Problem set 1

Juan José Rincón, Juanita Chacón, Andrés Opina

Link GitHub: https://github.com/JJR9903/Problem-set1-BigData-ML-Uniandes/

Introducción

En este Problem Set se busca predecir el ingreso de los ciudadanos de Bogotá como el resultado del procesamiento de la Gran Encuesta Integrada de Hogares (GEIH) para 2018 con herramientas de econometría, Big Data y Machine Learning. Este documento se encuentra dividido en cuatro secciones: La primera consiste en el procesamiento de los datos, su carga, descripción, limpieza y la descripción de la base final que se utilizar, la segunda parte utiliza un modelo econométrico que usa la edad para explicar el ingreso individual. la tercera parte contiene un modelo econométrico que involucra el género para explotar las brechas salariales y explicar el ingreso individual, y la última parte consiste en comprobar la validación de los modelos intuidos por medio del uso de herramientas propias de Machine Learning.

Parte 1.2

- **Carga de datos y descripción inicial**

Los sets de datos utilizados como insumos consisten en una serie de diez archivos parciales de la GEIH, realizada por el Departamento Nacional de Estadística (DANE) en 2018. En esta base cuenta con información de ciudadanos de Bogotá únicamente. Estos datos se obtuvieron por medio de web scraping, el cual se le realizo a diez archivos en formato html, con esto se obtuvo una data set con 32,177 observaciones, una para cada hogar encuestado en Bogotá, en donde la unidad de medida es por individuo y con 178 variables.

En relación con la calidad de los datos, la base trae información sobre mercado laboral, educación y características básicas de las personas. Sin embargo, debido a la conformación de las preguntas la data presenta problemas de missing values. Por ejemplo, la variable P6580 se refiere a la pregunta "¿el mes pasado recibía bonificaciones?", entonces sólo aquellas preguntas que respondan sí podrán contestar preguntas subsiguientes como la P6580s1 "¿Cuánto recibí por bonificaciones?".

Limpieza y descripción de los datos

- Limpieza

Teniendo en cuenta la finalidad del análisis se entiende que las variables de interés en este caso son aquellas relacionadas al mercado laboral y las condiciones propias de la persona. Así pues, serían las variables de edad, sexo, educación, tipo de sistema de seguridad social, empleo, edad profesional de empleo, diferentes tipos de ingreso (laboral, rentas, pensiones, etc.). La base con la que se cuenta tiene la construcción en agregado de diferentes variables útiles para el inicio.

La limpieza de los datos se inició con filtrar a los ciudadanos que son mayores de 18 años y que se encontraban dentro de la categoría de ocupados, por temas de eficiencia se dejó únicamente las variables de interés y se hizo las respectivas codificaciones para las variables categóricas con el fin de hacer uso de ellas de mejor manera durante las predicciones.

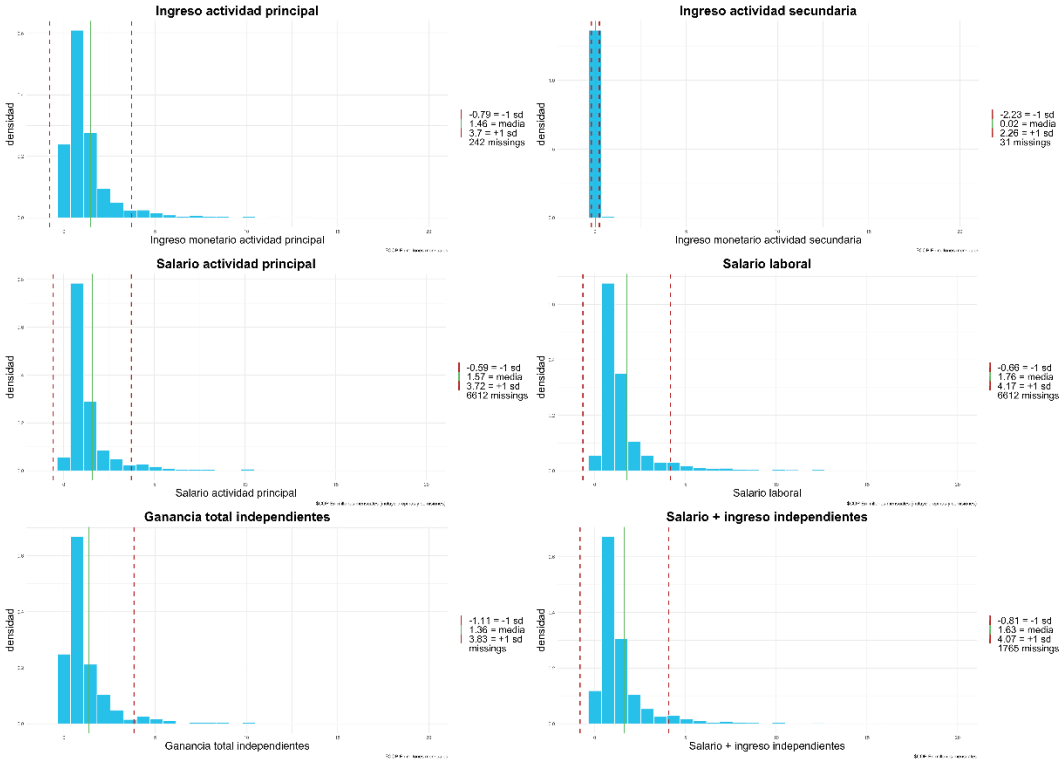
- Descripción.

