Universidad de los Andes

Maestría en Economía Aplicada

Andres Felipe Martinez - 202121008

Angela Paola Morales Guio – 201015503

Oscar Cortes - 200222692

Repositorio: https://github.com/paolamguio/Problem_-Set_1

Problem Set 1: Predicting Income

MECA 4107

# 1 Introduction

In the public sector, accurate reporting of individual income is critical for computing taxes.However, tax fraud of all kinds has always been a significant issue. According to the InternalRevenue Service (IRS), about 83.6% of taxes are paid voluntarily and on time in the US. One of the causes of this gap is the under-reporting of incomes by individuals. An income predicting model could potentially assist in flagging cases of fraud that could lead to the reduction of the gap. Furthermore, an income prediction model can help identify vulnerable individuals and families that may need further assistance.

The objective of the problem set is to apply the concepts we learned using "real" world data. For that, we are going to scrape from the following website: https://ignaciomsarmiento.github.io/GEIH2018 sample/. This website contains data for Bogotá from the 2018 GEIH.

## 1.1 General Instructions

The main objective is to construct a predictive model of individual income

$$Income = f(X) + u \tag{1}$$

Where *Income* is the income that an individual receives, and $X$ is a matrix that includespotential predictors. In this problem set, we will focus on $f(X) = X\beta$

***1.*** *Data acquisition*

(a) Scrape the data that is available at the following website https://ignaciomsarmiento. github.io/GEIH2018 sample/.

(b) Are there any restrictions to accessing/scraping these data?

Así es, no fue posible verificar la base de datos de forma fácil y rápida ya que era necesario obtener las tablas de cada chunk, sin embargo, no fue sencillo identificarlas ya que las tablas no se encontraban en ese enlace, sino que se obtenían por medio de otro enlace atado al principal. Adicionalmente, la información de la GEIH fue dividida en 10 chunks lo que complicaba aún más obtener la información.

(c) Using pseudocode describe your process of acquiring the data

Inicio del proceso

1. Se limpia el entorno de trabajo en R
2. Se llaman las librerías necesarias para la totalidad del proceso: pacman, tidyverse, rvest, datasets y data.table.
3. Para identificar la url requerida, se inspecciona el código html de la página web https://ignaciomsarmiento.github.io/GEIH2018_sample , evidenciando que en cada enlace de los chunks de esta página web no se observan las tablas directamente como html table, por lo tanto, se encuentra en el código html el enlace que si contiene la tabla directamente para cada chunk.
4. Una vez identificada la url en la que se encontraba cada una de las tablas, se procede a crear un loop para cargar la información de cada tabla.

   La url correspondiente a la primera tabla es la siguiente: https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_1.html

   4.1. Para realizar el loop se crea el valor nombrado "url" que contiene el link anterior sin el numero de la página con el fin de que el loop se encargue de llamar página por página y traer cada una de las 10 bases de datos, así:

```
url <- "https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_"
```

   4.2. Se crea un data frame nombrado "data" que va a servir para guardar la información de cada tabla generada por el loop, de la 1 a la 10.

   4.3. Se ejecuta el loop pegando el valor creado con cada iteración "i" del 1 al 10. Posteriormente, se carga cada tabla en la lista "data" creada así:

```
data <- data.frame()
for (i in 1:10) {
  url_i <- paste0(url, i, ".html")
  tablas <- url_i %>%
    read_html() %>%
    html_table() %>% .[[1]]
  data <- rbind.data.frame(data, tablas)
}
```

5. Considerando que la primera columna de las 10 tablas no tiene nombre, se procede a renombrarla como "id" para poder unir las tablas y convertirlas a tibble.

```
colnames(data)[1] <- "id"

data <- as_tibble(data)
```

   Fin del proceso

Nota: El proceso anterior se encuentra en el archivo 1_scraping.R incluido en la carpeta 2. Scripts.

**2.** *Data Cleaning*. In this problem set, we will focus only on employed individuals olderthan eighteen (18) working in Bogotá. In this section, you are going to focus on cleaning and describing the data.

En el proceso de data cleaning se seleccionó la variable ocupados la cual define 1= ocupado; 0= no

ocupado, de acuerdo con los datos y el directorio, se deduce que esta variable corresponde a las personas que tiene trabajo independientemente de las características de este. Esto quiere decir que cuando la variable ocupado toma el valor de 1 el individuo está empleado, por otro lado cuando la variable toma el valor de 0 el individuo puede estar desempleado y/o también puede estar inactivo, por esta razón la variable "dsi" que considera los desempleados y los no desempleados, no es una variable idónea para elegir las personas empleadas ya que los no desempleados no necesariamente tienen un trabajo sino que pueden estar inactivos.

El objetivo del trabajo se enfoca en los individuos empleados (ocu=1) y mayores de 18 años. Se elige la variable "ocu" ya que relaciona todas las personas que tienen una ocupación y por lo tanto devengan un ingreso salarial, el enfoque en analizar el ingreso de los individuos que se encuentran empleados y esta variable recoge esa información.

Inicialmente, la base general de la GEIH contiene 32177 observaciones y una vez realizando el filtro mencionado anteriormente se reduce a 16397 observaciones.

- The data set include multiple variables that can help explain individual income.Guided by your intuition and economic knowledge, choose the most relevant and perform a descriptive analysis of these variables. For example, you can include variables that measure education and experience, given the implications of the human capital accumulation model (Becker, 1962, 1964; and Mincer (1962, 1975).

De acuerdo con los aportes de Mincer (1958), Schultz (1960) y Becker (1964) a la Teoría Económica sobre capital humano, las variables que explican el ingreso individual son los años de escolaridad y la experiencia profesional[1], sin embargo, existen variables omitidas en la ecuación de Mincer y que controlan el ingreso individual como lo son: la habilidad del individuo[2] (ya que aún teniendo el mismo nivel de educación, los individuos no son iguales entre sí, existiendo características distintivas en el mercado laboral), el género, la raza, las horas de trabajo, el estrato, el tamaño de la empresa, el tipo de trabajo si es formal o informal y el tipo de ocupación.

De acuerdo con lo anterior, seleccionamos las siguientes variables:

- **Ingreso total**

  Para seleccionar la variable de ingreso se analizaron las variables "y_total_m" e "ingtot". Por un lado, "y_total_m" considera el ingreso salarial y el ingreso total de los independientes, por otro lado, "ingtot" relaciona el ingreso total que agrupa el ingreso salarial, arriendos, dividendos, pensiones, ingresos en especie, entre otros.

  De acuerdo con el análisis realizado, se encontró que la variable "ingtot" realiza imputación a los registros faltantes[3] y considera los ingresos no laborales, estas dos características de "ingtot" pueden afectar la

[1] La experiencia profesional no es tenida en el análisis dado que la GEIH no cuenta con información relacionada con los años trabajados en total sino con el tiempo que ha durado el individuo en el cargo actual, la literatura ha utilizado como proxy de la experiencia la experiencia potencial. Esta surge de restarle a la edad de la persona los años que ha estudiado y, además, le resta cinco (5) años más en representación de los años de primera infancia en los que no se estudió ni se trabajó. Teniendo en cuenta que para su cálculo se basa en los años de educación, podría generarse una correlación alta entre la experiencia profesional y los años de escolaridad, generando problemas de multicolinealidad, por esta razón consideramos que las variables a incluir edad y años de escolaridad, recogen la experiencia potencial por lo que no es necesario incluirla.

[2] De acuerdo con Rosen (1992), las habilidades del individuo se traducen en su capacidad innata, la cual afecta al valor marginal de la educación y al entorno familiar que afecta a las oportunidades de financiación.

[3] De acuerdo con la descripción de la variable en el archivo nacional de datos del DANE para la GEIH.

inferencia estadística a realizar pues el objetivo a analizar es el ingreso laboral de los individuos ocupados sin realizar imputaciones a los datos [4]. Teniendo en cuenta lo anterior, se considera que la variable "y_total_m" recoge el ingreso laboral de los individuos ocupados sin imputar ingresos adicionales, variable escogida para el análisis.

- **Educación**

Considerando la revisión de literatura, Mincer (1958) menciona que existe una relación positiva entre los años de educación y el salario, Becker (1975) propone analizar las actividades que influyen en el ingreso, coincidiendo con Mincer en el argumento de incluir la educación, este autor plantea ver la escolarización como una inversión y por lo tanto se esperan retornos a la educación. De acuerdo con un estudio realizado por Tarazona y Remolina (2017), la tasa de retorno de la educación para Colombia es de 9,1%, indicador de cuánto puede crecer el salario por un año adicional de educación.

Se analizará el efecto de la variable "edu" en el ingreso laboral esperando que a mayor nivel de educación mayor será el ingreso laboral, esta variable se creó teniendo en cuenta las variables "maxEducLevel" y "p6210s1[5]", asignando el ultimo grado aprobado a cada nivel educativo en el cual se encuentra el individuo, así:

**Tabla 1.** Criterios variable Edu

| maxEducLevel | Nivel | Edu (años de educación finalizados) |
|---|---|---|
| 1 | Ninguno | 0 |
| 2 | Preescolar [6] | 2 |
| 3 | Primaria incompleta | p6210s1 + 2 |
| 4 | Primaria completa (5 años) | 7 |
| 5 | Secundaria incompleta (6-10 años) | p6210s1 + 2 |
| 6 | Secundaria completa (11 años) | 13 |
| 7 | Terciaria | p6210s1 + 13 |

Fuente: Elaboración propia

- **Oficio**

Se considera importante incluir esta variable pues se puede controlar por cada una de las ocupaciones que existen, el ingreso de los individuos, bajo el supuesto que los oficios con los ingresos más bajos pueden tener requisitos mínimos de formación, lo que lleva a responsabilidades limitadas para los individuos, lo contrario ocurre con los oficios que requieren de formaciones avanzadas o especializadas de los individuos, siendo esta formación valorada a través del aumento potencial de los ingresos.

