Big Data

Solución del Problem Set 1

Presentado por: Carlos Avilán y Francisco Ortiz

Enlace GitHub: https://github.com/faor10/ProblemSet1/

1. **Data acquisition**

- **Are there any restrictions to accessing/scraping these data?**

La principal restricción para acceder y poder realizar el scraping de los datos fue que el website https://ignaciomsarmiento.github.io/GEIH2018 sample/ es una web dinámica, por lo cual el proceso requirió de algunos pasos adicionales ya que los datos no estaban fijados en la web.

- **Using pseudocode describe your process of acquiring the data**

Pseudocode (proceso de adquisición de datos):

Ingreso a Google Chrome;

Activación de la opción *inspeccionar (click derecho - inspeccionar);*

Acceso al primer enlace PS1 (ignaciomsarmiento.github.io) en Chrome;

En la opción Red de la consola se obtiene la URL de donde se extrae la primera tabla https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_1.html;

Se repite el proceso b y d para los otros 9 enlaces;

En R, un loop extrae la información de cada una de las 10 tablas y se guardan en un vector;

Cada una de las 10 tablas del vector se convierten en variables tipo data.frame();

Se elimina la primera columna de cada data.frame() (esta solo contiene el número de cada columna);

Con el comando *rbind* se unen cada una de las 10 variables para conformar la base de datos total;

2. Data cleaning

El primer paso para la limpieza y organización de la base de datos fue realizar una inspección general. Se encuentran 32.177 observaciones y 177 variables, donde la mayoría del tipo de variables son numéricas y solo existe una variable de texto la cual corresponde a *dominio*.

Así mismo se definen las variables que eran categóricas y estas se convierten a tipo factor para poder usarlas en R.

Seguidamente se hace una limpieza para cumplir con la condición de que los individuos trabajaran, fueran mayores de 18 y vivieran en Bogotá. Para filtrar los individuos que solo trabajaran se utilizó la variable *Oc* del DANE. Según la definición del DANE, esta variable es la utilizada para medir el empleo en Colombia y se define como las personas que durante el período de referencia se encontraban en una de las siguientes situaciones:

1. Trabajó por lo menos una hora remunerada en dinero o en especie en la semana de referencia
2. Los que no trabajaron la semana de referencia, pero tenían un trabajo.
3. Trabajadores familiares sin remuneración que trabajaron en la semana de referencia por lo menos 1 hora.

La variable para definir que los individuos fueran mayores de 18 años fue *age.* Así mismo, al realizar la inspección se encontró en la variable *dominio* que todas las observaciones estaban en BOGOTA, con lo cual se infirió que todos ellos hacían cumplían con las condiciones.

Después de realizar este filtrado quedaron 16.542 observaciones, casi la mitad de las encontradas inicialmente.

Para poder proseguir con el filtrado y la organización de datos, se definieron las variables para desarrollar la actividad. Estas de alguna manera deben explicar el ingreso individual para cada uno de los individuos. Las siguientes fueron las variables:

Variable	Descripción
ingtot	Ingreso total por persona que resulta de sumar cada una de las fuentes de ingresos tanto observadas como imputadas.
age	Edad del individuo
Sex	Género del individuo. 1. Hombre 0. Mujer
oficio	Oficio que ejerce la persona
maxEducLevel	Máximo nivel de educación alcanzado por la persona. 1. Ninguno 2. Preescolar 3. Primaria incompleta 4. Primaria completa 5. Secundaria incompleta 6. Secundaria completa 7. Terciaria. 9 NA
totalHoursWorked	Total de horas trabajadas en la semana anterior
estrato1	Estrato de energía para las 13 a.M., y sextil de icv para otras cabeceras y resto

Tabla 1. Descripción de las variables escogidas

Posteriormente se revisaron los valores faltantes para cada variable. La única variable con NAs fue maxEducLevel con un 0,0031% del total, un valor bastante mínimo, por lo cual se decidió imputar estos datos con la moda de la variable.

En el caso de los datos atípicos se encontraron datos lejanos a la gran masa de datos para la variable de ingreso *ingtot*. Al realizar una revisión exhaustiva, se verifica que los datos atípicos no son datos sin sentido, sino que corresponden a datos reales pero extremos (explicada por la desigualdad de ingresos en Colombia). Así mismo, se encontraron 265 valores en 0, que de alguna u otra manera pueden ser datos correctos pero que el uso de estos en nuestros modelos puede generar problemas, especialmente en el uso de los logaritmos. Para la corrección de estos se decidió reemplazarlos por el ingreso mínimo que puede recibir una persona, el cual está en 15.000 COP. Otras posibles soluciones como reemplazar esos valores por la media de todos los ingresos de la base de datos, o por el ingreso de otra persona del mismo hogar (una persona del mismo hogar puede ganar mucho más que otra, ejemplo: la cabeza del hogar comparado con los hijos) no lo consideramos pertinente ya que inflaría sustancialmente el resultado y perdería la naturaleza de la observación, la cual radica en un ingreso extremadamente bajo.

Finalmente, con la limpieza y transformaciones pertinentes obtenemos las siguientes estadísticas descriptivas para las variables numéricas continuas:

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Ingtot	15.000	8.00E+05	1.051.159,83	1.769.379,01	1.723.158,25	85.833.333,33
age	18	28	38	39.4361625	50	94
totalHoursWorked	1	40	48	47.402672	50	130

Tabla 2. Estadísticas descriptivas de las variables continuas.