Variables Ingreso

La base de datos cuenta con numerosas variables de ingreso, las cuales están desagregadas por tipos, en una primera sección se encuentra las variables de ingreso en “crudo” y por otra parte se encontraban otras previamente construidas. También la base de datos cuenta con cada tipo de variable de ingreso, por ingreso mensual o por horas, sin embargo, dado que en colombia el común denominador es hablar de ingresos y salarios mensuales, se conservaron solo las variables se encuentran en millones de pesos colombianos mensuales.

Con el fin de realizar un análisis más detallado se trabajó con las variables previamente construidas y limpias.

Grafica 1. Distribución variable ingreso.



Los anteriores histogramas muestran como sin importar cual categoría de ingreso sea la distribución se encuentra concentrada a la cola izquierda. Uno de los problemas que se vio fue el alto nivel de missing que presentaron. Se observa que en promedio la media de la concentración de la población para las variables es muy similar entre sí. Por lo que es posible establecer patrones de comportamiento del mercado laboral sin importar el tipo de actividad (esto para los ocupados, mayores de 18 y pertenecientes a la ciudad de Bogotá)

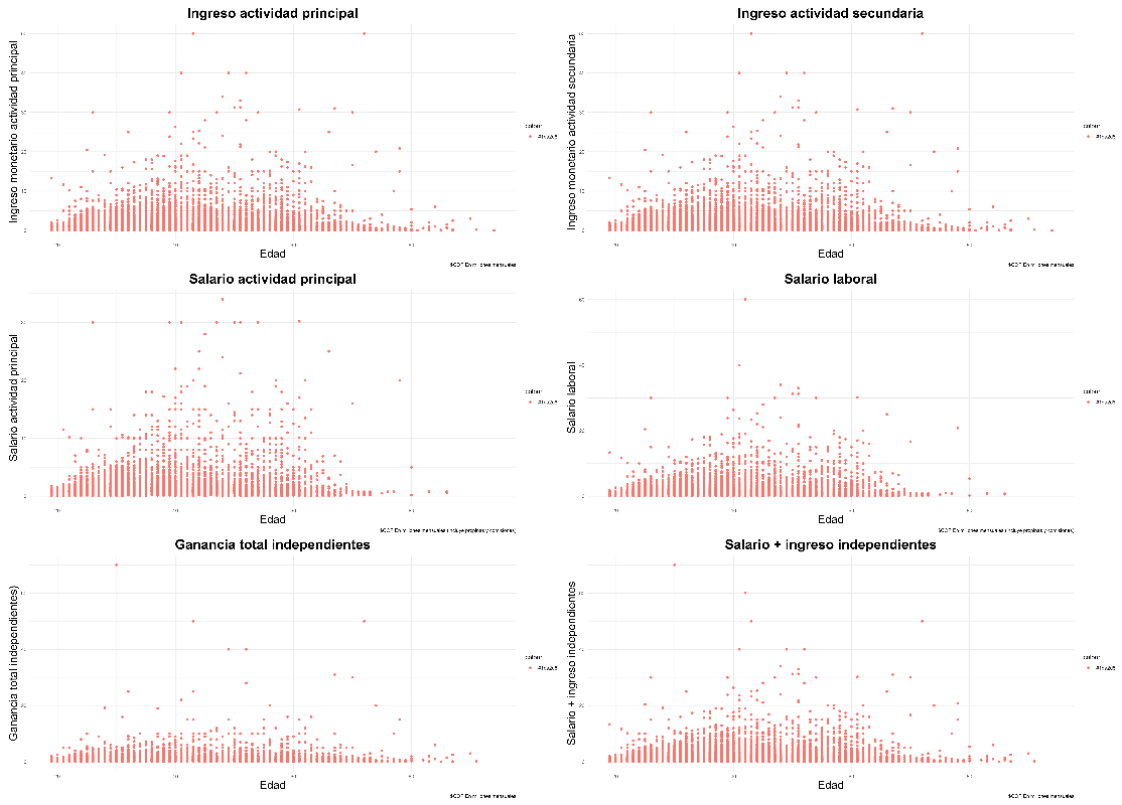
Variable ingreso en relación con la edad