De acuerdo con esto, se considera que dependiendo del oficio pueden existir diferencias en los ingresos. Esta variable será importante para el objetivo de inferencia estadística en la brecha salarial de género y también para realizar pronósticos de ingreso.

La base de datos cuenta con la variable "oficio" la cual está dividida en 82 ocupaciones, esta variable se incluirá en la regresión como un factor lo que implica una dummy para cada oficio.

[4] Estos impactos en la inferencia estadística se verán en detalle en un ejercicio adicional realizado para el punto 4 de análisis "The earnings GAP."

[5] Variable que refleja el grado escolar aprobado, esto, de acuerdo con la descripción de la variable en el archivo nacional de datos del DANE para la GEIH.

[6] Se tiene como supuesto que las personas con máximo nivel educativo escolar en Preescolar realizaron dos años educativos.

- **Edad**

  De acuerdo con la literatura, la edad es un factor importante que incide en la generación de ingreso, un año adicional genera incrementos en los ingresos hasta un punto determinado en el cual un año adicional de edad genera efectos marginales negativos en los ingresos.
  Por lo anterior, se incluye en el análisis las variables "age" y edad al cuadrado "age2" para modelar estos rendimientos marginales decrecientes y de esta manera poder encontrar el peak age para el ingreso.

- **Sexo**

  Según el DANE (DANE, 2020) para el año 2019 en Colombia de acuerdo con la GEIH, la brecha salarial general entre hombres y mujeres es de media de 12,9%, adicionalmente, existen numerosos trabajos y estudios que concluyen que las mujeres reciben salarios por debajo del que perciben los hombres, esto hace que sea muy importante incluir la variable sexo "female" para identificar las posibles brechas salariales de género.

  "female" =1 mujer =0 hombre

- **Horas trabajadas:**

  Se incluye esta variable ya que dependiendo de las horas que trabaje un individuo puede generar o no mayores ingresos comparados con otros individuos que trabajen menos horas y que tengan características similares, por lo tanto, horas trabajadas, variable "totalHoursWorked" puede ayudar a controlar el efecto de la brecha salarial que se busca estimar.

- **Estrato**

  Se crean dos variables dummy a partir de la variable "estrato1" que informa el estrato de energía, la variable "estrato_medio" recoge los estratos medio 3 y 4 y la variable "estrato_alto" corresponde a los estratos altos 5 y 6, estas variables[7] se incluyen en el análisis ya que puede estar evidenciando diferencias salariales por la pertenencia o no a un estrato socioeconómico, se espera que los estratos altos generen mayores ingresos comparados con los estratos bajos.

- **Tamaño empresa**

  Se considera importante incluir esta variable porque puede controlar el ingreso dependiendo del tamaño de la empresa en donde trabaja el individuo, pueden existir diferencias en los ingresos por tamaño de empresa. Esta variable será importante para el objetivo de inferencia estadística en la brecha salarial de género y también para realizar pronósticos de ingreso.

  La base de datos cuenta con la variable "sizeFirm" la cual está dividida en 5 categorías: cuenta propia, de 2-5 trabajadores, 6-10 trabajadores, de 11 a 50 trabajadores y más de 50 trabajadores, esta variable se incluirá en la regresión como un factor lo que implica una dummy para 4 categorías.

- **Formal:**

[7] No se incluye la variable estrato bajo correspondiente a los estratos 1 y 2 porque generaría multicolinealidad perfecta entre estas variables dummy creadas.

El tener un trabajo formal implicaría mayores ingresos para un individuo respecto a un individuo con trabajo informal, por lo tanto se incluye la variable "formal" para controlar por ese factor el ingreso laboral.

- **Jefe del Hogar:**

  Esta variable se crea con base en la información reportada en la variable "p6050", toma el valor de 1 cuando el individuo es jefe del hogar, se considera relevante incluirla en el análisis ya que se espera que las personas jefe de hogar trabajen más al tener como motivación mayores responsabilidades por el rol en la familia y esto puede generar impactos en el ingreso

  - Note that there are many observations with missing data. I leave it to you to find a way to handle these missing data. In your discussion, describe the steps that you performed cleaning de data, and justify your decisions.

1. Partimos de una base de datos de la GEIH con 32.177 observaciones, en la cual la variable de ingreso seleccionada "y_total_m" cuenta con 17309 NA.
2. Al realizar el filtro de personas mayores de 18 años y ocupados, se tiene un total de 16.397 observaciones y para "y_total_m" se reduce la cantidad de NA a 1.765.

   Adicionalmente, se observa en la base que existen valores que se consideran no consistentes para el "y_total_m", los cuales son $84 y $97, se decide convertir estos valores a NA, aumentando la cantidad de NA a 1.767.
3. Se eliminan los 1.767 valores NA encontrados porque se considera que para realizar el análisis de inferencia no se deben imputar datos pues pueden afectar las conclusiones.

Nota: El proceso anterior se encuentra en el archivo 2_data_cleaning.R incluido en la carpeta 2. Scripts.

- At a minimum, you should include a descriptive statistics table, but I expect tables and figures. Take this section as an opportunity to present a compelling narrative to justify and defend your data choices. Use your professional knowl-edge to add value to this section. Do not present it as a "dry" list of ingredients.

**Tabla 2**. Estadísticas descriptivas

| Estadisticas | nbr.val | nbr.null | nbr.na | min | max | range | median | mean | std.dev |
|---|---|---|---|---|---|---|---|---|---|
| **age** | 14.629,00 | - | - | 19,00 | 91,00 | 72,00 | 37,00 | 39,07 | 13,11 |
| **female** | 14.629,00 | 7.695,00 | - | - | 1,00 | 1,00 | - | 0,47 | 0,50 |
| **estrato1** | 14.629,00 | - | - | 1,00 | 6,00 | 5,00 | 2,00 | 2,52 | 0,99 |
| **estrato_medio** | 14.629,00 | 8.430,00 | - | - | 1,00 | 1,00 | - | 0,42 | 0,49 |
| **estrato_alto** | 14.629,00 | 14.039,00 | - | - | 1,00 | 1,00 | - | 0,04 | 0,20 |
| **formal** | 14.629,00 | 5.766,00 | - | - | 1,00 | 1,00 | 1,00 | 0,61 | 0,49 |
| **sizeFirm** | 14.629,00 | - | - | 1,00 | 5,00 | 4,00 | 4,00 | 3,22 | 1,65 |
| **maxEducLevel** | 14.629,00 | - | - | 1,00 | 7,00 | 6,00 | 6,00 | 5,95 | 1,21 |
| **edu** | 14.629,00 | 97,00 | - | - | 28,00 | 28,00 | 13,00 | 13,40 | 4,33 |
| **oficio** | 14.629,00 | - | - | 1,00 | 99,00 | 98,00 | 45,00 | 50,17 | 28,07 |
| **totalHoursWorked** | 14.629,00 | - | - | 1,00 | 130,00 | 129,00 | 48,00 | 47,65 | 15,12 |
| **jefe_hogar** | 14.629,00 | 7.522,00 | - | - | 1,00 | 1,00 | - | 0,49 | 0,50 |
| **y_total_m** | 14.629,00 | - | - | 5.000,00 | 70.000.000,00 | 69.995.000,00 | 996.556,50 | 1.626.617,46 | 2.440.446,29 |

Fuente: Elaboración propia

De la tabla 2 se infiere que del total de 14.629 observaciones finales de individuos mayores de 18 años y ocupados, el 47% corresponde a mujeres el 53% restante a hombres, edad promedio de 39 años y 13 años de estudio en promedio. El 61% de los individuos cuentan con un trabajo formal y el 49% son jefe

del hogar.

Entre los estratos 1 y 2 se encuentra el 54% de los individuos, en los estratos 3 y 4 el 42%, el 4% entre los estratos 5 y 6, con un promedio de ingresos de $1.626.617.

**Tabla 3.** Ingreso por genero de acuerdo con el nivel educativo

| Nivel Educativo | Hombre | Mujer |
|---|---|---|
| **Sin educación** | $ 698.217,10 | $ 434.099,90 |
| **Primaria** | $ 940.441,30 | $ 635.466,20 |
| **Media** | $ 1.033.321,30 | $ 663.663,70 |
| **Secundaria** | $ 1.172.380,10 | $ 881.530,90 |
| **Universitaria** | $ 2.822.667,60 | $ 2.359.281,90 |
| **Ingreso promedio Total** | $ 1.741.044,00 | $ 1.499.633,00 |
| **Años de educación promedio** | 13,14 | 13,68 |

Fuente: Elaboración propia

La tabla 3 demuestra que la intuición de la brecha salarial entre hombres y mujeres es real, los ingresos promedio incrementan a medida que el nivel educativo es mayor, sin embargo, el incremento de los ingresos en los hombres es mayor al de las mujeres, existiendo esa brecha salariar por nivel educativo que lleva a que en promedio el ingreso de los hombres sea de $1.741.044 y el de las mujeres $1.499.633, brecha media que representa el 14% de los ingresos promedio. Esto, iría en contravía de los años de estudio de las mujeres, los cuales en promedio son más que los de los hombres.