Según la tabla 2, es interesante anotar que la media de los ingresos de las personas observadas se encuentra en $1.723.158 COP. La edad media está en 38 años y el promedio de horas trabajadas está en 48 horas a la semana. Datos muy acordes a la realidad laboral de la ciudad de Bogotá.

Para las variables categoricas se encuentra lo siguiente:

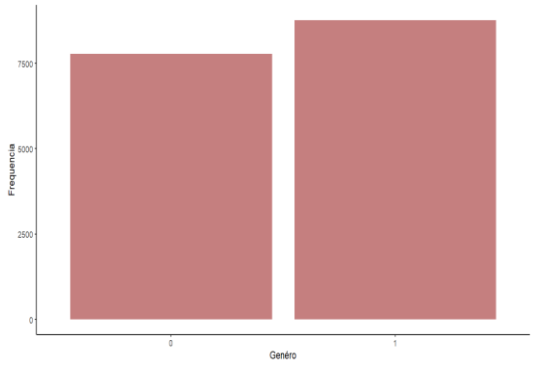

Ilustración 1. Variable género

Como se observa en la ilustración 1, en la variable Sex, se encuentran que 8.767 son hombres y 7.775 mujeres.

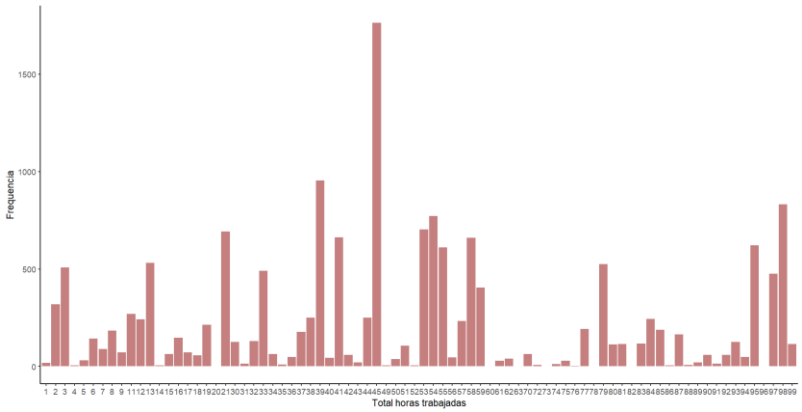

Ilustración 2. Variable oficio

Como se observa en la ilustración 2, la variable oficio, indica que el oficio que más predominante que ejercen las personas es el 45: vendedores, ambulantes, a domicilio, de loterías y periódicos, mercaderistas. Lo analizado indica una fuerte presencia de informalidad en las personas evaluadas.

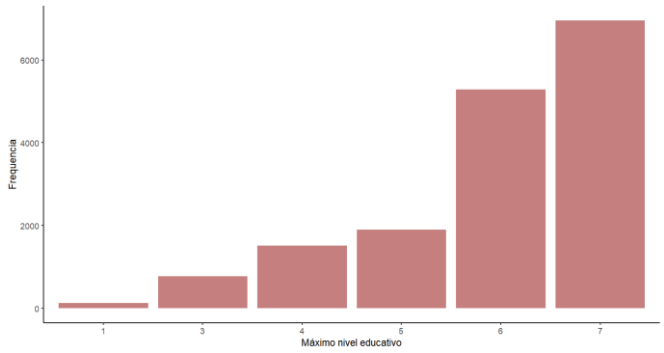

Ilustración 3. Máximo nivel educativo

En la ilustracion3, en variable MaxEducLevel, la máxima educación alcanzada por las personas se encuentra en la terciaria seguida de la secundaria completa, es decir, más de 11 mil personas han completado su educación media.

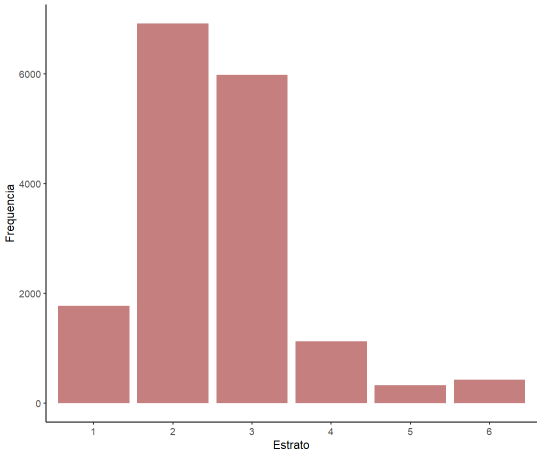

Ilustración 4. Estrato de las personas

Como se observa en la ilustración 4, en la variable estrato1 de las personas observadas, el predominante es el estrato 2 seguido del 3. Cerca del 80% de las observaciones corresponde a estos dos estratos.

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; and the wage rate tends to remain stable or decline slightly after age 50.