Con el fin de identificar la relación entre el ingreso y la edad se plateo la siguiente gráfica.

Como se puede ver en la gráfica 2. Todas las gráficas muestran la misma forma funcional en relación con la distribución de los datos de la interacción, en donde se ve claramente que existe una tendencia crecientemente marginal, donde el punto máximo es alcanzado en su mayoría en la edad media de la vida laboral.

Teniendo en cuenta lo anterior y con el fin de identificar la implicación de ciertas características en el mercado laboral, se tomó un proxy de esto el ingreso, así pues, se buscó una variable que tuviera una construcción robusta sobre los ingresos, donde el salario fuera implicado para los fines de este trabajo tanto para personas formales, informales e independientes. En este sentido las dos variables que cumplían con la descripción son ingreso de la primera actividad y salario + ingreso de independientes, y por lo tanto como se ve en la gráfica 2, estas son las variables wn las que se ve una relación cuadrática más clara entre la edad y el ingreso.

Grafica 2. Comparaciones variables ingreso según la edad.



Uno de los problemas encontrados frente a la variable elegida fue la presencia de missing values, sin embargo, estos se rellenaron con la media de la variable condicional en las siguientes características: estrato, sexo, edad, nivel educativo alcanzado, y oficio. Puesto que de esta forma se logra una mayor precisión en el proceso generador de los datos de ingreso que solamente utilizando la media muestral. Adicionalmente como corresponden a una variable en unidad de valor monetario se tomó la decisión de construir la normalización log linealizandola.

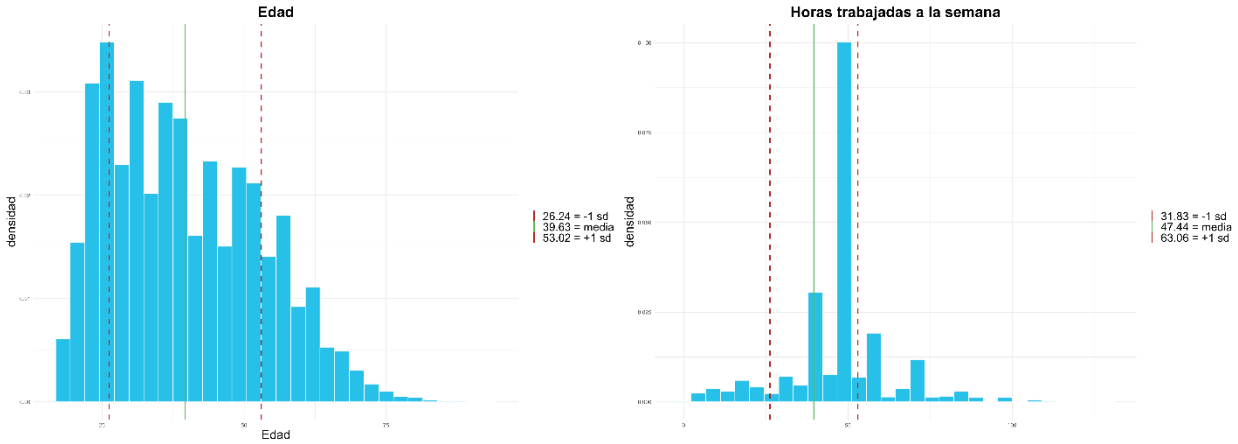
Edad y horas trabajadas

Al revisar cual es la edad de la población encuestada que se encuentra en la categoría de ocupados se observa que el mínimo de edad es de 18 años y el máximo es de 80, adicionalmente, la mayor concentración de la población se encuentra entre los 26 y los 53 años. Donde la media de la edad de trabajo está en 39 años.

En relación a las horas trabajadas el grueso de la población trabaja entre 31.83 - 63.03 horas a la semana, siendo 47.44 horas la media de horas trabajadas por la población Bogotana. Vale la pena

hacer la salvedad de los posibles sesgos que se puedan encontrar en esta, principalmente por la forma en la que es recolectada la información en la GEIH.