**Gráfico 1.** Dispersión entre años e ingreso total

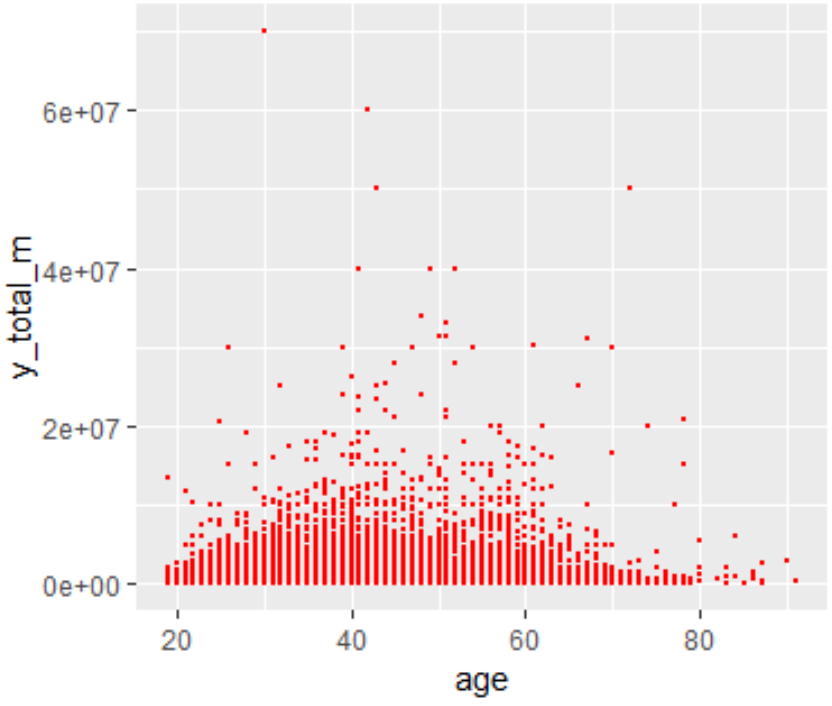


**Gráfico 2**. Dispersión entre años e ingreso total por género
1=Mujer 0=Hombre

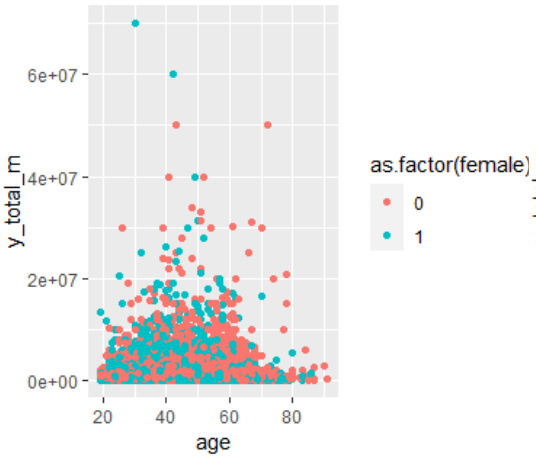
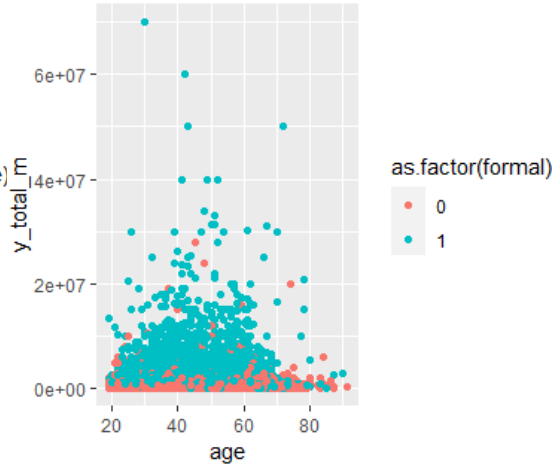
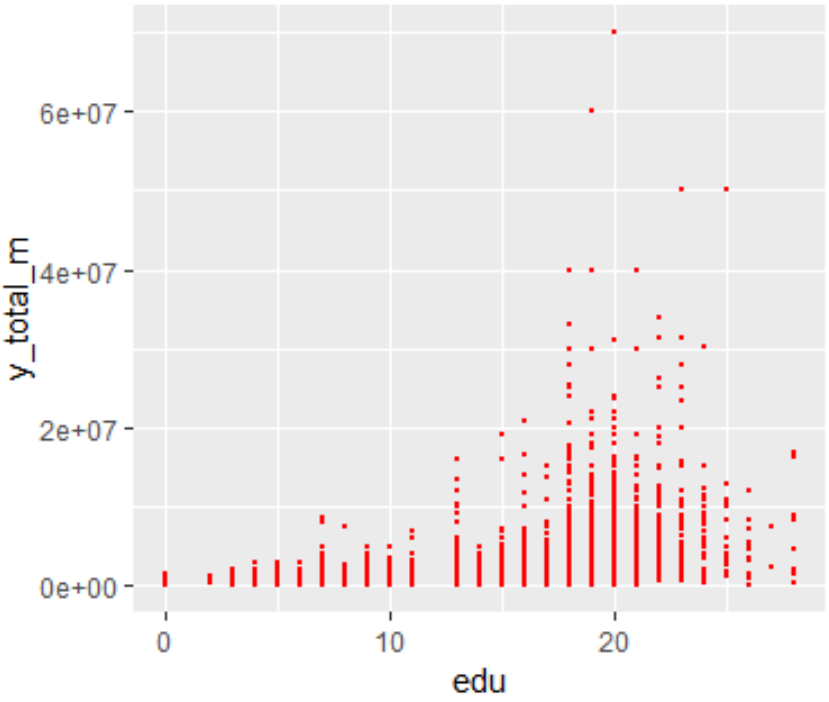



**Gráfico 3.** Dispersión entre años e ingreso total por tipo de trabajo
1=Formal 0=Informal




Estas gráficas muestran la relación entre edad y el ingreso laboral, se puede evidenciar que existen rendimientos marginales decrecientes pues en un punto se incrementa la edad y el ingreso se disminuye, lo que implica una relación cuadrática entre la edad y el ingreso. Por un lado, la Gráfica 2 presenta esta relación por género, encontrando que existe una brecha salarial pues los hombres tienen en promedio ingresos más altos. Por otro lado, la Gráfica 3 muestra la relación por sector laboral, se puede encontrar que las personas que trabajan en el sector formal tienen en promedio mayores ingresos que las personas que trabajan en el sector informal.

**Gráfico 4.** Dispersión entre años de educación e ingreso total




Esta gráfica muestra la relación entre los años de educación y el ingreso laboral, se puede evidenciar que esta relación es directa pues las personas con mayores años de educación presentan mayores ingresos.

**Gráfico 5.** Distribución años de educación

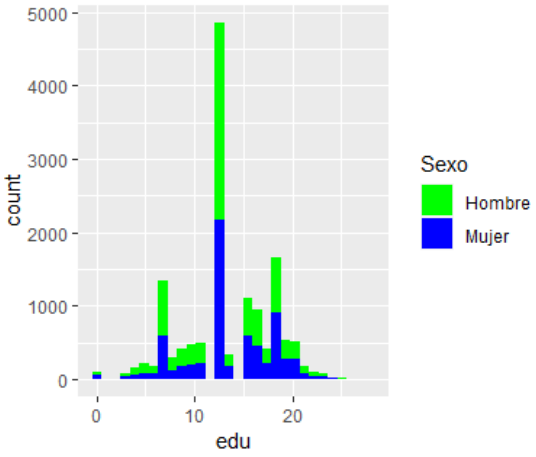
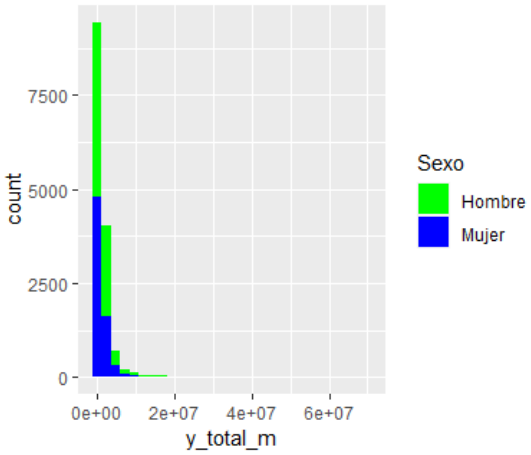


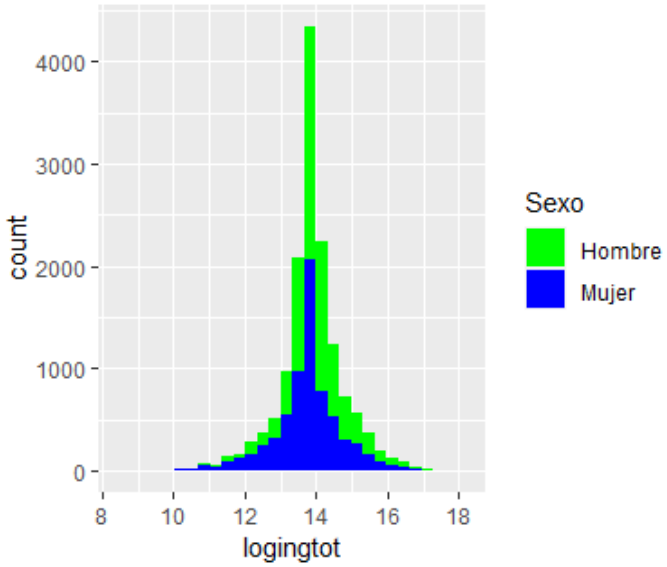

**Gráfico 6**. Distribución del ingreso laboral




Estas gráficas muestran la distribución por género de los años de educación y el ingreso laboral, para los años de educación la mayoría de personas se encuentran en la media de la distribución, presentan 13,4 años de educación. En cuanto al ingreso laboral, esta distribución es más asimétrica, lo cual indica que no es una distribución normal, para ajustar esta serie se realizará una transformación utilizando el logaritmo del ingreso, esta variable será utilizada para la estimación de la brecha salarial por género.

