0.017)	-0.212*** (0.009)	
Categor_oficio4	0.802*** (0.029)	-0.093*** (0.016)	
Categor_oficio5	1.383*** (0.022)	-0.180*** (0.012)	
residuals			-0.236*** (0.013)
Constant	13.512*** (0.014)	0.668*** (0.008)	-0.000 (0.006)
Observations	19,801	19,801	19,801
R2	0.207	0.040	0.018
Adjusted R2	0.206	0.040	0.018
Residual Std. Error	0.870 (df = 19796)	0.490 (df = 19796)	0.862 (df = 19799)
F Statistic	1,288.343*** (df = 4; 19796)	207.363*** (df = 4; 19796)	355.776*** (df = 1; 19799)
Note: ***p<0.01; **p<0.05; *p<0.1			

Con el ejercicio anterior, se comprueba el teorema FWL, ya que se obtuvo el mismo coeficiente  $\beta_2$  que acompaña la variable asociada al género femenino. Sin embargo, se presentan diferencias en cuanto al R2. Los errores estándar permanecen relativamente bajos. Como se ve en la figura siguiente, los coeficientes son iguales así como los errores estándar, por lo cual, se comprueba el teorema FWL.

Dependent variable:				
	lningtot (1)	(2)	sex_female (3)	residuals (4)
sex_female	-0.236*** (0.013)			
Categor_oficio2	0.082*** (0.021)	0.155*** (0.021)	-0.310*** (0.012)	
Categor_oficio3	0.167*** (0.017)	0.217*** (0.017)	-0.212*** (0.009)	
Categor_oficio4	0.780*** (0.029)	0.802*** (0.029)	-0.093*** (0.016)	
Categor_oficio5	1.340*** (0.022)	1.383*** (0.022)	-0.180*** (0.012)	
residuals				-0.236*** (0.013)
Constant	13.669*** (0.016)	13.512*** (0.014)	0.668*** (0.008)	-0.000 (0.006)
Observations	19,801	19,801	19,801	19,801
R2	0.221	0.207	0.040	0.018
Adjusted R2	0.220	0.206	0.040	0.018
Residual Std. Error	0.862 (df = 19795)	0.870 (df = 19796)	0.490 (df = 19796)	0.862 (df = 19799)
F Statistic	1,120.283*** (df = 5; 19795)	1,288.343*** (df = 4; 19796)	207.363*** (df = 4; 19796)	355.776*** (df = 1; 19799)
Note: ***p<0.01; **p<0.05; *p<0.1				

**1.4.4.3 ¿Cómo debemos interpretar el coeficiente  $\beta_2$ ? ¿Qué tan bueno es este modelo en ajuste de muestra? ¿Se reduce la brecha? ¿Es esta**

**evidencia de que la brecha es un problema de selección y no un  
"problema de discriminación"?**

El coeficiente  $\beta_2$  se debe interpretar como el porcentaje promedio de diferencia entre el ingreso de hombres y mujeres, es decir, se mantiene la misma interpretación del punto 1.4.1. Se nota que cuando se incorpora la variable de “Categoría Oficio” este porcentaje disminuye con respecto al obtenido en el punto 1.4.1 donde la brecha era mayor (-24.1% vs -23.6%). Si bien la brecha disminuyó al incorporar la variable oficio, esta persiste, por lo cual, sí existe evidencia que hay discriminación en contra de las mujeres al momento de los ingresos que estas perciben.

**1.5. Predicción de ganancias.**

**1.5.1. Divida la muestra en dos muestras: una muestra de entrenamiento (70%) y una muestra de prueba (30%). No olvide establecer una semilla (en R, `set.seed(10101)`, donde 10101 es la semilla).**

En el Script de R, se realiza lo mencionado anteriormente, para establecer los datos asociados a “train” y a “test”, incluyendo la semilla propuesta para que se pueda reproducir el resultado sin inconvenientes, que ya se había usado en los puntos anteriores.

