Problem Set 1: Predicting Income

Sofia Charry Tobar
s.charry@uniandes.edu.co

Laura Manuela Rodriguez Morales
lm.rodriguezm@uniandes.edu.co

Nicol Valeria Rodríguez Rodríguez
nv.rodriguezr1@uniandes.edu.co

Brahyan Alexander Vargas Rojas
ba.vargas@uniandes.edu.co

El repositorio del ejercico es: https://github.com/nvrr2028/Taller-1-BDML.git

1. Introducción

La brecha fiscal es un de los principales problemas del mundo, siendo esta la diferencia entre la cantidad de impuestos que el gobierno puede recaudar contra la cantidad de impuestos que realmente recauda. En Colombia, según Bonet-Morón y Ayala-García (2016), la brecha fiscal en el 2014 fue de aproximadamente 46 billones de pesos, lo que demuestra un gran desperdicio de eficiencia tributaria. Es importante reducir esta brecha debido a que se debe mantener un sistema tributario justo, eficiente y efectivo, para garantizar el mayor recaudo posible para el estado y poder brindar servicios esenciales a la población.

Dado lo anterior, se realizará un estudio de predicción de ingresos, el cual podría potencialmente ayudar a acumular casos de fraude que lograría la reducción de esta brecha. Además, un modelo de predicción de ingresos puede ayudar a identificar a las personas y familias vulnerables que pueden necesitar más ayuda. Así mismo, con este método el recaudo podría ser más eficiente y justo, para que el gobierno tenga mayor capacidad para efectuar mejores estrategias que estimulen el desarrollo social y económico del país, reduciendo las brechas de desigualdad y creando nuevas oportunidades para la población. Los modelos implementados en este estudio contienen estimaciones de la brecha salarial de género, la maximización del ingreso de los individuos por edad y un modelo predictor del ingreso. En términos generales, el estudio concluye que en Colombia hay brechas salariales por discriminación por género, y que el modelo propuesto puede ser un buen predictor del salario, ya que es efectivo prediciendo muestra dentro del grueso de la población pero no tanto en observaciones atípicas.

2. Data

a. Describe the GEIH briefly, including its purpose, and any other relevant information.

La Gran Encuesta Integrada de Hogares (GEIH), implementada y publicada por el Departamento Administrativo Nacional de Estadística (DANE), es una encuesta realizada a los hogares en Colombia del tipo corte transversal y con periodicidad mensual. La GEIH tiene como objetivo suministrar información sobre el mercado laboral, los ingresos, la pobreza y las condiciones sociodemográficas de la población en el territorio nacional (DANE, 2016). Por lo tanto, a partir de la GEIH se obtienen indicadores como la tasa de desempleo, la tasa global de participación, la informalidad, el ingreso, el nivel de educación, etc. Adicionalmente, la GEIH cuenta con una cobertura geográfica para el total nacional, las cabeceras, los centros poblados, y las zonas rurales dispersas. En específico, la muestra mensual de la GEIH incluye 20 669 hogares, 18 790 viviendas y 1 879 segmentos.

De igual forma, la GEIH ha sufrido múltiples actualizaciones metodológicas, siendo la más reciente de ellas la del Marco 2018. Lo anterior hace parte del proceso continuo de mejora en la recolección, el procesamiento y la divulgación de estadísticas por parte del DANE.

Finalmente, la GEIH es útil para la construcción de un modelo que permita la predicción del ingreso de los individuos y, así, la brecha en el pago de impuestos, debido a que está enfocada en la caracterización laboral de la población en Colombia. Dicho de otra manera, esta encuesta es representativa a nivel nacional e incluye variables como el nivel de educación, la edad, el tipo de trabajo, el ingreso, entre otras, que son fundamentales para el análisis requerido.

b. Describe the process of acquiring the data and if there are any restrictions to accessing/scraping these data.

Los datos de la GEIH para Bogotá de 2018 se obtuvieron a través de *web scrapping*, técnica que permite extraer datos, texto, etc. de páginas web de forma automática, eficiente, rápida y replicable. El anterior proceso requiere tener en cuenta aspectos legales y éticos, de acuerdo con la configuración de las páginas a *scrappear*, el tipo de datos no anonimizados, la legislación del país, entre otros factores.

Los datos se obtienen del siguiente link https://ignaciomsarmiento.github.io/GEIH2018_sample/. Al ingresar a la página, los datos se encuentran almacenados en 10 diferentes *chunks*, por lo que el objetivo de utilizar *web scrapping* es importar la información de cada link y unificarla en una base, que permita el desarrollo de los ejercicios. Esta técnica es especialmente útil, dado que abrir cada link es un proceso lento (5-10 min por link) que

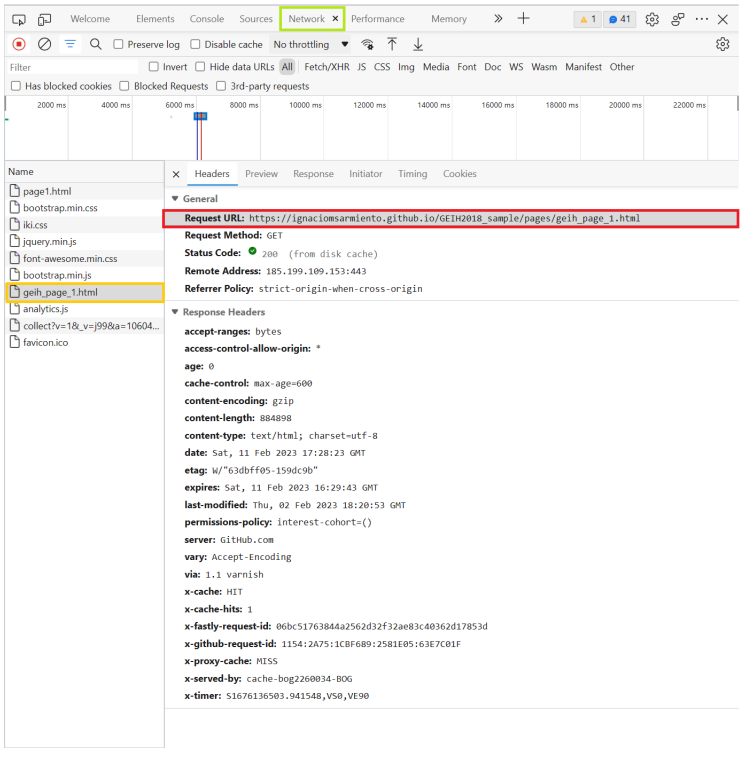
dificulta la obtención de los datos.

En primer lugar, si se lee el html asociado a cada *chunk*, para luego usar el comando `html_table()` para extraer la tabla asociada al link, el resultado es una lista vacía. Lo anterior implica que el comando no encontró tablas en el link.

```
> #Prueba inicial
> url <- "https://ignaciomsarmiento.github.io/GEIH2018_sample/page1.html"
> url <- read_html(url)
> my_table <- url %>% html_table()
> my_table
list()
```

Por lo tanto, la extracción de la información se debe realizar inspeccionando la página web. En específico, al abrir la ventana de *Herramientas del desarrollador* y dirigirnos al menú *Network* (que nos permite ver las diferentes páginas que a su vez está llamando el link), se encuentra el objeto *geih_page_1.html* que contiene el url directo a la tabla de los datos que se va a scrappear.

Gráfico 1



Luego, dado que existe un patrón común entre los vínculos, se realiza un *loop* en R que

permita recorrer cada uno de los links y extraer la información de interés. Por último, se construye una base de datos conjunta.

Gráfico 2

```
# Loop para obtener la información de los 10 chuncks.
links <- list()
for (i in 1:10) {
  links[[i]] <- import(paste("https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_",i, ".html", sep=""))
}
# Juntar la base de datos por filas.
data <- list.rbind(links)
```

c. Describe the data cleaning process

El análisis de la brecha en el pago de impuestos en los ingresos está centrado en la población mayor a 18 años y empleada. Asimismo, es necesario considerar que nuestra variable de interés (ingreso por hora) puede tener datos vacíos o nulos, lo cual no sería informativo para la investigación. De esta manera, en un primer momento, la base de datos es filtrada de acuerdo con la edad, la situación ocupacional, los *missing values* y otros valores poco informativos. Lo anterior da a lugar a una base con 9892 observaciones y 179 variables.

