Universidad de los Andes

Facultad de Economía

Big Data y Machine Learning para Economía

Predicción de Ingresos y Brecha Salarial en Bogotá

Profesor: Ignacio Sarmiento

Autor: [Lina María Alvarez y Sara Rocio Rojas]

Bogotá, Colombia

March 4, 2025

1 Introducción

El adecuado reporte del ingreso individual es un elemento clave para el diseño y ejecución de políticas públicas, en especial en lo relacionado con la tributación y la asignación de recursos para programas sociales. La evasión fiscal y la subdeclaración de ingresos siguen representando desafíos sustanciales. Según el Servicio de Impuestos Internos (IRS, 2022), aproximadamente el 83.6% de los impuestos en Estados Unidos se pagan voluntaria y puntualmente, lo que indica la existencia de una brecha de evasión que afecta la capacidad del Estado para financiar bienes y servicios públicos. En América Latina, la situación es aún más compleja debido a la alta informalidad laboral y la falta de mecanismos eficientes para verificar los ingresos reportados (CEPAL, 2021).

El problema de la subdeclaración no solo afecta la capacidad fiscal de los gobiernos, sino que también distorsiona la información económica, lo que dificulta la focalización de políticas sociales para los sectores más vulnerables. A esto se suma la persistente brecha salarial de género, un fenómeno ampliamente documentado en la región. Según el Banco Mundial (2022), en América Latina las mujeres perciben, en promedio, el 70% del salario de los hombres, incluso cuando poseen niveles educativos y experiencia laboral comparables. En el caso colombiano, esta brecha es más pronunciada en mujeres con hijos, quienes experimentan una penalización salarial de hasta 7 puntos porcentuales adicionales (ONU Mujeres, 2022).

Este estudio emplea información de la Encuesta Integrada de Hogares (GEIH) 2018 para Bogotá, obtenida mediante técnicas de web scraping a partir de la base de datos pública *GEIH 2018 Sample*. A partir de estos datos, se explorará la relación entre edad e ingresos, se analizará la brecha salarial de género y se implementarán modelos de predicción de ingresos individuales. La investigación se basa en tres preguntas fundamentales: cómo varía el salario a lo largo del ciclo de vida laboral, hasta qué punto persisten diferencias salariales de género después de controlar por educación y experiencia, y cuáles son los principales determinantes de los ingresos en el mercado laboral de Bogotá.

2 Datos y Proceso de Obtención

Los datos utilizados en este estudio provienen de la Gran Encuesta Integrada de Hogares (GEIH), administrada por el Departamento Administrativo Nacional de Estadística (DANE). Dado que el acceso directo a los microdatos requiere autorización específica, se utilizó una base pública disponible en la web, que contiene una muestra representativa de la población ocupada en Bogotá. La recopilación de estos datos implicó la implementación de técnicas de web scraping debido a la falta de una API estructurada para su descarga directa.

Para la extracción, se analizó la estructura del sitio web y se identificaron múltiples páginas donde los datos estaban almacenados en formato HTML. Se utilizó la librería `rvest` en R para automatizar la recolección de información, accediendo de manera secuencial a las páginas y extrayendo las tablas relevantes. Posteriormente, los datos fueron consolidados en un único `data.frame` y exportados a formatos CSV y Excel para su procesamiento.

El proceso presentó desafíos técnicos, como la variabilidad en la estructura HTML de algunas páginas y la presencia de valores faltantes en variables clave. En particular, se detectaron observaciones con salarios reportados como cero y registros con información

incompleta en educación y ocupación. Para garantizar la calidad del análisis, se implementó un riguroso proceso de limpieza y transformación de datos, el cual se detalla en la siguiente sección.

3 Limpieza y Transformación de Datos

Con el propósito de obtener un conjunto de datos adecuado para el análisis, se llevaron a cabo diversas etapas de limpieza. Inicialmente, se eliminaron registros duplicados y variables irrelevantes. Posteriormente, se restringió la muestra a individuos mayores de 18 años que reportaron estar ocupados en el mercado laboral, de acuerdo con la variable `ocu`. Además, se excluyeron observaciones con valores faltantes en salario y aquellas con ingresos iguales a cero, dado que estos datos no aportaban información relevante para el estudio.

Asimismo, se realizaron transformaciones clave sobre las variables más relevantes. El salario mensual fue convertido a su logaritmo natural para mejorar la interpretación econométrica y reducir la asimetría en su distribución. Se incluyó la variable `age_sq`, correspondiente al cuadrado de la edad, para capturar efectos no lineales en la relación entre edad e ingresos. También se recodificó el nivel educativo en categorías ordenadas y se generaron grupos etarios para facilitar el análisis descriptivo.

En términos metodológicos, estos procedimientos aseguran que los datos sean representativos y permitan una interpretación robusta de los resultados. La aplicación de filtros y transformaciones garantiza que el análisis no se vea afectado por inconsistencias en la recolección de información. En la siguiente sección, se presentará un análisis detallado de las estadísticas descriptivas de las principales variables incluidas en el estudio.

4 Marco Conceptual

El estudio de los determinantes del ingreso y la brecha salarial ha sido ampliamente abordado en la literatura económica y laboral. A lo largo de los años, diversos factores han sido identificados como fundamentales para explicar las diferencias en los salarios de los individuos en el mercado de trabajo. Entre estos factores, la edad, el nivel educativo, el género, el estrato socioeconómico y la informalidad laboral juegan un papel determinante en la configuración de los ingresos laborales y la movilidad económica de la población.

