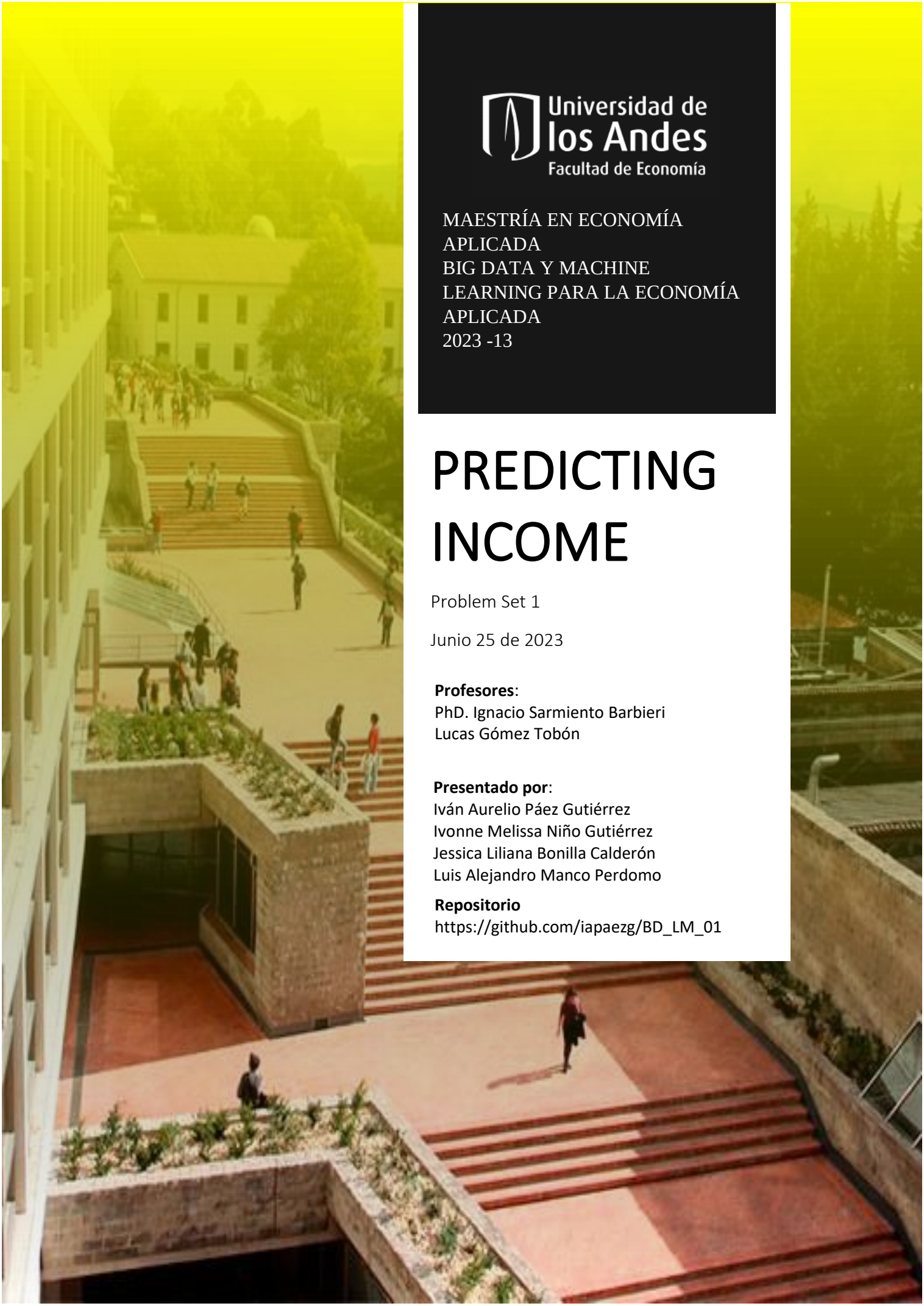


Universidad de los Andes
Facultad de Economía

MAESTRÍA EN ECONOMÍA APLICADA
BIG DATA Y MACHINE LEARNING PARA LA ECONOMÍA APLICADA
2023 -13

PREDICTING INCOME

Problem Set 1

Junio 25 de 2023

Profesores:
PhD. Ignacio Sarmiento Barbieri
Lucas Gómez Tobón

Presentado por:
Iván Aurelio Páez Gutiérrez
Ivonne Melissa Niño Gutiérrez
Jessica Liliana Bonilla Calderón
Luis Alejandro Manco Perdomo

Repositorio
https://github.com/iapaezg/BD_LM_01

Wage prediction for the analysis of tax fraud and labor inequality

Páez Iván[1], Niño Melissa[2], Bonilla Jessica[3], Manco Luis[4]

Facultad de Economía de la Universidad de los Andes

Palabras clave:
Salario
Impuestos
Género
Política pública
Modelo predictivo

I. Introducción

El fraude fiscal, entendido como la evasión al pago de impuestos, es un tema ampliamente estudiado por sus implicaciones en bienestar social, confianza de mercados y desarrollo económico. La disminución del recaudo de dineros públicos puede comprometer el desarrollo de programas y proyectos encaminados a aumentar el bienestar social, el desarrollo empresarial y al pago de deuda nacional. Implicando entonces que los sistemas fiscales nacionales deben anticipar disminuciones en este recaudo, para generar estrategias de captura efectiva de recursos, que garanticen el flujo deseado de estos propendiendo por aspectos como la distribución equitativa de las cargas impositivas (Babilla, 2023).

Diferentes estudios han empleado modelos econométricos para explicar la variación del fraude en sistemas fiscales, estudiando determinantes tales como cargas impositivas, aplicación de subsidios o periodicidad de recaudos. Recientemente se han desarrollado nuevas metodologías permitiendo analizar también determinantes menos técnicas y más comportamentales, como el nivel de confianza de los contribuyentes sobre la administración de los recursos públicos (Haaland & Olden, 2022), sugiriendo que el fraude va más allá de la operatividad de los sistemas fiscales.

Con la llegada de nuevas fuentes de datos y mejora de las capacidades computacionales, el análisis del fraude se ha beneficiado con nuevos abordajes implementando Big Data. Yi et al (2023) emplearon datos del mercado de capitales estadounidense para desarrollar un marco de detección de fraude empleando *machine learning*, sobre la información de 146.045 empresas, reduciendo los costos y tiempos de validación de fraude. Mao et al (2022) desarrollaron una red neuronal gráfica (heterogénea) para conectar las empresas chinas que cotizan en bolsa y sus empresas relacionadas para validar que a través de estas no hagan fraude. Vanhoeyveld et al (2020) considerando que tradicionalmente la detección de fraude fiscal se caracteriza por muy pocos datos etiquetados (fraude conocido/casos legales) que no son representativos de la población debido al sesgo de selección de la muestra, utilizaron técnicas de detección de anomalías no supervisadas para comparar compañías similares y determinar

1 Biólogo. Universidad Nacional de Colombia. ia.paez10@uniandes.edu.co
2 Abogada. Universidad del Rosario. i.nino@uniandes.edu.co
3 Administradora Ambiental. Fundación Universitaria Empresarial de la Cámara de Comercio de Bogotá jl.bonillac1@uniandes.edu.co
4 Ingeniero Ambiental. Universidad Distrital Francisco José de Caldas. l.manco@uniandes.edu.co

desviaciones financieras que sugirieran fraude, usando la información fiscal del 2014 de empresas belgas.