Luego, dado que la motivación del estudio es construir un modelo que nos permita predecir los ingresos de los individuos en Bogotá a partir de un conjunto de variables explicativas, se filtra por los datos vacíos presentes en el set de indicadores seleccionados. Por lo tanto, se obtiene una base de datos con 9891 observaciones y 10 variables.

d. A descriptive analysis of the data.

Variables explicativas seleccionadas

El conjunto de variables seleccionadas para el análisis del ingreso por hora de las personas en Bogotá se encuentra descrito en el Tabla 1. Cada variable indica lo siguiente:

- **ing_hr:** variable continua que representa el ingreso laboral nominal por hora, incluyendo todas las ocupaciones, propinas y comisiones.
- **maxEducLevel:** variable categórica sobre el máximo nivel de educación alcanzado.
  - 1: Ninguno.
  - 2: Preescolar.
  - 3: Primaria incompleta.
  - 4: Primaria completa.

    - 5: Secundaria incompleta.
    - 6: Secundaria completa.
    - 7: Terciaria
    - 9: N/A.
- **age:** variable continua que representa la edad.
- **sexo:** variable dicotómica que representa el sexo (1 = Hombre, 0 = Mujer).
- **totalHoursWorked:** variable continua que representa el número total de horas trabajadas la semana anterior.
- **formal:** variable binaria que toma el valor de 1 si el trabajador si cotiza a seguridad social. Es una proxy de formalidad.
- **estrato1:** variable categórica para el estrato socieconómico. Toma valores entre 1 a 6.
- **fulltime:** variable binaria que toma el valor de 1 si el trabajador trabajó más de 40 horas la semana pasada. Es una proxy del tipo de contrato.
- **relab:** variable categórica para el tipo de ocupación.
    - 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 remuneracin en empresas o negocios de otros hogares.
    - 8: Jornalero o peón.
    - 9: Otro.
- **sizeFirm:** variable categórica para el tamaño de la empresa.
    - 1: Independiente.
    - 2: 2-5 empleados.
    - 3: 6-10 empleados.
    - 4: 11-50 empleados.
    - 5: >50 empleados.

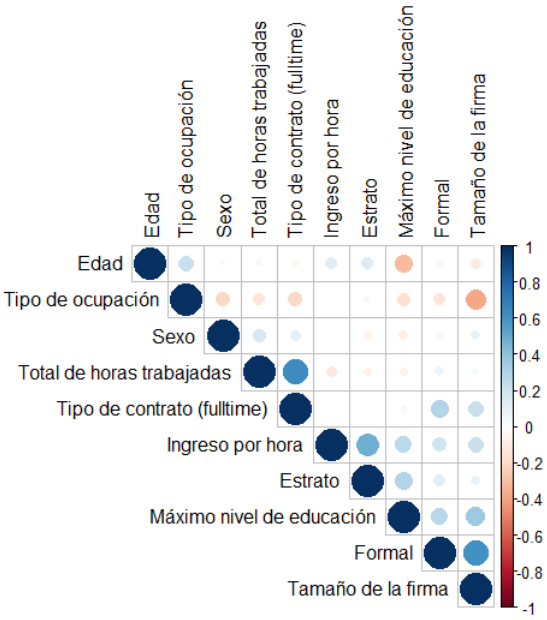
Como se mencionó previamente, la base utilizada cuenta con 9891 observaciones. Con respecto a nuestra variable dependiente, podemos observar que en promedio una persona en Bogotá recibe un ingreso de $8 822 por hora, siendo el ingreso mínimo y máximo de $326 y $350 583, respectivamente. Adicionalmente, el individuo promedio en Bogotá completó la secundaria, tiene alrededor de 36 años, trabaja aproximadamente 48 horas a la semana, suele ser formal, pertenece a los estratos 2 y 3 y es empleado/a en una empresa particular con entre 11 y 50 trabajadores.

Tabla 1: Resumen descriptivo de la base de datos

Statistic	N	Mean	St. Dev.	Min	Max
ing_hr	9,891	8,822.721	12,886.720	326.667	350,583.300
maxEducLevel	9,891	6.098	1.106	1	7
age	9,891	36.239	12.025	18	86
totalHoursWorked	9,891	48.343	12.250	1	130
formal	9,891	0.767	0.422	0	1
sex	9,891	0.503	0.500	0	1
estrato1	9,891	2.509	0.975	1	6
fulltime	9,891	0.915	0.279	0	1
relab	9,891	1.172	0.511	1	8
sizeFirm	9,891	3.921	1.333	1	5

Por otro lado, en el Gráfico 3 se puede observar que el ingreso por hora cuenta con una correlación positiva con el estrato socioeconómico (0,48), el máximo nivel de educación (0,26), la formalidad (0,20), el tamaño de la firma (0,22) y la edad (0,14). Por el contrario, presenta una correlación negativa con el total de horas trabajadas (-0,12). Finalmente, el ingreso por hora no tendría una correlación significativa al 5 % con el sexo, el tipo de ocupación y el tipo de contrato (fulltime).

Gráfico 3: Mapa de correlaciones del conjunto de variables seleccionadas

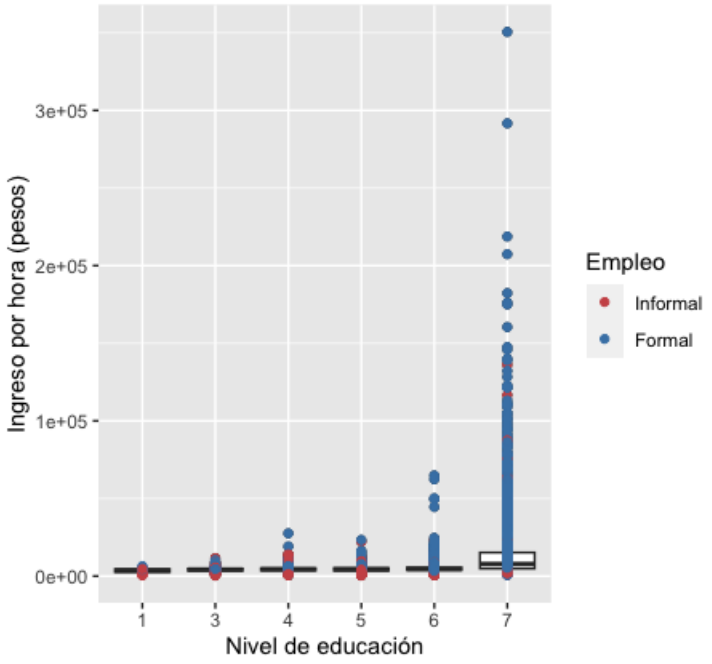


Nota: El tamaño de los círculos indica la magnitud de la correlación entre las variables. Además, se excluyen los coeficientes de correlación que no sean signficativos al 5 %. Fuente: cálculos de los autores.

Máximo nivel de educación

En el Gráfico 4 podemos ver como se distribuyen el ingreso de los hogares encuestados en la GEIH y su años de educación. En términos generales podemos ver una relación creciente en el ingreso. Es decir, entre más años de educación tenga el individuo mayor va a ser su ingreso. Además, el Gráfico 4 muestra la cantidad de personas que están vinculadas al sector formal, se puede evidenciar que mayoritariamente, las personas que pertenecen al sector formal y tienen una educación alta tienen más ingresos. El Gráfico 4 tiene información importante ya que para nuestro modelo de predicción se puede se evidencia claramente que la educación y sector laboral son importantes para la predicción del salario.

Gráfico 4: Boxplot Ingreso por hora de acuerdo al nivel de educación

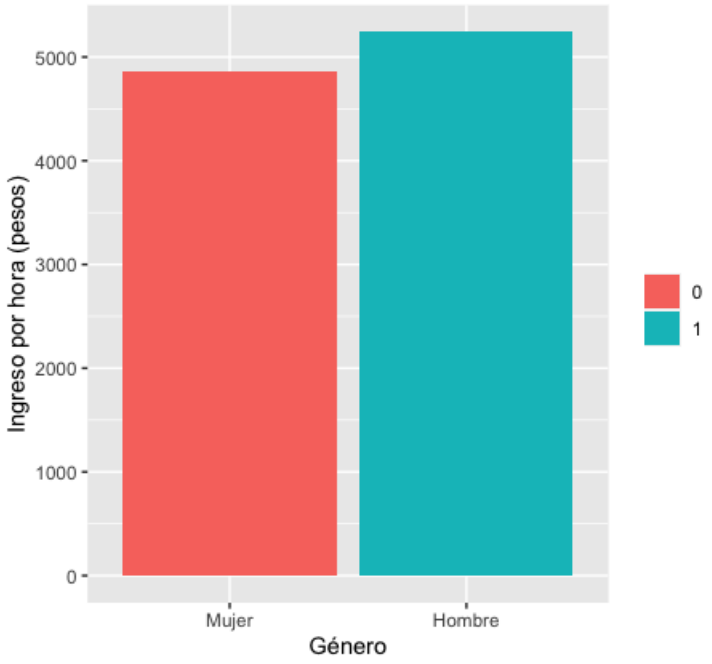


Fuente: cálculos de los autores.