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

La variable escogida para representar los ingresos totales de los trabajadores fue "ingtot" y representa el ingreso total de los mismos. En primer lugar, se escogió esta variable ya que, a comparación de las demás variables relacionadas con los ingresos, representa los ingresos totales de los individuos pues otras variables representan ingresos parciales o fraccionados del ingreso total (según el DANE es el resultado de sumar cada una de las fuentes de ingresos tanto observadas como imputadas). Así mismo, como se describió en la sección anterior esta variable no tiene *missing values*.

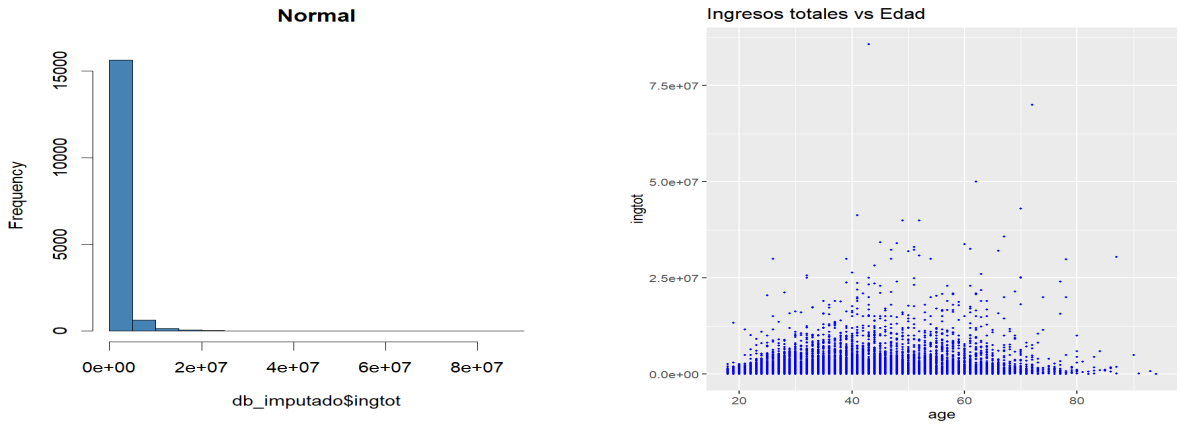

Ilustración 5. Ingresos totales

Estadísticas descriptivas variable ingtot					
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
15.000	800.000	1.051.160	1.769.379	1.723.158	85.833.333

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

Income = β1 + β2Age+ β3Age2 + u

Resultados regresión ingresos-edad	
Dependent variable: ingtot	
age	**91,143.460*** **
	(8,886.416)
age2	**-799.261*** **
	(102.852)
Constant	**-436,662.900** **
	(178,347.200)
Observations	16,542
R2	0.017
Adjusted R2	0.017
Residual Std. Error	2,652,732.000 (df = 16539)
F Statistic	144.382*** (df = 2; 16539)

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

La regresión por MCO nos muestra que todas las variables explicativas del modelo son significativas al 1% puesto que su p-valor es menor a 0,01. Así mismo, El estadístico F de significancia global del modelo es significativo al 1%, es decir que el modelo tiene una significancia explicativa global. Se puede apreciar que uno de los coeficientes relacionados con la edad es positivo mientras que el otro es negativo, lo cual es coherente con la forma funcional cuadrática del modelo.

• **How good is this model in sample fit?**

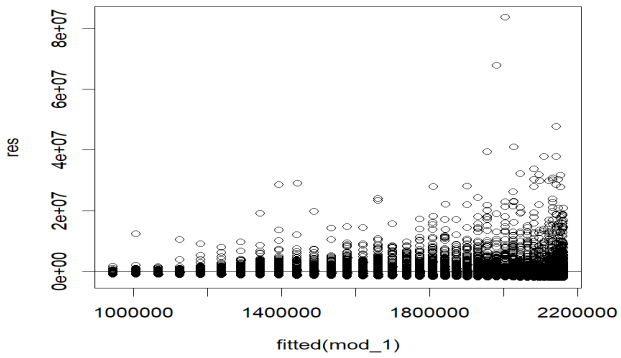

Ilustración 6. Gráfica de residuales

Para poder saber qué tan bueno es este modelo en el ajuste de la muestra, se realizó una gráfica de los residuales frente a los valores predichos del modelo (Ilustración 6). En esta gráfica se puede observar que el modelo ***Income = β1 + β2Age+ β3Age2 + u*** no se ajusta del todo bien a la muestra puesto que existen diferencias entre los residuales y valores predichos diferentes a cero, siendo esto una muestra de que el ajuste no es el adecuado. De la misma manera, si analizamos el R-cuadrado de la regresión podemos ver que este fue de 0,017, es decir, que las variables explicativas explican en 1,7% la varianza de la variable explicativa siendo este valor muy bajo e ilustrando el bajo ajuste del modelo en la muestra.

Algunas de las razones del por qué el modelo no se ajusta correctamente a la muestra pueden ser: variables omitidas y mala especificación del modelo.