Considerando estos avances en materia de detección del fraude en firmas, en este documento se realiza una aproximación a un modelo de predicción de sueldos de personas naturales en Colombia, que permita la identificación rápida de personas con un reporte inferior de ingresos, lo cual es uno de los principales mecanismos de evasión en el país, al responder a una cultura tributaria en la cual los sujetos tienen una aversión natural a desprenderse de parte de sus ingresos porque estos no generarían aumento en su bienestar (Hoyos Londoño, 2020). Este tipo de evasión representó una cifra cercana al 0,7% del PIB nacional (La República, 2022) la cual similar a la cartera de ciencia y tecnología del 2022 (Lopez, 2022).

Considerando el abordaje propuesto, el análisis de datos requerido para el desarrollo de este modelo, permite evidenciar desigualdades salariales entre hombres y mujeres en el país, siendo este un aporte a la literatura para la identificación de perfil de edad por ingresos y brecha salarial en economías en desarrollo partiendo de la ecuación minceriana clásica (Mittag 2019; Demirgüç-Kunt et al. 2021; Männasoo 2022; Groisman 2014; Ahmed et al 2014). Esto que podría redundar en análisis como el de Cremer & Roeder (2019) en el cual a partir de la identificación de brecha salarial se reconoce que la tributación de renta de mujeres casadas no refleja su verdadera productividad implicando una distorsión en la distribución de cargas en el sistema fiscal.

Este estudio se desarrolló con 19.529 registros de datos de personas, obtenidos después de la limpieza de la base de datos de la Encuesta General de Hogares del 2018 realizada por el Departamento Nacional de Estadística (DANE). La base de datos filtrada y limpiada contiene las variables de estrato, género, edad, oficio, educación terciaria, máximo nivel educativo, cotización a pensión, empleado, formalidad del empleo, población económicamente activa, horas totales trabajadas en la última semana, ingresos laborales, ingresos no laborales, posición en el hogar, relación laboral.

II. Datos

La base de datos usada corresponde a la Encuesta General de Hogares (GEIH) realizada por el DANE en el año 2018 que contiene la información de ingresos por hogares y unidad de gasto para 32.177 individuos pertenecientes a 10.567 hogares de 10.403 viviendas localizadas en la ciudad de la ciudad de Bogotá, sin discriminar por rango de edad.

Se utilizaron técnicas de web scraping para obtener los datos de la GEIH del sitio web: https://ignaciomsarmiento.github.io/GEIH2018_sample/, página dinámica que divide la información en 10 "chunks" o fragmentos de datos. En la *Uniform Resource Locator* (URL) original no se encuentra la información requerida. Una vez se carga la página por completo, se puede localizar el enlace específico de los datos a extraer. El web scraping se llevó a cabo utilizando el paquete "rvest" de R. Se identificó un patrón en la estructura de la URL de cada fragmento de datos, con lo cual se importó la información mediante un bucle uniendo los 10 fragmentos en un único *data frame*.

Considerando que en la evaluación de la brecha salarial por género es necesario revisar el impacto de las labores de cuidado en las horas trabajadas por mujeres, se creó una variable *dummy* para todos los individuos que son jefes de hogar y sus parejas en la que asignó el valor 1 si tienen hijos y 0 en caso contrario. Esta variable fue denominada *tiene_hijos.* Adicionalmente, se creó una variable que recoge la cantidad de hijos por hogar y, se asignaron por pareja.

Posteriormente, con el objetivo de estudiar los ingresos totales (ingresos laborales y no laborales observados e imputados en la muestra original) por individuo se seleccionaron aquellos individuos empleados, desocupados e inactivos con edad mayor o igual a 18 años, después de esto, las observaciones resultantes fueron 24.568 individuos. Seguidamente, se identificaron las variables de

interés para determinar su relación con el ingreso: *estrato, género, edad, oficio, educación terciaria, máximo nivel educativo, cotización a pensión, empleado, formalidad del empleo, población económicamente activa, horas totales trabajadas en la última semana, ingresos laborales, ingresos no laborales, posición en el hogar, relación laboral.*

Después del análisis de estas variables, se eliminaron las siguientes: *educación terciaria* debido a que la variable *máximo nivel educativo* provee un mayor nivel de detalle acerca de la educación del individuo y, *cotización a pensión* ya que ésta es que sirve de base para determinar si el empleo es formal o no, por lo que la información se subsume en la variable *formalidad del empleo,* que se mantiene. Por último, se eliminaron las variables de ingresos discriminados (laborales y no laborales) ya que la variable ingresos totales consolida esta información.

Se remplazaron los datos faltantes en las variables oficio y formalidad por cero y se mantuvieron. Las observaciones de individuos con ingresos iguales a 0 fueron eliminados de la base de datos, también fue eliminado un individuo que reportaba N.A. en la variable *máximo nivel educativo* toda vez que el objetivo es predecir el ingreso en función de las demás variables de interés. A su vez, se identificaron y eliminaron los *outliers* definidos como aquellas observaciones alejadas en tres desviaciones estándar de la media, después de lo cual se obtuvieron 19.529 observaciones finales. Luego se creó la variable *ingreso por hora* que fue calculada como el ingreso total dividido por la cantidad de semanas promedio en el mes (4.28 semanas), para validar estos resultados se calculó la misma variable tomando la cantidad total de horas trabajadas que el individuo reportó en la encuesta. Sin embargo, se mantuvo el primer cálculo.

Para efectos de estimar la ecuación minceriana clásica, se generó la variable de experiencia potencial, se incluyó dentro de las variables el último grado de educación alcanzado. Con esto se calculó la experiencia potencial, tomando la edad del individuo menos los 5 años que se asumen como los primeros años previos a la escolaridad y los años totales de educación. En algunos estudios se usa la variable edad como un proxy de experiencia, por lo anterior, con el objeto de mitigar los efectos de correlación entre las variables edad y experiencia se hace uso de la experiencia potencial.

En la siguiente Tabla 1 se describen las variables de interés

Tabla 1. Variables de interés

Variable	Descripción
Estrato	Tomado de la base del estrato socioeconómico del servicio de energía. 1 representa el estrato socioeconómico más bajo y el 6 el más alto
Edad	Edad en años
Género	1 = Hombre, 0 = Mujer
Oficio	Categoría de oficio por ocupación de individuo. Ver descripción en https://ignaciomsarmiento.github.io/GEIH2018_sample/labels.html
Máximo nivel educativo	1= ninguno, 2 = prescolar, 3 = primaria incompleta (hasta 4°), 4 = primaria completa, 5 = secundaria incompleta (hasta 10°), 6 = secundaria completa, 7 = terciaria
Educación	Último grado de educación alcanzado expresado en años.
Empleado	1 = Desempleado, 0 = Empleado
Formalidad del empleo	1 = Si cotiza a pensión, 0 = Si no cotiza a pensión
Población económicamente activa	1 = Si está ocupado o en búsqueda de empleo 0 = Si no
Log ingreso por hora	Log del ingreso total dividido por la cantidad de semanas promedio en el mes (4.28 semanas)
Ingresos totales	Ingreso total imputado y observado de fuente laboral y no laboral del mes
Tiene_hijos	1 = Si tiene hijos, 0 = No tiene hijos

Variable	Descripción
Hijos	Número de hijos para el jefe de hogar y su pareja
Experiencia potencial	Edad en años - (5 años + años de educación)
Relación laboral	1 = Obrero o empleado de empresa particular, 2 = Obrero o empleado del gobierno, 3 = Empleado doméstico, 4 = Trabajador por cuenta propia, 5 = Patrón o empleador, 6 = Trabajador familiar sin remuneración, 7 = Trabajador sin remuneración en empresas o negocios de otros hogares, 8 = Jornalero o peón, 9 = Otro
Posición en el hogar	Jefe, cónyuge, hijos, otros

Las variables expuestas recogen el componente de capital humano en la productividad, particularmente en la literatura se ha afirmado reiteradamente que existe una relación entre variables como el nivel educativo, la educación, la ocupación, el sexo, el sector de actividad, la zona de residencia, la experiencia que explican la heterogeneidad en salarios. Bajo esta perspectiva, de acuerdo con Schultz la decisión de inversión en capital humano, como lo es la inversión en educación, está asociada a la concepción de que individuos más escolarizados son más productivos y por ello, devengan mayores salarios. Sin embargo, este postulado fue modulado por Bourdieu, quien indicó que dicho efecto de la educación sobre el salario depende de la no existencia de desigualdades previas marcadas por diferencias de género, clase y etnia (Araujo, 2015).