La edad y la experiencia laboral han sido tradicionalmente consideradas como factores clave en la determinación del salario. Desde la teoría del capital humano desarrollada por Becker (1964), se ha argumentado que los ingresos laborales tienden a aumentar con la edad debido a la acumulación de habilidades y conocimientos específicos del trabajo. Sin embargo, diversos estudios han encontrado que esta relación no es lineal, sino que sigue una trayectoria en forma de U invertida, donde los salarios aumentan en las primeras etapas de la vida laboral, alcanzan un punto máximo y luego tienden a estabilizarse o decrecer ligeramente a medida que los trabajadores envejecen (Mincer, 1974). En este sentido, la edad no solo está asociada con el nivel de ingresos, sino también con la estabilidad y la capacidad de negociación en el mercado de trabajo (Fedesarrollo, 2023).

Otro de los determinantes más relevantes en la estructura salarial es el nivel educativo. La literatura ha demostrado que los trabajadores con mayor formación académica tienden a recibir salarios más altos debido a la mayor productividad y competitividad en

el mercado laboral (Psacharopoulos Patrinos, 2018). En América Latina, la educación ha sido identificada como un factor clave para la reducción de la desigualdad y la promoción del crecimiento económico (CEPAL, 2021). Sin embargo, aunque la relación entre educación e ingresos es positiva, también se ha evidenciado que el retorno de la educación puede variar según el contexto económico y social de cada país (Zárate Fernández, 2019). En Colombia, estudios recientes han mostrado que las personas con educación universitaria pueden ganar hasta el doble que aquellas con educación secundaria, lo que refleja la importancia del capital humano en la determinación del ingreso (DANE, 2023).

La brecha salarial de género continúa siendo un fenómeno persistente en la mayoría de los mercados laborales. En América Latina, las mujeres ganan en promedio un 17

Otro factor clave en la estructura salarial es el estrato socioeconómico. En Colombia, el sistema de estratificación ha sido utilizado como un indicador de las condiciones de vida de los hogares y ha demostrado estar estrechamente relacionado con las oportunidades laborales y los niveles de ingreso (DANE, 2023). Estudios han encontrado que las personas provenientes de estratos bajos enfrentan mayores barreras para acceder a empleos formales y bien remunerados, lo que contribuye a la persistencia de la desigualdad económica en el país (López Peña, 2021). A su vez, la movilidad social intergeneracional sigue siendo un desafío, ya que las diferencias en acceso a educación de calidad y redes de contactos influyen en la probabilidad de alcanzar mejores niveles salariales (Bourguignon, 2015).

Finalmente, la informalidad laboral constituye uno de los principales retos en la configuración de los ingresos en América Latina. La Organización Internacional del Trabajo (OIT, 2022) estima que alrededor del 50

En conclusión, la determinación de los ingresos laborales es un fenómeno complejo que depende de múltiples factores estructurales y sociales. La edad, el nivel educativo, el género, el estrato socioeconómico y la informalidad son variables clave para entender la dinámica salarial en Bogotá y, en general, en los mercados laborales de América Latina. Comprender estas relaciones permite no solo identificar las principales brechas existentes, sino también diseñar políticas públicas orientadas a la reducción de la desigualdad y la mejora de las condiciones laborales en la región.

5 Estadística Descriptiva

El análisis de la estadística descriptiva permite entender la distribución de los ingresos laborales en Bogotá y la manera en que varían según diferentes características sociodemográficas. En esta sección, se presentan estadísticas descriptivas de las principales variables del estudio, incluyendo salario, género, edad, nivel educativo, afiliación a seguridad social e informalidad laboral. Estos análisis proporcionan una base fundamental para la posterior modelación de los determinantes del ingreso y la brecha salarial.

5.1 Distribución del Salario

El salario es la variable central de este estudio, ya que representa la principal fuente de ingresos de los trabajadores y está determinado por una combinación de factores individuales y estructurales. En la muestra analizada, el salario promedio mensual es de COP 1,566,234, con una mediana de COP 900,000. La dispersión de los salarios es considerablemente alta, con una desviación estándar de COP 2,158,107, lo que indica una variabilidad importante en los ingresos percibidos. El salario mínimo registrado en la muestra es de COP 10,000, mientras que el máximo alcanza los COP 34,000,000.

Table 1: Estadísticas descriptivas del salario

Estadística	Salario (COP)
Media	1,566,234
Mediana	900,000
Desv. Est.	2,158,107
Mínimo	10,000
Máximo	34,000,000

5.2 Diferencias Salariales por Género

La distribución del salario por género evidencia la presencia de una brecha salarial en detrimento de las mujeres. En promedio, los hombres tienen un salario mensual de COP 1,651,645, mientras que las mujeres perciben un ingreso promedio de COP 1,480,209. La mediana salarial para los hombres es de COP 1,000,000, mientras que para las mujeres es de COP 800,000. Además, la dispersión del salario es mayor en las mujeres, con una desviación estándar de COP 2,298,418, en comparación con COP 2,003,456 en los hombres.

Estos resultados reflejan una desigualdad persistente en el mercado laboral, donde las mujeres tienden a recibir menores remuneraciones que los hombres, incluso cuando desempeñan funciones similares. Esta brecha salarial podría explicarse por múltiples factores, como la segregación ocupacional, diferencias en la cantidad de horas trabajadas, barreras en el acceso a puestos mejor remunerados y la penalización por responsabilidades de cuidado. Estos hallazgos coinciden con la evidencia previa sobre la disparidad salarial de género en Colombia y América Latina, donde las mujeres suelen enfrentar mayores obstáculos para acceder a ingresos equivalentes a los de sus pares masculinos.

