Universidad de los Andes
Facultad de Economía
Big Data y Machine Learning para Economia Aplicada
PhD. Ignacio Sarmiento-Barbieri

Merit Tejeda: 202210104
Celin Hernández: 202210067
Estefania Laborde:

Problem Set 1: Predicting Income

1. Introduction

In the public sector, accurate reporting of individual income is critical for computing taxes. However, tax fraud of all kinds has always been a significant issue. According to the InternalRevenue Service (IRS), about 83.6% of taxes are paid voluntarily and on time in the US.[1]One of the causes of this gap is the under-reporting of incomes by individuals. An income predicting model could potentially assist in flagging cases of fraud that could lead to thereduction of the gap. Furthermore, an income prediction model can help identify vulnerable individuals and families that may need further assistance.

The objective of the problem set is to apply the concepts we learned using "real" world data. For that, we are going to scrape from the following website: https://ignaciomsarmiento.github.io/GEIH2018_sample/. This website contains data for Bogotá from the 2018 *"Medición de Pobreza Monetaria y Desigualdad Report"* that takes information from the GEIH.

1.1. General Instructions

The main objective is to construct a model of individual hourly wages

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

where w is the hourly wage, and X is a matrix that includes potential explanatory variables/predictors. In this problem set, we will focus on $\mathbf{f}(X) = X\beta$.

The final document, in .pdf format, must contain the following sections:

1. *Introduction*. The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.

Introducción

El propósito de este trabajo es encontrar dentro del conjunto de variables seleccionadas, cuales de ellas determinan o aportan al comportamiento del salario por hora. Para abordar la relación entre el salario por hora y sus variables explicativas seguiremos lo planteado por Mincer (1974), el cual es el

[1] See https://www.irs.gov/newsroom/the-tax-gap

paper seminal para el análisis de los ingresos y su distribución en el contexto del mercado laboral.

Siguiendo la propuesta minceriana, Cardoso et all (2016) analizan la brecha salarial entre hombres y mujeres para el caso de Portugal, concluyendo que la clasificación entre empresas y puestos de trabajo puede explicar alrededor de dos quintas partes de la brecha salarial de género. Galassi & Andrada (2009), estiman la relación entre educación e ingresos por regiones geográficas de Argentina y encuentran que existe una relación inversa entre nivel de desarrollo y rendimiento de la educación. Adicionalmente, Oxa & Laoyza (2017), estiman modelos de Mincer para Bolivia, evalúando el rol de capital en la determinación de los ingresos, estos autores señales que la tendencia a la baja de las tasas de retorno a la escolaridad podría explicarse por al aumento de la proporción de población con mayores años de educación formal, lo cual estaría dando una mayor homogeneidad educativa en la población dada la universalización de la educación formal.

Como fuente de datos se dispone de la Gran Encuesta Integrada de Hogares - GEIH 2018, la cual tiene como unos de sus principales objetivos la recopilación de datos del mercado laboral e ingresos de los hogares. Se utilizarán datos de la GEIH 2018 solo para la ciudad de Bogotá, por su robustez estadística en términos de cobertura geográfica, diseño metodológico y su disponibilidad de datos del mercado laboral, la GEIH es idónea para el presente trabajo. Los resultados encontrados siguen los postulados por la literatura económica, donde los rendimientos de la educación tienen un efecto positivo en los ingresos de los trabajadores y la edad y experiencia (lineal) aporta positivamente al ingreso, pero tienen un efecto decreciente en forma cuadrática. Adicionalmente, existe brechas salariales entre ingresos de hombre y mujer, cuyas estimaciones son estadísticamente significativas, pero no tienen mucha significancia económica, ya que la dummy de sexo, no tiene mayor efecto en la media del salario por hora.

2. Data.2 We will use data for Bogot´a from the 2018 “Medición de Pobreza Monetaria y Desigualdad Report” that takes information from the GEIH.

The data set contains all individuals sampled in Bogota and is available at the following website https://ignaciomsarmiento.github.io/GEIH2018 sample/. To obtain the data, you must scrape the website. In this problem set, we will focus only on employed individuals older than eighteen (18) years old. Restrict the data to these individuals and perform a descriptive analysis of the variables used in the problem set. Keep in mind that in the data, there are many observations with missing data or 0 wages. I leave it to you to find a way to handle this data.

When writing this section up, you must:

a) Describe the data briefly, including its purpose, and any other relevant informa tion.

Para estimar la relación entre el salario por hora y las variables explicativas, se utilizarán datos de la Gran Encuesta Integrada de Hogares - GEIH 2018, implementada por Departamento Administrativo Nacional de Estadísticas de Colombia (DANE), la cual es una base de datos de corte transversal. La GEIH tiene como eje central la recopilación de datos de la estructura mercado laboral y los ingresos de los hogares, con una muestra total con cobertura a nivel nacional (DANE, 2019). Para efectos de este taller, se utilizarán datos de la GEIH 2018, solo para la ciudad de Bogotá, que contiene variables originales y construidas, las que permiten identificar y describir características importantes de los

[2] This section is located here so the reader can understand your work, but probably it should be the last section you write. Why? Because you are going to make data choices in the estimated models. And all variables included in these models should be described here.

individuos, empresas y sectores productivos en el ámbito del mercado de trabajo.

Por su robustez estadística en términos de cobertura geográfica, diseño metodológico y su especialidad en compilar datos del mercado laboral, la GEIH es idónea para estimar la relación entre el salario y el conjunto de variables explicativas. Es importante indicar que aun con las fortalezas de esta encuesta, la base presenta una serie de falencias, en vista de la presencia de muchos valores perdidos en algunas variables y datos atípicos en otras, por lo que previo a las estimaciones econométricas, es pertinente y necesario la implementación de un proceso de revisión y depuración de los datos, a fin de contar con información más consistente.