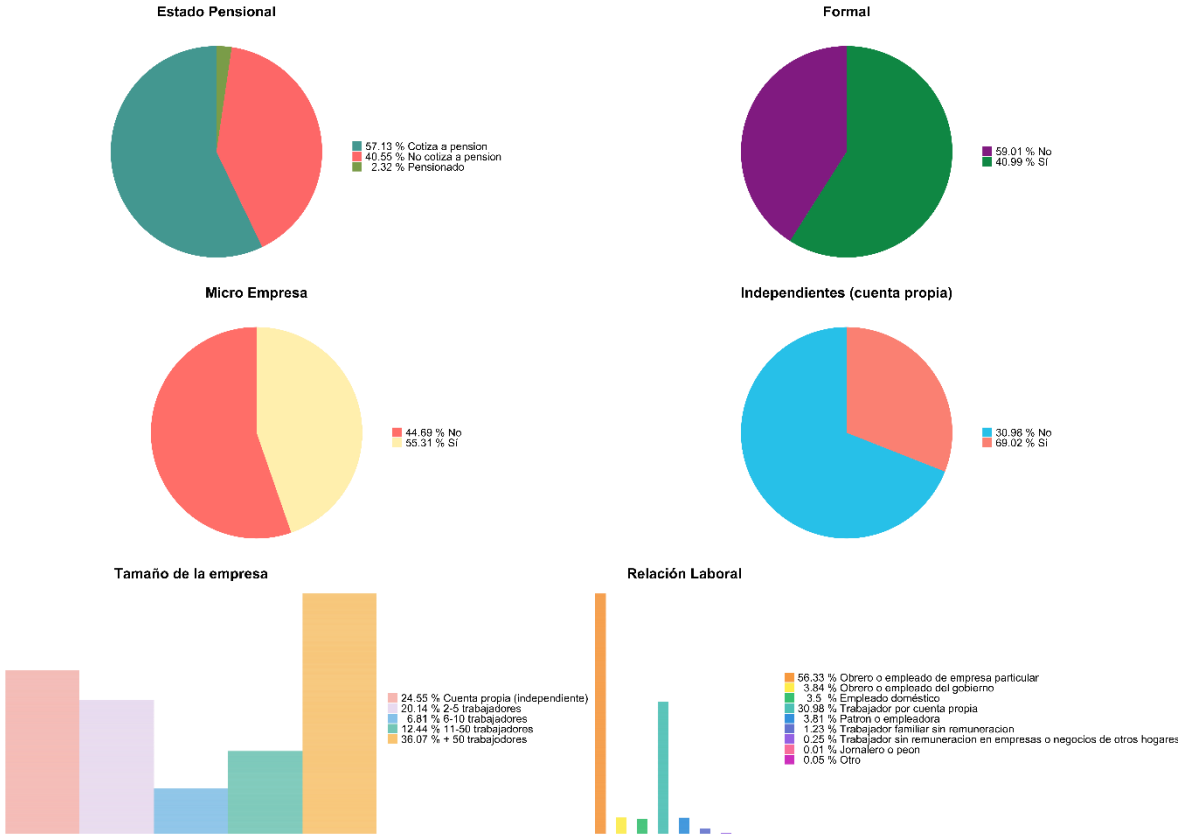
Grafica 3. Densidad poblacional en edad y horas trabajadas.



Mercado laboral

En relación con el mercado laboral, el 57,13 % de la población cotiza a pensión. Además, el 59.1% de la población es formal, el 31% de la población ocupada pertenece al grupo de cuenta propia. Mientras que la mayoría de la población es empleada de una empresa particular (56,33%). Del total de empresas que se encuentran registradas en la ciudad la mayoría son IPMES con más de 50 empleados (36.07%).

Grafica 4. Variables Mercado laboral.



Caracterización Individual

Del total de la muestra las personas que se encuentran en la categoría de ocupados el 52,95 por ciento pertenece al género masculino frente a un 47,05 % que pertenece al género femenino, la mayoría de la población se encuentra concentrada en los estratos 2 y 3; 41,65% y 36,26% respectivamente. La mayor concentración de la población ocupada tiene un nivel de educación máximo superior o universitario.