Ahora bien, Guataquí et altri (2018) encontraron que las variables definitorias del capital humano presentan una distribución diferente para asalariados y trabajadores por cuenta propia. En el caso de los asalariados las diferencias en el ingreso provienen mayoritariamente de características educativas, en comparación con los trabajadores por cuenta propia. Las variables de edad, experiencia e ingresos por hora y totales muestran efectos no lineales sobre el salario, lo cual se tuvo en cuenta en la definición de la forma funcional del modelo. La edad, experiencia y sus valores al cuadrado corresponden con el ciclo de vida al ingreso: presenta tasas crecientes en los primeros años y decrece a partir de cierta edad.

a. Análisis descriptivo de los datos

Considerando los datos de la base de datos limpia, se tienen las gráficas de estadísticas descripticas de la Figura 1. (a) El 50% de los 19.529 individuos de la muestra analizada se encuentran entre los 30 y 50 años con una mediana de 40 años y media de 43 años. (b) La educación de los individuos se determina a partir del último grado de educación alcanzado expresado en años. Según los datos proporcionados, la media de esta variable es de 6.63 años. Esto significa que el promedio de los encuestados ha finalizado la educación secundaria.

Por su parte, (c) la variable de *máximo nivel educativo* representa la categoría máxima de educación alcanzada medida en niveles de educación, el 40% de los encuestados cuentan con educación terciaria, y el 30% cuentan con educación superior o universitaria, categorías con mayor frecuencia. (d) Por otra parte, se puede observar que el promedio de los individuos se sitúa en el estrato 2, seguido del estrato 3, 1, 6 y 5.

(e) La experiencia potencial mide el número de años de experiencia de los individuos. De acuerdo con los datos analizados, se observa que el 50% de los individuos tiene una experiencia por debajo de los 28 años. Además, las observaciones se centran entre 17 y 43 años con una media de 31 años. (f) La variable ingreso por hora se expresa en su forma logarítmica para facilitar el análisis e interpretación de los datos y representa los ingresos en pesos colombianos por hora; se observa que la media de esta variable es 8.57, equivalente a 7.915 pesos por hora en la escala original. Además, se identificó que el ingreso mínimo por hora es de 277 pesos por hora, mientras que el máximo es de 95.729 pesos por hora.

Del total de la muestra (19.529), el 51% corresponde a 9.937 hombres y el 49% restante equivale a 9.592 mujeres. Lo anterior, evidencia que los hombres tuvieron mayor participación en comparación con las mujeres en la muestra analizada

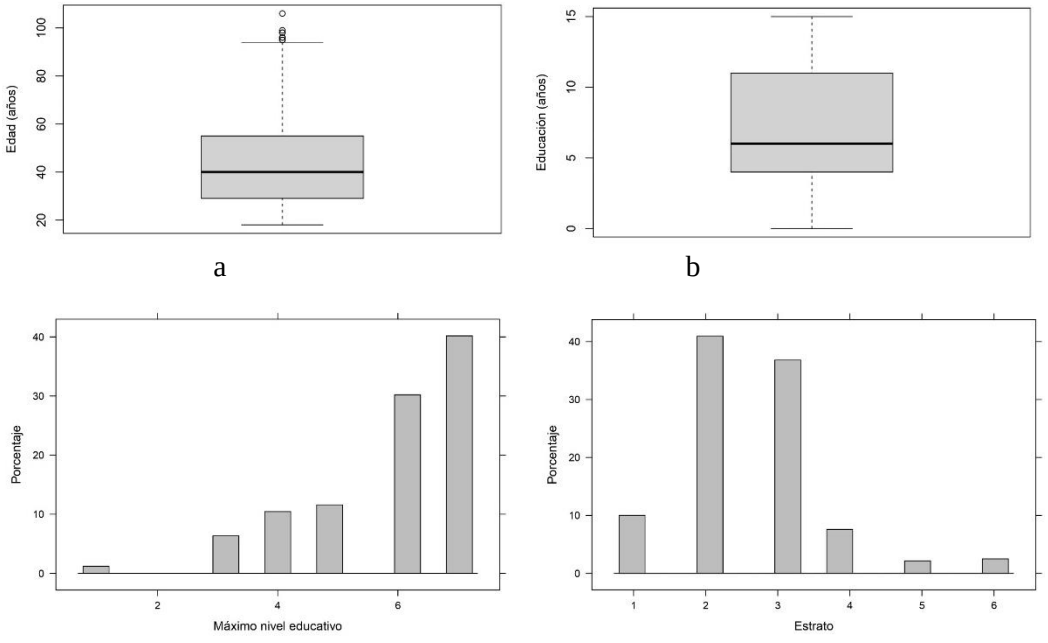
(g) El número de hijos para el jefe de hogar y su pareja puede variar desde 0 hasta 8. Los valores en cero representan los casos en los que los individuos no tienen hijos. A partir del histograma, se observa que más del 53% de los individuos no tiene hijos, seguidos por aquellos que tienen 1 hijo, 2 hijos, 3, 4 y 8 hijos. (h) Por otro lado, se analiza la distribución detallada del número de hijos según el género. Se observa que la mayoría de los hombres y mujeres no tienen hijos. De las 9.592 mujeres encuestadas el 50% no tienen hijos, el 26% tienen un hijo, 18% tiene dos hijos y el 6% restante tienen entre 3 y 8 hijos (ver histograma de la izquierda, 0= Mujeres). De los 9.937 hombres encuestados el 56% no tienen hijos, el 20% tienen un hijo, 18% tiene dos hijos y el 6% restante tienen entre 3 y 7 hijos (ver histograma 1= Hombres).

(i) El tipo de relación laboral que mayoritariamente reportan los encuestados corresponde a Obrero o empleado de empresa particular (categoría 1 superior al 40%), seguido por Trabajador por cuenta propia (4) que representa más del 20% de los individuos de la muestra.

Figura 1. Análisis descriptivo de la base de datos

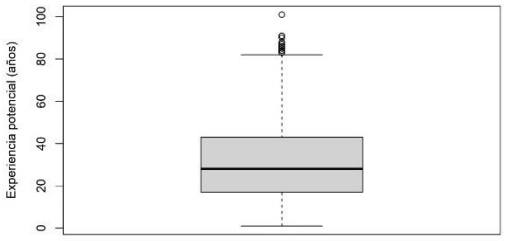
```
── Variable type: factor ──────────────────────────────────────────────────────
  skim_variable n_missing complete_rate ordered n_unique top_counts
1 estrato1              0             1 FALSE          6 2: 7994, 3: 7192, 1: 1956, 4: 1482
2 relab                 0             1 FALSE         10 1: 9293, 4: 4998, 0: 3376, 2: 630
3 t_hijo                0             1 FALSE          2 0: 10297, 1: 9232

── Variable type: numeric ─────────────────────────────────────────────────────
  skim_variable n_missing complete_rate   mean     sd    p0   p25   p50   p75  p100 hist
1 sex                   0             1  0.509  0.500  0     0     1     1     1   ▇▁▁▁▇
2 age                   0             1 42.9   16.7   18    29    40    55    106 ▇▇▆▃▁
3 educ                  0             1  6.63   3.47   0     4     6    11    15  ▅▇▃▆▁
4 exp                   0             1 31.3   17.3    1    17    28    43    101 ▇▇▅▂▁
5 ln_income             0             1  8.57   0.893  5.63  8.24  8.50  9.02  11.5 ▁▂▇▂▁
```