De la base de datos GEIH 2018 y siguiendo lo planteado por Mincer (1974), se seleccionaron un conjunto de variables, las cuales se presentan en la siguiente ecuación:

$$\text{Ln}(w_i) = \beta_0 + \beta_1(Educ_i) + \beta_2(Exp_i) + \beta_3(Exp_i)^2 + \gamma'(Z_i) + \varepsilon_i \qquad (2)$$

Donde:

Log$(\boldsymbol{w_i})$: Es el logaritmo natural del salario real por hora
Educ: Es el nivel educativo de los individuos seleccionados

1 → Ninguna Educación,
2 → Preescolar,
3 → Primaria Incompleta,
4 → Primaria Completa,
5 → Secundaria Incompleta,
6 → Secundaria Completa,
7 → Universitaria

Exp: Se mide por cuánto tiempo se lleva trabajando en esta empresa, negocio, industria, oficina, firma o finca de manera continua.

Para medir la variable experiencia, que está incluida en la ecuación de Mincer (1974), se utilizó como indicador, los resultados de la siguiente pregunta: ¿cuánto tiempo lleva ... Trabajando en esta empresa, negocio, industria, oficina, firma o finca de manera continua? Es importante notar que la variable de experiencia no es observable en la información de encuestas de hogares en Colombia, motivo por el cual es usual construir la experiencia potencial a partir de la edad y la escolaridad, Guataquí, et all (2009).

$\boldsymbol{Z_i}$: Es un vector de controles que incluye las siguientes variables:

- ***Sexo***: Es una variable binaria, $1 =$ hombre y $0 =$ Mujer
- ***Edad***: Es el número de años edad de los Individuos reportados en la GEIH 2018
- ***Horas trabajadas***: Numero de horas trabajadas a la semana
- ***Tamaño de la Empresa***: Medida por el número de empleados por empresa

1 → trabajo solo,
2 → 2 a 3 personas,
3 → 4 a 5 personas,
4 → 6 a 10 personas,
5 → 11 a 19 personas,
6 → 20 a 30 personas,
7 → 31 a 50 personas,

8 → 51 a 100 personas,
9 → 101 a más personas

- ***Sector***: 1 si el trabajador labora en el sector Formal y 0 para el Informal
- ***Estrato***: El estrato socioeconómco de los individuos en la GEIH 2018

1 → Bajo-bajo,
2 → Bajo,
3 → Medio-bajo,
4 → Medio,
5 → Medio-alto,
6 → Alto

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

La obtención de la base de datos se realizó mediante un web scraping de tablas de una lista de urls y luego se combinaron todas estas tablas descargadas en un solo dataframe. Un supuesto para poder importar la base de datos, es que todas las urls proporcionadas contienen tablas con la misma estructura. En este caso PS1 (ignaciomsarmiento.github.io) contiene 10 tablas con las mismas dimensiones.

c) Describe the data cleaning process and

Al importar la base de datos se cuenta con una tabla total que contiene 32,177 observaciones y 178 variables, previo al análisis econométrico, se realizó el siguiente proceso de filtrado, depuración, eliminación e imputación para los valores atípicos, el cual se describe a continuación:

- Excluir de la base de datos los individuos con edades menores a 18 años, con lo cual, la base de datos se redujo en un 25%.
- Eliminar los registros de los individuos que entran en la categoría de desempleados y población económicamente inactiva, alcanzando con esta exclusión un 50% de la base de datos originales.
- Se identificó la variable dependiente (y_salary_m_hu), el salario - real por hora (habitual) - principal occ. (incluye propinas y comisiones), en la revisión de esta variable se encontró que un 40.32% de sus registros son valores perdidos (NA), los cuales fueron excluidos, con lo cual, se obtiene una base de datos de 9,785 observaciones, equivalente a un 30.4% de la base de datos original.
- De las 178 variables disponibles en el set de datos, se conservaron aquellas con un % de valores perdidos (NA) menor al 30%, este proceso dio como resultado 43 variables a analizar.
- Se construyeron dos (2) variables adicionales, remuneraciones extras en forma monetaria y en especie, a partir de la suma de un conjunto de beneficios en términos monetarios y en especie que recibe el trabajador; sin embargo, estas variables presentan una alta variabilidad, asociada a la presencia de muchos valores extremos; por tanto, serán excluidas del análisis de regresión.
- Se analizó e imputó valores para las siguientes variables:
    - *maxEducLevel:* Para los valores perdidos de esta variable, que representaban el 1% del total de las observaciones, se imputó el valor modal.
    - *totalHoursWorked:* Al revisar esta variable se identificó la presencia de valores extremos, el análisis en este caso fue el siguiente, de acuerdo con la normativa colombiana, la jornada de trabajo comprende 48 horas semanales, que a su vez coincide con el promedio de las horas trabajadas; en ese sentido, se calculó un factor umbral para el total de horas trabajadas, que es igual a la media de horas + 2 desviaciones estándar, esta es una regla empírica muy usada en estadísticas, donde se considera que los valores se encuentran dentro de una banda con

dos veces la desviación típica respecto a la media, el valor obtenido se imputó para todos los valores por encima del umbral.

- *p6426(experiencia):* En la revisión de esta variable se detectaron muchos valores extremos, en este sentido, se identificó en la base de datos que la edad máxima es 86, al cual le restamos 18, que son los años de inicio en el mercado laboral, se determinó que el máximo de años de experiencia laboral de un individuo puede alcanzar es 68, este valor fue imputado para todas las observaciones mayores a 68.

d) A descriptive analysis of the data. At a minimum, you should include a descriptive statistics table with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a "dry" list of ingredients.