**1.5.1.1. Estime un modelo que solo incluya una constante. Este será el punto de referencia.**

Se estableció el modelo que incluye solo una constante, empleando la regresión con el número 1, es decir:

$$\log(\text{ingtot}) = 1 \quad (1)$$

Al obtener el resultado de correr el código en R, se obtiene que el coeficiente es igual a: 13., que sería el promedio del logaritmo natural del ingreso total. Se esperan estos valores con gran orden de magnitud, teniendo en cuenta que se está trabajando con el logaritmo natural. Esto se puede observar en la imagen a continuación:

Dependent variable:	
lningtot	
Constant	13.870*** (0.008)

### 1.5.1.2. Estime nuevamente sus modelos anteriores

Se realiza la estimación de los modelos anteriores asociados a:

$$\log(\text{income}) = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + u \quad (1)$$

$$\log(\text{ingreso}) = \beta_1 + \beta_2 \text{Sex}_{\text{female}} + u \quad (2)$$

$$\log(\text{ingreso}) = \beta_1 + \beta_2 \text{Sex}_{\text{female}} + \theta(\text{Clasificacion}_{\text{oficio}_{\text{similar}}}) + u \quad (3)$$

Con la base “train” y se calcula con base en el “test” el error cuadrático medio.

A continuación se presenta el resultado de la regresión con base en la base de datos: “train”:

Dependent variable:			
	(1)	lningtot (2)	(3)
age	0.045*** (0.003)		
Age2	-0.0005*** (0.00003)		
sex_female		-0.249*** (0.017)	-0.242*** (0.015)
Categor_oficio2			0.097*** (0.025)
Categor_oficio3			0.182*** (0.020)
Categor_oficio4			0.795*** (0.035)
Categor_oficio5			1.367*** (0.026)
Constant	12.943*** (0.057)	13.993*** (0.012)	13.658*** (0.020)

Se observa que las regresiones tienen unos coeficientes diferentes a los obtenidos en los puntos anteriores, esto porque la regresión se obtiene de un data set diferente al usado en los puntos anteriores. Los errores estándar, permanecen de baja dimensión (menores a 0.1). El único valor que permaneció igual, fue el coeficiente asociado a la edad al cuadrado ( $\text{age}^2$ ).

**1.5.1.3. En las secciones anteriores, los modelos estimados tenían diferentes transformaciones de la variable dependiente. En este punto, explora también otras transformaciones de tus variables independientes. Por ejemplo, puede incluir términos polinómicos de ciertos controles o interacciones de estos. Pruebe al menos cinco (5) modelos que aumentan en complejidad.**

Para realizar este numeral, se tomaron otros modelos y otras variables que se habían calculado en el punto 1.2 para que sean incorporadas al modelo y así tener más opciones con las cuales realizar la predicción.

Los modelos seleccionados se presentan a continuación (incluyendo los que se trabajaron a lo largo del Problem Set):

$$\log(ingtot) = 1 \quad (1)$$

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

$$\log(ingreso) = \beta_1 + \beta_2 Sex_{female} + u \quad (3)$$

$$\log(ingreso) = \beta_1 + \beta_2 Sex_{female} + \theta(Clasificacion_{oficio_{similar}}) + u \quad (4)$$

Los siguientes modelos, son los nuevos propuestos y que incrementan en complejidad:

$$\log(ingreso) = \beta_1 + \beta_2 Sex_{female} + \beta_3 ocu + u \quad (5)$$

$$\log(ingreso) = \beta_0 + \beta_1 ocu + \beta_2 age + \beta_3 exp + \beta_4 sex_{female} * age^2 + u \quad (6)$$

$$\log(ingreso) = \beta_1 + \beta_2 sex_{female} + \beta_3 ocu + \beta_4 exp + \beta_5 age + \beta_6 age^2 + u \quad (7)$$

$$\log(ingreso) = \beta_1 + \beta_2 sex_{female} + \beta_3 ocu + \beta_4 exp + \beta_5 age + \beta_6 age^2 + \beta_7 age^3 + u \quad (8)$$

$$\log(ingreso) = \beta_1 + \beta_2 sex_{female} + \beta_3 ocu + \beta_4 exp + \beta_5 age + \beta_6 age^2 + \beta_7 age^3 + \beta_8 age^4 + u \quad (9)$$

Los anteriores son los modelos propuestos, en donde se trató de aumentar la complejidad de la variable “age”, que ya era cuadrática en el modelo principal.