En la gráfica 7 se evidencia el ajuste en el ingreso al utilizar el logaritmo de esta variable respecto al género, la cual muestra que esta serie es más simétrica respecto al ingreso y es más similar a una distribución normal.

**Gráfico 7.** Distribución logaritmo del ingreso laboral




En la gráfica 7 se evidencia el ajuste en el ingreso al utilizar el logaritmo de esta variable respecto al género, la cual muestra que esta serie es más simétrica respecto al ingreso y es más similar a una distribución normal.

**Gráfico 8.** Box-Plot por estrato

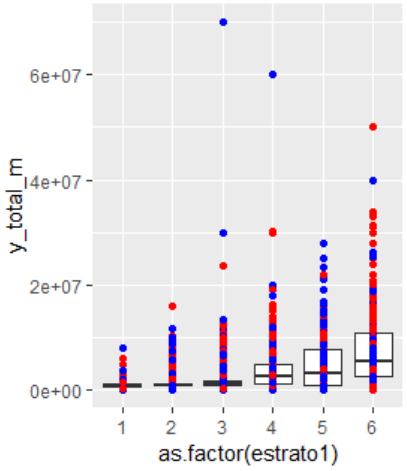
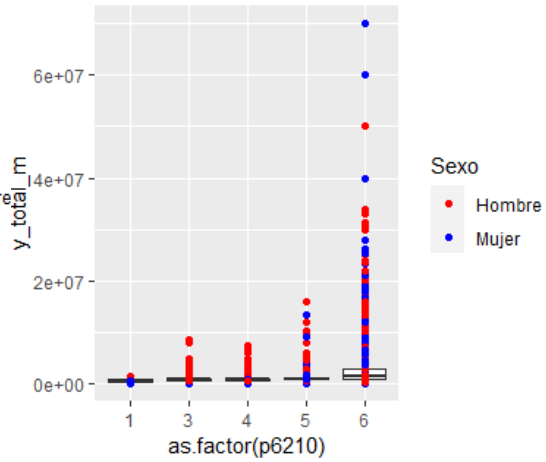
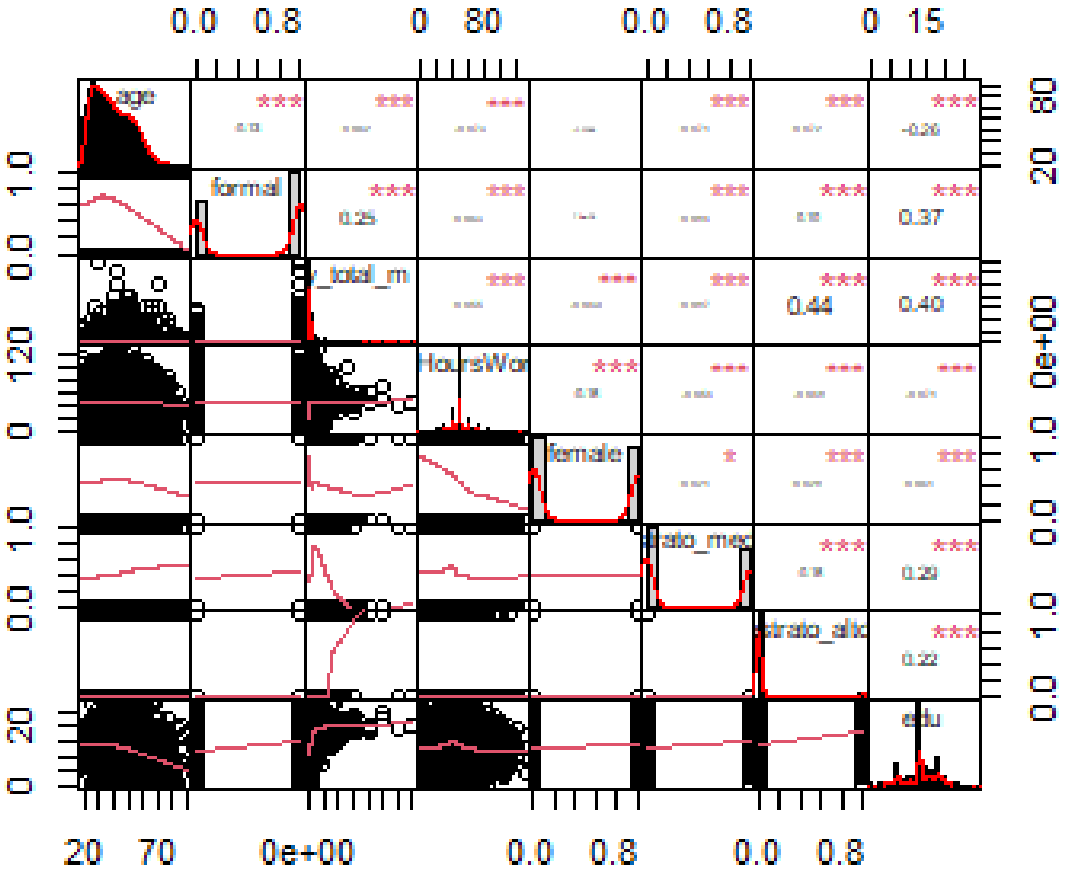



**Gráfico 9.** Box-Plot por nivel educativo



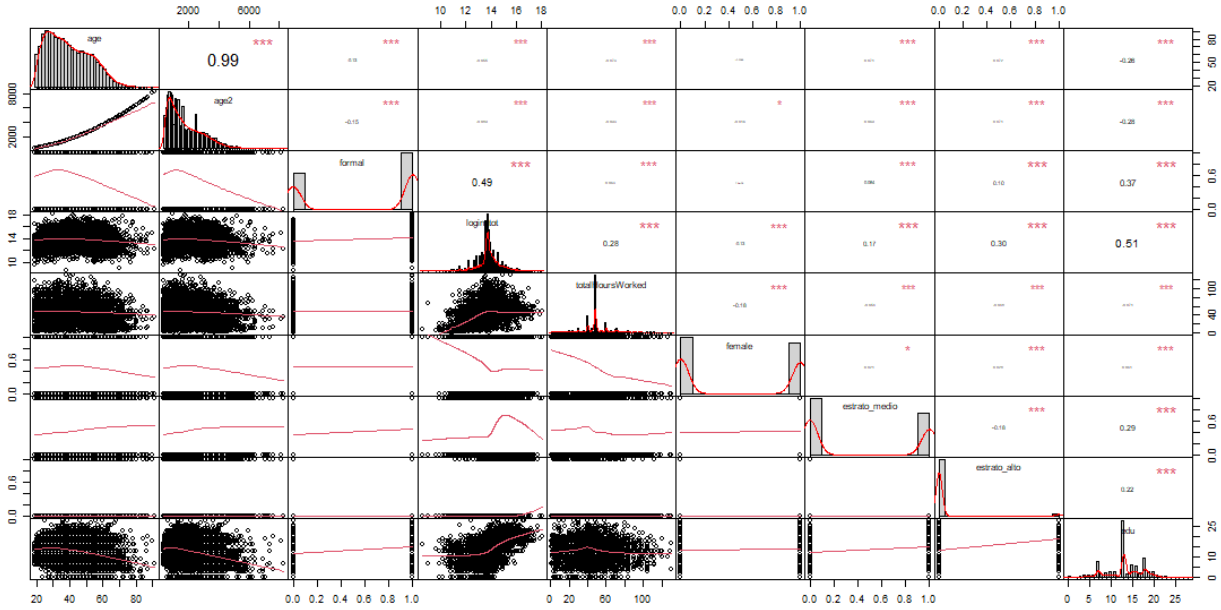

Estas gráficas de caja muestran el ingreso por estrato socioeconómico y nivel educativo, se puede evidenciar que las personas que se encuentran en los estratos bajos presentan menores ingresos, en los estratos medios tienen ingresos mayores comparados con los estratos bajos y en los estratos altos se presentan los mayores ingresos. En cuanto al nivel educativo, las personas que no tienen educación presentan los menores ingresos, las personas que tienen nivel educativo primaria y secundaria tienen mayores ingresos comparados con las personas que no tienen educación, las personas que tienen nivel educativo universitaria presentan los mayores ingresos.

**Gráfico 10.** Correlación entre variables



La gráfica de correlaciones entre las variables muestra la relación entre la variable ingreso laboral y las demás variables, se evidencia que todas las correlaciones son estadísticamente significativas y las variables que presentan mayor correlación con el ingreso son estrato alto, años de educación y la formalidad del trabajo con un coeficiente de correlación de 0,44, 0,40 y 0,25 respectivamente, esto quiere decir que individuos con mayores años de educación, que hagan parte de los estratos 5 y 6 y su trabajo sea formal, presentan mayores ingresos respecto a los demás individuos. Respecto a la variable de género "female" se presenta un coeficiente de correlación con los ingresos de -0,049, lo que soporta la brecha salarian entre hombres y mujeres.