Sexo

En el Gráfico 5 se muestra la media de ingresos por hora por género, es importante acotar el género del individuo encuestado ya que puede varias su ingresos. Especialmente como vemos en el Gráfico 5 los Hombres tienen un ingreso superior al de las mujeres, es de importancia valora estas características para crear un buen modelo de estimación.

Gráfico 5: Media del ingreso por género



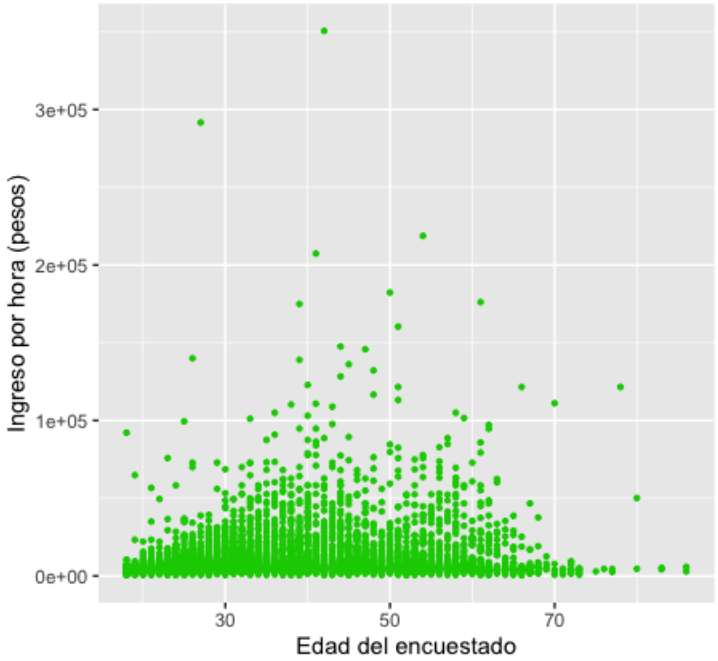
Fuente: cálculos de los autores.

Edad y Total de horas trabajadas

En el Gráfico 6 se evidencia la relación del ingreso y la edad del individuo. En este se ve que los individuos encuestados de la GEIH trabajan constantemente a largo de su vida. Esto gráficos puede ser un buen predictor del ingreso ya que se podría conocer cuál es el rango de ingresos respecto a la edad del individuo,el cual es diferente para todos los encuestados.

Edad y Total de horas trabajadas

Gráfico 6: Gráfico de dispersión de la edad de los individuos encuestados

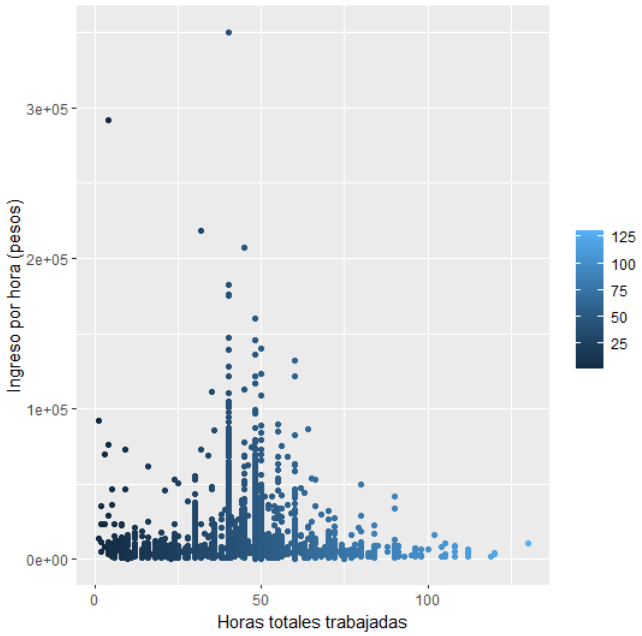


Fuente: cálculos de los autores.

Horas trabajadas

Al analizar el ingreso por hora de acuerdo con el número se horas trabajadas, se puede observar que los individuos que reportan haber trabajado alrededor de 50 horas en la semana anterior, suelen contar con los mayores niveles de ingreso por hora. Además, también se evidencia que, conforme el número de horas se va acercado a 0 y 100, el pago por hora se reduce (ver Gráfico 7). En consecuencia, incluir la variable de total de horas trabajadas la semana anterior, podría ser relevante para predecir el ingreso por hora de los individuos.

Gráfico 7: Ingreso por hora de acuerdo con el número total de horas trabajadas la semana anterior

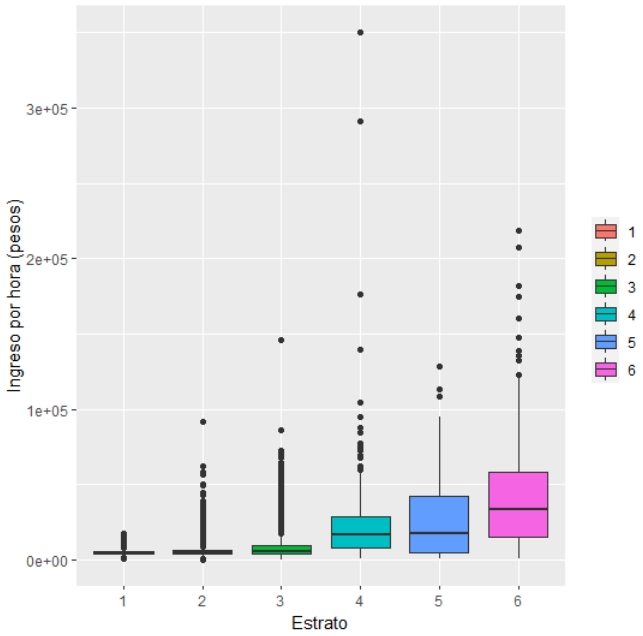


Fuente: cálculos de los autores.

Estrato socioeconómico

De acuerdo con el Gráfico 12, el ingreso por hora incrementa a medida de que aumenta el estrato socioeconómico. La mediana del ingreso por hora para los estratos 1, 2, 3, 4, 5 y 6 es de $4 525, $4 679, $5 688, $17 127, $ 17 243 y $ 33 299, respectivamente, por lo que los tres estratos más bajos reciben menos de la mitad de lo que reciben los tres estratos más altos por hora. Adicionalmente, los ingresos atípicos por hora reportados en los estratos 1, 2 y 3 tienen un sesgo al alza. Por otro lado, para los estratos altos se presenta un sesgo positivo en la distribución, por lo que hay una mayor dispersión de las observaciones hacia valores más altos del ingreso. Por último, el ingreso atípico que sobresale del gráfico pertenece al estrato 4, con un valor superior a $300 000.

Gráfico 8: Boxplot ingreso por hora de acuerdo con el estrato

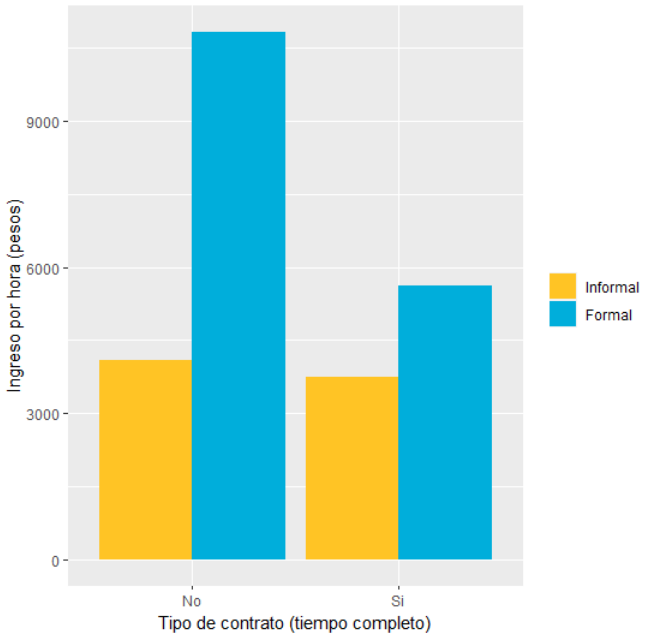


Fuente: cálculos de los autores.