Table 2: Estadísticas descriptivas del salario por género

Género	Media Salario (COP)	Mediana Salario (COP)	Desv. Est. (COP)	Observaciones
Mujeres (0)	1,480,209	800,000	2,003,456	4,875
Hombres (1)	1,651,645	1,000,000	2,298,418	4,910

5.3 Relación entre Edad y Salario

El análisis de la edad permite observar cómo los salarios evolucionan a lo largo del ciclo de vida laboral. En la muestra analizada, se encuentra que el grupo de edad entre 35 y 44 años presenta el mayor salario promedio (COP 1,906,857), mientras que los trabajadores más jóvenes (18-24 años) tienen los ingresos más bajos (COP 928,074). Esto es consistente con la teoría del capital humano, según la cual los trabajadores acumulan habilidades y experiencia con el tiempo, lo que se traduce en mayores ingresos. Sin embargo, después de los 55 años, el salario promedio tiende a disminuir ligeramente.

Table 3: Estadísticas descriptivas del salario por edad

Grupo de Edad	Media Salario (COP)	Mediana Salario (COP)	Desv. Est. (COP)	Observaciones
18-24	928,074	800,000	630,982	1,683
25-34	1,430,138	1,000,000	1,371,453	3,280
35-44	1,906,857	1,100,000	2,591,706	2,315
45-54	1,836,327	900,000	2,923,044	1,583
55+	1,895,579	800,000	3,021,978	924

5.4 Impacto del Nivel Educativo en los Ingresos

El salario promedio aumenta con el nivel educativo. Los trabajadores con el menor nivel educativo tienen un ingreso promedio de COP 638,970, mientras que aquellos con educación superior alcanzan, en promedio, COP 2,369,089. Esta diferencia refleja la importancia del capital humano en la determinación del ingreso y la necesidad de políticas que promuevan el acceso a educación superior.

Table 4: Estadísticas descriptivas del salario por nivel educativo

Nivel Educativo	Media Salario (COP)	Mediana Salario (COP)	Desv. Est. (COP)	Observaciones
1	638,970	781,242	351,490	45
2	807,730	781,242	350,708	324
3	848,763	781,242	450,044	685
4	831,706	781,242	361,941	921
5	939,124	800,000	521,628	3,366
6	2,369,089	1,100,000	2,966,671	4,413

5.5 Afiliación a Seguridad Social

El acceso a la seguridad social influye en los ingresos laborales. En esta muestra, los trabajadores afiliados al régimen contributivo tienen un ingreso promedio de COP 1,662,948, mientras que aquellos en el régimen subsidiado ganan, en promedio, COP 2,283,150. Llama la atención que los afiliados al régimen subsidiado reporten salarios más altos, lo que puede estar relacionado con inconsistencias en la declaración de ingresos o con características particulares del grupo de estudio.

Table 5: Estadísticas descriptivas del salario por afiliación a seguridad social

Afiliación	Media Salario (COP)	Mediana Salario (COP)	Desv. Est. (COP)	Observaciones
1	1,662,948	970,000	2,246,137	8,199
2	2,283,150	1,900,000	2,140,382	284
3	665,906	720,000	326,318	733
9	8,050,000	8,050,000	9,828,784	2

5.6 Visualización de los Datos

Para complementar el análisis descriptivo, se presentan a continuación las visualizaciones de los datos, que permiten observar las distribuciones y comparaciones entre las variables clave.

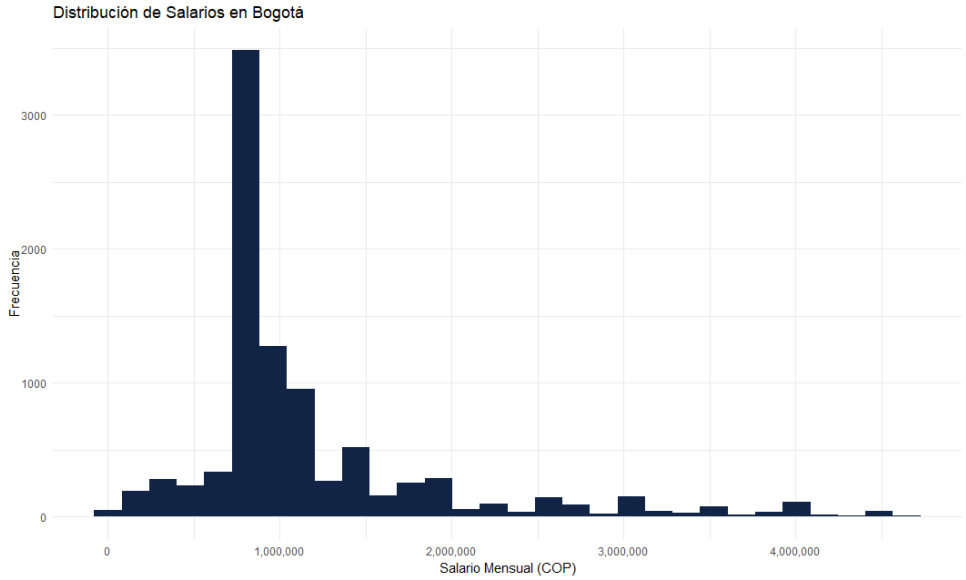

Figure 1: Distribución de Salarios en Bogotá

El histograma de salarios muestra la concentración de ingresos en niveles bajos y medios, con una mayor densidad en torno a los COP 3 millones, lo que indica una distribución sesgada a la derecha.