**Gráfico 11.** Correlación entre variables (ln ingreso)




Nota: Los gráficos 10 y 11 se encuentran como archivo PDF en la carpeta 4. Views

La gráfica anterior, refleja las correlaciones entre las variables control y nuestra variable objetivo en unidades de logaritmo, se evidencia que todas las correlaciones son estadísticamente significativas y las variables que presentan mayor correlación con el ingreso son los años de educación, la formalidad del trabajo y el hacer parte del estrato alto con un coeficiente de correlación de 0,51, 0,49 y 0,30 respectivamente, esto quiere decir que individuos con mayores años de educación, que hagan parte de los estratos 5 y 6 y su trabajo sea formal, presentan mayores ingresos respecto a los demás individuos. Respecto a la variable de género "female" se presenta un coeficiente de correlación con los ingresos de -0,13, lo que soporta la brecha salarian entre hombres y mujeres.

Adicionalmente se incluye la variable edad al cuadrado "age2" porque esta será necesaria para calcular el peak age, presenta una correlación con el logaritmo del ingreso de -0,089.

3. *Age-earnings profile*. A great deal of evidence in Labor economics suggests that the typical worker's age-earnings profile has a predictable path: Wages tend to be low when the worker is young; they rise as the worker ages, peaking at about age 50; andthe wage rate tends to remain stable or decline slightly after age 50.

   - In the data set, multiple variables describe income. Choose one that you believeis the most representative of the workers' total earnings, justifying your selection.

   - In the data set, multiple variables describe income. Choose one that you believeis the

most representative of the workers' total earnings, justifying your selection.

En la base de datos de la GEIH para Bogotá en el año 2018 se encontraron las siguientes variables relacionados con el ingreso de los trabajadores:

***Tabla 1.*** Variables "Age-earnings profile"

| Variable | Descripción |
|---|---|
| Ingtotob | Ingreso monetario primera actividad, ingreso segunda actividad, ingreso en especie, ingreso monetario des ocupados e inactivos e ingresos provenientes de otras fuentes no laborales |
| Ingtotes | Ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos imputadas a los registros faltantes |
| Ingtot | Ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos tanto observadas como imputadas |
| y_ingLab_m | labor income salaried - nominal monthly - all occ. (includes tips and commission |
| y_salary_m | salary - nominal monthly - principal occ. (includes tips and commissions) |
| y_total_m | income salaried + independents total - nominal monthly |

Fuente: Elaboración propia

En primer lugar, descartamos las tres (3) primeras variables (Ingtot, Ingtotes y Ingtotob) puesto que estas incluyen ingresos de personas desocupados e inactivos.

Después nos concentramos en revisar lo que pide el numeral, lo cual corresponde a las ganancias totales de los trabajadores, y puesto que esto incluye tanto los ingresos laborales como los no laborales, seleccionamos la variable y_total_m.

- Based on this estimate using OLS the age-earnings profile equation:

$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u \qquad (2)$$

- How good is this model in sample fit?

A continuación, se presenta los resultados del modelo de regresión realizado en R con los datos filtrados y seleccionados de la GEIH:

***Tabla 2.*** Estadísticas regresión

| | Dependent variable: |
|---|---|
| | y_total_m |
| age | 124.164.,00*** |
| | (9.118,354) |
| age2 | -1.332,245*** |
| | (106,343) |
| Constant | -962.128,300*** |

(181,878.000)

| | |
|---|---|
| Observations | 14.629 |
| R2 | 0,014 |
| Adjusted R2 | 0,014 |
| Residual Std. Error | 2.422.949,000 (df = 14626) |
| F Statistic | 107,016*** (df = 2; 14626) |
| Note: | *p<0,1; **p<0,05; ***p<0,01 |

Si bien el modelo arroja que las variables de edad y edad al cuadrado son relevantes para explicar el ingreso con un nivel de significancia del 1% y el modelo demuestra que hay dependencia (F Estadístico) también con un nivel de significancia del 1%, el grado de ajuste del modelo (R2) es muy bajo esto derivado a que hay otras variables que explican el ingreso de las personas tal como lo plantea Jacob Mincer (1974) que también depende de los años de escolaridad.

El modelo arroja que el ingreso estimado de una persona se puede obtener multiplicando su edad por 124.116 menos el valor de multiplicar su edad al cuadrado por 1.331,928 y finalmente restando un valor constante de 962.128,300.

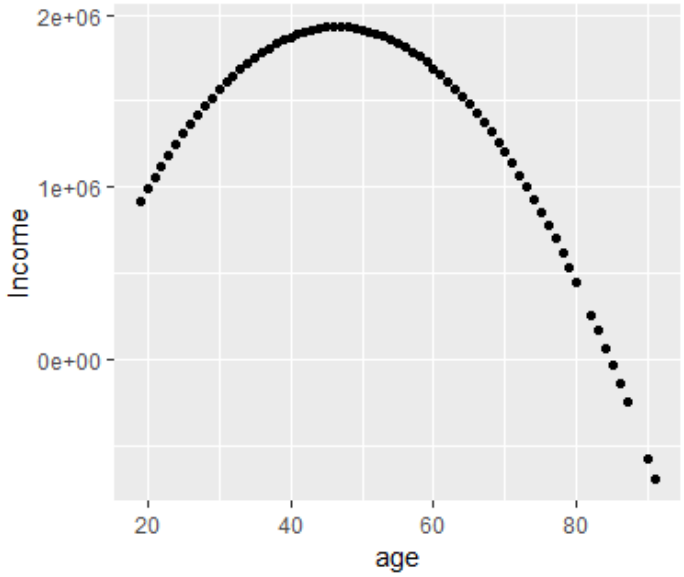
Esta estimación muestra que la edad y la edad al cuadrado con significativas, lo que soporta la teoría de que existen efectos marginales decrecientes de la edad sobre el ingreso

Adicionalmente varios autores han planteado la relación entre salario y género, puesto que hay una diferencia entre los salarios que reciben los hombres y las mujeres bajo las mismas condiciones de edad y educación (Nordin & Persson, 2010). Igualmente, si se mira si el trabajo es formal o no, como lo muestra la siguiente gráfica, se presenta una diferencia significativa para los que tienen un empleo formal frente a los que no lo tienen.

**Gráfica 1.** Edad contra Ganancias con filtro por trabajo formal=1


- Plot the predicted age-earnings profile implied by the above equation.

**Gráfica 13.** Modelo de predicción del ingreso a partir de la edad.




En la anterior gráfica se observa como el modelo tiene una forma cóncava, es decir, los ingresos aumentan cada vez menos con la edad hasta cierta edad determinada, a partir de la cual comienzan a decrecer los ingresos cada vez en mayor proporción con el aumento de la edad.

- What is the "peak age" suggested by the above equation? Use bootstrap to calculate the standard errors and construct the confidence intervals.

Con el modelo se obtienen los siguientes valores de los coeficientes:

- b1 = -962128,3
- b2 = 124164,9
- b3 = -1332,245

$$peak_{age} = -\frac{\beta 1}{2 * \beta 2}$$

Se obtiene que la edad pico es 46,59 años

```
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = dbIncome, statistic = eta_mod.fn, R = 1000)

Bootstrap Statistics :
   original       bias           std. error
t1* 46,59986 0,08728924   0,8982211
```

Por lo tanto, si el error estándar (σ) es 0,8982211 años el intervalo de confianza para la edad pico es:

$$\text{Intervalo peak}_{age} = 46{,}59 \pm 1{,}96 * \sigma = (44{,}83934\ ,\ 48{,}36037)$$

**4.** *The earnings GAP*. Most empirical economic studies are interested in a single low dimensional parameter, but determining that parameter may require estimating additional "nuisance"

parameters to estimate this coefficient consistently and avoid omitted variables bias. Policymakers have long been concerned with the gender earningsgap.

- Estimate the unconditional earnings gap

$$log(Income) = \beta_1 + \beta_2 Female + u \quad (3)$$

- How should we interpret the $\beta_2$ coefficient? How good is this model in samplefit?

**Tabla 6.** Estadísticas regresión

| Dependent variable: | |
|---|---|
| | logingtot |
| female | -0,239*** |
| | (0.015) |
| Constant | 13,990*** |
| | (0.010) |
| Observations | 14.629 |
| R2 | 0,018 |
| Adjusted R2 | 0,018 |
| Residual Std. Error | 0.880 (df = 14627) |
| F Statistic | 269,451*** (df = 1; 14627) |
| Note: | *p<0,1; **p<0,05; ***p<0,01 |

De acuerdo con la tabla anterior, las mujeres tienen en promedio un ingreso que es 23% menor, comparado con el de los hombres, ceteris paribus.

El modelo arroja que la variable que define el género "female" es relevante para explicar el ingreso con un nivel de significancia del 1%, se demuestra que hay dependencia (F Estadístico) también con un nivel de significancia del 1%, el grado de ajuste del modelo (R2) es bajo, lo que implica que la variable "female" explica en un 1,8% la variabilidad del ingreso total y el 98,2% restante es explicado por el término de error, lo que lleva a inferir que existen variables omitidas que también pueden explicar el ingreso de las personas.

- Estimate and plot the predicted age-earnings profile by gender. Do men and women in Bogotá have the same intercept and slopes?

Para estimar el perfil de ingresos edad por género, se aplica la siguiente regresión:

$$log(Income) = \beta_1 + \beta_2 Female + \beta_3 age + \beta_4 age2 + \beta_5 age*female + u$$

En esta regresión incluye la edad y la edad al cuadrado para considerar los efectos marginales decrecientes de la edad, adicionalmente se incluye la interacción entre la edad y el género con el fin de obtener el efecto de la edad por género, lo cual permitirá analizar si existen diferencias por género del impacto de la edad en el ingreso.

