Problem Set 1: Predicting Income

David Salamanca, Gabriel Perdomo, Sergio Díaz

September 16, 2024

Course: [Machine Learning & Big Data]
Professor: [Ignacio Sarmiento]
Submission Date: September 16, 2024
Institution: [Universidad de los Andes]

Git Hub repository link

1 Introduction

En Colombia, la comprensión de los patrones de ingresos es crucial para abordar la desigualdad y mejorar la eficiencia de la política fiscal. El mercado laboral colombiano está marcado por disparidades, no solo en términos de ingresos entre diferentes grupos socioeconómicos, sino también en aspectos como la brecha salarial de género y la informalidad laboral. La falta de transparencia en la declaración de ingresos por parte de los individuos contribuye a esta desigualdad, complicando la asignación equitativa de recursos y el diseño de intervenciones efectivas. Un modelo de predicción de ingresos puede ser una herramienta valiosa para identificar y comprender estos patrones, facilitando la detección de posibles casos de subreporte de ingresos y la identificación de grupos vulnerables que puedan requerir apoyo adicional.

En este trabajo, se busca aplicar técnicas de análisis de datos y modelado predictivo para construir un modelo de ingresos utilizando información proveniente de la Gran Encuesta Integrada de Hogares (GEIH) de 2018 para la ciudad de Bogotá En el presente trabajo, no solo se busca obtener un entendimiento más profundo sobre los ingresos de los colombianos, sino que también se realizará un análisis predictivo al respecto para identificar y comprender mejor estos factores y sus implicaciones en la brecha salarial de género.

2 Data Description

2.1 Resumen datos

Los datos utilizados provienen del estudio de "Medición de Pobreza Monetaria y Desigualdad" realizado en Bogotá en 2018. Este conjunto de datos es parte de la Gran Encuesta Integrada de Hogares (GEIH) que recopila información detallada sobre el mercado laboral y las condiciones socioeconómicas de los hogares en Bogotá. El propósito de los datos es proporcionar una visión amplia de las características de la población, incluyendo aspectos demográficos, educativos, laborales, y de ingresos. Con el objetivo en mente de utilizar estos datos para construir un modelo que prediga los salarios por hora de los individuos en función de diversas variables explicativas (como edad, nivel educativo, género, tipo de trabajo, entre otros). La base de datos inicialmente cuenta con 32177 observacion y 177 variables.

2.2 Acquisiscion base de datos

Para la adquisición de los datos utilizados en este estudio, el equipo implementó un proceso de web scraping utilizando Selenium, con el objetivo de automatizar la descarga de la información. El sitio proporciona datos divididos en múltiples páginas o "chunks". Para extraer los datos, se configuró un navegador Chrome automatizado, que permitió la navegación a través del sitio. Una vez cargada cada página, se empleó BeautifulSoup para analizar y extraer el contenido, identificando y consolidando los datos relevantes en

un DataFrame. El proceso fue repetido para cada página del sitio web, lo que permitió al equipo extraer y unificar todos los chunks de datos en un solo archivo.
El acceso al sitio web es público y no se encontraron restricciones técnicas significativas, ademas de una alta latencia para cargar cada pagina que contenia un chunk. Para obtener la información, el proceso de scraping fue lento debido al gran volumen de datos y como estaban distribuidos en varias páginas ademas selenium no suele ser el metodo mas rapido de scraping.

Link Datos Scrapeados

2.3 Limpieza de datos

Dado el contexto de la pregunta del taller, el objetivo es desarrollar un modelo de predicción de ingresos. Para lograrlo, es fundamental trabajar con datos limpios y precisos.
Para realizar el analisis se seleccionaron solo las personas mayores de 18 años que estaban empleadas, con el fin de centrarse en aquellos que participan activamente en el mercado laboral como mayores de edad. Luego, se eliminaron las observaciones que no tenían información sobre el salario total o por hora, ya que esta es la variable principal del análisis. Igualmente filtramos bastantes varibles en base a su relevancia para el estudio y la completitud de sus datos, es decir que tantas observaciones faltantes tenia cada variable. Sin embargo, en lugar de eliminar todas las filas con datos faltantes, se adoptó un enfoque más cuidadoso. Se conservo aquellas variables que tuvieran hasta un 20 porciento de datos faltantes. Este límite fue establecido para mantener un tamaño de muestra suficientemente grande, mantener la consistencia de la distribución, Preservar la variabilidad de la muestra, y evitar perder información valiosa.
Para los valores faltantes se optó por reemplazarlos con el promedio observacional. Este enfoque busca minimizar la pérdida de información sin introducir sesgos significativos.
Para los otros filtros se seleccionaron las variables más relevantes para el análisis en base a una matriz de correlacion con la variable dependiente, algunas variables se excluyeron debido a que no eran pertinentes para el objetivo del análisis. Despues del proceso de limpieza de datos, se redujo la muestra a 12826 observaciones con 29 variables relevantes para construir el modelo de prediccion de ingresos.