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

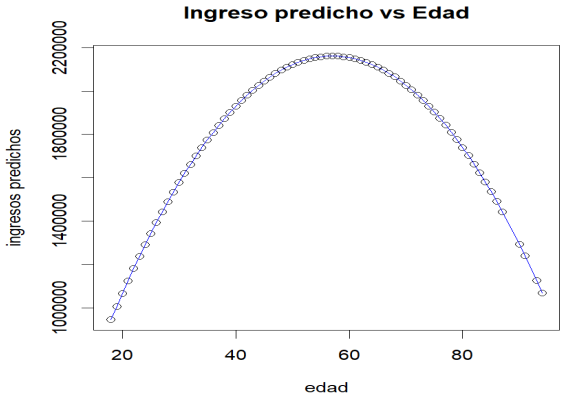

Ilustración 7. Ingreso predicho vs Edad

La ilustración 7 nos muestra los valores predichos de los ingresos totales de los individuos de Bogotá frente a la edad de estos. Se puede analizar en la gráfica que la relación entre valores predichos de los ingresos y la edad es coherente con la forma funcional cuadrática del modelo, en la cual se espera que el ingreso total alcance un máximo a cierta edad y empiece a disminuir paulatinamente a medida que la edad aumenta.

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

Para poder calcular "la edad pico" de ingresos en nuestro modelo se tuvo que encontrar la elasticidad ingreso-edad a partir de derivar nuestra función inicial e igualar a cero en busca de encontrar la edad máxima. A continuación, se muestra el procedimiento realizado:

$$Income = B_1 + B_2 Age + B_3 Age^2 + u$$

$$\frac{dy(Income)}{dx\ (age)} = B_2 + 2B_3 Age = 0$$

$$Age = \frac{-B_2}{2B_3} = \frac{-\mathbf{91{,}143.460}}{2 * -\mathbf{799.261}} = 57\ años$$

De acuerdo con este procedimiento, "la edad pico" de ingresos en nuestro modelo es de 57 años estando acorde con lo que expone la literatura de una edad pico de ingresos cerca a los 50 años.

Ahora bien, dado que la forma funcional del modelo es cuadrática, los coeficientes estimados con respecto a la edad no se pueden interpretar directamente y es necesario encontrar el efecto marginal ingreso-edad por medio de derivar nuestro modelo. Así mismo, se va a utilizar la edad media de la muestra para poder encontrar este efecto marginal

$$Income = B_1 + B_2 Age + B_3 Age^2 + u$$

$$\frac{dy(Income)}{dx\ (age)} = B_2 + 2B_3 meanAge = 28.103{,}87$$

Este efecto marginal se puede interpretar como que, manteniendo todas las demás variables constantes, en promedio un año más de edad implica que el ingreso aumente en 28.103,7 pesos.

Para encontrar los errores estándar asociados al efecto marginal ingreso-edad e intervalos de confianza, usamos el procedimiento Bootstrap y obtuvimos los siguientes resultados:

Bootstrap Statistics :		
original	**bias**	**std. error**
t1* 28103.87	**35.86716**	**1491.037**

Estos resultados están acordes a nuestras estimaciones, puesto que se puede apreciar que el efecto marginal por medio de Bootstrap es el mismo que encontramos con anterioridad (28.103,87) y podemos ver que el error estándar es de 1.491,037.

Con este error estándar, procedemos a calcular un intervalo de confianza del 95% de probabilidad del efecto marginal ingreso-edad como se puede observar a continuación:

$$Efecto\ Marginal\ ingreso_{edad} \pm 1.96 * SE$$

$$28.103{,}87 \pm 1.96 * 1.491{,}037$$

$$(25.262{,}37\ ,\ 30.945{,}37\ )$$

Este intervalo de confianza se puede interpretar como que en promedio un año más de edad de un individuo de Bogotá implica que sus ingresos aumenten en el rango del intervalo $(25.262{,}37\ ,\ 30.945{,}37\ )$.

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 earnings gap.

• Estimate the unconditional earnings gap: log(Income) = β1 + β2Female + u

Regresión Log ingresos-mujer

	Dependent variable: log_ingreso
	---
mujer	**-0.378*** **
	(0.030)
Constant	**13.927*** **
	(0.021)

Observations	16,542
R2	0.009
Adjusted R2	0.009
Residual Std. Error	1.950 (df = 16540)
F Statistic	154.943*** (df = 1; 16540)

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

La regresión por MCO nos muestra que todas las variables explicativas del modelo son significativas al 1% puesto que su p-valor es menor a 0,01. Así mismo, El estadístico F de significancia global del modelo es significativo al 1%, es decir que el modelo tiene una significancia explicativa global.

• How should we interpret the β2 coefficient? How good is this model in sample fit?

El coeficiente asociado a la variable dummy de mujer (toma el valor de 1 si el individuo es mujer y 0 si es hombre) se puede interpretar como que, manteniendo todas las demás variables constantes, en promedio las mujeres ganan 37,8% menos en ingresos que a comparación de los hombres.