Frente al sistema pensionario de la ciudad de esta población el 40,55 % de las personas no cotizan a salud frente a un 57,13 que si lo hace. Por otro lado, el 74,79% de las personas ocupadas de la muestra se encuentra en el régimen contributivo de salud, y el 14,16% en el régimen subsidiado. En esta variable también nos encontramos con un buen porcentaje de missing values (8,5%), los cuales fueron rellenados utilizando el mismo algoritmo que para las variables de ingreso (pero en este caso en vez de la media condicional, el valor más frecuente).

Grafica 5. Características individuales

Parte 2 Age-earnings profile

Para lograr establecer el perfil de ingreso por edad es clave tener en cuenta la variable a la cual se va a hacer referencia al hablar de ingreso.

Teniendo en cuenta la información de la tabla 1, la Grafica 2 y la sección de mercado laboral se seleccionó la variable ln_y como income de interés. Esta variable hace referencia a la normalización de la variable ingreso_total_m que tiene 14,632. Observaciones y un máximo y mínimo en 84.000 - 70,000,000.000 Haciendo referencia a el ingreso total de las personas en pesos y aunque muestra la misma forma funcional de las demás variables, esta es la suma de los salarios con los ingresos independiente por lo que tiene no solo da una mayor información acerca del mercado laboral en general, sino que también, muestra una mayor desviación en las observaciones.

Tabla 1. Estadísticas descriptivas variables ingreso

Variable	N	Media	St.	Min	Max
Edad	16,155	1,457,754.000	2,244,763.000	0.000	50,000,000.000
Ingreso de la actividad principal	16,366	19,294.830	219,840.500	0	10,000,000

Ingreso actividad secundaria	9,785	1,566,234.000	2,158,107.000	10	34,000,000
Salario actividad principal	9,785	1,757,076.000	2,413,729.000	30,000.000	60,100,000.000
Salario laboral	4,847	1,362,411.000	2,472,288.000	84.000	70,000,000.000
Salario + ingreso	14,632	1,626,340.000	2,440,279.000	84.000	70,000,000.000
Ganancia total	16,397	47.441	15.616	1	130

Así pues, para ver la influencia de la edad en el ingreso laboral se plantea el siguiente modelo:

Modelo I

$$ln\_y = \beta1 + \beta2Age + \beta3Age2$$

Tras realizar la estimación se encuentra que el coeficiente β_1 es positivo, lo que indica que un año de edad incrementa marginalmente el ingreso de un individuo que se encuentra en la categoría ocupado, de igual manera, se encuentra que este resultado es significativo al 1%. Además, se observa una desviación de 0.03 por lo si bien esta es grande no implica mucho en termino de cambio marginal entre edad. Al revisar el coeficiente β_2, este tiene un signo negativo, indicando que existe una concavidad en la función polinómica de segundo grado, tal que se puede encontrar un punto máximo del ingreso en función de la edad. (tabla 2)

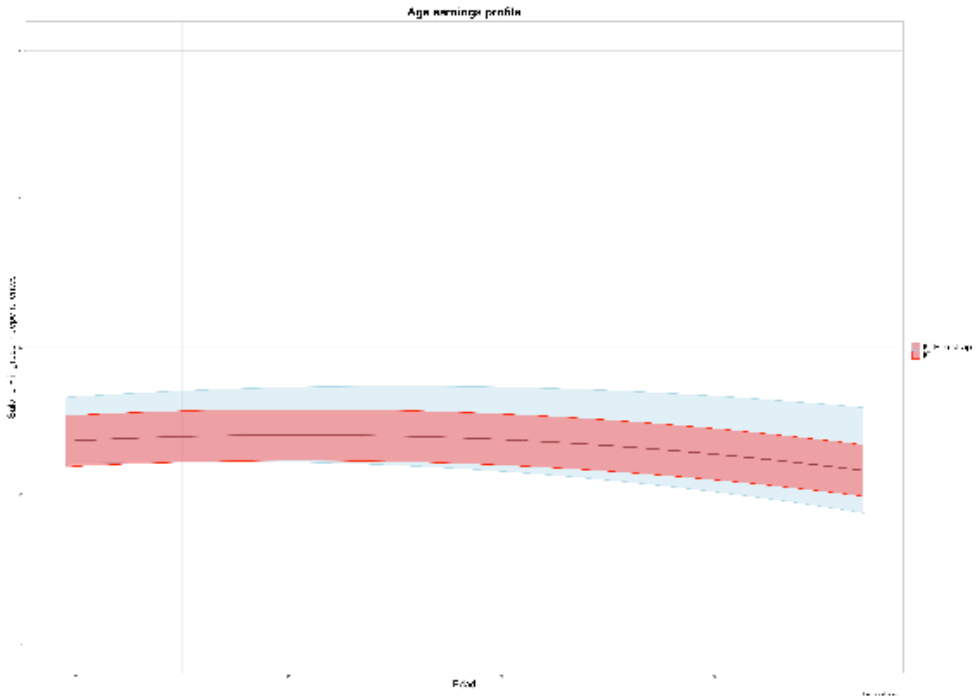
Para analizar β_2 vale la pena hacer la salvedad de que este tiene un mínimo en 19 años, por lo que no es que haya un ingreso base negativo, sino que como la edad comienza en un valor positivo, este intercepto nunca va a tomar el valor del ingreso en una estimación de variable dependiente, sino que va a ser una parte del ajuste del modelo a los datos teniendo en cuenta los otros dos estimadores y el grupo poblacional en estudio (ocupados mayores de edad). (tabla 2)