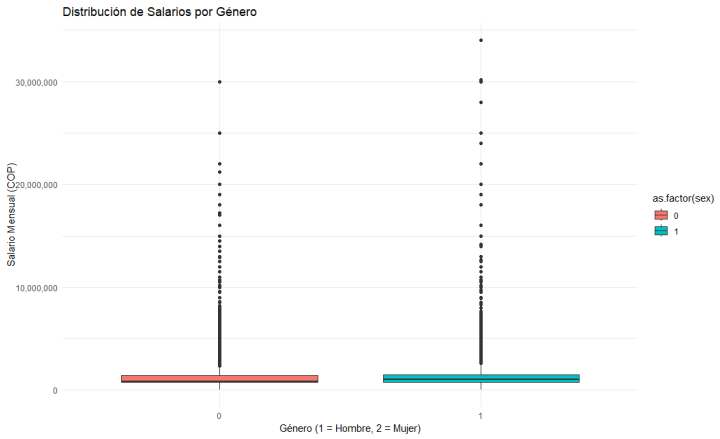

Figure 2: Distribución de Salarios por Género

El boxplot de salarios por género revela una clara disparidad en los ingresos, con una mediana salarial más alta para los hombres en comparación con las mujeres. La dispersión de los salarios también es mayor en los hombres, lo que sugiere una mayor variabilidad en sus ingresos.

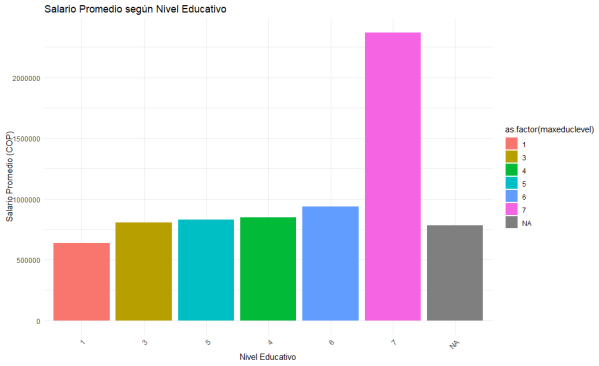

Figure 3: Salario Promedio según Nivel Educativo

El gráfico de barras muestra que el salario promedio aumenta con el nivel educativo. Las personas con estudios universitarios presentan los mayores ingresos, mientras que aquellos con educación primaria tienen los salarios más bajos.

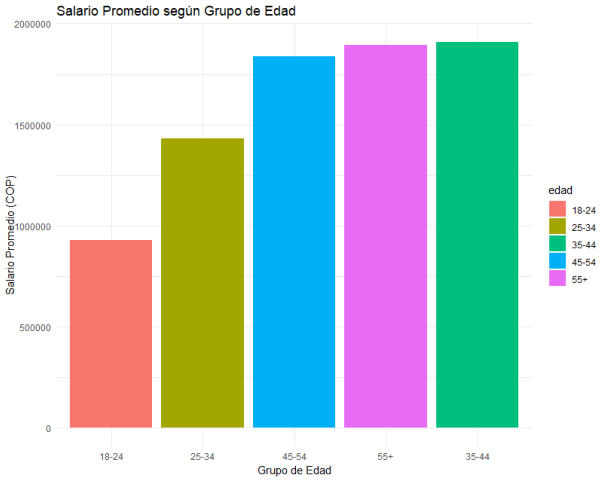

Figure 4: Salario Promedio según Grupo de Edad

La relación entre edad e ingresos sigue una curva en forma de U invertida, con los salarios más altos en el grupo de 35 a 44 años. A partir de los 55 años, los ingresos tienden a disminuir.

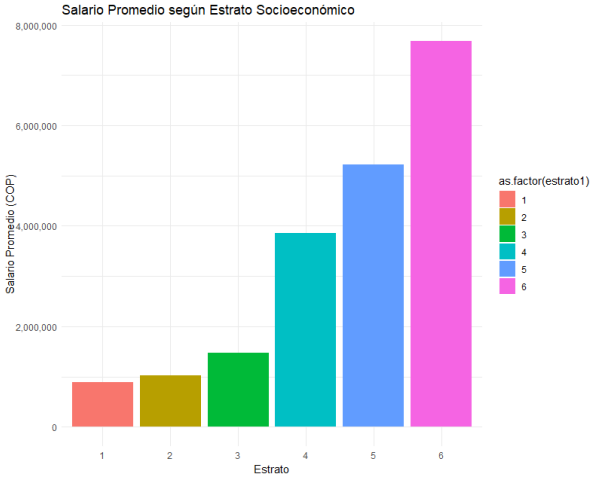

Figure 5: Salario Promedio según Estrato Socioeconómico

Se observa que los trabajadores de estratos más altos tienen ingresos considerablemente mayores, lo que refleja una fuerte segmentación del mercado laboral y diferencias en oportunidades de empleo.

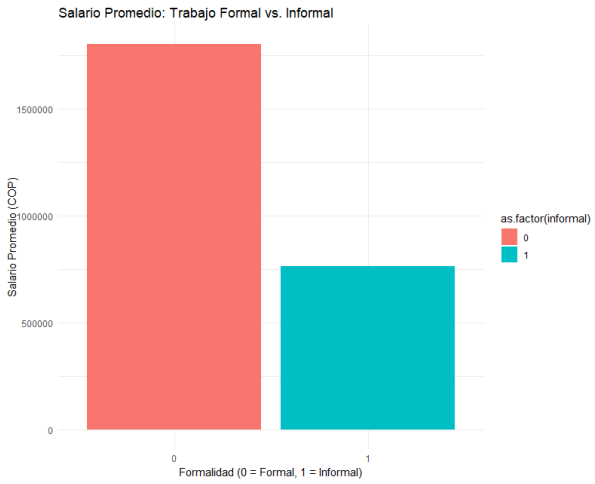

Figure 6: Salario Promedio: Trabajo Formal vs. Informal

La diferencia en ingresos entre trabajadores formales e informales es notable, con los primeros obteniendo, en promedio, un salario significativamente mayor. Esto resalta la vulnerabilidad económica de quienes se encuentran en el sector informal.