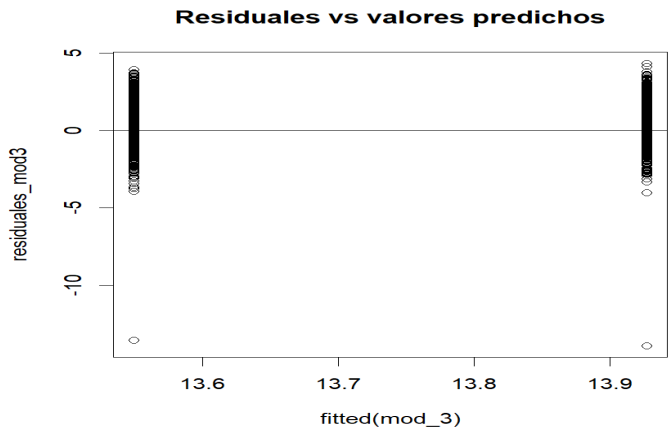

Ilustración 8. Residuales vs valores predichos

Para poder saber qué tan bueno es este modelo en el ajuste de la muestra, se realizó una gráfica de los residuales frente a los valores predichos del modelo ***log(Income) = β1 + β2Female + u*** . En la ilustración 8 se puede observar que el modelo no se ajusta bien a la muestra puesto que existen grandes diferencias entre los residuales y valores predichos al ser esas diferencias diferentes a cero, siendo esto una muestra de que el ajuste no es el adecuado. De la misma manera, si analizamos el R-cuadrado de la regresión podemos ver que este fue de 0,009, es decir, que las variables explicativas explican en 0,9% la varianza de la variable explicativa siendo este valor muy bajo e ilustrando el bajo ajuste del modelo en la muestra.

Algunas de las razones del por qué el modelo no se ajusta correctamente a la muestra pueden ser: variables omitidas y mala especificación del modelo.

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

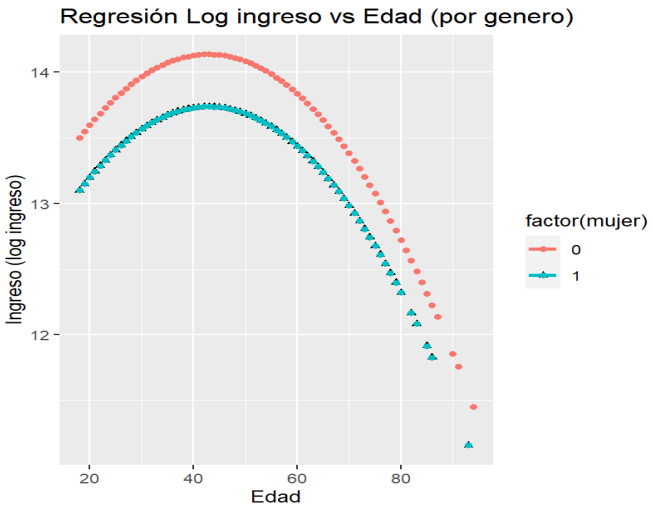

Ilustración 9. Valores predichos edad-ingresos por genero

De acuerdo con la ilustración 9 del logaritmo natural de los ingresos predichos contra la edad de los individuos por género podemos analizar de que definitivamente el sexo del individuo

afecta el ingreso de acuerdo con su edad. Específicamente, la gráfica ilustra que los valores predichos de los ingresos-edad es diferente entre hombres y mujeres, pues las pendientes e interceptos de las líneas de regresión difieren. En el caso de los hombres, la linea de regresión (línea rosada) tiene una pendiente y un intercepto mayor que el de las mujeres (linea azul), estando esto relacionado con que los ingresos que reciben los hombres son mayores a los de las mujeres.

• 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?

Regresión lineal ingresos-edad-genero-edad*genero

	Dependent variable: log_ingreso
age	0.098***
	(0.007)
age2	-0.001***
	(0.0001)
mujer	0.244***
	(0.094)
mujer_age	-0.016***
	(0.002)
Constant	11.908***
	(0.138)
Observations	16,542
R2	0.023
Adjusted R2	0.023
Residual Std. Error	1.936 (df = 16537)
F Statistic	99.240*** (df = 4; 16537)

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

Para poder calcular "la edad pico" de ingresos se construyó un modelo que tuviera la interacción entre género y edad con el fin de encontrar la elasticidad ingreso-edad por género. Esto a partir de derivar nuestra función e igualarla a cero en busca de encontrar la edad máxima. A continuación, se muestra el procedimiento realizado:

En el caso de las mujeres:

$$\log(income) = B_1 + B_2 Age + B_3 Age^2 + B_4 Age * Female + u$$

$$\frac{dy(\log(income))}{dx\ (age)} = B_2 + 2B_3 Age + B_4 Female = 0$$

$$Age\_Female = \frac{-B_2 - B_4}{2B_3} = \frac{-0.098 + 0.016}{2 * -0.001} = \ 38.5 \ años$$

En el caso de los hombres:

$$\log(income) = B_1 + B_2 Age + B_3 Age^2 + B_4 Age * Female + u$$

$$\frac{dy(\log(income))}{dx\ (age)} = B_2 + 2B_3 Age = 0$$

$$Age\_Male = \frac{-B_2}{2B_3} = \frac{-0.098}{2 * -0.001} = \ 46.2 \ años$$

Los picos de edad de ingresos para mujeres son a los 38,5 años mientras que el de los hombres es a los 46,2 años. Estos resultados están acordes con nuestra gráfica de los ingresos predichos con respecto a la edad por género y, así mismo, ilustra que los hombres tienen una mayor fuente de ingresos por más tiempo a comparación de las mujeres.