Tipo de contrato (tiempo completo o no) y formalidad

El ingreso por hora reportado por los individuos que trabajan más de 40 horas a la semana (contrato tiempo completo) es inferior al ingreso que reciben las personas que laboran menos de 40 horas, sin importar si el trabajador es formal o informal (ver Gráfico 9). Analizando por formalidad, entendida como aquel trabajador que cotiza a seguridad social, los trabajadores formales cuentan con un ingreso por hora mediano superior al recibido por los trabajadores informales, diferencia que es particularmente alta para las personas que trabajan menos de 40 horas a la semana.

Gráfico 9: Ingreso por hora mediano de acuerdo con el tipo de contrato y formalidad

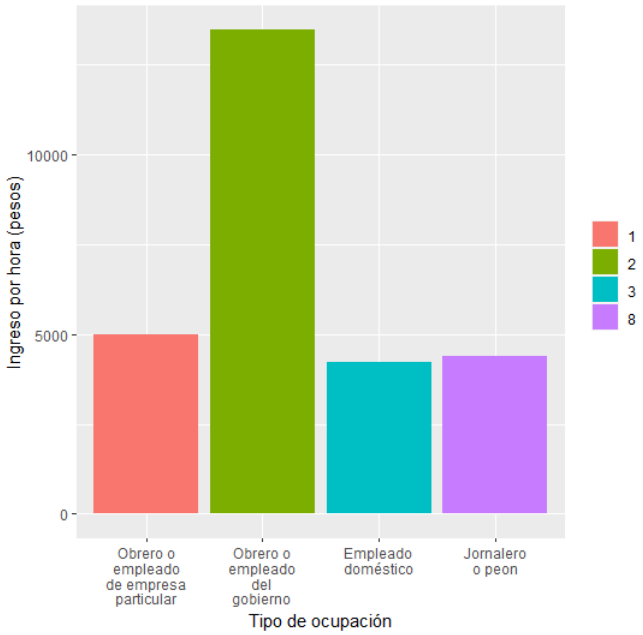


Nota: Se define que una persona cuenta con un contrato de tiempo completo si reporta que trabaja usualmente más de 40 horas a la semana. Fuente: cálculos de los autores.

Tipo de ocupación

Con respecto al ingreso mediano por hora desagregando por el tipo de ocupación (ver Gráfico 10), las personas que reportan ser empleados del gobierno presentan el ingreso más alto entre las diferentes posiciones ocupacionales. De la misma forma, l@s emplead@s domestic@s reciben el menor ingreso mediano por hora.

Gráfico 10: Ingreso por hora mediano de acuerdo con el tipo de ocupación

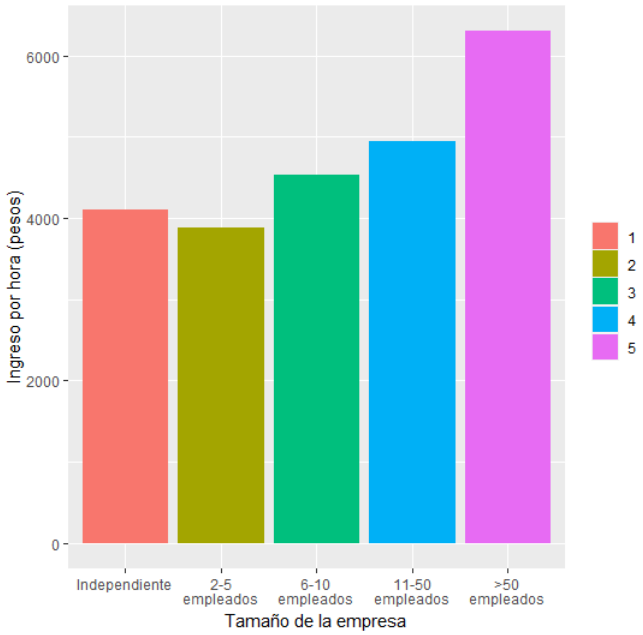


Fuente: cálculos de los autores.

Tamaño de la empresa

Finalmente, el Gráfico 11 permite analizar el ingreso mediano por hora de acuerdo con el tamaño de la empresa. Podemos observar que el ingreso de los trabajadores incrementa progresivamente conforme las empresas aumentan su tamaño, siendo los trabajadores en organizaciones con más de 50 empleados los que devengan el mayor nivel de ingreso. Vale la pena resaltar que los trabajadores independientes cuentan con un ingreso mediado por hora superior al recibido por los trabajadores en una empresa con 2 a 5 empleados.

Gráfico 11: Ingreso por hora mediano de acuerdo con el tamaño de la empresa



Fuente: cálculos de los autores.

3. Age-wage profile

En este punto buscamos estimar la relación entre la edad y los ingresos de las y los Colombianos según la muestra de la GEIH. Para esto corremos el siguiente modelo:

$$log(w) = \beta_0 + \beta_1 Age + \beta_2 Age^2 + u \tag{1}$$

Donde $log(w)$ representa el logaritmo de el salario por hora del individuo, Age y Age^2 representan la edad y la edad al cuadrado respectivamente, insertamos la edad al cuadrado porque queremos evaluar si los retornos de la edad al salario varían a medida que aumentan los años y u que representa el error.

a. Regression table.

A continuación, se encuentran los resultados de la regresión:

Tabla 2

	Variable dependiente:
	log(w)
Age	0.067***
	(0.004)
Age^2	−0.001***
	(0.00004)
Constante	7.374***
	(0.068)
Observaciones	9,892
R^2	0.044
R^2 ajustada	0.044
Error estándar residual	0.711 (df = 9889)
Estadístico F	228.437*** (df = 2; 9889)
Nota:	*p<0.1; **p<0.05; ***p<0.01

b. Iinterpretation of the coefficients and its significance.

La edad y la edad cuadrática explican el logaritmo natural del salario con un nivel de significancia del 1 %. La forma funcional es cóncava puesto que el intercepto es positivo, el coeficiente que acompaña la edad es positivo y el que acompaña la edad cuadrática es negativo, lo cual indica que hay una edad que maximiza el logaritmo natural del salario. Los coeficientes estimados indican que el efecto de la edad sobre el logaritmo natural del salario depende del valor que tome la edad y es, específicamente, 0,067 - 0,001*edad. Por lo anterior, tener un año más de edad siempre está asociado con un aumento de un 6,7 % en el salario, aunque también tener un año más de edad genera una reducción del salario en un 0,1 % de la edad que tiene el individuo. En este caso, no tiene sentido interpretar el intercepto, ya que este nos indica cual sería el salario si la persona tiene 0 años.

c. Discussion of the model's in sample fit.

El ajuste del modelo es de un 4,4 %, por lo que la edad y su término cuadrático explican un 4,4 % de la variación del logaritmo del salario. Bajo el supuesto de que un R^2 por debajo de 0,04 indica que el modelo no explica satisfactoriamente la variación de la variable dependiente, es posible decir que el modelo no es satisfactorio para explicar la variación del logaritmo del salario. Lo anterior podría implicar que las variables utilizadas, la edad y la

edad al cuadrado, no explican en gran medida la variación de los salarios, sin embargo, no podemos concluir que el modelo no se ajusta a los datos debido a que es necesario recordar que esta medida es relativa y puede tener variaciones debido a la cantidad de las variables y no exactamente a un mejor ajuste del modelo.

d. A plot of the estimated age-earnings profile implied by the above equation, including a discussion of the peak age with its respective confidence intervals.

Como la introducción al punto nos explica existe evidencia que indica una trayectoria usual para la relación entre salario y edad, esta se comporta como una parábola y alcanza su punto máximo alrededor de los 45 años. Para obtener este valor en la edad vamos a maximizar nuestra función objetivo. Ya que sabemos que el máximo punto de la función es en la tangente de la recta.

$$\begin{aligned} \text{máx}\{Age\} &= log(w) - \beta_0 - \beta_1 Age - \beta_2 Age^2 - u \\ \frac{\partial log(w)}{\partial Age} &= -\beta_1 - 2\beta_2 Age = 0 \\ -\beta_1/(2\beta_2) &= Age_p \end{aligned}$$

De esta forma encontramos el valor de la edad que máximiza la trayectoria de los ingresos.