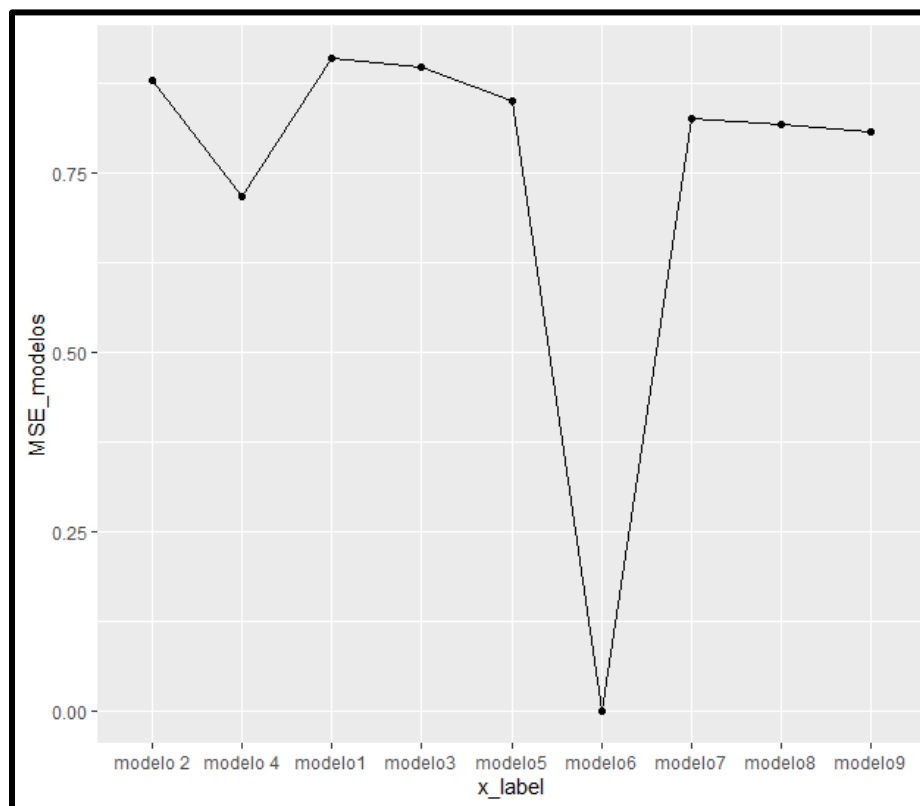


**1.5.1.4. Informe y compare el error de predicción promedio de todos los modelos que estimó anteriormente. Discuta el modelo con el error de predicción promedio más bajo.**

Para cada modelo mencionado en el numeral anterior se calculó el error cuadrático medio para las predicciones. Estas se muestran en la tabla a continuación.

x_label	MSE_modelos
modelo6	1.362766e-05
modelo 4	7.183074e-01
modelo9	8.073089e-01
modelo8	8.181464e-01
modelo7	8.260628e-01
modelo5	8.518219e-01
modelo 2	8.803824e-01
modelo3	8.980601e-01
modelo1	9.107985e-01

Al graficar los MSE, se obtiene lo siguiente:



Con base en lo anterior, se puede observar que el modelo 6 es el que tiene menor error cuadrático medio (MSE). Los modelos 7, 8 y 9 tienen los polinomios de orden 2 hasta 4 para la variable “age”, con lo cual se constata que entre mayor orden y mayor cantidad de variables que se consideren en el modelo, el MSE disminuye, sin embargo, vuelve a aumentar porque ya dichas variables desmejoran la predicción.

El modelo 6 es el siguiente:

$$\log(\text{ingreso}) = \beta_0 + \beta_1 \text{ocu} + \beta_2 \text{age} + \beta_3 \text{exp} + \beta_4 \text{sex}_{\text{female}} * \text{age}^2 + u \quad (6)$$

En donde se constata que, la interacción entre la variable dummy: sex\_female con la edad al cuadrado (age2) se consideraría importante para determinar si el age al cuadrado se tiene en cuenta en la regresión cuando sex\_female es igual a 1, y mejora considerablemente el MSE del modelo.

**1.5.1.5. Para el modelo con el error de predicción promedio más bajo, calcule la estadística de apalancamiento para cada observación en la muestra de prueba. ¿Hay valores atípicos, es decir, observaciones con un alto apalancamiento que impulsen los resultados? ¿Son estos valores atípicos personas potenciales que la DIAN debería investigar, o son simplemente el producto de un modelo defectuoso?**

Se realizó el análisis del “leverage” o apalancamiento, bajo la siguiente expresión:

$$\alpha = \frac{u_j}{1 - h_j}$$

Donde  $\alpha$  es el parámetro de “leverage”, obtenido del residuo del modelo, dividido entre la diagonal de la matriz Px.

Para el alfa, se determinó como criterio del grupo revisar cuántas observaciones eran mayores a 1 o menores a -1, para tener una idea de la magnitud de este valor, y si hay valores que sean atípicos o con gran apalancamiento y por ello están teniendo influencia en la regresión.