2.4 Variables y tablas descriptivas

A continucacion se presenta una descripcion mas detallada de las variables utilizadas, y las estadisticas descriptivas de estas mismas. Ademas de algunas graficas complementarias para entender el contexto de los datos y la distribucion de ingresos.

La tabla proporciona las características de la población laboral analizada. La edad promedio es de 38.46 años, con un rango de 18 a 91, lo que sugiere una fuerza laboral que incluye tanto jóvenes como adultos mayores. En cuanto al sexo, hay una ligera mayoría masculina (55%), lo que puede ser relevante para el análisis de disparidades de género en el mercado laboral. Los trabajadores reportan, en promedio, 62.11 horas trabajadas por semana en todas sus actividades (p6426), con una alta variabilidad (desviación estándar de 87.08), lo que indica que algunos trabajadores tienen cargas laborales significativamente altas, probablemente combinando múltiples empleos. El ingreso por hora (y_total_m_ha) muestra una alta dispersión, con una media de 8,607.14 y un rango que va desde 0.47

Table 1: Estadisticas descriptivas

Statistic	Mean	St. Dev.	Min	Max
age	38.46	12.76	18	91
sex	0.55	0.50	0	1
p6426	62.11	87.08	0	720
p7040	1.96	0.18	1	2
p7070	21,662.55	241,012.80	0.00	10,000,000.00
y_total_m_ha	8,607.14	13,571.75	0.47	350,583.30
maxEducLevel	5.98	1.18	1	7
totalHoursWorked	50.44	12.51	1	130
formal	0.64	0.48	0	1
cuentaPropia	0.27	0.44	0	1
microEmpresa	0.39	0.49	0	1
y_total_m	1,728,744.00	2,531,112.00	97.00	70,000,000.00
college	0.33	0.47	0	1

hasta 350,583.30, reflejando disparidades salariales significativas.
En términos de educación, el nivel educativo máximo alcanzado (maxEducLevel) tiene una media de 5.98 en una escala de 1 a 7, indicando que la mayoría de los trabajadores han alcanzado un nivel educativo considerable, como educación técnica o universitaria, lo que puede influir positivamente en las oportunidades de empleo y el nivel de ingresos. La variable p7040 indica si el trabajador está afiliado a una caja de compensación familiar (1 = Sí, 2 = No), y con una media de 1.96, muestra que la mayoría no está afiliada, lo que puede afectar el acceso a ciertos beneficios sociales. p7070 se refiere a los ingresos netos provenientes de una actividad secundaria, con una media de 21,662.55, evidenciando que una parte de la población laboral depende de ingresos adicionales a su empleo principal. En promedio, los trabajadores trabajan 50.44 horas semanales (totalHoursWorked), lo que sugiere jornadas laborales extensas. La formalidad del empleo (variable formal, 0 = No, 1 = Sí) muestra que el 64% de los trabajadores tiene un empleo formal, lo cual es significativo ya que estos empleos suelen ofrecer mayor seguridad y beneficios. Además, el 27% de los trabajadores son trabajadores por cuenta propia (cuentaPropia), lo que implica un nivel de independencia laboral, aunque a menudo con menos seguridad laboral y beneficios sociales. El tamaño de la empresa (sizeFirm), con una media de 3.35 en una escala de 1 a 5, sugiere que muchos trabajadores están empleados en empresas pequeñas o medianas, lo que puede influir en la estabilidad y las oportunidades de crecimiento profesional.
El ingreso total mensual en pesos (y_total_m) presenta una amplia variabilidad, con una media de $1,728,744.00 y un rango maximo de $70,000,000, reflejando una alta desigualdad en los ingresos laborales. Esta variabilidad puede estar influenciada por factores como el nivel educativo, la formalidad del empleo y la ocupación. El 33% de los trabajadores en la muestra han completado educación universitaria (college), lo que generalmente se asocia con mejores oportunidades laborales y salarios más altos.
Otras variables categóricas no incluidas directamente en la tabla, como el estrato socioeconómico,, y relacion laboral.