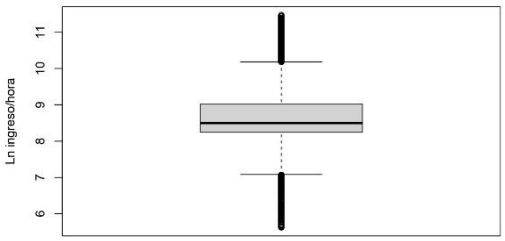


a b

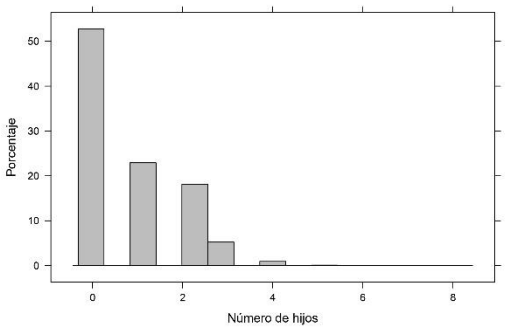
c d



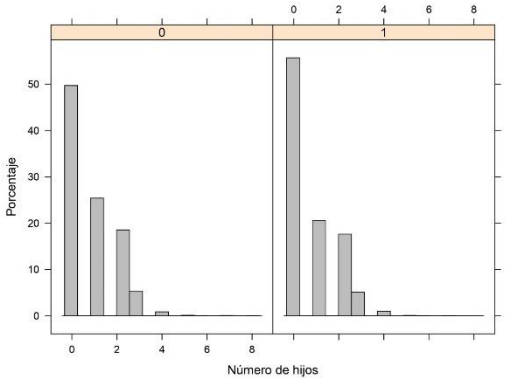
e



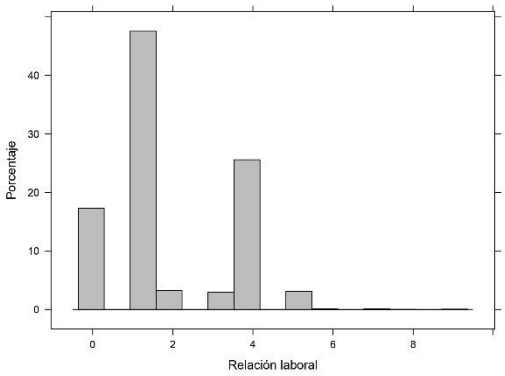
f



g



h



i

III. Revisión de literatura

a. Perfil salarial por edad

La ecuación minceriana propone un modelo para analizar cómo la inversión en capital humano determina la relación entre ingreso y edad de un individuo a lo largo de su vida, a través de la cual se busca obtener el nivel óptimo de escolaridad y la tasa de retorno de la educación. Este punto óptimo de escolaridad se obtiene igualando el costo de oportunidad de un año de escolaridad adicional con el valor descontado del ingreso marginal obtenido del incremento de escolaridad (Rojas *et* altri, 2.000).

De acuerdo con Rojas *et* altri (2.000) la educación desplaza la curva de salario y edad, sin alterar su pendiente. La relación entre salario y edad está soportada en que la experiencia del individuo es creciente, su estado de salud y la cercanía a la edad de retiro, por lo anterior, la edad determina el ciclo de vida del salario percibido, en el que el salario incrementa con la edad hasta cierta edad a partir de la cual empieza a reducirse.

b. Brecha salarial por género

La brecha salarial por género debe ser explicada por factores de educación, escolaridad y por aquellos que discriminatorios que se asocian a mecanismos de discriminación basados en el género (Araujo, 2015), producto de los roles reproductivos y al cumplimiento de obligaciones de cuidado de las mujeres que reducen sus opciones ocupacionales y disminuyen su tiempo y esfuerzo dedicado al trabajo remunerado y a la acumulación de capital humano (Tyrowicz *et* altri, 2018) que se traduce en menores recompensas o aumentos salariales asociados a la experiencia.

Cremer & Roeder (2018) señalan que las mujeres que han interrumpido su carrera por completo o se han dedicado al cuidado infantil a tiempo parcial, sufren una sanción salarial que persiste décadas después cuando comienzan a cuidar de sus padres como consecuencia de "*la pena de la buena hija*", que conlleva a una brecha entre los salarios y las verdaderas productividades de hombres y mujeres. Con fundamento en lo anterior, estos autores proponen que el impuesto general sobre la renta o el ingreso debe diseñarse para neutralizar las distorsiones derivadas de la brecha salarial de género.

c. Perfil salarial por edad y género:

Como lo menciona Tyrowicz et altri (2018), algunas teorías explican que las diferencias en la acumulación de capital humano para hombres y mujeres ajustadas a la edad presentan una forma de U invertida, mientras que las teorías sociales explican que la brecha va incrementando a medida que la edad aumenta asociado a fenómenos como interrupciones en la trayectoria laboral y tiempo dedicado al trabajo doméstico no remunerado. Sin embargo, los autores llaman la atención a que en la muestra estudiada las cohortes subsiguientes de mujeres están gradualmente mejor educadas que los hombres, su fertilidad disminuye o la maternidad se retrasa y el acceso a los centros de atención, lo cual se manifiesta en la tendencia aparente hacia una igualación de la división de mano de obra dentro de los hogares. A pesar de ello, la penalización del hecho de ser mujer en las diferencias de salario continúa aumentando con la edad, incluso después de la etapa reproductiva, lo que les permite concluir que la edad y el género son dos desventajas simultaneas dentro del mercado laboral.

IV. Resultados

a. Perfil salarial por edad

En la Tabla 2 se presenta la regresión lineal de la Ecuación 1.

$$\log Ingreso = \beta_1 + \beta_2\, Edad + \beta_3\, Edad^2 + u \quad \text{(Ecuación 1)}$$

Tabla 2. Perfil salarial por edad

	Variable dependiente:
	Modelo ingreso hora con la edad
Edad	0.0420***
	(0.0020)
Edad^2	-0.0004***
	(0.00002)
Constante	7.7049***
	(0.0437)
Observaciones	19,529
Error estándar residual	0.8825 (df = 19526)

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

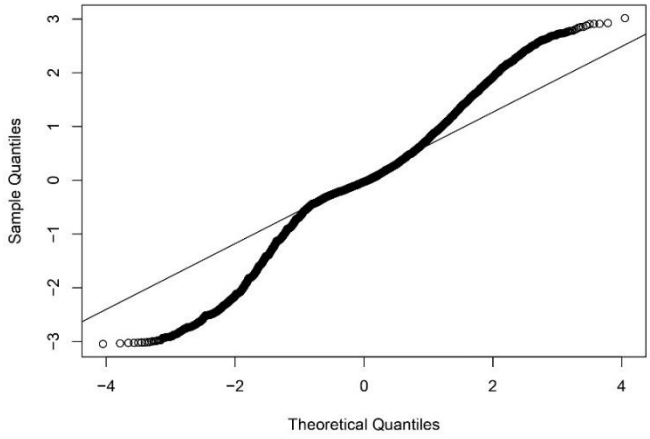
Los coeficientes de la regresión representan el impacto que tienen las variables *edad* y e*dad*^{2} con una significancia del 1% en la variable independiente, que en este caso es el logaritmo de ingreso por hora. El coeficiente de edad indica que, en promedio, cuando un individuo aumenta un año de vida, su ingreso aumenta en 4.2%. Este coeficiente se interpreta como un cambio porcentual en el ingreso por hora, por cada año adicional de edad. Además, se proporciona el error estándar asociado a este coeficiente, que es una medida de la precisión del estimado (0.0020). A su vez, el coeficiente de e*dad*^{2} indica una relación negativa entre la edad al cuadrado y el ingreso por hora. Esto implica que después de cierto punto, la edad comienza a tener un impacto negativo en el ingreso, lo que sugiere que hay un punto máximo en la relación entre la edad y el ingreso. Este coeficiente indica que por cada año de vida del individuo al cuadrado su ingreso se disminuye en 0.04% (error estándar de 0.00002)

La constante no genera ningún análisis más allá de corresponder a la intersección que define la relación entre dos variables (*edad,* e*dad*^{2} *y log(ingreso))*.