Ahora bien, para construir los intervalos de confianza vamos a seguir la siguiente ecuación:

$$CI_j = \bar{x}_j \pm z * \sigma_j \tag{2}$$

Valores que obtenemos al correr el Boostrap.

Tabla 3

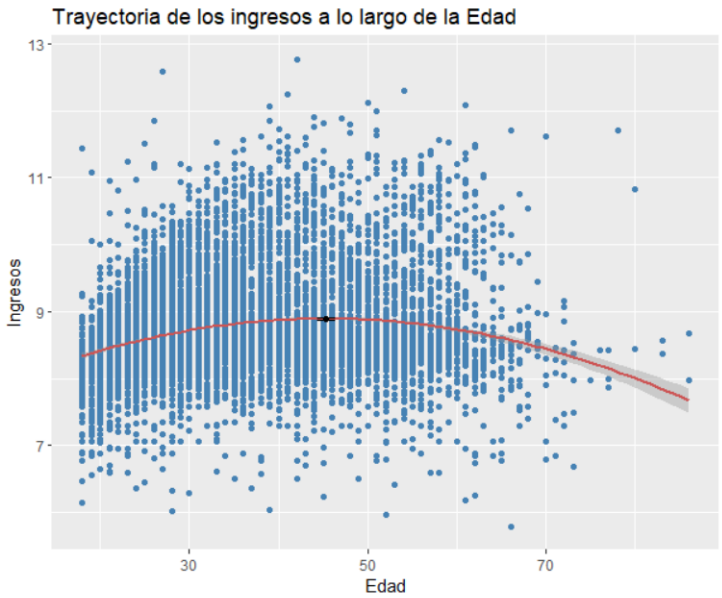
Peak Age	45.31
Bias	0.015
SE	0.67

Ya con nuestros valores encontados podemos observar la trayectoria de los ingresos a lo largo de la edad, alcanzando su punto máximo demarcado en negro, con su respectivo intervalo de confianza, el cuál es realmente estrecho.

Tabla 4: Boostrap PeakWage

Peak Wage (en log)	8.892
Bias	0.000
SE	0.011

Gráfico 12: Peak-age



Fuente: cálculos de los autores.

4. The gender earnings GAP

a. Begin by estimating and discussing the unconditional wage gap.

En la Tabla 5 se presentan los resultados de estimar por mínimos cuadrados ordinarios (MCO) una regresión lineal del logaritmo del ingreso mensual (***ing_m***) y por hora (***ing_hr***) con la dummy *female* (1 = Mujer, 0 = Hombre), tal como se representa en las siguientes ecuaciones:

$$ing\_m = \beta_0 + \beta_1 female - u \quad (1)$$
$$ing\_hr = \beta_0 + \beta_1 female - u \quad (2)$$

En cuanto al análisis de los coeficientes, se tienen los siguientes resultados:

- **ing_m (1):** Una mujer recibe un ingreso mensual inferior en un 14,7 % al devengado por un hombre, manteniendo todo lo demás constante. Además, la variable *female* es significativa al 1 %.
- **ing_hr (2):** Una mujer recibe un 4,5 % menos de ingreso por hora frente a un hombre, manteniendo todo lo demás constante. Además, la variable *female* es significativa al 1 %.

Con respecto al ajuste de ambos modelos, el R^2 registra valores inferiores al 1 %, lo que indica que los modelos explican menos del 1 % de la variabilidad de la variable dependiente. Por consiguiente, *female* como una variable individual no permite explicar de forma adecuada ni el salario por hora ni el salario mensual.

Tabla 5: Brecha salarial de género no condicionada

	Dependent variable:	
	ing_m	ing_hr
	(1)	(2)
female	-0.147^{***}	-0.045^{***}
	(0.015)	(0.015)
Constant	14.088^{***}	8.747^{***}
	(0.011)	(0.010)
Observations	9,891	9,891
R^2	0.009	0.001
Adjusted R^2	0.009	0.001
Residual Std. Error (df = 9889)	0.762	0.727
F Statistic (df = 1; 9889)	91.527^{***}	9.354^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Lo anterior hace necesario el uso de otras variables que permitan mejorar el desempeño del modelo.

b. Equal Pay for Equal Work?.

La Tabla 6 presenta los resultados de estimar la brecha salarial de género para el ingreso mensual, utilizando tres diferentes regresiones: (i) brecha salarial no condicionada, que solo incluye como regresor a *female*, (ii) brecha salarial condicionada, que incluye como variables

explicativas a *female*, *maxEducLevel*, *age*, *age2* (edad al cuadrado), *formal*, *fulltime* y *relab*, y (iii) brecha salarial condicionada siguiendo el teoreman FWL, que plantea una regresión entre *wageResidF* (error del modelo lineal $Y \sim X_2$[1]) y *femaleResidF* (error del modelo lineal ${X_1}^2 \sim X_2$). La regresión (2) en la tabla solo muestra *female* para facilitar la comparabilidad entre modelos, no obstante, la regresión completa se puede consultar en el Anexo 15. Además, la Tabla 7 comprende los resultados de emplear la técnica de muestreo *bootstrap* para el coeficiente de *female* en las regresiones (2) -Modelo original- y (3) -Modelo FWL-.

En primer lugar, se evidencia que el teorema FWL se satisface, en la medida que el coeficiente y el error estándar asociado *female* y *femaleResidF* son idénticos (ver Tabla 7). En segundo lugar, de acuerdo con los resultados de las regresiones (2) y (3), una mujer recibe un ingreso mensual inferior a un hombre en un 13,4 %, *ceteris paribus*, lo cual es significativo al 1 %. Lo anterior difiere en términos de magnitud a lo obtenido en la brecha salarial no condicionada (14,7 %), producto de la inclusión de variables explicativas informativas sobre el nivel de ingreso de los individuos en (2) y (3).

Con respecto al ajuste, el modelo (2) para la estimación de la brecha salarial de género condicionada explica en un 47,4 % la variabilidad del ingreso mensual, por lo que logra una mejora importante con relación a lo encontrado en (1).

Tabla 6: Estimaciones de la brecha salarial para el ingreso mensual

	Dependent variable:		
	ing_m		wageResidF
	Brecha salarial no condicionada	Brecha salarial condicionada	Brecha salarial condicionada FWL
	(1)	(2)	(3)
female	-0.147^{***}	-0.134^{***}	
	(0.015)	(0.012)	
femaleResidF			-0.134^{***}
			(0.012)
Constant	14.088^{***}	11.389^{***}	0.000
	(0.011)	(0.099)	(0.006)
Observations	9,891	9,891	9,891
R^2	0.009	0.474	0.013
Adjusted R^2	0.009	0.473	0.013
Residual Std. Error	0.762 (df = 9889)	0.556 (df = 9877)	0.555 (df = 9889)
F Statistic	91.527^{***} (df = 1; 9889)	685.134^{***} (df = 13; 9877)	132.063^{***} (df = 1; 9889)

*p<0.1; **p<0.05; ***p<0.01

[1] X_2 incluye *maxEducLevel*, *age*, *age2* (edad al cuadrado), *formal*, *fulltime* y *relab*
[2] X_1 incluye *female*

Tabla 7: Bootstrap brecha salarial condicionada para el ingreso mensual

	Modelo original	Modelo FWL
Coeficiente	-0.134105608239695	-0.134105608239699
Sesgo	0.0004555001	8.962507e-05
Errores estándar	0.0116767004720925	0.0116767004720929

Por otro lado, la Tabla 8 presenta los resultados de estimar la brecha salarial de género para el ingreso por hora, utilizando las tres mismas regresiones descritas en el ingreso mensual. La regresión (2) en la tabla solo muestra *female* para facilitar la comparabilidad entre modelos, no obstante, la regresión completa se puede consultar en el Anexo 16. Asimismo, la Tabla 9 contiene los resultados de emplear la técnica de muestreo *bootstrap* en el coeficiente de *female* para las regresiones (2) -Modelo original- y (3) -Modelo FWL-.