Al analizar la distribución de observaciones por estrato, se observa una clara concentración en los estratos medios-bajos. Específicamente, el estrato 2 cuenta con aproximadamente el 30% de las observaciones, mientras que el estrato 3 comprende cerca del

35%. En conjunto, estos dos estratos representan alrededor del 65% de la muestra total. Los estratos más altos, como el estrato 5 y el estrato 6, tienen una representación significativamente menor, con menos del 10% de las observaciones cada uno. Esta distribución sugiere que una gran proporción de la población laboral se encuentra en los segmentos socioeconómicos más vulnerables, lo cual tiene importantes implicaciones para las políticas públicas y las estrategias de intervención, ya que estos grupos suelen tener acceso más limitado a recursos como la educación, el empleo formal y la seguridad social.

En cuanto a la relación laboral (relab), esta variable proporciona una visión detallada de los tipos de vínculo que los individuos tienen con su trabajo. La muestra incluye diversas categorías como "Empleado", "Trabajador Independiente", "Empleador", y "Ayuda Familiar sin Remuneración". Los datos muestran que la categoría de Empleado es la más común, abarcando aproximadamente el 50% de la muestra. Esto indica que una parte significativa de la población laboral tiene un empleo relativamente más estable, potencialmente con acceso a beneficios laborales y seguridad social.

Sin embargo, la presencia de un considerable 30% de Trabajadores Independientes indica una economía donde una proporción significativa de la fuerza laboral opera fuera de las relaciones laborales tradicionales. Estos trabajadores suelen tener menores niveles de seguridad laboral y acceso limitado a beneficios sociales. Otras categorías, como "Empleador" y "Ayuda Familiar sin Remuneración", están menos representadas, sumando menos del 20% en total, pero son cruciales para comprender la estructura ocupacional, ya que reflejan la presencia de pequeñas unidades económicas y actividades familiares no remuneradas.

2.4.1 graficas descriptivas

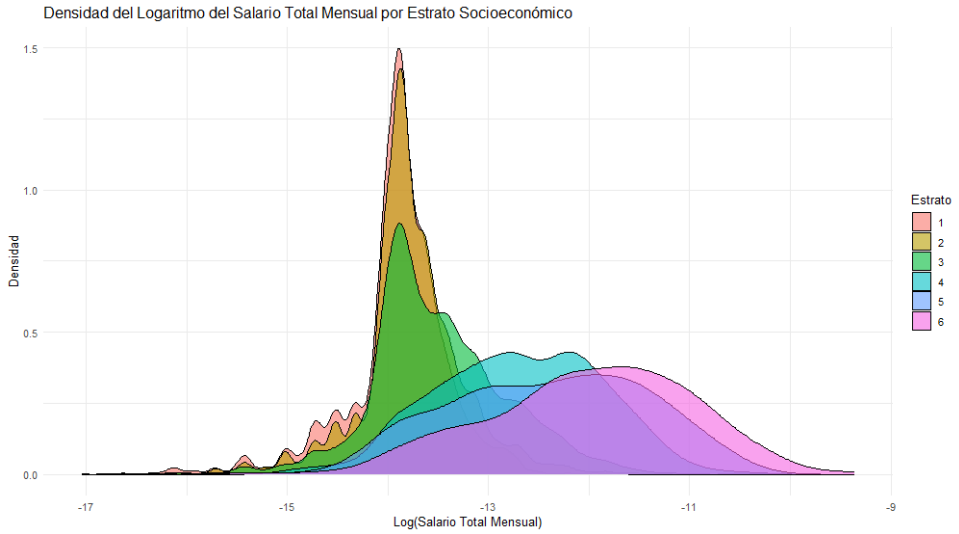


Figure 1: Este gráfico de densidad muestra cómo se distribuyen los salarios totales mensuales en escala logarítmica según los estratos socioeconómicos, del 1 al 6. Se observa una clara diferenciación en la concentración salarial, donde los estratos más bajos (1 y 2) presentan una mayor densidad en niveles salariales más bajos, mientras que los estratos superiores (5 y 6) tienen distribuciones salariales que se desplazan hacia valores más altos. Esta gráfica es relevante para analizar la desigualdad económica y cómo el estrato socioeconómico está relacionado con los niveles de ingresos, evidenciando la disparidad económica entre los diferentes grupos sociales.