Finalmente, el error estándar residual (0.8825) indica que tan bien se ajustan los datos a la recta de la regresión En la figura 2 se puede observar que, aunque los errores no se ajustan normalmente y se presentan valores extremos, gran parte de los datos sí se encuentran cerca de la recta. Por otro lado, se emplea el Q-Q Plot (Quantile-Quantile Plot) en la figura 3 para evaluar si los residuos se ajustan a una distribución normal, con lo cual se detectan valores atípicos, puntos en los extremos que se alejan de la línea diagonal en el Q-Q Plot.

Figura 2. Errores residuales del modelo

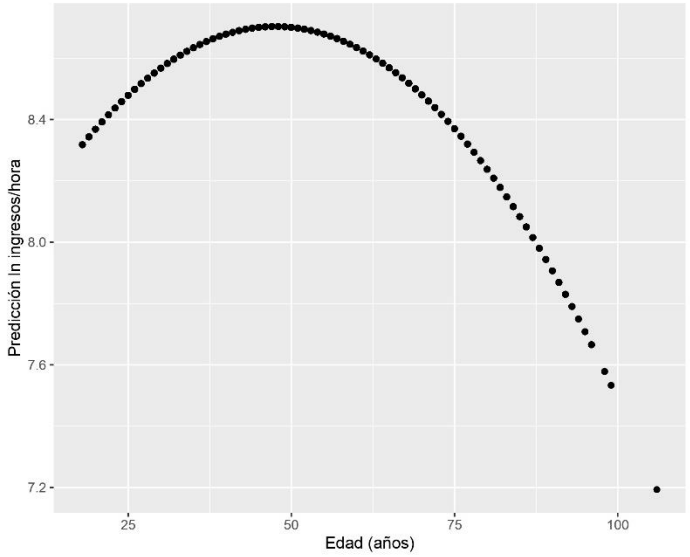
Figura 3. Q-Q Plot de los errores del modelo



Perfil estimado de ingresos por edad e intervalos de confianza

La figura 4 indica que, a medida que la edad de un individuo aumenta, el ingreso marginal es decreciente, lo que implica que, a partir de cierta edad, es menos probable que el individuo experimente aumentos significativos en los ingresos. Por ejemplo, consideremos dos individuos: uno de 20 años y otro de 60 año; es probable que el individuo de 20 años experimente un aumento considerable en sus ingresos a lo largo de su vida hasta cerca de los 50 años, mientras que el segundo, tendrá un crecimiento nulo o incluso decrecimiento en sus ingresos a medida que envejece. Estas conclusiones son consistentes con la literatura revisada, pues concuerdan con que, la base de la relación salario y edad se encuentra en la experiencia del individuo, su estado de salud y la edad de retiro, por lo que, al acercarse a la edad de retiro la relación se torna negativa.

Figura 4. Relación edad-ingresos



Perfil salarial por edad: estimación edad pico

Ahora bien, se construyen los intervalos de confianza para aproximar la distribución del estimador a partir de los datos disponibles usando el método Bootstrap, el cual ayuda a caracterizar la viabilidad de cada una de las variables, es decir, permite obtener estimaciones más precisas de los intervalos de confianza utilizando la semilla 2023 y R=1000, es decir, que se generaron 1.000 muestras Bootstrap a partir de los datos originales, los cuales arrojan las estadísticas de la tabla 3.

Tabla 3. Resultados de bootstrap

	Original	*Varianza*	*Error estándar*	*IC*
t1	7.7049375892	-5.336027e-04	4.345893e-02	0.0851795
t2	0.0420029261	4.283343e-05	2.069112e-03	0.00405546
t3	-0.0004417979	6.671647e-07	- 2.209662e-05	4.330938e-05

Las estadísticas reportadas incluyen el estimador original, la varianza, el error estándar e intervalo de confianza para cada variable. El estimador original es el valor calculado a partir de los datos originales. La varianza representa la dispersión de los valores en las muestras bootstrap, y el error estándar es una medida de la precisión del estimador. Por último, se presentan los intervalos de confianza (IC) para cada variable (*t1*, *t2*, *t3*).

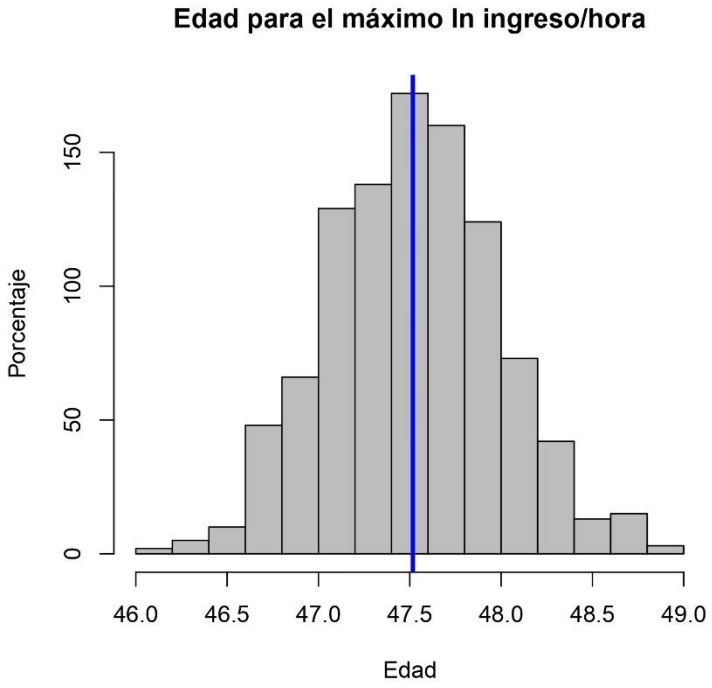
Dado que el objetivo es maximizar la función de la Ecuación 1 se decidió que, dentro de todos los puntos de la distribución, se escogió la media de cada variable para que dentro del bootstrap se pudiera generar la *edad pico*. Finalmente, se calcula con la siguiente fórmula $Edad_{pico} = -\beta_1/(2\beta_2)$ obtenida del siguiente procedimiento

$$\log Ingreso\ x\ hora = \beta_0 + \beta_1\ Edad + \beta_2\ Edad^2 + u$$

$$\frac{d\ Ingreso\ x\ hora}{d\ edad} = \beta_1 + 2\beta_2\ Edad = 0\ (Para\ el\ máximo)$$

$$Edad_{pico} = -\frac{\beta_1}{2\beta_2}$$

El resultado de esta *edad pico* es de 47.52 años aproximadamente, como se representa en la siguiente gráfica. Siguiendo los supuestos de la ecuación minceriana, este es el punto óptimo en el que se logra el mayor salario en función de la edad de los individuos.



b. Brecha salarial por género

A continuación, se presenta la estimación y discusión de la brecha salarial incondicional de género partiendo del siguiente modelo *log*(ingreso)= $\beta_1 + \beta_2$Género+ *u,* en donde *log(ingreso)* es el logaritmo del ingreso por hora, *Género* es una variable dicótoma que toma el valor de 0 si el individuo es mujer y 1 si el individuo es hombre, y *u* son los errores de la regresión.