Se hará un análisis descriptivo de los ingresos reportados por los individuos en la GEIH 2018

Tabla 1: Estadísticas Descriptivas de los Ingresos Reportados en la GEIH 2018

Ingresos	N	Mean	St. Dev.	Mínimo	Máximo
Salario Mensual	9,785.0	1,566,234.0	2,158,107.0	10,000.0	34,000,000.0
Ingreso Monetario 1	9,785.0	1,739,064.0	2,314,670.0	0.0	40,000,000.0
Ingreso Monetario 2	9,771.0	16,861.5	216,344.8	0.0	10,000,000.0
Ingreso Total	9,785.0	1,884,787.0	2,519,838.0	40,000.0	41,333,333.0
Rem. Adicionales en Especie	9,740.0	85,637.7	163,287.8	0.0	6,300,000.0
Rem. Adicionales Monetarias	9,767.0	1,139,646.0	3,285,939.0	0.0	130,000,000.0
Salario por Hora Real	9,785.0	7,984.3	11,629.4	151.9	291,666.7
Ingreso Laboral Mensual	9,785.0	1,757,077.0	2,413,729.0	30,000.0	60,100,000.0
Ingreso Laboral por Hora	9,785.0	8,868.2	12,917.1	326.7	350,583.3

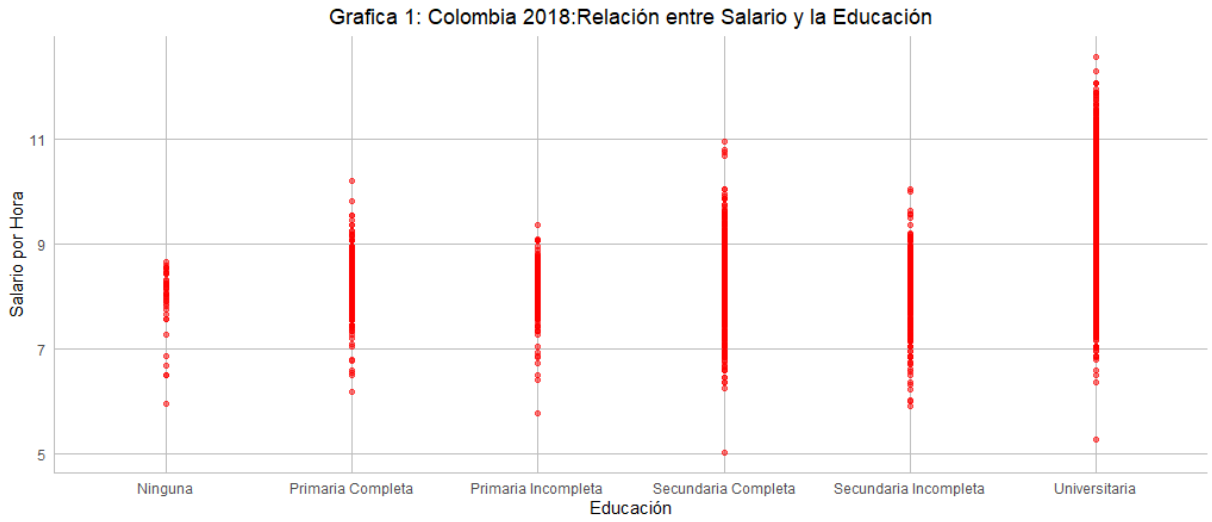
La GEIH 2018 reporta una serie de indicadores que reflejan la fuente de ingresos de los hogares, a excepción del salario por hora real, los demás ingresos están expresados en términos nominales, es decir, su valor describe la composición de efecto precios y volumen. Para obtener el salario por hora real, la encuesta contempla la construcción de un deflactor que permite ajustar el gasto y eliminar las diferencias atribuibles a los precios relativos, el valor de este índice de precios depende de una canasta de bienes que representa los gastos comunes de todos los dominios (DANE, 2019).

Las estadísticas descriptivas presentadas en la tabla 1, denotan una variabilidad alta en estos ingresos, atribuible a la presencia de valores extremos en los datos y a la heterogeneidad de los ingresos, según edad, sexo, estrato, nivel educativo, entre otros. Al observar los ingresos promedios, un aspecto a destacar es que todas las categorías de ingresos están por encima del salario mínimo, que para 2018 era de $ 781,242.0, lo cual pueda reflejar un importante diferencial de ingresos si comparamos con otras regiones de Colombia, según datos del DANE, Bogotá tiene uno de los PIB per cápita más alto de Colombia[3] y dado que solo estamos utilizando información de esta departamento, algunas de las conclusiones de este trabajo puedan carecer de validez externa, al no considerar la dinámica de ingresos de otras zonas del país.

La teoría del capital humano, desarrollada principalmente por Gary Becker (1964), nos indica una relación entre el nivel educativo e ingresos, los datos reportados en esta encuesta se pueden representar en la

3/ https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-departamentales

siguiente gráfica:



Tal y como se esperaría teóricamente, existe una relación directa y creciente entre el ingreso y la educación, la grafica 1 describe que el menor nivel de ingreso salarial por hora (en forma logarítmica) lo reciben los individuos que carecen de educación; en cambio, las personas con educación universitaria presentan los salarios más altos de Bogotá. Adicionalmente, es importante destacar que el cambio en el salario de secundaría completa a universitario es mayor que el cambio observado de primaria completa a secundaría completa, es decir, que un año adicional es educación universitaria, tiene un mayor efecto de un año adicional de educación secundario y primaria.

La siguiente tabla presenta la distribución porcentual de los ingresos, según los niveles educativos en Bogotá

Tabla 2: Estadísticas Descriptivas de Ingreso por Nivel Educativo

Nivel Educativo	Salario por hora	Salario Mensual	Ingreso Total
Ninguno	0.19	0.19	0.20
Primaria Completa	3.66	3.79	3.97
Primaria Incompleta	1.62	1.71	1.76
Secundaria Completa	19.41	20.63	20.95
Secundaria Incompleta	4.81	5.00	5.07
Universitaria	70.31	68.69	68.04

Del total de salarios por hora recibidos por los trabajadores en Bogotá, un 70.3% se concentró en individuos con educación universitaria, similar relación porcentual se observa en los salarios mensuales y en los ingresos totales; definitivamente, los datos indican con claridad y precisión que la educación y sus retornos juegan un rol crucial en el mejoramiento de los ingresos en Bogotá.

No obstante, para un análisis más detallado de los salarios e ingresos de los hogares en Bogotá, es pertinente observar otras dimensiones o incluir otras variables que nos un espectro mas amplio de los determinantes del ingreso, en tal sentido, la siguiente tabla presenta los mismos indicadores de ingreso, considerando otras variables adicionales y definiendo un rango de edades que nos ilustran algunas diferencias importantes.