Se determinó que el 22,42% de los datos tenían valores mayores a 1 o menores a -1, teniendo en cuenta que el mejor de los casos es que sean menores, para que el  $u_j$ , si está sobre la regresión, sea cero y el alfa también lo sea. El resultado del cálculo se presenta en la figura siguiente.

	<b>x</b>
<b>1</b>	22.42424

Adicionalmente, se determinó el mayor y el mínimo leverage para esta regresión. Los datos se presentan en la figura siguiente:

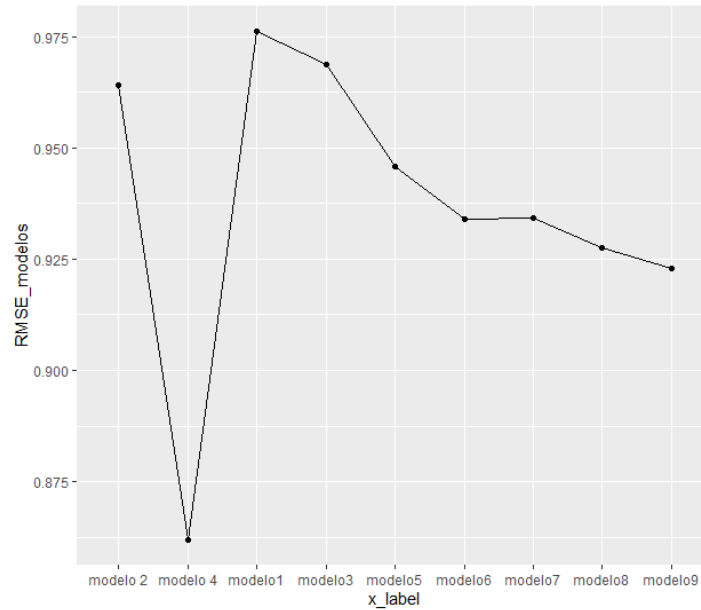
```
> max(alphass)
[1] 3.668015
> min(alphass)
[1] -5.299713
```

El mayor valor fue 5.3 (en magnitud), por lo cual, podemos concluir que los datos escogidos y la regresión realizada no están siendo afectados por datos con un gran leverage (mayor a 10, por ejemplo), por lo cual, podemos concluir que para este modelo, no se evidencian datos atípicos que afecten la regresión.

**1.5.2. Repita el punto anterior pero use la validación cruzada K-fold. Comente las similitudes/diferencias del uso de este enfoque.**

Para los modelos presentados en el numeral 1.5.1.3, se realizó la validación cruzada K-fold, con  $k=5$  (valor determinado a priori) con la función que se encuentra en la librería “caret” de R. Con base en este cálculo, se obtendrán los RMSE que arroja cada regresión.

Se presenta la gráfica del RMSE de cada modelo para validar los resultados. En este caso, se obtiene que el modelo con menor RMSE es el modelo 4. Si bien en el punto 5.1.3 se realizó el cálculo del MSE, el RMSE es solo obtener la raíz cuadrada, por lo cual, los órdenes de magnitud se mantienen y es válido el análisis al determinar, que contrario a lo obtenido en el punto 5.1.3 donde el modelo de menor MSE fue el modelo6, en este caso fue el modelo 4.



Se presenta la tabla donde se clasifican los RMSE por modelos:

x_label	RMSE_modelos
modelo 4	0.8619700
modelo9	0.9228124
modelo8	0.9274051
modelo6	0.9339405
modelo7	0.9340982
modelo5	0.9457839
modelo 2	0.9640007
modelo3	0.9687223
modelo1	0.9761657

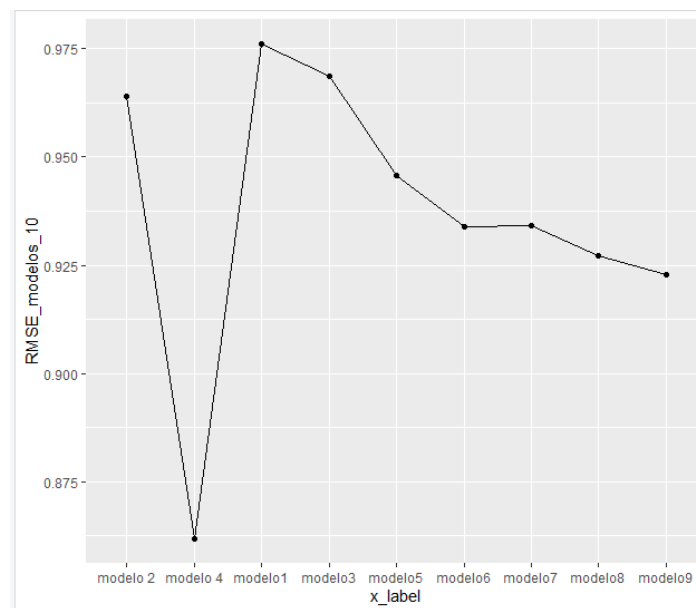
El modelo 4 es el siguiente:

$$\log(\text{ingreso}) = \beta_1 + \beta_2 \text{Sex}_{\text{female}} + \theta(\text{Clasificacion}_{\text{oficio}_{\text{similar}}}) + u \quad (4)$$