6 Punto 3

En esta sección, se presenta la estimación del modelo cuadrático que relaciona el salario con la edad y su cuadrado. Esta especificación permite capturar la no linealidad en la relación edad-ingresos, reflejando la acumulación de experiencia y la posible disminución del salario en edades avanzadas. El modelo estimado es el siguiente:

$$\log(\text{Salario}) = \beta_0 + \beta_1 \text{Edad} + \beta_2 \text{Edad}^2 + \varepsilon \tag{1}$$

6.1 Resultados de la Regresión

La tabla 6 muestra los coeficientes estimados del modelo cuadrático:

Table 6: Resultados de la regresión cuadrática del salario

Variable	Estimación	Error Estándar	Valor t	p-valor
Intercepto	7.7580	0.0455	170.58	$< 2 \times 10^{-16}$
Edad	0.0786	0.0038	20.49	$< 2 \times 10^{-16}$
Edad^2	-0.0007	0.00003	-22.16	$< 2 \times 10^{-16}$

6.1.1 Interpretación de los Coeficientes

El coeficiente de la variable edad es positivo y significativo ($\beta_1 = 0.0786$), lo que indica que los salarios aumentan con la edad en las primeras etapas de la vida laboral. Sin embargo, el coeficiente del término cuadrático de la edad es negativo ($\beta_2 = -0.0007$), lo que sugiere que existe un punto en el cual los ingresos dejan de crecer y comienzan a decrecer, reflejando la típica forma de U invertida en la relación edad-ingresos.

La edad en la que el salario alcanza su valor máximo se obtiene mediante la siguiente expresión:

$$\text{Edad Pico} = -\frac{\beta_1}{2\beta_2} = -\frac{0.0786}{2 \times (-0.0007)} = 43.31 \text{ años} \tag{2}$$

Esto implica que el salario promedio en Bogotá alcanza su punto más alto alrededor de los 43 años, tras lo cual comienza a disminuir.

6.1.2 Ajuste del Modelo

El modelo estimado tiene un R^2 de 0.0477, lo que indica que la edad y su cuadrado explican aproximadamente el 4.77% de la variabilidad en los ingresos. Aunque este valor es relativamente bajo, es importante destacar que la edad es solo uno de los múltiples factores que determinan el salario. Factores adicionales como educación, experiencia específica del trabajo, sector económico y género también influyen significativamente en los ingresos.

6.2 Intervalo de Confianza de la Edad Pico

Para evaluar la precisión de la estimación de la edad en la que se alcanza el máximo salario, se aplicó la técnica de bootstrap con 1000 replicaciones. El intervalo de confianza al 95% se obtuvo utilizando el método percentil, proporcionando un rango más robusto para la estimación de la edad pico.

El intervalo de confianza calculado es el siguiente:

$$IC_{95\%} = [42.41, 44.38] \tag{3}$$

Esto significa que, con un 95% de confianza, la edad en la que el salario promedio alcanza su punto máximo se encuentra entre 42.41 y 44.38 años. Este resultado confirma que el pico salarial estimado en 43.31 años es estadísticamente sólido.

En la Figura 7, se observa la curva estimada de ingresos por edad, junto con la edad pico destacada y su intervalo de confianza.

6.2.1 Gráfico del Perfil Edad-Salario

Para visualizar la relación estimada entre la edad y el salario, se presenta la siguiente figura:

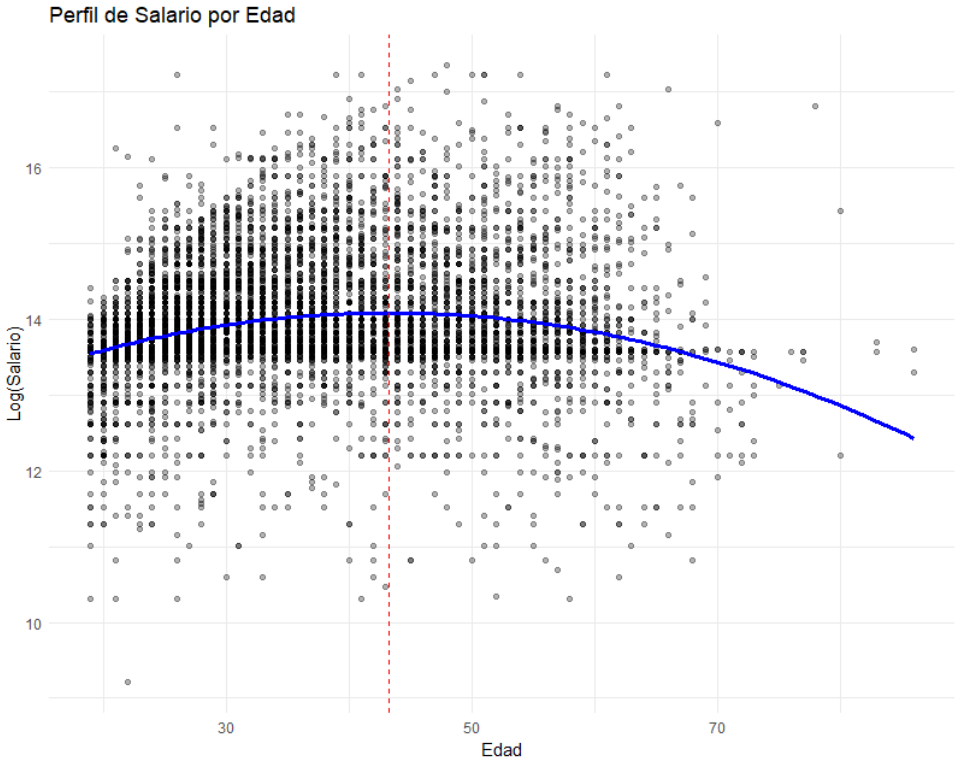

Figure 7: Relación estimada entre edad y salario

La curva muestra cómo los ingresos aumentan con la edad hasta alcanzar un máximo alrededor de los 43 años, tras lo cual comienzan a disminuir gradualmente.