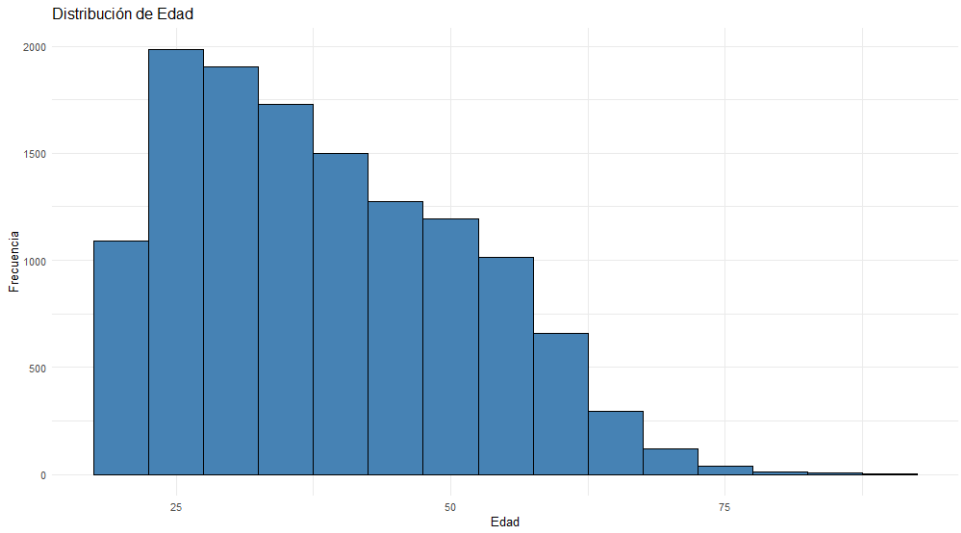


Figure 2: El histograma muestra la distribución de la edad en la muestra, con una mayor concentración de individuos en el rango de 20 a 40 años. La distribución decrece gradualmente a medida que la edad aumenta, con pocos trabajadores mayores de 65 años. Esto sugiere una población laboral predominantemente joven y de mediana edad, lo que puede tener implicaciones en las políticas laborales y las estrategias de capacitación profesional.

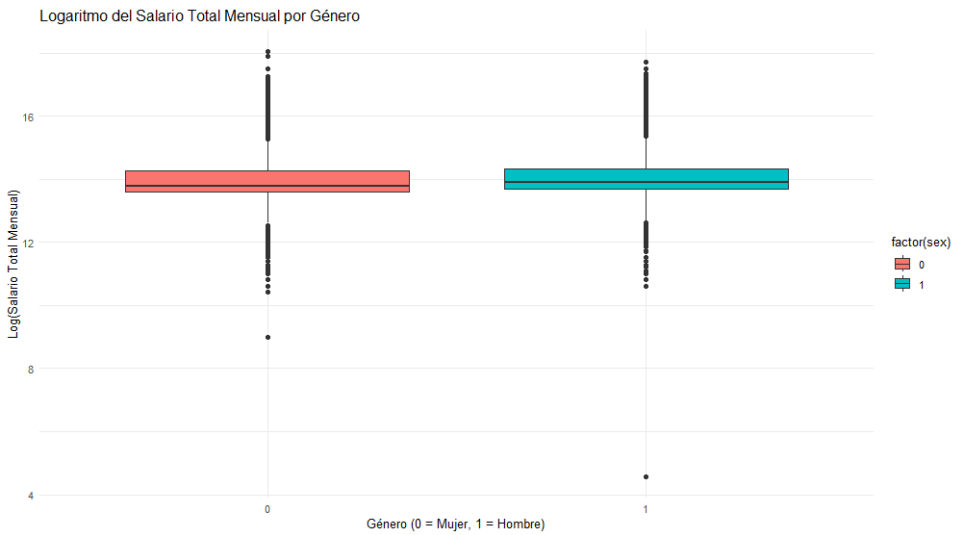


Figure 3: El diagrama de cajas muestra la distribución del salario total mensual en escala logarítmica por género. Se observa que las medianas salariales son similares entre hombres y mujeres aunque el promedio masculino es ligeramente mayor, ademas hay una mayor dispersión en los ingresos de los hombres, como indican los valores atípicos. Esto sugiere que, aunque las medianas sean comparables, existen desigualdades en la distribución salarial, con una mayor variabilidad de ingresos en la población masculina.