Para encontrar los errores estándar asociados al efecto marginal ingreso-edad por género e intervalos de confianza, usamos el procedimiento Bootstrap y obtuvimos los siguientes resultados:

En el caso de las mujeres: (se utilizó la edad promedio de las mujeres en nuestra muestra)

Bootstrap Statistics :		
original	**bias**	**std. error**
t1* -0.00143758	**-5.376217e-05**	**0.002163139**

Estos resultados nos muestran que el efecto marginal ingreso-edad para las mujeres es de -0,00143, es decir, manteniendo las demás variables explicativas constantes, en promedio con un año más de edad el ingreso de las mujeres disminuye en 0,143% con respecto a los hombres.

Con este error estándar, procedemos a calcular un intervalo de confianza del 95% de probabilidad del efecto marginal ingreso-edad para las mujeres como se puede observar a continuación:

$$Efecto\ Marginal\ ingreso\_edad\_mujeres \pm 1.96 * SE$$

$$-0.00143758 \pm 1.96 * 0.002163139$$

$$(\ -0{,}005706096\% \ , 0{,}002830935\%)$$

Este intervalo de confianza se puede interpretar como en promedio un año más de edad en las mujeres de Bogotá implica que sus ingresos cambien porcentualmente en el rango del intervalo $(-0{,}005706096\% \ , \ 0{,}002830935\%)$ frente a los hombres.

En el caso de los hombres: (se utilizó la edad promedio de los hombres en nuestra muestra)

Bootstrap Statistics :		
original	bias	std. error
t1* 0.01414982	4.626327e-06	0.001515723

Estos resultados nos muestran que el efecto marginal ingreso-edad para los hombres es de 0,00141, es decir, manteniendo las demás variables explicativas constantes, en promedio con un año más de edad el ingreso de los hombres aumenta en 1,41% con respecto a las mujeres.

Con este error estándar, procedemos a calcular un intervalo de confianza del 95% de probabilidad del efecto marginal ingreso-edad para los hombres como se puede observar a continuación:

$$Efecto\ Marginal\ ingreso\_edad\_hombres \pm 1.96 * SE$$

$$0.01414982 \ \pm 1.96 * \ 0.001515723$$

$$(\ 0{,}01117025\% \ , 0{,}01712938\% \ )$$

Este intervalo de confianza se puede interpretar como en promedio un año más de edad en los hombres de Bogotá implica que sus ingresos cambien porcentualmente en el rango del intervalo $(0{,}01117025\% \ , 0{,}01712938\% \ )$ frente a las mujeres.

Los intervalos de confianza entre hombres y mujeres no se superponen, siendo muestra de esto que, en promedio, la diferencia de ingresos al aumentar un año entre hombres y mujeres es significativa.

• 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 job characteristics (X).

(a) Estimate the conditional earnings gap log(Income) = β1 + β2Female + θX + u

Regresión Log ingresos-mujer-oficio

Dependent variable: log_ingreso	
mujer	**-0.358*** **
	(0.034)
.	
.	
.	
.	
.	
Constant	**15.208*** **
	(0.423)
Observations	16,542
R2	0.117
Adjusted R2	0.113
Residual Std. Error	1.844 (df = 16461)
F Statistic	27.391*** (df = 80; 16461)

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

El modelo planteado para este punto fue el siguiente:

$$\log(Income) = B_1 + B_2 Female + B_3 Oficio + u$$

La variable de control seleccionada fue *"oficio"*, la cual representa el trabajo en el que se estaba desenvolviendo cada individuo en el momento de realizar la encuesta.

La regresión por MCO nos muestra que todas las variables explicativas del modelo son significativas al 1% puesto que su p-valor es menor a 0.01. Así mismo, el estadístico F de significancia global del modelo es significativo al 1%, es decir que el modelo tiene una significancia explicativa global.

(b) Use FWL to repeat the above estimation, where the interest lies on β2. Do you obtain the same estimates?

Para aplicar el Teorema FWL es necesario encontrar los residuales de la regresión entre el logaritmo del ingreso frente a la variable explicativa Oficio y, así mismo, se deben encontrar los residuales de la regresión entre el género del individuo (*female*) y el *Oficio*.

$$Residuales_1 \sim \log(ingreso) = B_1 + B_2 Oficio + u$$

$$Residuales_2 \sim \text{Female} = B_1 + B_2 Oficio + u$$

Con la identificación de los residuales, realizamos la siguiente regresión para estimar el coeficiente B2.

$$B_2 \sim Residuales_1 = -1 + Residuales_2$$

$$B_2 = -0.358$$

Este coeficiente estimado por medio del teorema FWL es igual al estimado por medio de OLS.

(c) How should we interpret the β2 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"?

El coeficiente se interpreta como que, manteniendo las demás variables explicativas constantes, el ingreso de las mujeres es en promedio 35,8% menor en comparación al ingreso de los hombres. Si comparamos el efecto marginal ingreso-genero de este modelo (-35,8%) con interacción frente al modelo sin interacción (-37,8%) podemos analizar que la brecha disminuye ligeramente. Esto podría explicarse como que la brecha de ingresos entre genero si puede relacionarse con un problema de selección de las variables explicativas del modelo, puesto que puede que haya variables omitidas. Sin embargo, estas brechas de ingresos entre hombres y mujeres siguen siendo un problema de discriminación porque en todas las regresiones realizadas las mujeres tienen un menor salario que los hombres.