Al igual que en los resultados para el ingreso mensual, el teorema FWL se cumple para las variables *female* y *femaleResidFhr* (ver Tabla 9). Luego, una mujer devenga un ingreso por hora inferior en comparación a un hombre en un 9,9 %, manteniendo todo lo demás constante. Este resultado es significativo al 1 %. Del mismo modo, el coeficiente de *female* en la regresión de la brecha salarial condicionada es mas del doble que el coeficiente en el modelo de la brecha salarial no condicionada.

Por último, el modelo (2) logra un mejor ajuste que en el modelo (1), debido a que, según el estadístico R^2, el primero explica en un 41,5 % la variabilidad del ingreso por hora, mientras que el segundo explica menos del 1 %.

Teniendo en cuenta todo lo anterior, se podría concluir que la brecha salarial de género producto de la discriminación es del 13,4 % y 9,9 % para el ingreso mensual y el ingreso por hora en contra de las mujeres, respectivamente. Lo anterior es resultado del ejercicio de controlar por múltiples factores relacionados a la productividad de los trabajadores, por lo que el coeficiente de *female* permitiría capturar lo que está primordialmente relacionado con el género. Además, estos resultados van en línea con los cálculos del DANE para brecha salarial[3] mensual (12,1 %) para el total nacional en 2018. Por el contrario, nuestro resultado de la brecha salarial por hora difiere de la cifra oficial del DANE para la brecha salarial por hora, que para 2018 es del 3,2 % a favor de las mujeres (DANE, 2021). Vale la pena resaltar que el coeficiente asociado a relab3, que representa la categoría de empleado doméstico, es significativo al 1 % y toma un valor de -24,1 %. Esto podría ser relevante, debido a que es usual que la mayoría de esta población sean mujeres.

Sin embargo, el análisis previo sobre la brecha salarial de género cuenta con espacio de mejora. En particular, las regresiones consideradas no incorporan información sobre experiencia laboral, distribución del tiempo, presencia de menores, estado civil, ubicación geográfica,

[3]La brecha salarial de género se entiende como la diferencia promedio entre el ingreso laboral recibida por los hombres y las mujeres, como porcentaje del ingreso del hombre.

Tabla 8: Estimaciones de la brecha salarial para el ingreso por hora

	Dependent variable:		
	ing_hr		wageResidFhr
	Brecha salarial no condicionada	Brecha salarial condicionada	Brecha salarial condicionada FWL
	(1)	(2)	(3)
female	-0.045^{***}	-0.099^{***}	
	(0.015)	(0.012)	
femaleResidFhr			-0.099^{***}
			(0.012)
Constant	8.747^{***}	6.932^{***}	-0.000
	(0.010)	(0.099)	(0.006)
Observations	9,891	9,891	9,891
R^2	0.001	0.415	0.007
Adjusted R^2	0.001	0.414	0.007
Residual Std. Error	0.727 (df = 9889)	0.557 (df = 9877)	0.557 (df = 9889)
F Statistic	9.354^{***} (df = 1; 9889)	538.763^{***} (df = 13; 9877)	72.043^{***} (df = 1; 9889)

*p<0.1; **p<0.05; ***p<0.01

Tabla 9: Bootstrap brecha salarial condicionada para el ingreso por hora

	Modelo original	Modelo FWL
Coeficiente	-0.0992841381702638	-0.0992841381702663
Sesgo	0.0005964313	0.0003360888
Errores estándar	0.0117043486124327	0.011704348612433

sector económico, jefe de hogar, entre otras, que podrían ser relevantes en la predicción del salario por género. De igual forma, en la literatura se han desarrollado metodologías que enfocadas la estimación de la brecha salarial, como lo es el caso de la descomposición de Oaxaca (1973)-Blinder (1973) o de Machado y Mata (2005), que podrían proporcionar una mejor medición.

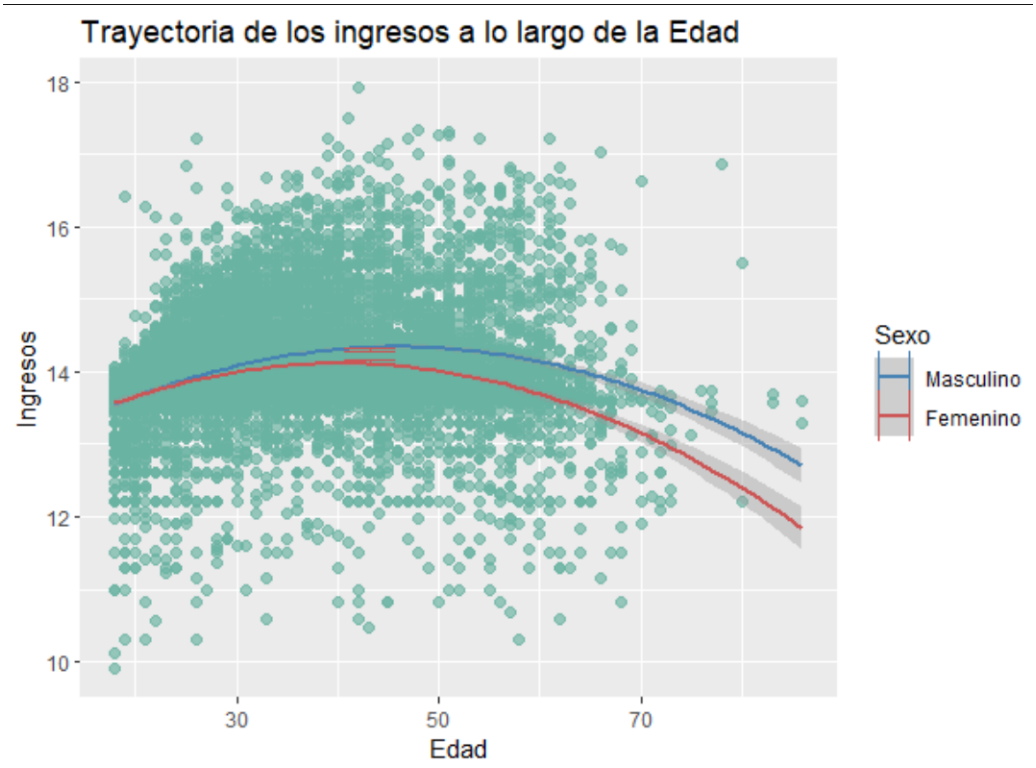
c. Plot the predicted age-wage profile and estimate the implied "peak ages" with the respective confidence intervals by gender.

Para finalizar nuestro análisis por sexo, generamos el Gráfico 13 que representa la trayectoria de los ingresos a través de la edad diferenciado por el sexo de la persona, siguiendo la ecuación:

$$log(w) = \beta_0 + \beta_1 Female + \beta_2 Age + \beta_3 Age^2 + u \tag{3}$$

Las variables corresponden al mismo uso que les hemos dado a lo largo del punto. En principio podemos observar que si se presenta el fenómeno donde a partir de los 43 años la curva empieza a descender para ambos sexos, sin embargo vemos que la trayectoria de las mujeres desciende con mayor rapidez que la trayectoria de los hombres y gracias a los intervalos de confianza nos damos cuenta que los valores en los ingresos si son diferenciales.

Gráfico 13: Peak-age según sexo



El proceso para construir el gráfico es el mismo que en el punto 3. Los resultados utilizados para los intervalos de confianza los encontramos mediante Boostrap y obtuvimos los siguientes resultados.

Tabla 10: Boostrap PeakAge

Peak Age	43.178
Bias	-0.005
SE	0.420

Tabla 11: Boostrap PeakWage Femenino

Peak Wage (en log)	14.137
Bias	-7.83817e-05
SE	0.014

Tabla 12: Boostrap PeakWage Masculino

Peak Wage (en log)	14.299
Bias	0.000
SE	0.014

5. Predicting earnings

a. Segmentación de la muestra en un subconjunto de entrenamiento y otro de validación

Para evaluar las predicciones del ingreso, se pondrá a prueba el rendimiento fuera de la muestra. Para esto se usará el enfoque de validación. En este enfoque, una parte aleatoria de los datos se designa como el conjunto de validación y el modelo se entrena con los datos restantes. Para este caso, se utilizará el 70 % de la muestra como conjunto de entrenamiento y el 30 % restante como la muestra de prueba. Con el conjunto de validación, se evaluará luego la capacidad de predicción del modelo en estudio fuera de muestra. A continuación, se mostrará como se realizo la separación de estos grupos.

```
> set.seed(10101)
> #use 70% of the dataset as a training set and 30% as a test set.
```

```
> sample <- sample(c(TRUE, FALSE), nrow(base2), replace=TRUE, prob=c(0.7,0.3))
> train  <- base2[sample, ]
> test   <- base2[!sample, ]
```

b. Performance de los modelos propuestos.