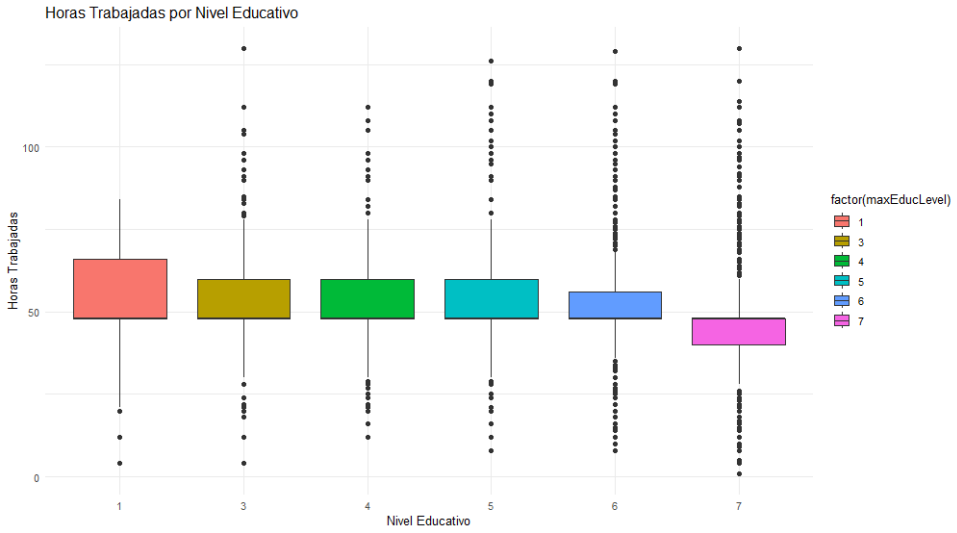


Figure 4: El diagrama de cajas muestra la relación entre las horas trabajadas y el nivel educativo. Se observa que los niveles educativos más bajos tienden a tener una mayor variabilidad en las horas trabajadas, mientras que los niveles más altos (6 y 7) presentan una distribución más centrada alrededor de la mediana. Esto indica que las personas con niveles educativos más altos podrían tener jornadas laborales más estables, posiblemente debido a empleos más formales y estructurados.

3 Perfil Edad Salario

Para este punto, se corrió el siguiente modelo con el fin de entender cómo afecta la edad al nivel de ingreso de un individuo.

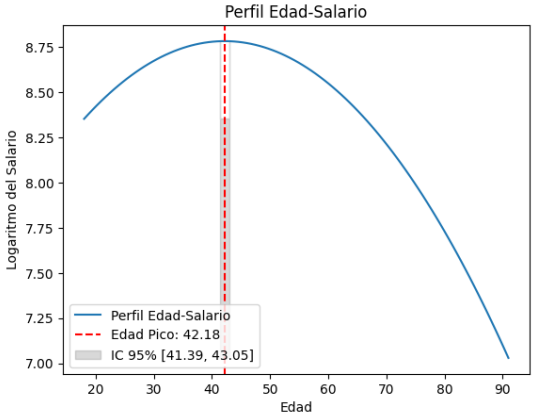
$$ln(salarioporhora) = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 + u$$

Despúes de correr el modelo obtenemos los siguientes resultados:

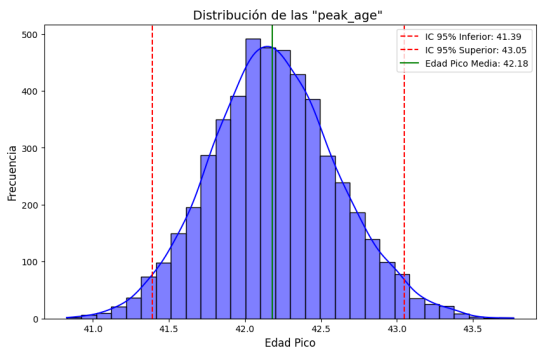
Variable Dependiente:	Log_Salario	
	Coeficiente	P-valor
const	7.4742	0.000
Edad	0.0620	0.000
$Edad^2$	-0.0007	0.000
Número de Observaciones: 12826		
Método: OLS		
R^2: 0.027		

La tabla enterior muestra que la edad sí tiene un efecto significativo sobre los salarios (Ambos coeficientes son altamente significativos, con p-valores menores a 0.05). Más aún, queda en evidencia la existencia de una peak-age por la relevancia de la edad al cuadrado, lo cual nos indica qeu presetna una forma concava puesto que al llegar a cierta edad, los salarios empiezan a mermar.
especificamente, la tabla nos indica que, todo lo demás constante, ante un aumento de un año en la edad, el salario aumenta aproximadamente en 6,2%. Por el lado de la edad al cuadrado, el coeficiente negativo sugiere que, aunque el salario aumenta con la edad, el ritmo de ese aumento disminuye a medida que la edad incrementa.
En términos de ajuste del modelo, el R-cuadrado es 0.027, indicando que solo el 2.7% de la variabilidad en el logaritmo del salario puede ser explicada por el modelo. Esto sugiere que, si bien la edad tiene un impacto significativo en los salarios, existen una gran cantidad de variables.