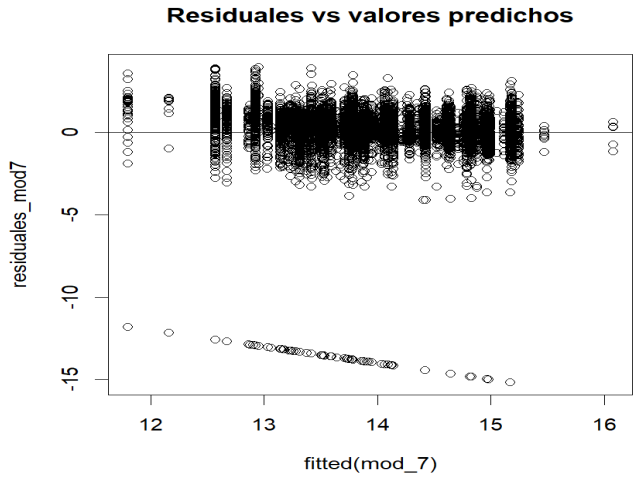

Ilustración 10. Residuales vs valores predichos

Para poder saber qué tan bueno es este modelo en el ajuste de la muestra, se realizó una gráfica de los residuales frente a los valores predichos del modelo. En la ilustración 10 se puede observar que el modelo ***log(Income) = β1 + β2Female+ β3Oficio + u*** no se ajusta bien a la muestra puesto que existen diferencias entre los residuales y valores predichos diferentes a cero, siendo esto una muestra de que el ajuste no es el adecuado. De la misma manera, si analizamos el R-cuadrado de la regresión podemos ver que este fue de 0,117, es decir, que las variables explicativas explican en 11,7% la varianza de la variable explicativa siendo este valor muy bajo e ilustrando el bajo ajuste del modelo en la muestra.

Algunas de las razones del por qué el modelo no se ajusta correctamente a la muestra pueden ser: variables omitidas, mala especificación del modelo y datos outliers.

5. **Predict earnings**

- **Estimate a model that only includes a constant.**
- **Estimate again your previous models**
- **At this point, explore other transformations of your independent variables also. Try at least (5) models that are increasing in complexity.**

Se generan las dos muestras: test y training, y se procede a estimar nuevamente los modelos obteniendo los siguientes resultados de MSE y RMSE

Modelo constante:

$$Income = B_1 + u$$

Modelo age earning:

$$Income = B_1 + B_2 Age + B_3 Age^2 + u$$

Modelo gap female:

$$\log(Income) = B_1 + B_2 Female + u$$

A continuación, explorando nuevas transformaciones y añadiendo nuevos elementos se proponen lo siguientes 5 modelos aumentando las variables de control, los polinomios, las interacciones y la complejidad de estos en cada uno:

Model 1:

$$Income = B_1 + B_2 Age + B_3 Age^2 + B_4 totalHoursWorked + B_5 maxEducLevel + B_6 Female$$

Model 2:

$$Income = B_1 + B_2 Age + B_3 Age^2 + B_4 totalHoursWorked + B_5 totalHoursWorked^2 + B_6 maxEducLevel + B_7 Female + u$$

Model 3:

$$Income = B_1 + B_2 Age + B_3 Age^2 + B_4 totalHoursWorked + B_5 totalHoursWorked^2 + B_6 maxEducLevel + B_7 Female + B_8 Female * totalHoursWorked + u$$

Model 4:

$$Income = B_1 + B_2 Age + B_3 Age^2 + B_4 totalHoursWorked + B_5 totalHoursWorked^2 + B_6 totalHoursWorked^3 + B_7 maxEducLevel + B_8 Female + B_9 Female * totalHoursWorked + u$$

Model 5:

$$Income = B_1 + B_2 Age + B_3 Age^2 + B_4 totalHoursWorked + B_5 totalHoursWorked^2 + B_6 totalHoursWorked^3 + B_7 totalHoursWorked^4 + B_8 maxEducLevel + B_9 Female + B_{10} Female * totalHoursWorked + B_{11} estrato1$$

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

Finalmente, obtenemos el MSE y RMSE para cada uno de los modelos:

Modelo	MSE	RMSE
Constante	7.688101e+12	2772742.51
Age earning	7.56546e+12	2750538.13
Gap female	3.424669	1.85058612
Modelo 1	6.502494e+12	2549998.824
Modelo 2	6.480542e+12	2545690.869
Modelo 3	6.471915e+12	2543995.873
Modelo 4	6.465092e+12	2542654.518
Modelo 5	5.041289e+12	2245281.497

Tabla 3. MSE y RMSE con enfoque de validación para cada modelo

Como se observa en la tabla el modelo 5 es el que tiene menor MSE. En este caso la idea era evitar un sobreajuste dentro de la muestra y poder predecir correctamente fuera de muestra.

- **For the model with the lowest average prediction error, compute the leverage statistic for each observation in the test sample.**

A continuación, realizamos la computación del *leverage statistic* y de los residuales en la muestra de prueba para poder encontrar posibles outliers y determinar que representan.