Tabla 2. Regresión modelo I.

Y	ln_y	*p<0.1; **p<0.05; ***p<0.01
Edad	0.071***	
	(0.003)	
Edad^2	-0.001***	
	(0.00003)	
Constante	12.599***	
	(0.061)	

Uno de los datos más relevantes en cuanto la edad y el ingreso es que la teoría indica que hay un máximo en el ingreso alrededor de los 50 años, con el fin de poder ver esto el nivel de validez externa de los datos se realizó un re muestreo por medio de bootstrap y se encontró, como muestra la gráfica, que en este caso existe un máximo en 48 años en el que e promedio el ingreso sería de $1’600.000. Se puede evidenciar que, con la información de la base de datos y el modelo planteado, con un 95% de confianza, la edad que maximiza el ingreso de los individuos se encuentre en la muestra en la media es de 12.175 - 15.63 (rojo), lo que es aproximadamente entre $200.000 y $3’600.000 pesos mensuales. al realizar el bootstrap para los intervalos de confianza se obtiene que en la media está entre 12.069 - 17.13 (azul).

Tabla 5. Intervalos de confianza y-edad.



Parte 3. The gender earnings GAP

Otro factor clave para analizar dentro del mercado laboral es el GAP que existe entre género y este cómo influye en los ingresos. Así pues, como se mencionó en la sección de descripción de variables, variables individuales, el 52,95 por ciento de la población ocupada pertenece al género masculino. Por lo tanto, para poder estimar esto se planteó el modelo:

$$Ln - y \ = \beta_1 + \beta_2 Femele + u$$

Tras su estimación se encontró que el ser de género femenino disminuye el salario marginalmente en 0,243 puntos esto pues el β_2 presenta un coeficiente positivo. Este valor además muestra tener una significancia al 1 %. (tabla 3)

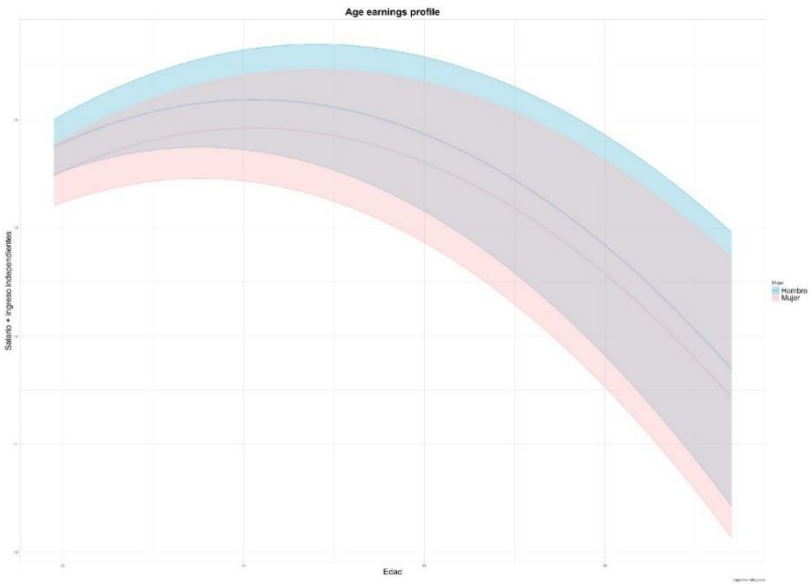
Tabla 3. Regresión modelo 2.