7 Brecha Salarial de Género

El análisis de la brecha salarial de género busca evaluar si existen diferencias significativas en los ingresos de hombres y mujeres en Bogotá, y si estas diferencias persisten después de controlar por factores laborales y personales relevantes. Para ello, se han estimado distintos modelos econométricos que permiten descomponer la brecha en su componente bruto y ajustado.

7.1 Brecha Salarial Incondicional

Se estimó un modelo de regresión simple para evaluar la diferencia bruta en los salarios entre hombres y mujeres:

$$\log(w) = \beta_1 + \beta_2 Female + u \tag{4}$$

donde la variable $Female$ es un indicador que toma el valor de 1 si el individuo es mujer. El coeficiente β_2 captura la diferencia porcentual en ingresos entre géneros sin considerar otras características.

Los resultados muestran que la brecha salarial incondicional es del 15.29% ($\beta_2 = -0.1529$, $p < 0.01$), lo que indica que, en promedio, las mujeres ganan significativamente menos que los hombres sin considerar otras variables explicativas.

7.2 Brecha Salarial Condicional

Para controlar por factores laborales y personales, se estimó un segundo modelo que incluye variables como educación, experiencia, tipo de empleo y sector laboral:

$$\log(w) = \beta_1 + \beta_2 Female + X\beta + u \tag{5}$$

donde X incluye controles como:

- Educación ($maxeduclevel$)
- Horas trabajadas ($totalhoursworked$)
- Tipo de ocupación y sector laboral ($microempresa$, $oficio$)
- Experiencia laboral ($p6210s1$)
- Características contractuales ($p6426$, $p6920$, $relab$)
- Edad (age)
- Estado civil ($p6050$)

Los resultados de la regresión condicional indican que la brecha salarial ajustada es del 9.74% ($\beta_2 = -0.0974$, $p < 0.01$), lo que sugiere que una parte de la disparidad en los ingresos puede explicarse por diferencias en educación, experiencia y tipo de empleo. Sin embargo, la brecha sigue siendo estadísticamente significativa .

7.3 Validación con el Método Frisch-Waugh-Lovell (FWL)

Para verificar que la brecha salarial no sea explicada por otras variables, se aplicó el método FWL , que implica tres pasos:

1. Regresión del salario sobre los controles y obtención de los residuos.
2. Regresión del género sobre los controles y obtención de los residuos.
3. Regresión de los residuos del salario sobre los residuos del género.

El coeficiente obtenido en el modelo FWL es de -0.0974 ($p < 0.01$), lo que confirma que la brecha salarial persiste incluso después de descontar los efectos de otras variables.

7.4 Estimación del Intervalo de Confianza con Bootstrapping

Para evaluar la robustez de la estimación, se realizó un bootstrapping con 1000 replicaciones , obteniendo el siguiente intervalo de confianza al 95%:

$$[-0.1210, -0.0758] \tag{6}$$

Esto significa que, con un 95% de confianza, la brecha salarial ajustada se encuentra entre 7.58% y 12.10%, confirmando que las mujeres ganan sistemáticamente menos que los hombres incluso después de ajustar por educación y experiencia.

El R^2 del modelo condicional es 0.6425, lo que indica que el 64.25% de la variabilidad en los ingresos es explicada por las variables incluidas. Esto confirma que las características laborales explican parte de la brecha, pero que el género sigue siendo un factor significativo.

Table 7: Regresión de la brecha salarial de género

Modelo	Intercepto	Coef. Female	Error Estándar	Valor t	R^2 Ajustado
Incondicional	13.9867	-0.1529	0.0151	-10.12	0.0103
Condicional	10.9477	-0.0974	0.0115	-8.49	0.6425
FWL	0.0000	-0.0974	0.0112	-8.68	0.0076

7.5 Perfil de Ingresos por Género

Para analizar la evolución de la brecha a lo largo del ciclo laboral, se presenta el siguiente gráfico:

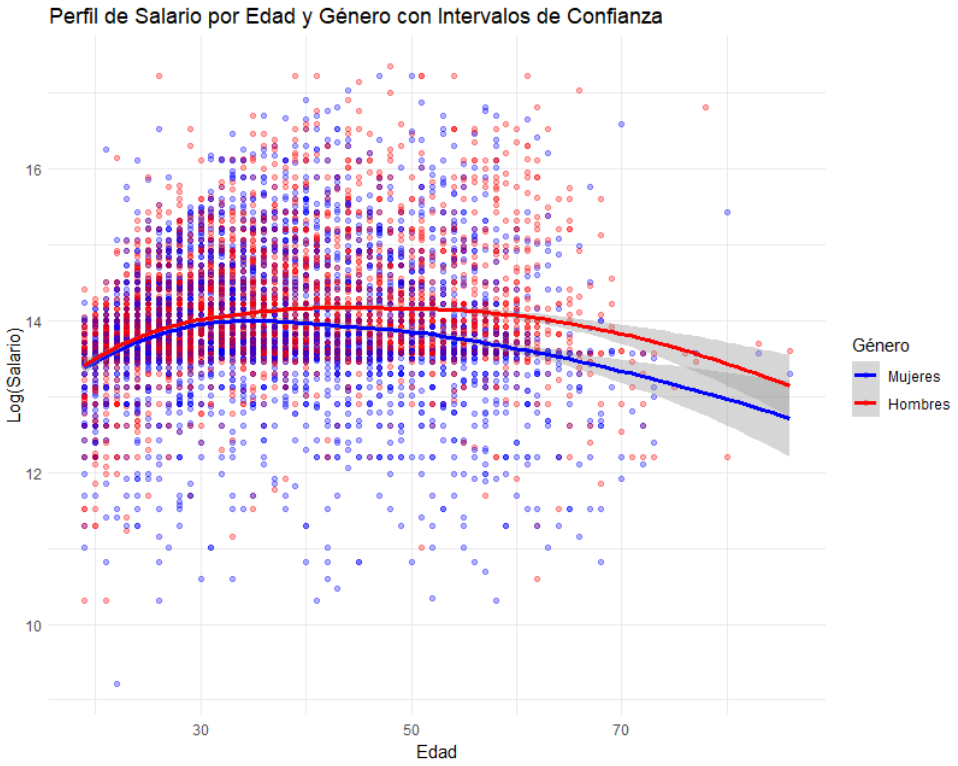

Figure 8: Perfil de Salario por Edad y Género con Intervalos de Confianza

Se observa que los salarios de las personas jovenes no tienen una marcada diferencia, sin embargo, la brecha salarial se empieza a evidenciar en personas ocupadas mayores de 30 años y se profundiza cuando aumenta la edad durante el ciclo laboral. Además, la edad pico de ingresos es ligeramente mayor para los hombres en comparación con las mujeres.

7.6 Conclusión

Los resultados indican que la brecha salarial de género persiste incluso después de controlar por educación, experiencia y sector laboral. La combinación de regresiones condicionales, el método FWL y bootstrapping confirma que las mujeres en Bogotá ganan menos que los hombres, lo que sugiere la presencia de desigualdades estructurales de género dentro del mercado laboral.

8 Punto 5: Evaluación de Modelos de Predicción

En esta sección se presenta la partición de la base de datos en conjuntos de entrenamiento y prueba, la estimación de modelos de predicción del ingreso, y la evaluación de su desempeño utilizando métricas como el error cuadrático medio (RMSE). Se comparan diferentes especificaciones del modelo para determinar cuál ofrece el mejor ajuste a los datos.

8.1 Partición de Datos

Para evaluar la capacidad predictiva de los modelos, la base de datos se dividió en dos subconjuntos:

- **Entrenamiento (70%)**: Usado para ajustar los modelos.
- **Prueba (30%)**: Usado para evaluar la precisión de las predicciones.

En la Figura 9 se observa la distribución de las observaciones en cada subconjunto.

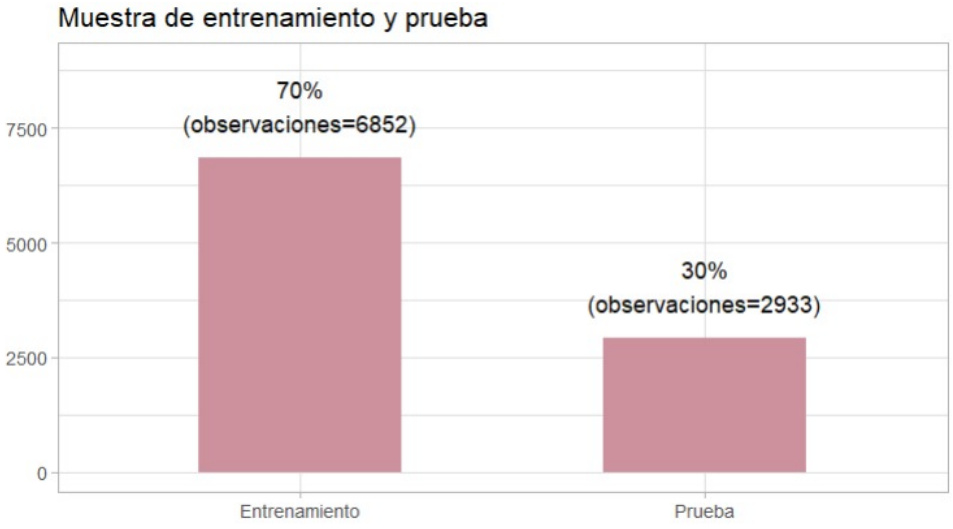

Figure 9: Muestra de entrenamiento y prueba

El conjunto de entrenamiento cuenta con 6,852 observaciones, mientras que el conjunto de prueba contiene 2,933 observaciones.

8.2 Ajuste de los Modelos y Predicciones

Se estimaron distintos modelos de regresión para predecir el ingreso, incluyendo modelos de mínimos cuadrados ordinarios (OLS) y el método de Frisch-Waugh-Lovell (FWL).