	Variable dependente:

	Modelo log de ingreso hora con el género
Género	0.2077***
	(0.0127)

Constante	8.4652***
	(0.0091)
Observaciones	19,529
Error estandar residual	0.8872 (df = 19527)
Nota:	*p<0.1; **p<0.05; ***p<0.01

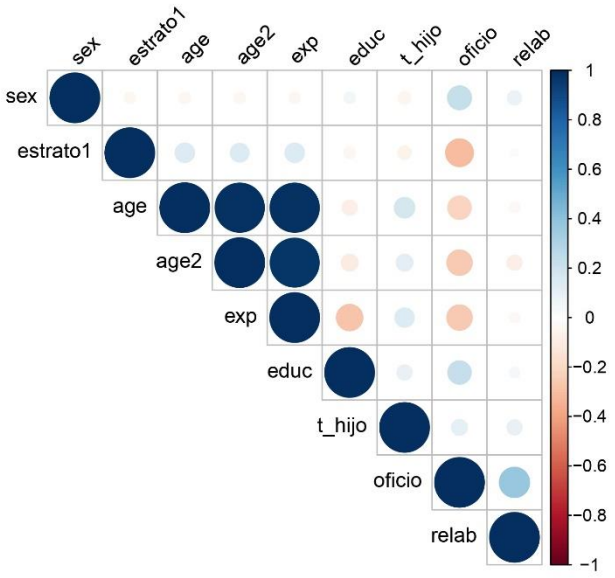
Los coeficientes de la regresión representan el impacto que tiene la variable género con una significancia del 1% en el logaritmo de ingreso por hora. El coeficiente de género indica que, cuando el individuo es hombre el salario aumenta en un 20,7 % en promedio, y cuando el individuo es mujer gana 20,7% menos que el hombre mostrando que las mujeres devengan menores salarios que los hombres en la ciudad de Bogotá para el 2018. Estos resultados son consistentes con la literatura sobre la existencia de una brecha salarial entre hombre y mujeres que responde a criterios discriminatorios asociados al género por los roles reproductivos de las mujeres y las obligaciones de cuidado que deben asumir en el hogar. Sin embargo, para confirmar la conclusión se hace necesario controlar por factores de educación y experiencia. Al presentar el coeficiente una desviación estándar de 0.0127 se puede inferir que tiene una variabilidad moderada.

Ahora, se presenta la estimación y discusión de la brecha salarial incondicional de género incorporando variables de control al siguiente modelo $log(\text{ingreso}) = \beta_1 + \beta_2 \text{Género} + \text{controles} + u$. A través de la incorporación de los controles se corrige el sesgo de selección de la decisión de ingreso al mercado y el relacionado con el oficio.

Después de revisar la literatura pertinente, se determinó que las variables de control son: educación, estrato, edad, edad ^2, experiencia, si tiene hijos o no, oficio y relación laboral.

$$\log(ingreso) = \beta_0 + \beta_1 Género + \beta_2 educ + \beta_3 estrato + \beta_4 edad + \beta_5 edad\text{^}2 + \beta_6 experiencia + \beta_7 tienehijos + \beta_7 oficio + \beta_9 relaciónlab + u$$

Para evaluar posibles eventos de multicolinealidad entre estas variables, se determinó gráficamente el coeficiente de correlación entre las variables usadas como controles, encontrando que la variable edad y experiencia presentan correlación perfecta, lo cual es consistente con el uso de la edad como un *proxy* de la experiencia.



Conservando la totalidad de las variables control antes indicadas, los resultados de la regresión previamente planteada son los siguientes:

	Variable dependiente:

	Modelo ingreso hora con el género
Género	0.1738***
	(0.0114)
Constante	6.9979***
	(0.0424)
Observaciones	19,529
Error estandar residual	0.8872 (df = 19527)
Controles	SI
Nota:	*p<0.1; **p<0.05; ***p<0.01

Cuando se incluyen los controles la brecha salarial entre hombres y mujeres se disminuye, aunque se mantiene una diferencia de 17.38 % producto de la cual, por el hecho de ser mujer, estas devengan 17,38% menos ingresos por hora que los hombres. Esto confirma que la diferencia en ingresos entre hombres y mujeres se debe a factores discriminatorios, por lo que no radica en el sesgo de selección.

Brecha salarial condicional

Teniendo en cuenta que se advirtió la existencia de multicolinealidad en el modelo antes propuesto, se corre nuevamente usando el teorema Frisch-Waugh-Lovell (FWL), con el que se busca eliminar los efectos de la correlación sin afectar los coeficientes de la regresión original.

Usando FWL

	Variable dependente:		
	OLS		Residuos_log_ingresos_controles
	(1)	(2)	(3)
Género	0.2077***	0.1738***	
	(0.0127)	(0.0114)	
Residuos_género_controles			0.1738***
			(0.0114)
Constante	8.4652***	6.9979***	-0.0000
	(0.0091)	(0.0424)	(0.0049)
Control	NO	SI	
Observaciones	19,529	19,529	19,529
Error residual estandar	0.8872 (df = 19527)	0.6831 (df = 19430)	0.6814 (df = 19527)
Nota:			*p<0.1; **p<0.05; ***p<0.01

De la forma en la que se puede observar de la tabla precedente, tras usar FWL los coeficientes y la suma total de los cuadrados de los errores no cambian, sin embargo, si se presenta una ligera variación en el error estándar que se debe al ajuste en los grados de libertad de la regresión.

Usando FWL con boostrap

Ahora bien, se construyen los intervalos de confianza para aproximar la distribución del estimador a partir de los datos disponibles usando el método Bootstrap, el cual ayuda a caracterizar la viabilidad de cada una de las variables, es decir, permite obtener estimaciones más precisas de los intervalos de confianza utilizando la semilla 2023 y R=1000. Después de ello, arrojan las siguientes estadísticas:

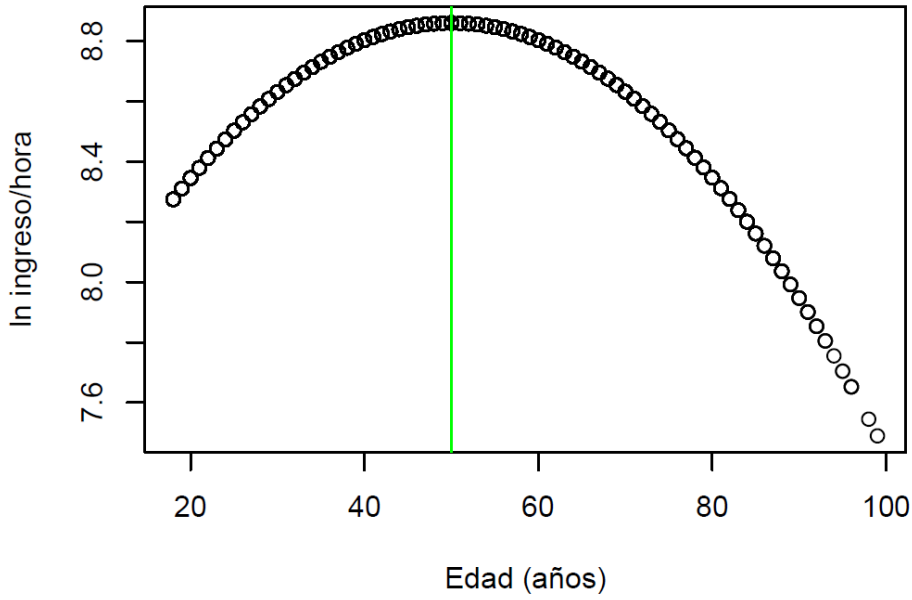
	Original	*Varianza*	*Error estándar*
t1	0.1738334	-0.0004896027	0.011984

Las estadísticas reportadas incluyen el estimador original, la varianza, y el error estándar. La varianza representa la dispersión de los valores en las muestras bootstrap, y el error estándar es una medida de la precisión del estimador para la variable *t1*.