Y	ln_y
Mujer	-0.243***

	(0.014)
Constante	14.016**
	(0.010)
*p<0.1;	**p<0.05; ***p<0.01

Al abordar con mayor profundidad en el modelo se observa que los máximos de edad para mujer son más pronto a lo de los hombres. Sin embargo, el intervalo de confianza para el género masculino es más pequeño lo que nos da el indicio de la existencia de mayor concentración de la población en términos de ingreso. Así pues, con una confianza del 95% se establece que el intervalo que corresponde al género masculino es de 12.8 - 14.2 (azul) frente al femenino que es de 12.6 - 14.0 (rojo). Teniendo un gap de 0,3 puntos entre ellos.

Grafica 6. Intervalos de confianza y-edad por género.



Comúnmente se habla de que, en igualdad de trabajo, hay igualdad de salario sin importar el género o características externas. Para comprobar esto se llevó a cabo la estimación del modelo

$$\ln(y) = \beta_1 Mujer + \alpha X$$

donde X es un vector con las siguientes variables: Edad, Estrato, maxEducLevel, oficio, Edad*maxEducLevel, Estrato*Edad

En el cual se busca observar la veracidad de la afirmación. De esta estimación se obtiene que factores como el estrato, la educación y el nivel de education sí tienen una influencia en el nivel de ingreso de las personas, así como las diferentes interacciones posibles entre ellas como edad y nivel de educación o estrato y edad. El género no es la excepción a ello. Pues la estimación arroja que existe diferencia en el ingreso ligado al género, es decir, aun teniendo en cuenta dos trabajos completamente

iguales existe una diferencia negativa 0.19 con respecto al género masculino. Si bien uno puede pensar en atribuir este resultado a la forma funcional el modelo, al realizar las "limpiezas" del mismo tanto por FWL como por boot se encuentra que el coeficiente de estimación se mantiene exactamente igual. (Tabla 4)

Tabla 4. Modelo controles, FWL y Boots

Modelo 2 con controles	**Long**	**FWL**	**Bootstrap**
Mujer	-0.19 ***	-0.19***	-0.19***
	(0.013)	(0.012)	(0.004)

Lo anterior muestra como en la práctica efectivamente existe una diferencia salarial con respecto al género, lo cual no se le puede atribuir a ningún otro factor diferencial. Además, se observa que este comportamiento del mercado laboral se mantiene a lo largo de todo el rango de edad de las personas ocupadas y la diferencia es más fuerte en los inicios de la vida laboral.

Parte 4 Predicting earnings

a) Se lleva a cabo una aleatorización para poder hacer la partición de la base de datos en dos, el 70% se vuelve un nuevo data frame denominado train, con el que se entrenan los modelos (se ajustan a estos datos) y otro con el 30% de la información denominado test, en el que se testean fuera de muestra los modelos de predicción.

b) Al explorar nuevos modelos para explicar la variación en el ingreso y tener en cuenta los manejados hasta el momento, se utilizan en total los siguientes modelos.

$$lny = \beta_1 mujer + u_1 \ (modelo\ 4.1)$$

En este primer modelo se utiliza el sexo para estimar una relación lineal con el logaritmo del ingreso, cabe decir que este modelo no cuenta con un intercepto y la variable es dicótoma, así que busca expresar las diferencias en ingreso por vía únicamente del sexo.

$$lny = \beta_0 + \beta_1 age + \beta_2 age^2 + u_2 \ (modelo\ 4.2)$$

El segundo modelo utiliza únicamente la edad como fuente de variación, pero aquí se encuentra una función cuadrática, siendo consistente con el punto dos del problem set, donde hay una relación tal que se encuentra una edad que maximiza el ingreso.

$$lny = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 mujer + u_3 \ (modelo\ 4.3)$$

Este tercer modelo contempla los efectos de los dos primeros modelos.

$$lny = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 mujer + \beta_4 lneduc + u_4 \ \ (modelo\ 4.4)$$

El cuarto modelo cuenta con las variaciones de los primeros tres modelos, pero también se le implica el logaritmo del máximo nivel de educación alcanzada, esta variable cobra sentido al entender que el nivel educativo puede tener rendimientos marginales decrecientes, pues mayor educación puede incrementar el ingreso, pero una vez se alcanza un nivel de profesionalidad y experiencia puede que seguir estudiando no implique un ingreso más alto.

$$lny = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 mujer + \beta_4 lneduc + \beta_5 estrato1 + u_5 \ \ (modelo\ 4.5)$$

El quinto modelo toma elementos de los otros modelos, pero le suma el efecto del estrato socioeconómico, pues el ingreso se puede relacionar con el lugar en el que las personas viven, lo cual tiene una dependencia con el estrato.