**Tabla 7.** Estadísticas regresión

| | Dependent variable: | |
|---|---|---|
| | logingtot | |
| | (1) | (2) |
| female | -0,239*** | 0,298*** |
| | (0,015) | (0,045) |
| age | | 0,091*** |
| | | (0,003) |

| | | |
|---|---|---|
| age2 | | -0,001*** |
| | | (0,00004) |
| female:age | | -0,014*** |
| | | (0,001) |
| Constant | 13,990*** | 12,214*** |
| | (0,010) | (0,068) |
| Observations | 14.629 | 14.629 |
| R2 | 0,018 | 0,079 |
| Adjusted R2 | 0,018 | 0,079 |
| Residual Std. Error | 0,880 (df = 14627) | 0.852 (df = 14624) |
| F Statistic | 269,451*** (df = 1; 14627) | 313.193*** (df = 4; 14624) |
| Note: | | *p<0.1; **p<0.05; ***p<0.01 |

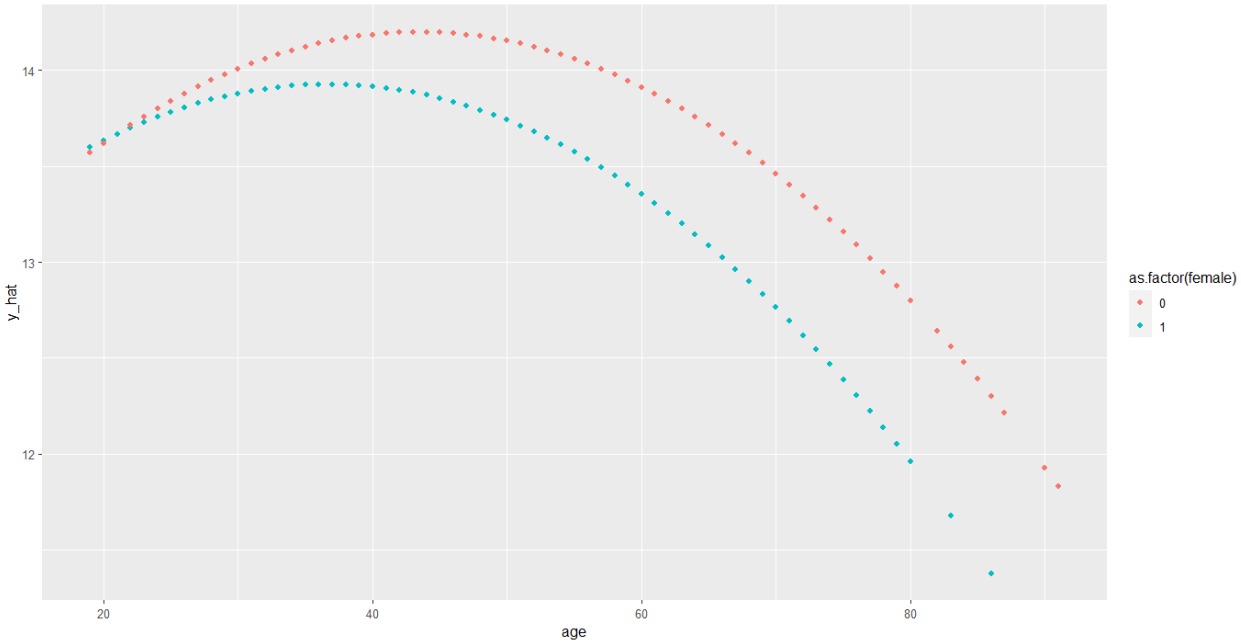
La tabla anterior presenta una regresión con 14629 observaciones, todas las variables son estadísticamente significativas a un nivel de significancia del 1%, existe dependencia global al 1%, sin embargo, el grado de ajuste del modelo sigue siendo muy bajo pues estas variables sólo explican en 7,9% a la variabilidad del logaritmo del ingreso debido a las razones explicada anteriormente para el modelo 1.

Se evidencia que el incremento de un año de edad de las mujeres, tiene un efecto negativo en el ingreso de 1,4% menos que el efecto en el ingreso de los hombres ante de un incremento de un año de edad. Para calcular el efecto marginal del ingreso por ser mujer se tiene en cuenta la siguiente ecuación:

$$Efecto\ marginal = \beta_2 + \beta_5 age$$

Teniendo en cuenta una edad promedio de 39 años, las mujeres tienen en promedio un ingreso menor en 25,7% comparado con el de los hombres. La regresión anterior permite predecir los valores del ingreso considerando que pueden existir diferencias en el pico de la edad por género para un máximo de ingresos laborales.

**Gráfico 13**. Perfil predicción ingresos-edad por género




En términos del intercepto, hombres y mujeres a la edad de 18 años empiezan a percibir ingresos con valores muy similares, aunque la mujer recibe un poco más, esta diferencia parece no ser significativa en el intercepto, sin embargo, a medida que aumenta la edad se empieza a crear la brecha salarial entre

hombres y mujeres lo que implica que las mujeres y los hombres en Bogotá presentan una pendiente diferente, de igual forma, esta gráfica evidencia que la brecha de ingresos por género se va incrementando a medida que aumentan los años de edad. Adicionalmente, el punto en el cual los ingresos comienzan a decrecer se da más rápido para las mujeres que para los hombres.

- What is the implied "peak age" by gender?. Use bootstrap to calculate the standard errors and construct the confidence intervals. Do these confidence intervals overlap?

Para hacer el cálculo del peak age por género se tendrán en cuenta las siguientes ecuaciones:

$$peak_{ageMale} = -\frac{\beta 2}{2 * \beta 4}$$

$$peak_{ageFemale} = -\frac{\beta 3 + \beta 5}{2 * \beta 4}$$

Estas ecuaciones se obtienen de igualar a cero la derivada del logaritmo del ingreso respecto a la edad por género para encontrar el punto máximo.

De acuerdo a la regresión estimada en el punto anterior, se obtienen los siguientes peak -age por género:

$$peak_{ageMale} = 43{,}5$$

$$peak_{ageFemale} = 36{,}7$$

Se procede a realizar el boostrap para obtener los errores estándar que servirán para el cálculo de los intervalos de confianza por género.

**Tabla 8.** Boostrap para hombre (df)

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = df, statistic = eta_mod.fn, R = 5000, female_bar = 0)

Bootstrap Statistics :

| | original | bias | std. error |
|---|---|---|---|
| t1* | 43,49387 | 0,01793892 | 0,3823748 |

Se realiza el cálculo de Bootstrap para 5.000 repeticiones obteniendo un sesgo de 0,01 y un error estándar de 0,38, esto quiere decir que al realizar la estimación del peak age para los hombres con 5.000 submuestras se presenta un sesgo relativamente bajo entre el valor estimado con la muestra total y el valor estimado promedio de las submuestras, por otro lado, el error estándar muestra que no hay mayor dispersión en los datos, generándose un intervalo de confianza entre 42,7 y 44,2 años, considerando la siguiente ecuación:

$$CI_{Male} = (peak_{ageMale} - 1{,}96 * std.error_{Male}\ , peak_{ageMale} + 1{,}96 * std.error_{Male}$$

**Tabla 9.** Boostrap para mujer (df).

ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = df, statistic = eta_mod.fn, R = 5000, female_bar = 1)

Bootstrap Statistics :

| | original | bias | std. error |
|---|---|---|---|
| t1* | 36,71082 | -0,01620718 | 0,4694526 |

Se realiza el cálculo de Bootstrap para 5.000 repeticiones obteniendo un sesgo de -0,01 y un error

estándar de 0,46, esto quiere decir que al realizar la estimación del peak age para las mujeres con 5.000 submuestras se presenta un sesgo relativamente bajo entre el valor estimado con la muestra total y el valor estimado promedio de las submuestras, por otro lado, el error estándar muestra que no hay mayor dispersión en los datos, generándose un intervalo de confianza entre 35,8 y 37,6 años, considerando la siguiente ecuación:

$$CI_{Female} = (peak_{ageFemale} - 1{,}96 * std.error_{Female}\ , peak_{ageFemale} + 1{,}96 * std.error_{Female}$$

De acuerdo con los intervalos de confianza del peak age para hombres y mujeres, se infiere que existen diferencias estadísticamente significativas entre estos debido a que los intervalos de confianza no de solapan.

De los puntos anteriores se concluye que persiste una brecha salarian entre hombres y mujeres, la cual se acentúa con el incremento en la edad; aunado a esto, existe evidencia estadística para afirmar que las mujeres encuentran el pico de edad del máximo ingreso más rápido que el de los hombres.

Ejercicio adicional:

El análisis de los puntos anteriores se realiza para dos bases de datos adicionales i) df2 incluye como variable ingreso "ingtot" la cual considera los ingresos no laborales y se eliminan los valores iguales a cero ii) df3 incluye la variable "y_total_m" pero a diferencia de la base de datos "df" utilizada en los puntos anteriores, a esta se le imputan datos a los NA iniciales considerando el ingreso promedio por vivienda y nivel educativo para las viviendas que no reportaron información. Se obtienen los siguientes peak age:

**Tabla 10.** Escenarios

| Base de datos | Descripción | Género | Peak age | Intervalo de confianza |
|---|---|---|---|---|
| **df** | Variable ingreso y_total_m *Se eliminan NA | Mujer | 36,7 | 35,8 - 37,6 |
| | | Hombre | 43,5 | 42,7 - 44,4 |
| **df2** | Variable ingreso ingtot *Se eliminan valores=0 | Mujer | 39,3 | 38 - 40,6 |
| | | Hombre | 48,2 | 46,8 - 49,7 |
| **df3** | Variable ingreso y_total_m *Se imputan valores de acuerdo a la vivienda y el nivel educativo | Mujer | 37,14 | 36,2 - 38,1 |
| | | Hombre | 44,12 | 43,3 - 44,9 |

Fuente: Elaboración propia

Cuando se considera la variable "ingtot" que incluye el ingreso no laboral, el peak age tanto para hombres como para mujeres se desplaza unos años más a la derecha, es decir que el tener en cuenta los ingresos no laborales como pensión, arriendos, dividendos, entre otros, se afecta la inferencia estadística pues los intervalos de confianza no se solapan con los intervalos de confianza calculados con la base de datos df.