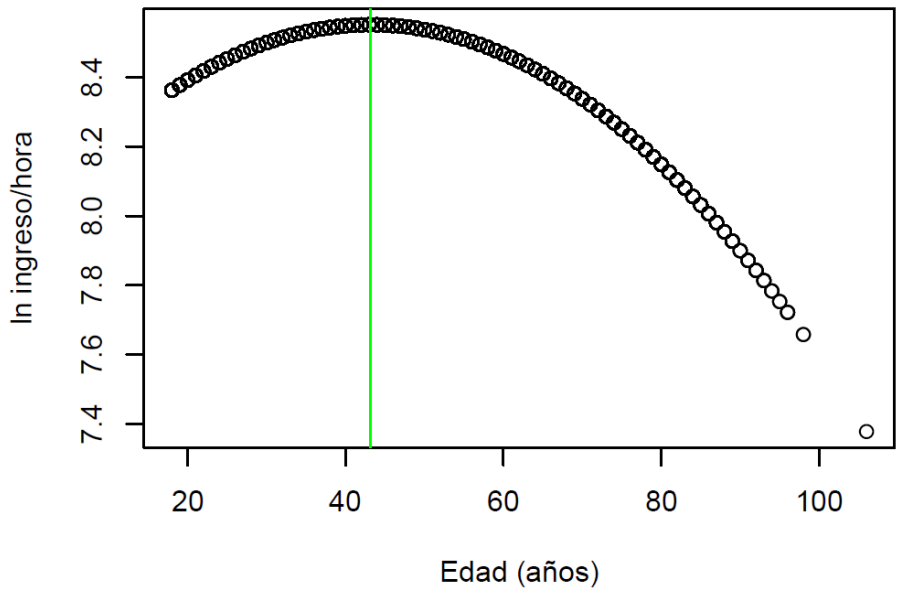
Comparando los modelos propuestos, el modelo usando los controles sobre la brecha salarial de género con FWL y la técnica Bootstrap presenta el menor de los errores estándar del parámetro para la variable género de los cuatro modelos analizados.

Perfil salarial por edad y género: estimación edad pico

Para estimar la edad pico se construyen los intervalos de confianza para aproximar la distribución del estimador a partir de los datos disponibles usando el método Bootstrap utilizando intervalos de confianza con la semilla 2023 y R=1000, es decir, que se generaron 1.000 muestras Bootstrap a partir de los datos originales, los cuales arrojaron que la edad pico para los hombres, esto es, la edad en la que los hombres reciben en promedio el máximo ingreso por hora es 50,016 años con un sesgo de 0,017, un error estándar de 0,58 y un intervalo de confianza de 95% entre 48,85 años y 51,14 años.



En el caso de las mujeres, siguiendo los mismos parámetros antes indicados, la edad pico para las mujeres, esto es, la edad en la que las mujeres reciben en promedio el máximo ingreso por hora es 43,18 años con un sesgo de –0,096, un error estándar de –0,096 y un intervalo de confianza de 95% entre 41,15 años y 45,43 años.



A pesar de que las mujeres alcanzan el ingreso por hora antes que los hombres en términos de edad, el valor de dicho ingreso es inferior al que perciben los hombres por la brecha antes indicada.

c. Predicción de ingresos

A continuación, se presenta una tabla con los perfiles edad y salarios predichos por género con sus correspondientes errores estándares.

	Ln Ingresos	
	(1) Hombre	(2) Mujeres
age	0.0572***	0.0257***
	(0.0026)	(0.0029)
age2	-0.0006***	-0.0003***
	(0.00003)	(0.00003)
Constant	7.4310***	7.9967***
	(0.0576)	(0.0653)
Observations	9,937	9,592
Residual Std. Error	0.8191 (df = 9934)	0.9284 (df = 9589)

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

En esta sección, se lleva a cabo una evaluación del rendimiento predictivo de las especificaciones planteadas en las partes anteriores y se proponen modelos adicionales con el objetivo de mejorar la capacidad de predicción. Para garantizar la reproducibilidad del ejercicio, se establece una semilla utilizando la función "set.seed(2023)". Se divide la base de datos en dos conjuntos: una muestra de entrenamiento, que representa el 70% de los datos, y otra muestra de prueba, que corresponde al 30% restante. Como resultado de esta división, la muestra de entrenamiento (train_data) consta de 13.670 observaciones, mientras que la muestra de prueba (test_data) está compuesta por 5.859 observaciones. Estas muestras permitirán ajustar los modelos de predicción utilizando los datos de entrenamiento y evaluar su desempeño utilizando los datos de prueba, lo que proporcionará una medida objetiva de su capacidad para predecir correctamente los resultados.

A continuación, se presentan los modelos utilizados previamente que se replicarán en esta sección para evaluar su rendimiento predictivo en los datos de prueba. Además, se incluirá un modelo simple sin covariables, que consiste únicamente en una constante, para establecer un punto de referencia:

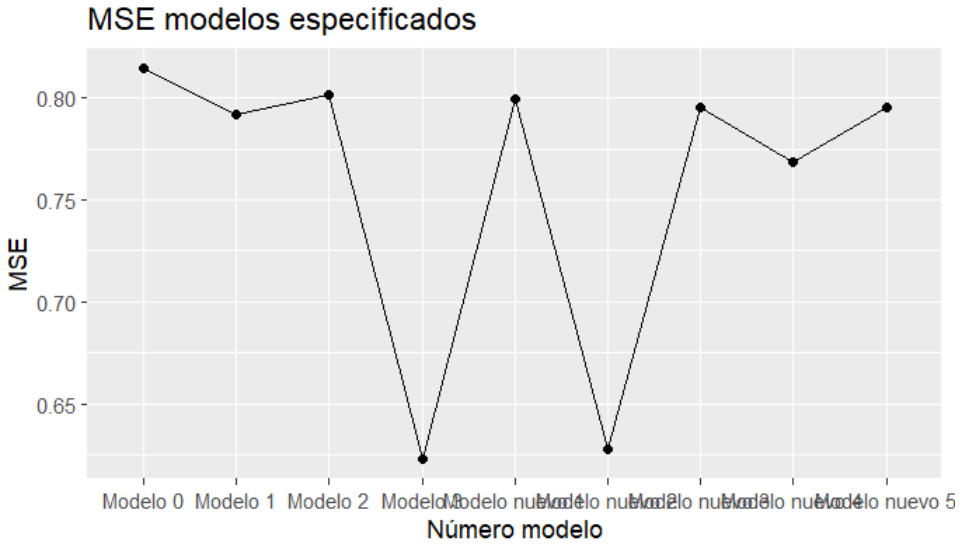
	Modelo
Modelo base, constante	*log*(ingreso)= 8.567+ *u* *(0.0076)*
Modelo previo 1	*Log*(ingreso)= β_1 +4.475Edad + -4.699*Edad* ^2 + *u* *(2.359)* *(2.422)*
Modelo previo 2	*Log*(ingreso)= β_1 +0.216Género+ *u* *(0.015)*
Modelo previo 3	Log(ingreso)=β_0 + 2.832 Género + 4.759 educ +3.618 estrato + 4.222edad+ (-4.907edad^2) *(1.391)* *(2.004)* *(7.069)* *(2.315)* *(2.373)* +1.095 tiene hijos + -2.323 oficio + (-3.415relaciónlab) + u *(1.454)* *(2.583)* *(4.686)*