8.3 Especificaciones de Modelos de Regresión

8.3.1 Modelo 1: Modelo Simple de Predicción

$$\log(\text{salary}) = \beta_0 + \beta_1 \cdot \text{dummy\_sex} + \varepsilon \tag{7}$$

8.3.2 Modelo 2: Modelo con controles

$$\begin{aligned}\log(\text{salary}) = \beta_0 &+ \beta_1 \cdot \text{dummy\_sex} + \beta_2 \cdot \text{totalhoursworked} + \beta_3 \cdot \text{maxeduclevel} \\ &+ \beta_4 \cdot \text{microempresa} + \beta_5 \cdot \text{oficio} + \beta_6 \cdot \text{p6210s1} + \beta_7 \cdot \text{p6426} \\ &+ \beta_8 \cdot \text{p6920} + \beta_9 \cdot \text{relab} + \beta_{10} \cdot \text{age} + \beta_{11} \cdot \text{p6050} + \varepsilon\end{aligned} \tag{8}$$

8.3.3 Modelo 3: Modelo con Términos Cuadráticos de Edad

$$\log(\text{salary}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2 + \varepsilon \tag{9}$$

8.3.4 Modelo 4: Modelo con Términos Cuadráticos de Horas Trabajadas

$$\log(\text{salary}) = \beta_0 + \beta_1 \cdot \text{totalhoursworked} + \beta_2 \cdot \text{totalhoursworked}^2 + \varepsilon \tag{10}$$

8.3.5 Modelo 5: Modelo con Términos Cuadráticos y Cúbicos de Edad

$$\log(\text{salary}) = \beta_0 + \beta_1 \cdot \text{age}^2 + \beta_2 \cdot \text{age}^3 + \varepsilon \tag{11}$$

La Figura 10 muestra la comparación entre los valores reales y las predicciones de los modelos.

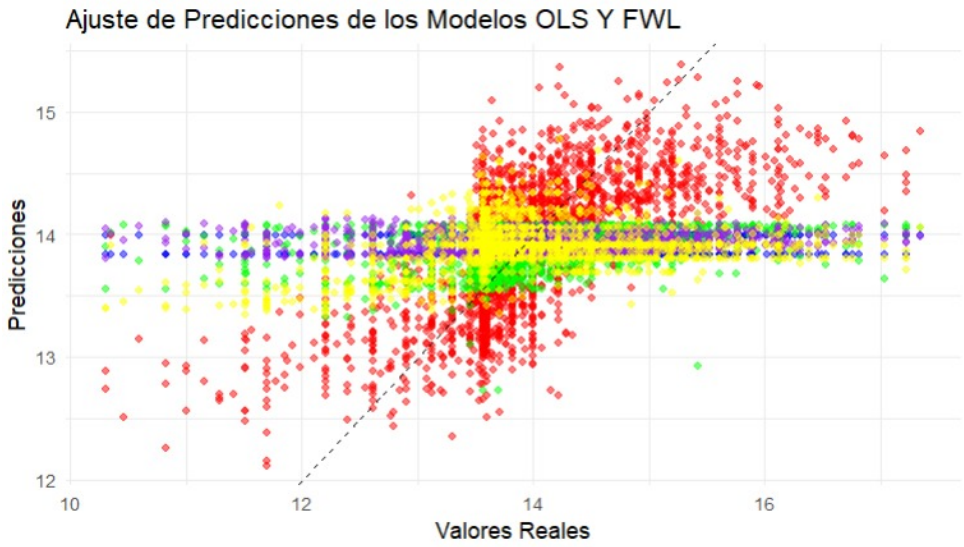

Figure 10: Ajuste de Predicciones de los Modelos OLS y FWL

Se observa que las predicciones siguen una tendencia alineada con los valores reales, aunque con dispersión en ciertos rangos de ingreso.

8.4 Evaluación del Desempeño de los Modelos

Para comparar la precisión de los modelos, se calculó el error cuadrático medio (RMSE) en el conjunto de prueba. La Tabla 8 presenta los resultados obtenidos para cada especificación.

Table 8: Comparación de RMSE entre Modelos

Modelo	Simple	Edad Pico	Edad Modificado	Con Controles	Total Horas Trabajadas
RMSE	0.76	0.75	0.76	0.57	0.75

Los resultados indican que el modelo que incluye controles adicionales es el que presenta el menor RMSE (0.57), lo que sugiere que incorpora mejor las variables relevantes para la predicción del ingreso. En contraste, los modelos más simples tienen un desempeño menos preciso, con RMSE cercanos a 0.75 y 0.76.

8.5 LOOCV

Realizamos una validación de los 2 modelos con el RMSE más bajos del punto anterior, es decir, el modelo con controles y el de edad pico. Con el método LOOCV se obtienen los resultados de la Tabla 9

Table 9: Comparación de RMSE con LOOCV

Métrica	Valor
RMSE LOOCV - Edad pico	0.7346
RMSE LOOCV - Con controles	0.5856

Al usar este método se reduce la posibilidad de sobreajuste y se obtiene un RMSE más estable y menos dependiente a la partición de los datos. Es así como se interpreta finalmente que el modelo que contiene los controles de caracterización laboral y explica el salario con la dummy del sexo, tiene un mejor RMSE y por lo tanto se toma como el mejor modelo.

9 Referencias

References

[1] Servicio de Impuestos Internos (IRS). (2022). *The tax gap*. IRS.

[2] Comisión Económica para América Latina y el Caribe (CEPAL). (2021). *Brechas de género en los mercados laborales de América Latina y el Caribe*. CEPAL.

[3] Banco Mundial. (2022). *The gender wage gap in Latin America and the Caribbean*. Banco Mundial.

[4] Organización de las Naciones Unidas Mujeres (ONU Mujeres). (2022). *Brechas de género en Colombia: Impacto de la maternidad en el mercado laboral*. ONU Mujeres.

[5] Banco Mundial. (2022). *The gender wage gap in Latin America and the Caribbean*. Banco Mundial.

[6] Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.

[7] Bourguignon, F. (2015). *The globalization of inequality*. Princeton University Press.

[8] Comisión Económica para América Latina y el Caribe (CEPAL). (2021). *Brechas de género en los mercados laborales de América Latina y el Caribe*. CEPAL.

[9] Departamento Administrativo Nacional de Estadística (DANE).