En cuanto a la base de datos df3, la cual realiza imputación de datos ala variable de ingreso "y_total_m", no se encuentran diferencias estadísticamente significativas entre estos intervalos de confianza y los calculados por la base de datos df.

Teniendo en cuenta lo anterior, se concluye que para el análisis realizado en el punto 4 se descarta la variable "ingtot" ya que el objetivo del análisis se enfoca en el ingreso laboral y esta variable genera ruido en la inferencia. Por otro lado, aunque para el caso de la df3 no hay diferencias en la inferencia estadística, no se considera necesario realizar imputación a los datos y por lo tanto se trabaja con la base de datos df que elimina los NA de la variable "y_total_m".

- *Equal Pay for Equal Work?* A common slogan is "equal pay for equal work". One way to interpret this is that for employees with similar worker and job characteristics, no gender earnings gap should exist. Estimate a conditional earnings gap that incorporates control variables such as similar worker and jobcharacteristics ($X$).

  (a) Estimate the conditional earnings gap $log(Income) = \beta_1 + \beta_2 Female + \vartheta X + u$

Para estimar (a), se aplica la siguiente regresión:

$log(Income) = \beta_1 + \beta_2 Female + \beta_3 age + \beta_4 age2 + \beta_5 edu + \beta_6 formal + \beta_7 factor(oficio) + \beta_8 factor(sizeFirm) + \beta_9 totalHoursWorked + \beta_{10} estrato\_medio + \beta_{11} estrato\_alto + u$

**Tabla 11.** Resultados regresión

| | Dependent variable: | | |
|---|---|---|---|
| | logingtot | | |
| | (1) | (2) | (3) |
| female | -0.239*** | 0.298*** | -0.173*** |
| | (0.015) | (0.045) | (0.011) |
| age | | 0.091*** | 0.048*** |
| | | (0.003) | (0.002) |
| age2 | | -0.001*** | -0.001*** |
| | | (0.00004) | (0.00003) |
| female:age | | -0.014*** | |
| | | (0.001) | |
| edu | | | 0.040*** |
| | | | (0.002) |
| … | | | |
| formal | | | 0.279*** |
| | | | (0.014) |
| Constant | 13.990*** | 12.214*** | 11.777*** |
| | (0.010) | (0.068) | (0.145) |
| Observations | 14,629 | 14,629 | 14,629 |
| R2 | 0.018 | 0.079 | 0.595 |
| Adjusted R2 | 0.018 | 0.079 | 0.592 |
| Residual Std. Error | 0.880 (df = 14627) | 0.852 (df = 14624) | 0.567 (df = 14538) |
| F Statistic | 269.451*** (df = 1; 14627) | 313.193*** (df = 4; 14624) | 237.096*** (df = 90; 14538) |

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

(b) Use FWL to repeat the above estimation, where the interest lies on $\beta_2$. Do you obtain the same estimates?

Con el fin de ejecutar The Frisch-Waugh-Lovell (FWL) Theorem, se corren las siguientes dos regresiones:

1. Regresión principal relacionada en (a), sin la variable objetivo "female"

   $log(Income) = \beta_1 age + \beta_2 age2 + \beta_3 edu + \beta_4 formal + \beta_5 factor(oficio) + \beta_6 factor(sizeFirm) + \beta_7 totalHoursWorked + \beta_8 estrato\_medio + \beta_9 estrato\_alto + u$

2. Regresión secundaria (3) en la que la variable objetivo "female" es explicada por las demás variables control

   $female = \beta_1 age + \beta_2 age2 + \beta_3 edu + \beta_4 formal + \beta_5 factor(oficio) + \beta_6 factor(sizeFirm) + \beta_7 totalHoursWorked + \beta_8 estrato\_medio + \beta_9 estrato\_alto + u$

**Tabla 12.** Resultados regresión FWL

| | Dependent variable: | | | |
|---|---|---|---|---|
| | logingtot | | | res_reg4 |
| | (1) | (2) | (3) | (4) |
| female | -0.239*** | 0.298*** | **-0.173*** ** | |
| | (0.015) | (0.045) | (0.011) | |
| age | | 0.091*** | 0.048*** | |
| | | (0.003) | (0.002) | |
| age2 | | -0.001*** | -0.001*** | |
| | | (0.00004) | (0.00003) | |
| *** | | | | |
| totalHoursWorked | | | 0.015*** | |
| | | | (0.0003) | |
| estrato_medio | | | 0.138*** | |
| | | | (0.011) | |
| estrato_alto | | | 0.854*** | |
| | | | (0.027) | |
| res_reg5 | | | | **-0.173*** ** |
| | | | | (0.011) |
| Constant | 13.990*** | 12.214*** | 11.777*** | -0.000 |
| | (0.010) | (0.068) | (0.145) | (0.005) |
| Observations | 14,629 | 14,629 | 14,629 | 14,629 |
| R2 | 0.018 | 0.079 | 0.595 | 0.015 |
| Adjusted R2 | 0.018 | 0.079 | 0.592 | 0.015 |
| Residual Std. Error | 0.880 (df = 14627) | 0.852 (df = 14624) | 0.567 (df = 14538) | 0.565 (df = 14627) |
| F Statistic | 269.451*** (df = 1; 14627) | 313.193*** (df = 4; 14624) | 237.096*** (df = 90; 14538) | 229.206*** (df = 1; 14627) |
| Note: | | | | *p<0.1; **p<0.05; ***p<0.01 |

De acuerdo con la tabla anterior, se cumple el teorema FWL debido a que el coeficiente de los residuales del modelo female contra controles, es igual al modelo original.

(c) How should we interpret the $\beta_2$ coefficient? How good is this model in sample fit? Is the gap reduced? Is this evidence that the gap is a selection problem and not a "discrimination problem"?

De acuerdo con la tabla presentada en el literal anterior, se puede inferir que el coeficiente $\beta_2$ significa que las mujeres tienen en promedio un ingreso que es 17,3% menor, comparado con el de los hombres, ceteris paribus. Sin embargo, respecto a la regresión original (punto 4 del problema-set), se evidencia disminución del coeficiente de género (para mujeres) sobre el ingreso laboral, por esto, se podría inferir que la regresión original que no tiene en cuenta las variables control estaba sesgada al generar un coeficiente de – 0,23 siendo que con las variables control dicho coeficiente disminuye a -0,17; así pues que la brecha entre el ingreso laboral de hombres y mujeres se reduce pero sigue siendo estadísticamente significativa. El grado de ajuste del modelo es mucho mejor ya que se evidencia que las variables de este modelo explican en un 60% la variabilidad del logaritmo del ingreso.

Por otro lado, independientemente de las variables control de características que pueden afectar los ingresos laborales, siguen existiendo brechas salariales entre hombres y mujeres, es decir, a pesar de que una mujer tenga el mismo nivel educativo, el mismo oficio, el mismo tipo de empresa y la misma edad, persiste la brecha salarial, no es un problema de selección sino de discriminación.

**5.** *Predicting earnings*. Now we turn to prediction. You built a couple of models in the previous section using your knowledge as an applied economist, the task here is to assess the predictive power of these models.

(a) Split the sample into two samples: a training (70%) and a test (30%) sample. Don't forget to set a seed (in R, set.seed(10101), where 10101 is the seed.)

i. Estimate a model that only includes a constant. This will be the benchmark.

***Tabla 33.*** Resultados del modelo de predicción con la constante únicamente

```
================================================
                         Dependent variable:
                     ---------------------------
                               logingtot
------------------------------------------------
Constant                      13.876***
                               (0.009)

------------------------------------------------
Observations                   10,242
R2                              0.000
Adjusted R2                     0.000
Residual Std. Error      0.897 (df = 10241)
================================================
  Note:               *p<0.1; **p<0.05; ***p<0.01
```

En la anterior table se observa que el modelo no tiene ningún grado de ajusta puesto que su R2 es cero.

ii. Estimate again your previous models

***Tabla 14.*** Resultados del modelo de predicción con la constante únicamente

```
=============================================================================================
                                        Dependent variable:
                    -----------------------------------------------------------------------
```

| | logingtot | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| age | | 0.079*** | 0.081*** |
| | | (0.004) | (0.004) |
| age2 | | -0.001*** | -0.001*** |
| | | (0.00005) | (0.00005) |
| female | | | -0.249*** |
| | | | (0.017) |
| Constant | 13.876*** | 12.451*** | 12.522*** |
| | (0.009) | (0.079) | (0.078) |
| Observations | 10,242 | 10,242 | 10,242 |
| R2 | 0.000 | 0.044 | 0.063 |
| Adjusted R2 | 0.000 | 0.044 | 0.063 |
| Residual Std. Error | 0.897 (df = 10241) | 0.877 (df = 10239) | 0.868 (df = 10238) |
| F Statistic | | 236.112*** (df = 2; 10239) | 230.435*** (df = 3; 10238) |
| Note: | | | *p<0.1; **p<0.05; ***p<0.01 |

En la anterior table se observa que todas las variables independientes explican la variable independiente (ingreso) con un nivel de significancia de 1%, adicionalmente el grado de ajuste va incrementándose a la medida que se aumenta la complejidad de los modelos (más variables).

iii. In the previous sections, the estimated models had different transformations of the dependent variable. At this point, explore other transformations of your independent variables also. For example, you can include polynomial terms of certain controls or interactions of these. Try at least five (5) models that are increasing in complexity.