$$lny = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 mujer\_age + \beta_4 mujer\_ed + \beta_5 regsalud + u_6 \ \ (modelo\ 4.6)$$

El sexto modelo implica la interacción de mujer con edad y la educación .

$$lny = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 mujer\_age + \beta_4 mujer\_ed + \beta_5 regsalud + u_6 \ \ (modelo\ 4.7)$$

El septimo modelo ya no toma el anterior como su base, sino que decide tomar el modelo tres como base y agregar la interacción de la dummy mujer con la edad y con la educación, ya que se considera que estas variables pueden afectar de manera distinta a los dos sexos ya que hay disparidades en términos de edad reproductiva, pensión, entre otras. Adicionalmente, se involucra el efecto del régimen de salud, pues este tiene una relación con el nivel adquisitivo de las personas.

$$lny = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 mujer + \beta_4 totalHoursWorked + \beta_5 totalHoursWorked^2 + \beta_6 oficio + \beta_7 formal + u_7 \ \ (modelo\ 4.8)$$

Este modelo implica la edad y el sexo, pero también el total de horas trabajadas por el individuo en su forma lineal y cuadrática, pues las horas laboradas pueden determinar el salario, pero puede haber un punto máximo dadas las labores y condiciones de contratación que se dan en Colombia, que pueden implicar que un número muy alto de horas laboradas es señal de un pago por hora bajo. Finalmente, también se encuentran las variables de área de oficio y formalidad porque el ingreso salarial puede variar fuertemente dependiendo de donde se desempeñe y el tipo de contratación que se tenga.

Al correr estos modelos con la muestra train y evaluar las predicciones con la muestra test se calculan los errores cuadráticos medios, esto, con el fin de utilizar esta métrica como el criterio para definir cuál es el modelo más preciso, dando como los menores errores de predicción a los modelos cinco y siete, ilustrado a continuación.

Tabla 5 . Errores cuadráticos medios

mse_41= 102.613702590387
mse_42= 0.772779761051467
mse_43= 0.754496014559723
mse_44= 0.642962693705686
mse_45= 0.514189296052766
mse_46= 0.638341357035027
mse_47= 0.609238780848605
mse_48= 0.67770235147024
mse_49= 0.397001651273943

c)

Tabla 6 . Influencia de las observaciones sobre los coeficientes del modelo

Statistic	N	Mean	St.	Dev.	Min
Mujer	4,919	0.000	0.0003	-0.002	0.002
Hombre	4,919	0.000	0.0002	-0.001	0.001

Tabla 7 . Influencia de las observaciones sobre los coeficientes del modelo

Statistic	N	Mean	St.	Dev.	Min
Intercepto	4,919	0.00000	0.002	-0.015	0.035
Edad	4,919	-0.00000	0.0001	-0.002	0.001
Edad2	4,919	0.000	0.00000	-0.00001	0.00002

Tabla 8. Influencia de las observaciones sobre la predicción

Modelo	N	Mean	St. Dev.	Min	Max
Modelo 4.1	4,919	-0.000	0.001	-0.010	0.008
Modelo 4.2	4,919	0.00003	0.028	-0.260	0.423
Modelo 4.3	4,919	0.00005	0.031	-0.265	0.414
Modelo 4.4	4,919	0.001	0.035	-0.277	0.416
Modelo 4.5	4,919	-0.0001	0.060	-0.549	0.647
Modelo 4.6	4,919	0.001	0.035	-0.288	0.398
Modelo 4.7	4,919	-0.00001	0.052	-0.611	0.382
Modelo 4.8	4,919	0.0003	0.042	-0.350	0.804

d)

Al realizar la validación de las estimaciones por el método de Leave-one-outcross-validation se obtuvieron los siguientes resultados de las predicciones para los dos modelos con mejor ajuste en el literal b.

Modelo 4.5:

RMSE	**R^2**	**MAE**	**Predicción promedio**
0.7118	0.3712	0.5009	13.9015

Modelo 4.6:

RMSE	**R^2**	**MAE**	**Predicción promedio**
0. 7801	0. 2447	0. 5491	13. 9014

Se observa que los resultados de las dos estimaciones tienen una predicción promedio muy cercana, por lo que se puede pensar que ambos modelos tienen un buen criterio en cuanto a la predicción, sin embargo, se puede ver que en términos de ajuste cuadrático el modelo más sencillo tiene un mayor R^2 y un menor RMSE y MAE, por lo que a la hora de elegir un modelo estas características pueden resultar relevantes.