Tabla 3: Ingresos por Rango de Edades, Educación, Estrato y Tamaño Empresas

Rango de Edad	Salario por hora	Salario Mensual	Ingreso Total	Nivel Educativo Medio	Estrato Predominante	Tamaño Empresas Predominante	% de Individuos
18-25	5,020.3	969,589.4	1,146,842.0	6	2	9	22.1
26-35	6,975.5	1,368,351.0	1,605,467.0	6	2	9	18.4
36-45	8,772.8	1,743,375.7	2,076,154.0	6	2	9	27.2
46-55	9,575.7	1,904,861.9	2,271,153.0	6	2	9	16.9
56-65	9,742.3	1,891,623.7	2,328,023.0	6	3	9	11.9
66-90	10,437.1	1,882,181.9	2,723,351.0	5	2	9	3.5

De la table 3 podemos observar que el ingreso es creciente en la edad; sin embargo, a medida que se incrementa la edad, el crecimiento del salario se vuelve de decreciente, tal y como se esperaría. Para relacionar el nivel educativo con los salarios y las edades, se calculó el promedio del nivel educativo para cada uno de los rangos de edades, en tal sentido, se observa que, en promedio, las personas en Bogotá se encuentran principalmente en la categoría 6, que los identifica con secundaria completa y predominan personas del estrato 2 y que trabajan en empresas grandes, de más de 100 trabajadores[4].

Los estratos 2 y 3 se denominan bajo y medio-bajo, respectivamente y corresponden a segmentos poblacionales que son beneficiarios de subsidios en los servicios públicos domiciliarios[5], la mayoría de personas encontradas en esta encuesta se ubican en esta escala, pero laboran en su mayoría en empresas grandes, lo cual es un factor positivo que debería de incidir en mejores remuneraciones. Otro aspecto a destacar es que el mayor % de los individuos se ubican en el rango de 36-45 años y el ingreso más alto lo ocupa el rango superior de edades, no obstante, este segmento poblacional es de menor participación en la estructura de edades (3.5%), pero concentra los salarios más altos.

Otra arista importante de la distribución de los ingresos esta relacionada con el género de los individuos.

Tabla 4: Estadísticas Descriptivas de Ingreso por Sexo

Sexo	Salario Real por Hora	Salario Nominal Mensual	Ingreso Laboral	Ingreso Total	Cantidad
Hombre	8,178.4	1,651,645.1	9,042.9	1,845,881.8	4,910
Mujer	7,788.7	1,480,209.2	8,692.3	1,667,633.6	4,875

Como se observa en la table 4, existe una cantidad levemente mayor de hombres y en promedio para cada una de estas líneas de ingreso, los hombres reciben una remuneración más alta, respecto a las mujeres, tema que se abordara con mayo detalle en el análisis de la brecha salarial que se presentará más adelante.

Finalmente, la tabla 5 presenta el análisis descriptivo de las variables a utilizar en las regresiones

4/ Para las variables estrato y tamaño de empresa de la tabla 3, se calculó la moda de sus datos.

5 https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-informacion/estratificacion-socioeconomica#preguntas-frecuentes

Tabla 5: Estadísticas Descriptivas Variables Seleccionadas

Variables	N	Mean	St. Dev.	Mínimo	Máximo
Horas trabajadas	9,785	47.989	10.928	1	72
Educación	9,785	6.098	1.009	1	7
Edad	9,785	36.438	11.937	19	86
Experiencia	9,785	20.249	15.813	0	68
Tamaño empresa	9,785	6.439	2.872	1	9
Estrato	9,785	2.513	0.977	1	6

Como se observa en la tabla anterior, la variabilidad de las variables finales, luego del proceso de filtrando y depuración, es bajo y los valores máximos se encuentra dentro de un marco lógico, dadas las características de los individuos en el mercado laboral. La edad promedio es de 36.4 años, como se vio en la tabla 3, este valor se encuentra en el rango de edades de mayor participación en nuestra base de datos final.

En cuanto a la variable experiencia, si bien, conceptualmente no proporcione una métrica precisa de la experiencia laboral; sin embargo, se utilizará como un indicador aproximado de la experiencia laboral, el promedio de esta variable es 20.2 años, lo cual podría ser un poco elevado, si consideramos que aproximadamente el 70% de los individuos de la GEIH 2018 son menores a 45 años de edad.

Para cada una de las ecuaciones siguientes se analizará la significancia estadística a partir de los valores de probabilidad o p_values; adicionalmente, se examinará la significancia económica del parámetro de interés, que dado que estamos construyendo regresiones semi-logaritmica, la significancia económica se define de la siguiente manera:

$$SE = \left(\left(e^{\beta_i} - 1 \right) * \frac{100}{promedio(w_i)} \right) * 100 \qquad (4)$$

Donde:

SE: Significancia Económica
β_i: Es el coeficiente de Interes en la regresión
w_i: Es la variable dependiente, el salario real por hora
e: Es el numero de euler o exponencial (2.178281829)

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

   In this subsection we are going to estimate the *Age-wage profile* for the individuals in this sample:

$$Log(w_i) = \beta_1 + \beta_2 Edad_i + \beta_3 Edad_i^2 + u_i \qquad (3)$$

When presenting and discussing your results, include:

- A regression table.

Tabla 6: Regresión del Perfil de Ingreso por Edad.

	Variable Dependiente
	log(w_hora)
Edad	0.058***
	(0.004)
Edad^2	-0.001***
	(0.000)
Constante	7.429***
	(0.07)
Observaciones	9,785
R^2	0.03

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

- An interpretation of the coefficients and it's significance.

Los resultados de la Tabla 6 indican que un incremento de un año en la edad se traduce en un aumento del 5.8% en el salario por hora y que este resultado es estadísticamente significativo. Sin embargo, al considerar dos años adicionales de edad, el aumento en el salario por hora es del 5.6%, lo que sugiere que conforme aumenta la edad, los ingresos individuales aumentan, pero cada año de edad tiene un efecto sobre los ingresos menor que el anterior, e incluso podría disminuir si el efecto cuadrático es significativo.