Tabla 13

Modelos	MSE
MSE modelo age2	0.5435
MSE Modelo fem	0.5435
MSE modelo 1	0.5435
MSE modelo 2	0.4733
MSE modelo 3	0.3295
MSE modelo 4	0.2427
MSE modelo 5	0.2672

Para nuestras estimaciones usamos el método de validación cruzada, el cual es una técnica importante en el aprendizaje, ya que permite evaluar la capacidad generalizada de un modelo. Unas de las principales razones por las cuales usamos el método de validación cruzada es para prevenir el sobre ajuste del modelo y hacer una evaluación robusta de nuestros modelos para así escoger el modelo con mejor comportamiento. Para esta sección se usara el modelo de validación cruzada para 7 modelos, se evaluaran los MSE de estos para elegir el modelos que mejor se comporte y así tener una predicción adecuada y precisa. Para ver en profundidad de los modelos escogidos ver el Anexo 17. La Tabla 13 expone el resultado mencionado anteriormente, el cual muestra que el modelo 4 es el modelo que menor MSE tiene.

c. Discusión de los resultados.

i. Métrica de performance elegida y su justificación

Para juzgar la capacidad de predicción de los modelos, se calculó el error cuadrático medio (MSE) de prueba de cada modelo ya que indica la capacidad predictiva de los modelos fuera de muestra. Dicho error señala cuál es el promedio de los errores cuadráticos de un modelo cuando predice sobre datos sobre los cuales nunca había sido aplicado. De esta forma, un error cuadrático medio (MSE) alto indica que el modelo predice que la variable dependiente tomaría un valor lejano del real, mientras que un MSE bajo indica valores predichos cercanos al real fuera de muestra.

ii. Modelo con menor error de predicción

El modelo que tiene la mejor capacidad de predecir correctamente fuera de muestra el salario es el siguiente modelo, llamado modelo 4:

$$\begin{aligned} ln(w_i) = totalHoursWorked_i + age_i + age_i^2 + formal_i + sex_i + maxEducLevel3_i \\ + maxEducLevel4_i + maxEducLevel5_i + maxEducLevel6_i + maxEducLevel7_i \\ + estrato12_i + estrato13_i + estrato14_i + estrato15_i + estrato16_i + sizeFirm2_i \\ + sizeFirm3_i + sizeFirm4_i + sizeFirm5_i + totalHoursWorked_i * formal_i. \end{aligned}$$

Donde la variable maxEducLevel3 toma el valor de 1 si el mayor nivel educativo de la persona es la primaria incompleta y 0 de lo contrario; la variable maxEducLevel4 toma el valor de 1 si el mayor nivel educativo de la persona es la primaria completa y 0 de lo contrario; la variable maxEducLevel5 toma el valor de 1 si el mayor nivel educativo de la persona es la secundaria incompleta y 0 de lo contrario; la variable maxEducLevel6 toma el valor de 1 si el mayor nivel educativo de la persona es la secundaria completa y 0 de lo contrario; y la variable maxEducLevel7 toma el valor de 1 si el mayor nivel educativo de la persona es la educación terciaria y 0 de lo contrario. Por otro lado, la variable sizeFirm2 toma el valor de 1 si la empresa en la que trabaja la persona tiene de 2 a 5 trabajadores y 0 de lo contrario; la variable sizeFirm3 toma el valor de 1 si la empresa en la que trabaja la persona tiene de 6 a 10 trabajadores y 0 de lo contrario; la variable sizeFirm4 toma el valor de 1 si la empresa en la que trabaja la persona tiene de 11 a 50 trabajadores y 0 de lo contrario; y la variable sizeFirm5 toma el valor de 1 si la empresa en la que trabaja la persona tiene más de 50 trabajadores y 0 de lo contrario. Por último, la variable estrato2 toma el valor de 1 si la persona vive en una vivienda de estrato 2 y 0 de lo contrario; la variable estrato2 toma el valor de 1 si la persona vive en una vivienda de estrato 3 y 0 de lo contrario; la variable estrato4 toma el valor de 1 si la persona vive en una vivienda de estrato 4 y 0 de lo contrario; la variable estrato5 toma el valor de 1 si la persona vive en una vivienda de estrato 5 y 0 de lo contrario; y la variable estrato6 toma el valor de 1 si la persona vive en una vivienda de estrato 6 y 0 de lo contrario.

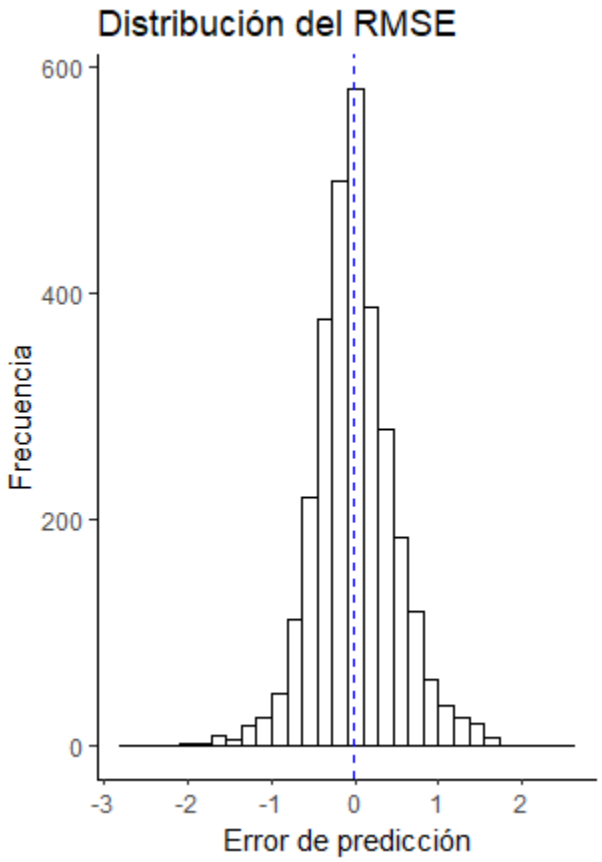
iii. Distribución de los errores de predicción del modelo con menor MSE de prueba

En el Gráfico 14 se muestra la distribución del error cuadrático del modelo con menor MSE fuera de muestra, que se asemeja a una normal. Hay una concentración alrededor del valor 0, lo cual indica que, para el grueso de las observaciones de prueba, el modelo es capaz de predecir correctamente el salario. Sin embargo, hay valores en las colas, especialmente en la cola derecha. Los valores en la cola derecha indican que, para valores de salarios excepcionalmente altos, el valor predicho del modelo dadas las características de esas observaciones es

mucho menor. Lo contrario sucede en la cola izquierda pues, para salarios excepcionalmente bajos, el modelo predice salarios mucho mayores.

El modelo es capaz de predecir el grueso de las observaciones fuera de muestra pero, para salarios muy altos y muy bajos, comete más errores de predicción. El hecho de que haya una alta desigualdad salarial en Colombia se refleja en que, en la muestra de prueba, hay observaciones con salarios excepcionalmente altos y otros con una alta desigualdad salarial. En este sentido, el modelo es bueno prediciendo salarios del grueso de la población más no es capaz de capturar con tal precisión la desigualdad salarial. Por lo tanto, es probable que no sean datos erróneos que la DIAN debería revisar sino que son valores atípicos que al modelo le cuesta predecir correctamente.

Gráfico 14: Distribución del RMSE para el modelo 4



Fuente: elaboración de los autores.

d. LOOCV.