**Tabla 15.** Resultados de los modelos de predicción

| | Dependent variable: | | | | | |
|---|---|---|---|---|---|---|
| | logingtot | | | | | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| age | | 0.079*** | 0.081*** | 0.070*** | 0.062*** | 0.068*** |
| | | (0.004) | (0.004) | (0.003) | (0.003) | (0.003) |
| age2 | | -0.001*** | -0.001*** | -0.001*** | -0.001*** | -0.001*** |
| | | (0.00005) | (0.00005) | (0.00004) | (0.00004) | (0.00004) |
| female | | | -0.249*** | -0.304*** | -0.288*** | 0.056 |
| | | | (0.017) | (0.015) | (0.014) | (0.043) |
| edu | | | | 0.109*** | 0.084*** | 0.084*** |
| | | | | (0.002) | (0.002) | (0.002) |
| formal | | | | | 0.610*** | 0.608*** |
| | | | | | (0.015) | (0.015) |
| age:female | | | | | | -0.009*** |
| | | | | | | (0.001) |
| Constant | 13.876*** | 12.451*** | 12.522*** | 11.121*** | 11.206*** | 11.030*** |
| | (0.009) | (0.079) | (0.078) | (0.070) | (0.065) | (0.068) |
| Observations | 10,242 | 10,242 | 10,242 | 10,242 | 10,242 | 10,242 |
| R2 | 0.000 | 0.044 | 0.063 | 0.314 | 0.409 | 0.413 |
| Adjusted R2 | 0.000 | 0.044 | 0.063 | 0.314 | 0.409 | 0.413 |
| Residual Std. Error | 0.897 (df = 10241) | 0.877 (df = 10239) | 0.868 (df = 10238) | 0.743 (df = 10237) | 0.689 (df = 10236) | 0.687 (df = 10235) |
| F Statistic | | 236.112*** (df = 2; 10239) | 230.435*** (df = 3; 10238) | 1,170.946*** (df = 4; 10237) | 1,417.569*** (df = 5; 10236) | 1,201.102*** (df = 6; 10235) |
| Note: | | | | | | *p<0.1; **p<0.05; ***p<0.01 |

En la tabla anterior, se observa que las variables independientes tienen un nivel de significancia del 1% en los modelos donde se emplean y a medida que se aumenta la complejidad del modelo se mejora el grado de ajuste, es decir, aumenta el R2. Se adiciona la variable oficio en polinomio de grado 8 para al modelo de predicciones, incrementando su complejidad y a la vez el grado de ajuste con un mayor R2.

**Tabla 16.** Resultados de proyección adicionando la variable Oficio con un polinomio de grado 8

| | Dependent variable: |
|---|---|
| | logingtot |
| age | 0,067*** |
| | (0,003) |
| age2 | -0.001*** |
| | (0.00004) |
| female | 0.037 |
| | (0.042) |
| edu | 0.059*** |
| | (0.002) |
| formal | 0.572*** |
| | (0.015) |
| poly(oficio, 8)1 | -13.873*** |
| | (0.830) |
| poly(oficio, 8)2 | 8.921*** |
| | (0.739) |
| poly(oficio, 8)3 | 3.266*** |
| | (0.680) |
| poly(oficio, 8)4 | -6.346*** |
| | (0.685) |
| poly(oficio, 8)5 | 2.362*** |
| | (0.682) |
| poly(oficio, 8)6 | 1.721** |
| | (0.676) |
| poly(oficio, 8)7 | -2.491*** |
| | (0.674) |
| poly(oficio, 8)8 | 0.179 |
| | (0.683) |
| age:female | -0.008*** |
| | (0.001) |
| Constant | 11.423*** |
| | (0.070) |
| Observations | 10,242 |
| R2 | 0.439 |
| Adjusted R2 | 0.439 |
| Residual Std. Error | 0.672 (df = 10227) |
| F Statistic | 572.419*** (df = 14; 10227) |
| Note: | *p<0.1; **p<0.05; ***p<0.01 |

iv. Report and compare the average prediction error of all the models that you estimated before. Discuss the model with the lowest average prediction error.

**Tabla 17.** Error Promedio al Cuadrado de las predicciones

| Modelo | MSE |
|---|---|
| 1 | 0,790 |
| 2 | 0,749 |
| 3 | 0,730 |
| 4 | 0,538 |
| 5 | 0,468 |
| 6 | 0,466 |
| 7 | 0,438 |

Según la tabla anterior, el mejor modelo es el modelo 7 ya que corresponde al menor MSE. Este modelo relaciona como variables control: educación, sexo, edad, sector formal del trabajo y oficio.

v. For the model with the lowest average prediction error, compute the lever- age statistic for each observation in the test sample. Are there any outliers, i.e., observations with high leverage driving the results? Are these out- liers potential people that the DIAN should look into, or are they just the product of a flawed model?

```
A tibble: 6 × 10
    age  age2 female   edu formal model7 logingtot oficio  alphas abs_alpha
  <int> <dbl>  <dbl> <dbl>  <int>  <dbl>     <dbl>  <int>   <dbl>     <dbl>
1    36  1296      0    13      1   14.2      14.3     39  0.363     0.363
2    51  2601      0    16      1   14.4      14.1     85 -0.194     0.194
3    45  2025      0     7      1   13.9      13.7     45 -0.0324    0.0324
4    39  1521      0    20      1   14.8      14.3     33  0.134     0.134
5    44  1936      1    10      0   13.2      12.6     57 -0.448     0.448
6    56  3136      1     7      1   13.4      13.6     55 -0.455     0.455

A tibble: 6 × 10
    age  age2 female   edu formal model7 logingtot oficio  alphas abs_alpha
  <int> <dbl>  <dbl> <dbl>  <int>  <dbl>     <dbl>  <int>   <dbl>     <dbl>
1    39  1521      0    15      0   13.7      13.5     53 -0.656     0.656
2    24   576      1    10      0   12.9      13.5     45  0.229     0.229
3    24   576      0    16      1   14.4      13.9      3  0.0564    0.0564
4    50  2500      1    13      1   14.2      14.0      3  0.0859    0.0859
5    24   576      1    13      1   13.7      13.8     53  0.447     0.447
6    36  1296      1    21      1   14.9      15.7     21  1.80      1.80
```

Con leaverage se puede identificar las personas que tiene mayor discrepancia con los ingresos reportados a través de los alphas en valor absoluto más grandes de las muestras y que la DIAN entre a revisar.

(b) Repeat the previous point but use K-fold cross-validation. Comment on simi- larities/differences of using this approach.

Para correr los modelos de predicción se establece un valor de partición (k-fold) de 5 cuyos resultados se presentan en la siguiente tabla en conjunto con los del numeral a.

**Tabla 4.** Error Promedio al Cuadrado de las predicciones
de Prueba-Entrenamiento y K-Fold

| Modelo | MSE | MSE_k-fold |
|---|---|---|
| 1 | 0,790 | 0.800 |
| 2 | 0,749 | 0.763 |
| 3 | 0,730 | 0.746 |
| 4 | 0,538 | 0.548 |
| 5 | 0,468 | 0.473 |
| 6 | 0,466 | 0.471 |
| 7 | 0,438 | 0.447 |

En la tabla anterior se presenta los errores medios estándar de los modelos de predicción del ingreso mediante la metodología de Prueba-Entrenamiento y de K-Fold. Sobre los resultados obtenidos se puede concluir que son muy similares los errores para el mismo modelo de predicción mediante ambas metodologías, siendo levemente menor los errores con la de prueba y entrenamiento.

(c) *LOOCV*. With your preferred predicted model (the one with the lowest average prediction error) perform the following exercise:

i. Write a loop that does the following:

- Estimate the regression model using all but the $i-th$ observation.
- Calculate the prediction error for the $i-th$ observation, i.e. $(y_i - \hat{y}_i)$
- Calculate the average of the numbers obtained in the previous step to get the average mean square error. This is known as the Leave-One-Out Cross-Validation (LOOCV) statistic.

A partir del siguiente código en R se obtiene el MSE de LOOCV

```
n<- nrow(df)
loocv<-matrix(rep(0,n),nrow=n,ncol=3)
loocv[,1]<-df$logingtot
colnames(loocv)<- c("Observacion","Prediccion","MSE")
for (i in 1:n) {
  reg_i<-lm(logingtot~age+age2+female+edu+formal+age:female+poly(oficio,8),
         data=df[-i,])#entrena con los datos menos la i observación
  loocv[i,2]<-predict(reg_i,newdata=df[i,]) #predice con la i observación
  loocv[i,3]<-(loocv[i,1]-loocv[i,2])^2
  }
loocv <- as.data.frame(loocv)
MSE_loocv <- mean(loocv$MSE)
MSE_loocv
```

El MSE es 0,4477996 que es igual al valor del MSE para K-Fold del numeral anterior, pero sin requerir tantos cálculos computacionales puesto que se requirió hacer $n^2$ iteraciones, siendo n=14,631 observaciones. Finalmente, es pertinente mencionar que la distribución aleatoria que conformó el grupo de entramiento (70% de la muestra) contra el grupo de testeo (30% de la muestra) fue por suerte la que mejores resultados dio al dar un MSE menor, pero esto es poco probable que ocurra.

ii. Compare the results to those obtained in the computation of the leverage statistic

Con las estadísitcas de leverage se pueden identificar los outliers de una manera más rápida y sencilla, sin embargo, se corre el riesgo que los modelos tengan errores altos en sus mediciones.