Es importante destacar que, en relación con el salario promedio que es 7,984.26 pesos por hora, un año adicional de edad solo representa una desviación del salario respecto a su media del 0.07%, mientras que el efecto de dos años adicionales no genera ningún cambio en relación con la media de los datos. Por su parte, el valor de la constante revela que el salario promedio, independientemente de la edad, asciende a 1,684.13 pesos por hora de trabajo.

- A discussion of the model's in sample fit.

El coeficiente de determinación R^2 es de 0.03, lo que indica que solamente alrededor del 3% de la variabilidad en el salario por hora puede explicarse mediante la edad y la edad al cuadrado. Este valor sugiere que existen otras variables adicionales que posiblemente tienen un impacto más significativo en nuestra variable dependiente además de la edad. Se considera que otros factores podrían estar contribuyendo de manera importante a la explicación del salario por hora como ser: el número de años de educación formal completada, los años de experiencia laboral, el sexo, la edad, la habilidad innata, así como la propia actitud de la persona hacia su trabajo, entre otras.

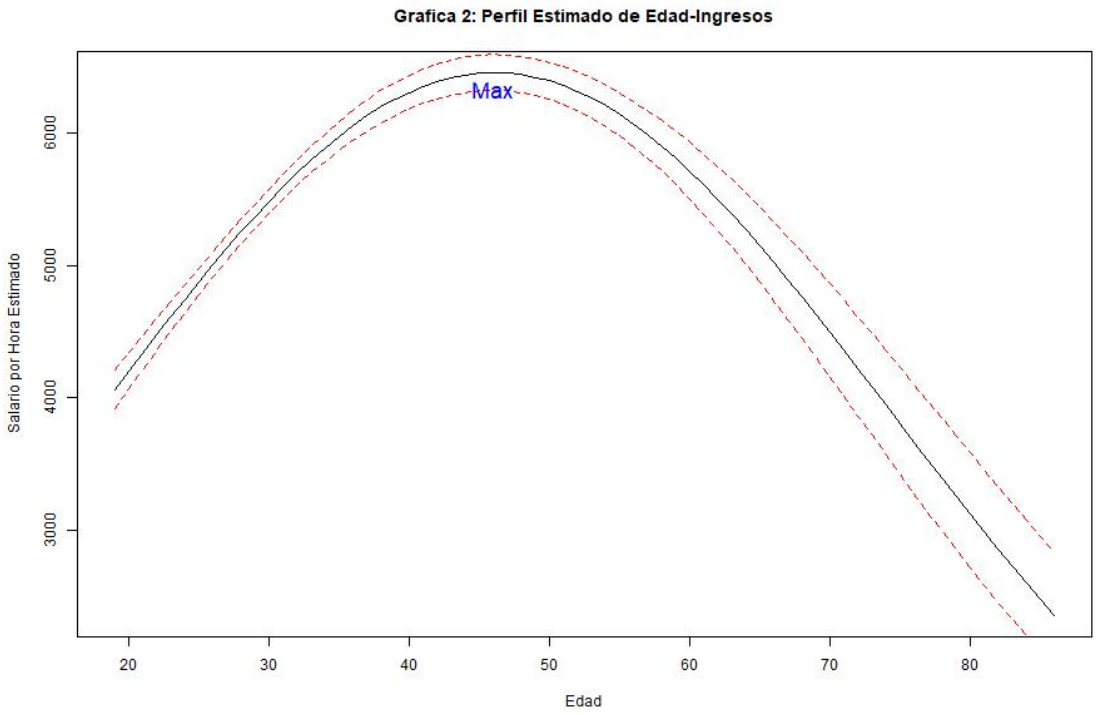
- A plot of the estimated age-earnings profile implied by the above equation. Including a discussion of the "peak age" with it's respective confidence intervals. (Note: Use bootstrap to construct the confidence intervals.)

De la ecuación 3 para obtener la edad máxima, optimizamos derivando respecto a la Edad:

$$\frac{\partial Log(w_i)}{\partial Edad_i} = \beta_1 + 2\beta_2 Edad_i = 0$$

$$2\beta_2 Edad_i = -\beta_1$$

$$Edad_i^* = \frac{-(\beta_1)}{2(\beta_2)}$$



La Gráfica 2 representa el perfil estimado del ingreso en relación con la edad, utilizando la ecuación anterior. Esta ecuación muestra una función cóncava para el salario por hora, lo cual se debe a que el coeficiente $\beta_2$es positivo mientras que el coeficiente β_3 es negativo.

En este contexto, la gráfica ilustra cómo el ingreso aumenta a medida que una persona cumple un año adicional de edad. No obstante, a partir de los 47 años, se observa una disminución en el ingreso. En otras palabras, esta edad marca el punto en el que el ingreso alcanza su valor máximo antes de comenzar a descender.

Es importante destacar que, aunque nuestra estimación puntual sugiere que la edad máxima es de 47 años, existe cierta incertidumbre en torno a este valor debido a la naturaleza de nuestros datos y el modelo utilizado. Por lo tanto, al realizar el ejercicio de Bootstrap para estimar la edad máxima, obtuvimos un intervalo de confianza del 95%, que abarca desde los 45 años hasta los 48 años.

Este intervalo de confianza indica que, con un alto nivel de confianza, se puede decir que la edad en el que se alcanza el salario máximo se encuentra en el rango de 45 a 48 años. Sin embargo, no podemos precisar un valor único dentro de este intervalo debido a la variabilidad en nuestros datos y modelo.

4. *The gender earnings GAP*. Policymakers have long been concerned with the genderwage gap, and is going to be our focus in this subsection.

a) Begin by estimating and discussing the unconditional wage gap:

$$Log(w_i) = \beta_1 + \beta_2 Female + u \quad (4)$$

where *Female* is an indicator that takes one if the individual in the sample isidentified as female.

Tabla 7: Regresión 1. Brecha Salarial Mujer

	Variable Dependiente
	log(w_hora)
Mujer	-0.047***
	(0.015)
Constante	8.648***
	(0.01)
Observaciones	9,785
R^2	0.001

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

A pesar de que los resultados de la regresión de la brecha salarial demuestran que ser mujer está asociado con una reducción del ingreso del 4.7%, y que es estadísticamente significativo, es importante destacar que la desviación con respecto a la media salarial generada por esta variable es relativamente baja, tan solo de -0.058%. Por otro lado, la constante refleja que el ingreso promedio, sin tener en cuenta la variable de género, es de 5,698.74 pesos colombianos.