Retomando la idea de la existencia de una edad tope por el componenete cuadrático, se utilizó bootstrap para construir la distribucion de la edad pico en todas las submuestras y la curva sobre el perfil edad salario.



(a) Perfil Edad-Salario



(b) Distribución de las Edades Pico Estimadas con Bootstrap

La figura (a) muestra una curva cuadrática que indica cómo los salarios, en términos logarítmicos, aumentan con la edad, alcanzando un punto máximo a los 42.18 años. Este comportamiento es consistente con los resultados ilustrados por el modelo. La línea roja marca esta edad pico, mientras que las líneas grises señalan el intervalo de confianza al 95% (41.39 - 43.05 años). La segunda figura (b) refuerza estos resultados mostrando la distribución de las edades pico estimadas mediante bootstrap, confirmando la robustez y estabilidad de la estimación, ya que la mayoría de las muestras se agrupan alrededor de los 42.18 años, con un rango de confianza que coincide con el identificado en el perfil edad-salario.

4 Unconditional Wage Gap

En este punto, estimaremos el siguiente modelo:

$$\log(w) = \beta_1 + \beta_2 \cdot \text{Female} + u$$

Con este modelo, intentaremos evidenciar el comportamiento de los salarios según el género. El análisis es posible gracias a la variable **Female**, que toma el valor de 0 cuando la persona es hombre y 1 cuando es mujer. Esta primera estimación nos permite observar el comportamiento de los salarios controlando únicamente por el género, sin incluir otras variables adicionales, las cuales serán consideradas en análisis posteriores.

Table 2: Resultados de la regresión incondicional: Brecha salarial

Variable	Coeficiente	Error Estándar	t-statistic	P-valor
Constante	14.0468	0.009	1550.480	0.000
Female	-0.1104	0.014	-8.150	0.000
Número de Observaciones: 12,826				
R-cuadrado: 0.005				
Método: OLS				

Los resultados obtenidos (ver tabla2) muestran que existe una **brecha salarial** significativa entre hombres y mujeres. Específicamente, las mujeres ganan aproximadamente un **11.04% menos** en comparación con los hombres. Esta diferencia salarial es estadísticamente significativa, como lo indica el valor p cercano a 0, lo que confirma la presencia de una disparidad salarial sin considerar otros factores explicativos.

Para determinar si la premisa de "igual salario para igual trabajo" se cumple, es crucial evaluar la brecha salarial entre géneros al ajustar por características del trabajador, el entorno empresarial y el tipo de vinculación al mercado laboral. Inicialmente, se utilizó el teorema de Frisch-Waugh-Lovell (FWL) para estimar la brecha salarial condicionada, ajustando por variables de control adicionales.

En la primera etapa del método FWL, se ajustaron los salarios (o la variable dependiente) por las variables de control, lo que permitió obtener los residuos de salario ajustados. En una segunda etapa, se ajustó la variable de interés (en este caso, el género) por las mismas variables de control para obtener los residuos de la variable de género.

La regresión final se realizó con estos residuos, resultando en la siguiente tabla de resultados:

	Coef.	Std Err	t	P¿—t—
const	4.231e-14	0.005	7.82e-12	1.000
residuo$_female$	-0.1881	0.011	-17.188	0.000

Table 3: Resultados de la regresión después de ajustar por características del trabajador y el entorno empresarial.

De acuerdo con los resultados obtenidos, a pesar de la inclusión de controles adicionales, la brecha salarial persiste y parece ser aún más pronunciada. En particular, el coeficiente para la variable de género indica que, incluso después de ajustar por las características relevantes, las mujeres ganan aproximadamente un 18.81% menos que los hombres, con alta significancia estadística.

Estos hallazgos sugieren que la premisa de que un trabajo idéntico debería tener un salario idéntico no se cumple en este contexto. La brecha salarial observada implica que otros factores, además de las características del trabajador y el entorno empresarial, podrían estar influyendo en la disparidad salarial entre géneros.