En donde se relaciona la variable de género y la construida de “Clasificación oficios”.

Se repetirá el ejercicio con k= 10 para ver si se obtienen los mismos resultados. La tabla con los RMSE ordenados para k=10 y la gráfica de RMSE por modelo, se presentan a continuación:

x_label	RMSE_modelos_10
modelo 4	0.8618962
modelo9	0.9227284
modelo8	0.9271779
modelo6	0.9338875
modelo7	0.9340762
modelo5	0.9457628
modelo 2	0.9639441
modelo3	0.9686485
modelo1	0.9760729



Si bien los valores tienen ligeras variaciones, se constata que el modelo 4 es el de menor RMSE.

**1.5.3. Con su modelo predicho preferido (el que tiene el promedio más bajo error de predicción) realice el siguiente ejercicio:**

**Escribe un bucle que haga lo siguiente:**

**1.5.3.1. Estimar el modelo de regresión utilizando todas las observaciones menos la  $i$  – ésima.**

**1.5.3.2. Calcular el error de predicción para la  $i$  – ésima observación, es decir,  $(y_i - \hat{y}_i)$**

**1.5.3.3. Calcular el promedio de los números obtenidos en el paso anterior para obtener el error cuadrático medio. Esto se conoce como la estadística de validación cruzada Leave-One-Out (LOOCV).**

Para realizar el punto anterior, y teniendo en cuenta la librería “caret”, no fue necesario realizar el loop con el for para calcular el LOOCV. Esto fue lo que se implementó en R para realizar la LOOCV para el modelo con menor RMSE o MSE.

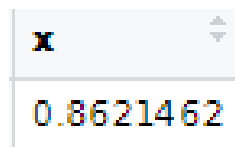
Como la aproximación de la validación k-fold tiene mayor confiabilidad que la del MSE, ya que esta última varía dependiendo de los data sets de “test” y training”, por lo cual, se realizará el LOOCV al modelo 4.

Se realizó el análisis de validación cruzada LOOCV, y el resultado del RMSE se muestra a continuación:

```
19801 samples
  2 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 19800, 19800, 19800, 19800, 19800, 19800, ...
Resampling results:

RMSE      Rsquared    MAE
0.8621462  0.2201189  0.6061062
```

A small, light gray rectangular window with a thin border. Inside the window, the text "0.8621462" is displayed in a monospaced font. Above the text, there is a small icon of a red 'X' and a small upward-pointing arrow.

Como el LOOCV es un caso especial de la validación cruzada del K-fold, se obtuvieron valores muy cercanos de RMSE entre un método y otro, por lo cual, se confirma que el modelo 4 tiene el menor RMSE.

**1.5.3.4. Compare los resultados con los obtenidos en el cálculo de la estadística de apalancamiento**

Como el apalancamiento del punto 1.5.1.3, se realizó con otro modelo (el modelo 6) se repetirá el ejercicio para el modelo 4.

Se determinó que el 23.13% de los datos tenían valores mayores a 1 o menores a -1, teniendo en cuenta que el mejor de los casos es que sean menores, para que el  $u_j$ , si está sobre la regresión, sea cero y el alfa también lo sea. El resultado del cálculo se presenta en la figura siguiente.



Adicionalmente, se determinó el mayor y el mínimo leverage para esta regresión. Los datos se presentan en la figura siguiente:

```
> max(alphas_m4$alphas_m4)
[1] 4.187301
> min(alphas_m4$alphas_m4)
[1] -5.275244
```

El mayor valor fue 5.3 (en magnitud), por lo cual, podemos concluir que los datos escogidos y la regresión realizada no están siendo afectados por datos con un gran leverage (mayor a 10, por ejemplo), por lo cual, podemos concluir que para este modelo, no se evidencian datos atípicos que afecten la regresión de este modelo 4, aún teniendo en cuenta que se realizó con toda la base de datos con la que se trabajó el presente taller.