Por otra parte, el coeficiente de determinación R^2 es de 0.001, lo que sugiere que solo alrededor del 0.01% de la variabilidad el salario por hora puede explicarse mediante la variable mujer. Esto indica que es necesario considerar otras variables de control, como la educación, experiencia laboral, tipo de empresa, etc., para determinar si la brecha salarial efectivamente existe y en qué medida se puede atribuir a estas variables.

b) *Equal Pay for Equal Work?* A common slogan is "equal pay for equal work". One way to interpret this is that for employees with similar worker and job characteristics, no gender wage gap should exist. Estimate a conditional earnings gap incorporating control variables such as similar worker and job characteristics.In this section, estimate the conditional wage gap:

i. First, using FWL

Tabla 8: Regresión Perfil de Ingreso por Edad: FWL

	Variable Dependiente	
	Log(w_hora)	Log(w_hora)

	(1)	(2)
Mujer	-0.139***	
	(0.011)	
Edad	0.041***	
	(0.003)	
Edad^2	-0.0004***	
	(0.000)	
Educación	0.140***	
	(0.006)	
Experiencia	0.012***	
	(0.001)	
Experiencia^2	-0.0002***	
	(0.000)	
Tamaño Empresa	0.053***	
	(0.002)	
Horas Trabajadas	-0.013***	
	(0.000)	
Sector	0.150***	
	(0.016)	
Estrato	0.284***	
	(0.006)	
Mujer_Resid		-0.139***
		(0.011)
Constante	0.284***	0.00
	(0.065)	(0.005)
Observaciones	9,785	9,785
R^2	0.492	0.017

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

Las variables que se emplean en el cálculo de la brecha salarial se fundamentan en la ecuación de Mincer (1974), la cual toma en consideración factores como la edad, nivel educativo, experiencia laboral, tamaño de la empresa, horas trabajadas, sector y estrato. Esta selección de variables se basa en una revisión exhaustiva de la literatura relacionada con el tema: Cardoso et al. (2016), Departamento Administrativo Nacional de Estadística(DANE,2020), López Lapo & Sarmiento Castillo (2019), Nazier (2017), Es importante destacar que, debido a limitaciones en la base de datos, no fue posible incluir algunas variables relevantes, como el estado civil y el número de hijos, en el análisis. Los resultados de esta estimación se encuentran en la Tabla 8 donde se puede observar que al controlar el salario por hora, por otras variables como condiciones laborales, el nivel educativo o edad se sigue capturando la presencia de una brecha salarial estadísticamente significativa ya que el coeficiente asociado a la mujer es -0.139 que indica una reducción del ingreso del 12%; en cuanto a su significancia económica dicha variable genera una desviación del salario promedio de tan solo -0.163%. Por otro lado, las variables de edad, educación y experiencia también resultan ser estadísticamente significativas y tienen el signo esperado según lo planteado

por Mincer. Sin embargo, llama la atención el comportamiento de las horas trabajadas, ya que contrariamente a lo esperado, muestran una reducción en el salario por hora a medida que aumentan las horas trabajadas. Esto podría indicar la presencia de errores en la digitación de la base de datos.

En cuanto al tamaño de la empresa, se observa un coeficiente con significancia estadística, lo cual indica que a medida que aumenta el tamaño de la empresa, el salario por hora se incrementa en un 5.3%. Este hallazgo sugiere que trabajar en empresas más grandes se asocia con salarios más elevado. Por otra parte, la variable sector que mide si un individuo trabaja en el sector formal o informal, muestra un coeficiente es positivo y estadísticamente significativo, nos revela que trabajar en el sector formal implica un efecto de 15.0% en el salario por hora. Finalmente, la variable estrato, con un coeficiente de regresión de 0.284, nos indica que un trabajador puede acceder a una mejor remuneración en la medida que proviene de un estrato socioeconómico más alto.

Al aplicar la metodología FWL (Frisch-Waugh-Lowell), se logra capturar de manera efectiva el efecto de la variable "Mujer_Resid" sin necesidad de incluir el resto de los controles, y este coeficiente sigue siendo igual a -0.122. Sin embargo, es importante notar que el R^2, que mide la capacidad de las variables explicativas para explicar la variabilidad del salario, es menor en comparación con el modelo estimado mediante MCO. Cuando se incorporan las variables explicativas, el R^2 es de 0.361, lo que indica que estas variables explicativas, respaldadas por la revisión de la literatura, explican el 36.1% de la variabilidad en el salario.

ii. Second, using FWL with boostrap. Compare the estimates and the standard errors.

Tabla 9: Regresión del Perfil de Ingreso por Edad: FWL y FWL Boostrap

	Variable Dependiente			
	Log(w_hora)	Log(w_hora_ Resid)	Log(w_hora)	Log(w_hora_ Resid)
	(1)	(2)	(3)	(4)
Mujer	-0.139***		-0.142***	
	(0.011)		(0.011)	
Edad	0.041***		0.042***	
	(0.003)		(0.003)	
Edad^2	-0.0004***		-0.0004***	
	(0.000)		(0.000)	
Educación	0.140***		0.147***	
	(0.006)		(0.006)	
Experiencia	0.012***		0.012***	
	(0.001)		(0.001)	
Experiencia^2	-0.0002***		-0.0002***	
	(0.000)		(0.000)	
Tamaño Empresa	0.053***		0.053***	
	(0.002)		(0.002)	
Horas trabajadas	-0.013***		-0.014***	
	(0.000)		(0.000)	

Sector	0.150***		0.151***	
	(0.016)		(0.016)	
Estrato	0.284***		0.277***	
	(0.006)		(0.006)	
Mujer_Resid		-0.139***		-0.142***
		(0.011)		(0.011)
Constante	6.269***	0.000	6.277***	0.000
	(0.065)	(0.005)	(0.065)	(0.005)
Observaciones	9,785	9,785	9,785	9,785
R^2	0.492	0.017	0.493	0.018