Para agregar precisión a las estimaciones, se realizó una iteración del modelo FWL incorporando el método Bootstrap. Los resultados mostraron una consistencia notable entre ambos métodos, con una variación mínima en los coeficientes y los errores estándar, lo que refuerza los hallazgos sobre la brecha salarial por género. A continuación se presentan los resultados obtenidos:

Table 4: Comparación de Resultados FWL y Bootstrap

Método	**Coeficiente Female**	**Error estándar**
FWL	-0.1881	0.0109
Bootstrap	-0.1880	0.0109

Como se puede observar, los resultados obtenidos mediante el método Bootstrap son muy similares a los resultados del método FWL, con una variación de solo 0.0001 en el coeficiente y sin cambios en el error estándar comparado con el error estandar promedio. Esta similitud en los resultados refuerza la validez de las conclusiones obtenidas sobre la brecha salarial de género, evidenciando que la brecha existente persiste incluso al utilizar un método alternativo para la estimación de los errores.

Ahora bien, como se observó en el punto 3, el comportamiento de los ingresos va variando conforme el tiempo va pasando, este tiene un comportamiento creciente hasta aproximadamente los 50 años en las mujeres y luego empieza a decaer. Esto anterior se denomina "la edad pico", siendo esta el objeto de estudio adicional para encontrar aún más comportamientos en los salarios de las mujeres.

Para analizar la relación entre la edad y el salario, así como el efecto del género en esta relación, estimamos el siguiente modelo de regresión:

$$\begin{aligned} \text{Salario}_i = \beta_0 + \beta_1 \text{Edad}_i + \beta_2(\text{Edad}_i^2) + \beta_3 \text{Female}_i + \\ \beta_4(\text{Female}_i \times \text{Edad}_i) + \beta_5(\text{Female}_i \times \text{Edad}_i^2) + \mathbf{X}_i\gamma + \epsilon_i \end{aligned} \tag{1}$$

Donde:

- Edad_i es la edad del individuo i.
- Edad_i^2 es el cuadrado de la edad del individuo i.
- Female_i es una variable dummy que indica el género del individuo (1 si es mujer, 0 si es hombre).
- $\text{Female}_i \times \text{Edad}_i$ es la interacción entre el género y la edad.
- $\text{Female}_i \times \text{Edad}_i^2$ es la interacción entre el género y el cuadrado de la edad.
- $\mathbf{X}_i$ representa un conjunto de variables de control adicionales, que incluyen el nivel educativo máximo, la formalidad del empleo, el tipo de empresa, el tamaño de la empresa y si la persona asistió a la universidad.
- ϵ_i es el término de error.

El modelo incondicional considera solo las variables relacionadas con la edad y el género, mientras que el modelo condicional incorpora variables de control adicionales para ajustar el impacto del género en el salario, proporcionando una visión más precisa del efecto de las características individuales y del entorno laboral en el salario.

A continuación se presentan las estimaciones y errores estándar para los modelos incondicional y condicional. El modelo incondicional incluye las variables básicas, mientras que el modelo condicional agrega controles adicionales para mejorar la precisión de las estimaciones.

Table 5: Comparación de Resultados: Modelos Incondicional y Condicional

Variable	**Modelo Incondicional**	**Error Estándar**	**Modelo Condicional**	**Error Estándar**
Constante	12.6389	0.079	10.9248	0.081
Edad	0.0714	0.004	0.0597	0.003
Edad al Cuadrado	-0.0008	4.68e-05	-0.0006	3.8e-05
Femenino	0.0046	0.124	-0.1475	0.100
Edad * Femenino	0.0035	0.006	0.0023	0.005
Edad al Cuadrado * Femenino	-0.0002	7.54e-05	-8.616e-05	6.1e-05

Controles adicionales en el modelo condicional: maxEducLevel, formal, microEmpresa, sizeFirm, college

De los resultados se puede evidenciar que cuando se incluyen variables de control adicionales, la brecha de salario de género es mayor, dado que el coeficiente de la variable Female es más negativo en el modelo condicional. Esto sugiere que, al controlar por variables adicionales como nivel educativo, tipo de empresa y tamaño, la brecha salarial de género se amplía en lugar de reducirse.