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

	Modelo
Modelo nuevo 1	*Log*(ingreso)= β_1 +0.222 Genero + (-0.0209) Relacionlab+ *u*
Modelo nuevo 2	Log(ingreso)=β_0 + 0.32 Género + 0.045 edad+(-0.0005) edad2 + 0.0022 educ + (-0.047) relacionlab+ 0.397estrato + (-0.0301 sex*estrato) + u
Modelo nuevo 3	*Log*(ingreso)= β_1 +(-0.04) Género + (-0.03) Relacionlab+ (-0.003)edad +0.006 genero*edad + *u*

Modelo nuevo 4	*Log*(ingreso)= β_1 +-1.19 Genero + (-4.808) edad+ (-5.29) edad2 *+ (*-5.27) relab*+ 4.96* (genero*edad) + u
Modelo nuevo 5	*Log*(ingreso)= β_1 +(-0.048) Genero + (-0.002)edad+ (0.005)educ+ *(*-0.023) relab*+ 0.0062* (genero*edad) + u

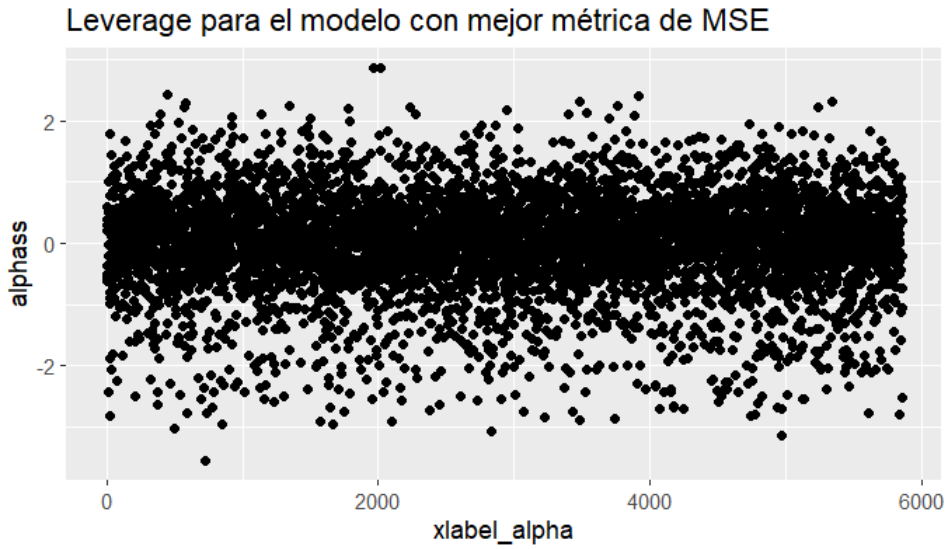
A continuación, se inserta la gráfica que evalúa el MSE de los 5 modelos propuestos.



Considerando lo anterior, se organizan los modelos desde el menor MSE hasta el mayor para determinar cuál presenta el mejor desempeño en predicción. Se observa que los modelos con menores MSE son el modelo 3 y el modelo nuevo 2.

Modelo estimado	**MSE**
Modelo 3	0.5927672
Modelo nuevo 2	0.5988026
Modelo nuevo 4	0.7360595
Modelo nuevo 3	0.7531596
Modelo nuevo 5	0.7532017
Modelo 1	0.7570223
Modelo nuevo 1	0.7582124
Modelo 2	0.7609430
Modelo 0	0.7695160

El modelo con menor MSE modelo 3, modelo que mejor predice el logaritmo del ingreso por hora de los trabajadores mayores de 18 años en Bogotá, es $log(\text{ingreso}) = \beta_1 + \beta_2 \text{Género} + \text{controles} + u$.



Los valores máximos y mínimos obtenidos fueron de 2.87 y -3.54, por lo cual, podemos concluir que, para este modelo, no se evidencian datos atípicos que afecten la regresión.

V. Conclusiones

La edad y el ingreso presentan una relación positiva y creciente hasta los 47.52 años aproximadamente, después de los cuales el ingreso empieza a reducirse a medida que los individuos envejecen. La brecha de ingresos asociada al género se reduce una vez se controla por variables adicionales asociadas a la educación, la experiencia, si se tienen hijos, entre otros; sin embargo, los hombres perciben 17,38% más ingresos que las mujeres por razones asociadas al género. Esto es coincidente con la literatura en la que se reporta la existencia de la brecha salarial por razones atribuibles a los roles reproductivos de las mujeres que conllevan a la destinación de tiempo parcial a la realización de labores remuneradas y a la interrupción de sus carreras por la maternidad o las tareas de cuidado de los padres.

Esta información resulta de utilidad para la determinación de criterios fidedignos que permitan identificar los eventos en los que la renta reportada por los individuos para el cumplimiento de sus obligaciones tributarias no coincide o discrepa respecto del ingreso esperado dada cuenta de las características específicas de edad y género con los controles propuestos.

VI. Anexos

Existe el repositorio (https://github.com/iapaezg/BD_LM_01) y contiene un README que ayuda al lector a navegar el repositorio e incluye instrucciones breves para replicar completamente el trabajo.

VII. Bibliografía

Araújo, A. (2015). La desigualdad salarial de género medida por regresión cuantílica: el impacto del capital humano, cultural y social. Revista Mexicana de Ciencias Políticas y Sociales, Volumen 60, Issue *223*, 287-315, ISSN 0185-1918, https://doi.org/10.1016/S0185-1918(15)72139-2.

Cremer, H., & Roeder, K. (2019). Income taxation of couples, spouses' labor supplies and the gender wage gap, *Economics Letters*,Volume 175, 71-75, ISSN 0165-1765, https://doi.org/10.1016/j.econlet.2018.12.015.

Guaraquí, J., García, A., & Rodríguez. (2009). Estimaciones de los determinantes de los ingresos laborales en Colombia con consideraciones diferenciales para asalariados y cuenta propia. Universidad del Rosario, facultad de Economía, Serie de documento de trabajo No 70. https://doi.org/10.48713/10336_10851

Rojas, M., Angulo, H., & Velázquez, I. (2000). Rentabilidad de la inversión en capital humano en México. *Economía Mexicana. Nueva Época*, vol. IX, núm. 2, 113 –142. Disponible en http://www.economiamexicana.cide.edu/num_anteriores/IX-2/01_MARIANO_ROJAS_113-142.pdf

Tyrowicz, J., Van der Velde, L & Van Staveren, L. (2018) Does Age Exacerbate the Gender-Wage Gap? New Method and Evidence From Germany, 1984–2014, Feminist Economics, 24:4, 108-130, DOI: 10.1080/13545701.2018.1503418

Haaland, I., & Olden, A. (2022). Fraud concerns and support for economic relief programs. *Journal of Economic Behavior and Organization*, *203*, 59–66. https://doi.org/10.1016/j.jebo.2022.08.026

Mao, X., Liu, M., & Wang, Y. (2022). Using GNN to detect financial fraud based on the related party transactions network. *Procedia Computer Science*, *214*(C), 351–358. https://doi.org/10.1016/j.procs.2022.11.185

Vanhoeyveld, J., Martens, D., & Peeters, B. (2020). Value-added tax fraud detection with scalable anomaly detection techniques. *Applied Soft Computing Journal*, *86*(May 2016), 105895. https://doi.org/10.1016/j.asoc.2019.105895

Yi, Z., Cao, X., Pu, X., Wu, Y., Chen, Z., Tamoor, A., Francis, A., & Li, S. (2023). Fraud detection in capital markets : A novel machine learning approach. https://doi.org/10.1016/j.eswa.2023.120760