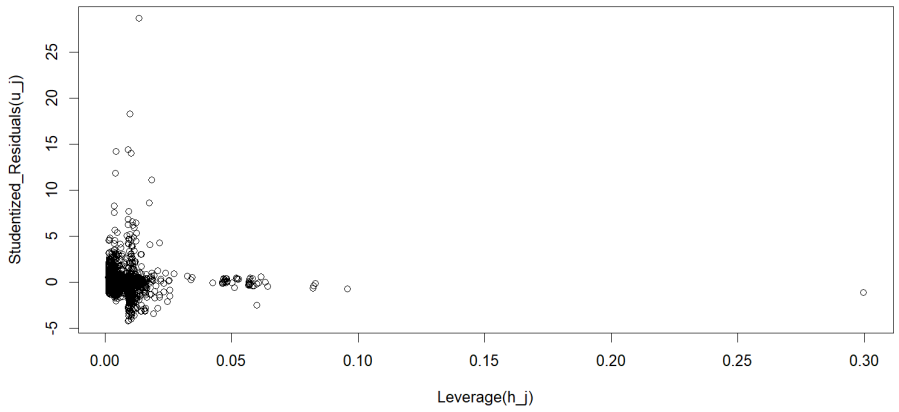

Ilustración 11. Leverage vs Studentized Residuals

En la ilustración 11, los *studentized residuals* nos ayuda a identificar posibles outliers. Al revisar las observaciones más alejadas con *studentized residuals* de 28, 18, 14 y 13. Encontramos que estos valores son ingresos que superan los 20 millones de COP mensuales y están alejados de lo que se considera un ingreso cercano a la media (1 millón 700 mil COP). Por ejemplo, para la observación con residual 28, la más alejada, el ingreso total es de 70 millones y el predicho es de 10.8 millones. Esto quiere decir que, aunque no se pueden considerar datos erróneos puesto que una persona si puede ganar este salario, son valores atípicos en el cual el modelo no se ajusta de la mejor manera. Ahora bien, indiscutiblemente se espera que la DIAN deba revisar que está pasando con estos ingresos tan elevados, sin embargo, esta no es la finalidad de la encuesta GEIH. Si las personas conocen de antemano que esta información la puede usar la DIAN para actividades fiscales, las personas podrían mentir en estas encuestas y por lo tanto la información seria sesgada.

En el caso del leverage se encuentra una observación con un valor cercano al 0.30. En este caso el valor que gana la persona es de 780 mil COP y el modelo sugiere que debería ganar 2 millones 800 mil COP. En este caso el leverage va a indicar una influencia negativa de la observación en particular sobre el modelo, sin embargo, no nos está indicando que es un valor atípico, puesto que un ingreso de 780 mil COP es sensato. Cabe aclarar que no se encuentran puntos con residuales y leverage altos al mismo tiempo, por lo que no resultan críticos las observaciones particulares encontradas.

- **Repeat the previous point but use K-fold cross-validation.**

Ahora realizamos el mismo procedimiento utilizando validación cruzada en K-partes:

Model	MSE	RMSE
Age Earning	6.98303E+12	2642542
Gap Female	3.790976444	1.947043
Modelo 1	5.93205E+12	2435580
Modelo 2	5.92807E+12	2434762
Modelo 3	5.94279E+12	2437783
Modelo 4	5.93844E+12	2436891
Modelo 5	4.53222E+12	2128900

Tabla 4. MSE y RMSE con enfoque de validación cruzada en K-partes

En este caso se realiza un punto intermedio entre el enfoque de validación y el LOOCV. Hay un punto medio el cual es óptimo, el K-partes, en este caso es 5. En este caso, la validación cruzada de K-partes nos ayuda a reducir a variabilidad del enfoque de validación haciéndolo más estable. Al comparar los resultados del MSE de la validación cruzada, se encuentra que son menores a los del enfoque de validación ya que estos últimos tienden a sobreestimar el error de predicción en la muestra de prueba. Nuevamente el modelo 5 es el que tiene menor MSE.

- **LOOCV. With your preferred predicted model (the one with the lowest average prediction error)**

Ahora con el modelo 5, realizamos el LOOCV, el cual se puede interpretar como una validación cruzada, pero de K=n (número de observaciones totales). En este caso mediante un loop que se repite i veces, estimamos la regresión sin la i-ésima observación y su correspondiente error de predicción $(y_i - \hat{y}_i)$. Finalmente calculamos la media de estos valores calculados para obtener el MSE. Esto se conoce como el Leave-One-Out Cross-Validation (LOOCV).

$$CV_n = \frac{1}{n}\sum_{i=1}^{n} MSE_i = 4.565283\text{e} + 12$$

Nótese que este valor es muy parecido al calculado con la validación cruzada en K-partes. Una diferencia bastante clara fue el tiempo de cómputo, ya que el loop se realiza para las n observaciones, se hicieron 16.542 regresiones. En un computador de características altas se demoró alrededor 20 minutos.

- **Compare the results to those obtained in the computation of the leverage statistic**

Ahora con un atajo gracias a la relación entre el MSE y el Leverage se puede hacer que el coste de LOOCV sea el mismo que el de una sola regresión. Al realizar el cálculo con R se encuentra:

$$CV_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{y_i}}{1 - h_i}\right)^2 = 4.565283\text{e} + 12$$

Como se observa en los resultados, ambos son iguales, la diferencia sustancial es el ahorro de procesamiento, al pasar de realizar 16.542 regresiones a solo una y ahorrarnos un tiempo de cerca de 20 minutos.