Nota 1: *p<0.1; **p<0.05; ***p<0.01
Nota 2: (1) y (2) con muestra única y (3) y (4) con Bootstrap

En la Tabla 9 se presentan los resultados del cálculo de la brecha salarial utilizando la metodología FWL con única muestra y con Bootstrap. Al comparar los resultados de ambas regresiones, se observa que, en la regresión con Bootstrap, el coeficiente asociado al género muestra un valor muy similar al obtenido en la regresión con única muestra, siendo ligeramente mayor en apenas 0.003 puntos porcentuales. Además, es importante destacar que este coeficiente es estadísticamente significativo en ambos casos. El coeficiente implica una reducción del salario del 14.2% en el caso de Bootstrap, en términos de significancia económica, genera una desviación del salario medio del -0.191%; es decir, mueve la media condicional aproximadamente en una quinta parte.

Es relevante mencionar que, en términos de los errores estándar de los coeficientes, no se observan diferencias estadísticamente significativas entre los modelos con y sin Bootstrap. En otras palabras, los resultados de ambos modelos son consistentes en términos de la significancia estadística de las variables explicativas. Al igual que en el modelo estimado con única muestra, en este caso, las demás variables explicativas siguen siendo estadísticamente significativas y están en consonancia con la teoría, con la excepción de las horas trabajadas. Por otro lado, el coeficiente de determinación R^2 es igual a 0.372, siendo ligeramente mayor en comparación con el caso de FWL con única muestra, con una diferencia de solo 0.011 puntos porcentuales.

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

$$Ln(w_i) = \beta_0 + \beta_1 Edad_i + \beta_2 Edad_i^2 + \beta_3 (Edad_i * Mujer_i) + \beta_4 \left(Edad_i^2 * Mujer_i\right) + u_i \quad (4)$$

Para obtener la edad máxima, optimizamos derivando respecto a la Edad:

$$\frac{\partial Ln(w_i)}{\partial Edad_i} = \beta_1 + 2\beta_2 Edad_i + \beta_3 (Mujer_i) + 2\beta_4 (Edad_i * Mujer_i) = 0$$

$$2\beta_2 Edad_i + 2\beta_4 (Edad_i * Mujer_i) = -\beta_1 - \beta_3 (Mujer_i)$$

$$Edad_i (2\beta_2 + 2\beta_4 Mujer_i) = -\beta_1 - \beta_3 Mujer_i$$

$$Edad_i^* = \frac{-(\beta_1 + \beta_3 Mujer_i)}{2(\beta_2 + \beta_4 Mujer_i)}$$

Dado que la variable dummy es: mujer $= 1$ y hombre $=0$; por tanto, se pueden obtener las siguientes edades máximas, según género:

$$Edad\ Max(Mujer)^{*} = \frac{-(\beta_1 + \beta_3)}{2(\beta_2 + \beta_4)}$$

$$Edad\ Max(Hombre)^{*} = \frac{-\beta_1}{2\beta_2}$$

Los resultados presentados en la Tabla 5 proporcionan la información sobre la ecuación anterior. En este análisis, se destaca que las variables edad y edad al cuadrado edad^2 son estadísticamente significativas, lo que sugiere que tienen un impacto en el salario por hora. Además, los signos de estos coeficientes son coherentes con la teoría económica. Específicamente, se observa que un año adicional de edad se asocia con un incremento del 6.2% en el salario por hora. Sin embargo, es interesante notar que este incremento disminuye ligeramente cuando se consideran dos años adicionales de edad, donde el salario por hora aumenta en un 6%; no obstante, el impacto de ambas variables en el salario por hora no resulta importante dado que solo generan una desviación de la media de aproximadamente 0.08%.

Por otra parte, las variables mujer y la interacción edad*mujer no muestran significancia estadística en relación con el salario por hora. Esto sugiere que, en este contexto particular, la pertenencia de género y la interacción lineal entre la edad y el género no tienen un impacto estadísticamente significativo en el ingreso.

Sin embargo, la interacción de edad^2*Mujer muestra significancia estadística con un coeficiente negativo de -0.0002. Este hallazgo sugiere una relación no lineal entre la edad y el salario por hora en el caso de las mujeres. Conforme la edad de las mujeres aumenta en algún punto, el salario por hora tiende a disminuir. Es importante señalar que, aunque esta interacción es estadísticamente significativa, su impacto en el salario por hora es relativamente pequeño, generando una desviación del salario medio de tan solo -0.00025%.

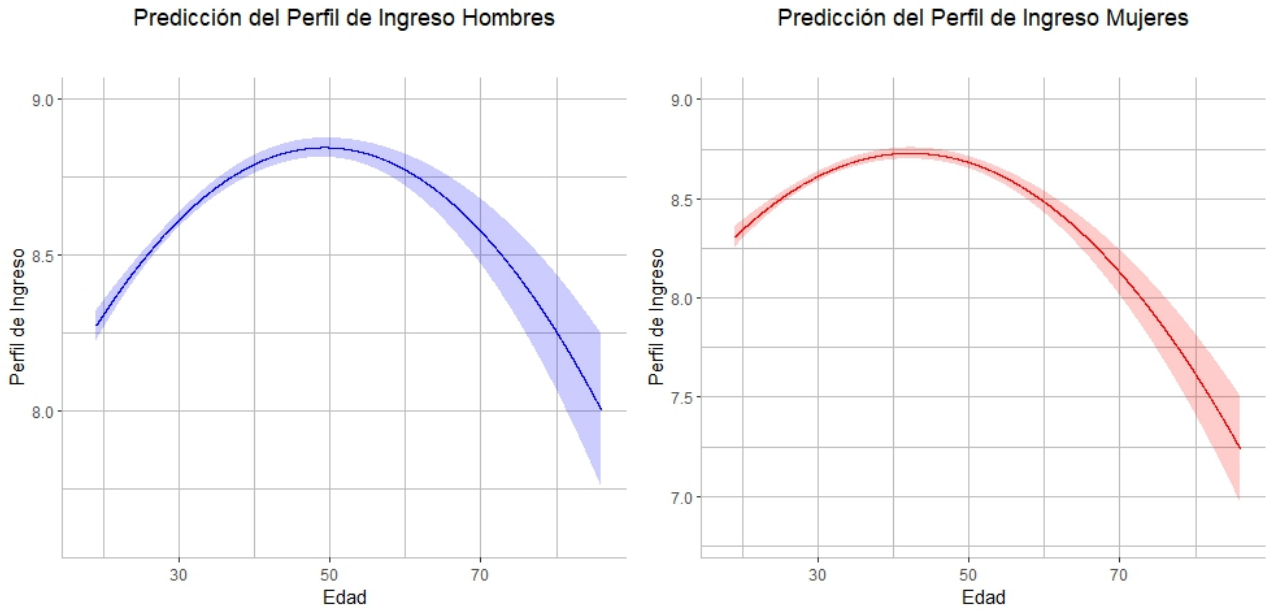
Tabla 10: Regresión del Perfil de Ingreso por Edad