Tabla 14

LOOCV	MSE
LOOCV Modelo 4	0.2327184
LOOCV Modelo 5	0.2593608

Para la siguiente sección, se seleccionaron los dos modelos con menor MSE (error cuadrático medio), que son los modelos 4 y 5 según la Tabla 13. Ahora, se evaluará el performance de los modelos usando el método de LOOCV (leave-one-out cross-validation), en el cual se utiliza una sola observación para la validación y el resto para el entrenamiento del modelo. En la Tabla 14, se encuentra el resultado de las validaciones por LOOCV. Se evidencia que la validación del modelo 4 paso de ser de 0.2431 MSE a 0.2354556, por debajo del error anterior. Por otra parte, en el modelo 5 el MSE paso de 0.2672 a 0.2593608, lo cual significó un cambio drástico respecto al enfoque de validación usado en el anterior inciso. Lo anterior indica que el modelo 5 no es capaz de predecir satisfactoriamente fuera de muestra, mientras que el modelo 4 sí. En términos generales el método LOOCV es más exhaustivo en la evaluación del modelo, ya que cada muestra se utiliza como conjunto de validación exactamente una vez, lo que brinda una evaluación más completa del rendimiento del modelo. Es por eso que el LOOOCV resulta ser más eficiente en términos de MSE pero no en términos computacionales, ya que este es un método muy costoso computacionalmente. Sin embargo, el método de LOOCV tiene una relación muy estrecha con la estadística de influencia, ya que cada vez que se hace una validación cruzada dejando una observación fuera de la muestra, se mide el impacto que tiene esa observación en la predicción del modelo, y como se repite n veces se puede obtener una medida de la estadística de influencia.

Referencias

Blinder, A. (1973). "Wage Discrimination: Reduced Form and Structural Estimates". En: *Journal of Human Resources* 8.4, págs. 436-455. URL: https://EconPapers.repec.org/RePEc:uwp:jhriss:v:8:y:1973:i:4:p:436-455.

Bonet-Morón, J. y Ayala-García, J. (2016). *La brecha fiscal territorial en Colombia*. Inf. téc. 235. Banco de la República. URL: https://www.banrep.gov.co/sites/default/files/publicaciones/archivos/dtser_235.pdf.

DANE (2016). *Ficha Metodológica Gran Encuesta Integrada de Hogares - GEIH*. Inf. téc. DANE. URL: https://www.dane.gov.co/files/investigaciones/fichas/empleo/ficha_metodologica_GEIH-01_V10.pdf.

— (2021). *Brecha salarial de género en Colombia en Colombia 2020*. Inf. téc. DANE. URL: https://www.dane.gov.co/files/investigaciones/notas-estadisticas/oct-2021-nota-estadistica-brecha-salarial-de-genero-en-Colombia.pdf.

Machado, J. A. F. y Mata, J. (2005). "Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression". En: *Journal of Applied Econometrics* 20.4, págs. 445-465. ISSN: 08837252, 10991255. URL: http://www.jstor.org/stable/25146370 (visitado 11-02-2023).

Oaxaca, R. (1973). "Male-Female Wage Differentials in Urban Labor Markets". En: *International Economic Review* 14.3, págs. 693-709. ISSN: 00206598, 14682354. URL: http://www.jstor.org/stable/2525981 (visitado 11-02-2023).

Anexo 15: Estimaciones de la brecha salarial para el ingreso mensual

	Dependent variable:		
	ing_m		wageResidF
	Brecha salarial no condicionada	Brecha salarial condicionada	Brecha salarial condicionada FWL
	(1)	(2)	(3)
female	-0.147^{***}	-0.134^{***}	
	(0.015)	(0.012)	
maxEducLevel3		0.219^{**}	
		(0.089)	
maxEducLevel4		0.232^{***}	
		(0.086)	
maxEducLevel5		0.274^{***}	
		(0.085)	
maxEducLevel6		0.373^{***}	
		(0.084)	
maxEducLevel7		0.902^{***}	
		(0.084)	
age		0.052^{***}	
		(0.003)	
age2		-0.001^{***}	
		(0.00004)	
formal		0.430^{***}	
		(0.015)	
fulltime		0.676^{***}	
		(0.022)	
relab2		0.442^{***}	
		(0.025)	
relab3		-0.241^{***}	
		(0.027)	
relab8		0.044	
		(0.556)	
femaleResidF			-0.134^{***}
			(0.012)
Constant	14.088^{***}	11.389^{***}	0.000
	(0.011)	(0.099)	(0.006)
Observations	9,891	9,891	9,891
R^2	0.009	0.474	0.013
Adjusted R^2	0.009	0.473	0.013
Residual Std. Error	0.762 (df = 9889)	0.556 (df = 9877)	0.555 (df = 9889)
F Statistic	91.527^{***} (df = 1; 9889)	685.134^{***} (df = 13; 9877)	132.063^{***} (df = 1; 9889)

Anexo 16: Estimaciones de la brecha salarial para el ingreso por hora

	Dependent variable:		
	ing_hr		wageResidFhr
	Brecha salarial no condicionada	Brecha salarial condicionada	Brecha salarial condicionada FWL
	(1)	(2)	(3)
female	-0.045^{***}	-0.099^{***}	
	(0.015)	(0.012)	
maxEducLevel3		0.170^{*}	
		(0.089)	
maxEducLevel4		0.212^{**}	
		(0.086)	
maxEducLevel5		0.251^{***}	
		(0.085)	
maxEducLevel6		0.367^{***}	
		(0.084)	
maxEducLevel7		0.946^{***}	
		(0.084)	
age		0.046^{***}	
		(0.003)	
age2		-0.0004^{***}	
		(0.00004)	
formal		0.454^{***}	
		(0.015)	
fulltime		-0.197^{***}	
		(0.022)	
relab2		0.366^{***}	
		(0.025)	
relab3		-0.180^{***}	
		(0.027)	
relab8		0.356	
		(0.558)	
femaleResidFhr			-0.099^{***}
			(0.012)
Constant	8.747^{***}	6.932^{***}	-0.000
	(0.010)	(0.099)	(0.006)
Observations	9,891	9,891	9,891
R^2	0.001	0.415	0.007
Adjusted R^2	0.001	0.414	0.007
Residual Std. Error	0.727 (df = 9889)	0.557 (df = 9877)	0.557 (df = 9889)

Anexo 17

	Dependent variable:				
	lnwage				
	(Modelo 1)	(Modelo 2)	(Modelo 3)	(Modelo4)	(Modelo 5)
totalHoursWorked:sex:age	0.00001				
	(0.00001)				
totalHoursWorked:sex		−0.011***			
		(0.001)			
sex:maxEducLevel3		0.133*			
		(0.077)			
sex:maxEducLevel4		0.252***			
		(0.065)			
sex:maxEducLevel5		0.233***			
		(0.060)			
sex:maxEducLevel6		0.339***			
		(0.052)			
sex:maxEducLevel7		0.971***			
		(0.048)			
totalHoursWorked			−0.010***		
			(0.001)		
age			0.013***	0.009***	0.009***
			(0.001)	(0.001)	(0.001)
sex			0.131***	0.107***	0.313***
			(0.014)	(0.012)	(0.023)
estrato12				0.030	0.033
				(0.020)	(0.021)
estrato13				0.134***	0.129***
				(0.021)	(0.022)
estrato14				0.763***	0.752***
				(0.031)	(0.033)
estrato15				0.932***	0.892***
				(0.049)	(0.051)
estrato16				1.260***	1.209***
				(0.043)	(0.045)
sizeFirm2				0.148***	
				(0.029)	
sizeFirm3				0.234***	
				(0.033)	
sizeFirm4				0.319***	
				(0.031)	
sizeFirm5				0.469***	
				(0.030)	
totalHoursWorked:formal				−0.015***	
				(0.001)	
totalHoursWorked:formal:sex					−0.005***
					(0.0005)
maxEducLevel3			0.186*	0.163*	0.138
			(0.103)	(0.090)	(0.094)
maxEducLevel4			0.207**	0.168*	0.171*
			(0.099)	(0.087)	(0.091)
maxEducLevel5			0.265***	0.218**	0.239***
			(0.099)	(0.086)	(0.091)
maxEducLevel6			0.413***	0.299***	0.360***
			(0.097)	(0.085)	(0.089)
maxEducLevel7			1.007***	0.677***	0.808***
			(0.097)	(0.086)	(0.090)
formal			0.494***	1.011***	0.582***
			(0.017)	(0.040)	(0.020)
Constant	8.716***	8.710***	7.681***	7.188***	7.210***
	(0.012)	(0.012)	(0.104)	(0.091)	(0.093)
Observations	6,866	6,866	6,866	6,866	6,866
R^2	0.0003	0.142	0.420	0.558	0.512
Adjusted R^2	0.0002	0.141	0.419	0.557	0.511
Residual Std. Error	0.723 (df = 6864)	0.670 (df = 6859)	0.551 (df = 6856)	0.481 (df = 6847)	0.506 (df = 6851)