	Variable Dependiente
	Log(w_hora)
Edad	0.062***
	(0.005)
Edad^2	-0.001***
	(0.000)
Mujer	0.009
	(0.139)
Edad: Mujer	0.004
	(0.007)
Edad^2: Mujer	-0.0002*
	(0.000)
Constante	7.324***
	(0.095)

Observaciones	9,785
R^2	0.044

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

Grafica 3



Por otra parte, al analizar el perfil de ingresos según el género, se destacan patrones interesantes. En el caso de los hombres, se observa un incremento en el salario por hora desde el inicio de su vida laboral hasta alcanzar su punto máximo a los 49 años. Sin embargo, a partir de este punto, se inicia una disminución gradual de los ingresos a medida que los hombres tienen un año adicional de edad. En contraste, el perfil de ingresos de las mujeres muestra una dinámica distinta. Si bien experimentan un incremento en sus ingresos durante los primeros años de su vida laboral, lo más relevante es que llegan al máximo nivel de ingresos a una edad más temprana, específicamente a los 42 años. Este hallazgo subraya una distinción significativa, ya que las mujeres alcanzan el máximo nivel de ingresos en un momento previo en sus trayectorias laborales en comparación con los hombres. Posteriormente, a partir de dicha edad, sus ingresos también tienden a decrecer con el paso de los años.

Al calcular el intervalo de confianza de la diferencia en edades máximas entre hombres y mujeres a partir de 1,000 muestras Bootstrap, se obtiene un intervalo de [3, 11], que proporciona el rango de diferencia en edades máximas según el perfil de ingreso entre hombres y mujeres en la población de Bogotá, con un nivel de confianza del 95%. Este intervalo indica que, según los datos y el análisis realizados en Bogotá, es probable que la diferencia en edades máximas esté dentro del intervalo de 3 a 11 años, aunque no se puede determinar con certeza el valor exacto de esta diferencia.

Es importante tener en cuenta que estos resultados y conclusiones se aplican específicamente a la población de Bogotá y no deben extrapolarse automáticamente a otras poblaciones o regiones. Dado que la encuesta se limitó a Bogotá para el año 2018, la validez externa de los hallazgos puede estar restringida a esta área geográfica y no se puede asumir para otros contextos.

When presenting and discussing your results, include:

- A regression table, with the estimates side by side of the conditional and unconditional wage
- gaps, highlighting the coefficient of interest. "Nuisance" controls, should not be included in the table but dutifully noted.[6]
- An interpretation of the "Female" coefficients, a comparison between the models, and the in-sample fit.
- A discussion about the implied peak ages and their statistical similarity/difference.
- A thoughtful discussion about the unconditional and conditional wage gap, seeking to answer if the changes in the coefficient are evidence of a selection problem,a "discrimination problem," a mix, or none of these issues.

5. *Predicting earnings. In the previous sections, you estimated some specifications with inference in mind. In this subsection, we will evaluate the predictive power of these specifications.*

a) Split the sample into two: a training (70%) and a testing (30%) sample. (Don'tforget to set a seed to achieve reproducibility. In R, for example you can use set. seed(10101), where 10101 is the seed.)

b) Report and compare the predictive performance in terms of the RMSE of all the previous specifications with at least five (5) additional specifications that explore non-linearities and complexity.

c) In your discussion of the results, comment:

   i. About the overall performance of the models.

   ii. About the specification with the lowest prediction error.

   iii. For the specification with the lowest prediction error, explore those observations that seem to "miss the mark." To do so, compute the prediction errors in the test sample, and examine its distribution. Are there any observations in the tails of the prediction error distribution? Are these outliers potential people that the DIAN should look into, or are they just the product of a flawed model?

d) LOOCV. For the two models with the lowest predictive error in the previous section, calculate the predictive error using Leave-one-out-cross-validation (LOOCV). Compare the results of the test error with those obtained with thevalidation set approach and explore the potential links with the influence statis- tic. (Note: when attempting this subsection, the calculations can take a long time, depending on your coding skills, plan accordingly!)

2. Additional Guidelines

- Turn a .pdf document in Bloque Neón.

[6] Tip: Look how applied papers construct their results tables. These papers usually present comparableresults in the same table with coefficients side by side, which helps the reader follow the discussion.

- The document must include a link to your GitHub Repository.
  - The repository must follow the template.
  - The README should help the reader navigate your repository. A good README helps your project stand out from other projects and is the first file a person sees when they come across your repository. Therefore, this file should be detailed enough to focus on your project and how it does it, but not so long that it loses the reader's attention. For example, Project Awesome has a curated list of interesting READMEs.
  - Include brief instructions to fully replicate the work.
  - The main repository branch should show at least five (5) substantial contributions from each team member.
  - The code has to be:
    - Fully reproducible.
    - Readable and include comments. In coding, like in writing, a good codingstyle is critical. I encourage you to follow the tidyverse style guide.
- Tables, figures, and writing must be as neat as possible. Label all the variables included. If you have something in your figures or tables, I expect they are addressed in the text. Tables must follow the AER format.