Al comparar los resultados de los modelos incondicional y condicional, se observa que el coeficiente de Female cambia significativamente. En el modelo incondicional, el coeficiente de Female es positivo y no significativo, mientras que en el modelo condicional es negativo y más significativo. Este cambio puede indicar una mezcla de problemas de selección y discriminación. Los cambios en coeficientes bajo los diferentes ajustes de los modelos (solo género, incorporación de controles y análisis de modelo condicional e incondicional) pueden explicar esto, ya que la discriminación es visible al igualar las condiciones de los trabajadores y solo cambiar el género. Sin embargo, este mismo análisis aplica para problemas de selección dado el aumento de los coeficientes cuando se agregan más controles a los modelos.

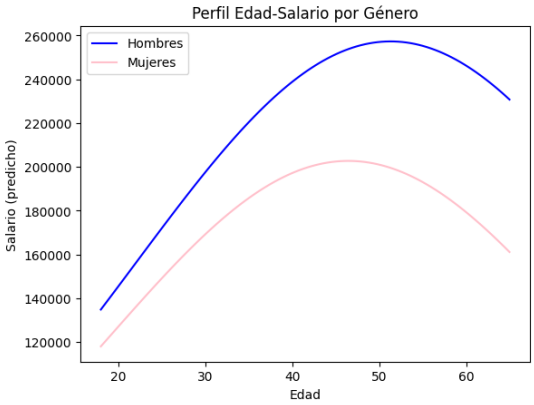


Figure 6: Perfil Edad-Salario por Género

	Edad de pico	IC al 95%
Hombres	51.38	[49.55, 53.74]
Mujeres	46.52	[44.90, 48.42]

Figure 7: Edad de pico por género

Además, para los datos de estudio, la edad pico de los hombres es mayor que la de las mujeres, lo que también puede reforzar los problemas de discriminación (ver Figura 6 y Tabla 7).

5 Predecir Ingresos

Con el objetivo de construir un modelo predictivo y en base a lo anterior se probaron 8 modelos de los cuales se presentan los 5 mejores en terminos del rmse. Estos parten de usar Backward Elimination de las variables que se tenían en la base de datos despues de añadir el cuadrado de las numericas y usar dummies para las categoricas(el primer nivel de cada dummy se eliminó). Luego, se utilizó Lasso para eliminar las que menos aportaban al modelo, y finalmente, se revisó la correlación entre las variables para quedarse con un conjunto de 17 variables que conforman el Modelo 0(el rmse de lasso es ligeramente menor al de los modelos seleccionados pero no se usó porque se queria implementar lo otros métodos para responder el punto y no dejarlo simpelente con el resultado de usar Lasso, por eso no se usa para nada más el modelo que se entrena con Lasso). A partir de estas 17 variables, se exploraron 6 especificaciones adicionales, incorporando interacciones y no linealidades para evaluar si estas mejoraban la capacidad predictiva y el ajuste del modelo a los datos.
Las especificaciones de los modelos son:

1. **Modelo 0:**

$$\begin{aligned}\text{Log\_salario} = \beta_0 &+ \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{p6426} + \beta_4 \text{p6870} \\ &+ \beta_5 \text{p7070} + \beta_6 \text{maxEducLevel} + \beta_7 \text{totalHoursWorked} \\ &+ \beta_8 \text{formal} + \beta_9 \text{cuentaPropia} + \beta_{10} \text{microEmpresa} \\ &+ \beta_{11} \text{estrato1\_3} + \beta_{12} \text{estrato1\_4} + \beta_{13} \text{estrato1\_5} \\ &+ \beta_{14} \text{estrato1\_6} + \beta_{15} \text{p6210\_6} + \beta_{16} \text{relab\_2} \\ &+ \beta_{17} \text{relab\_5} + \varepsilon\end{aligned}$$

2. **Modelo 1:**

$$\begin{aligned}\text{Log\_salario} = \beta_0 &+ \beta_1 \text{age} + \beta_2 \text{totalHoursWorked} + \beta_3 \text{sex} \\ &+ \beta_4 \text{p6210\_6} + \beta_5 \text{formal} + \beta_6 \text{totalHoursWorked}^2 \\ &+ \varepsilon\end{aligned}$$

3. **Modelo 2:**

$$\begin{aligned}\text{Log\_salario} = \beta_0 &+ \beta_1 \text{p7070} + \beta_2 \text{maxEducLevel} + \beta_3 \text{cuentaPropia} \\ &+ \beta_4 \text{p7070}^2 + \beta_5 \text{maxEducLevel}^2 + \varepsilon\end{aligned}$$

4. **Modelo 3:**

$$\begin{aligned}\text{Log\_salario} = \beta_0 &+ \beta_1 \text{age} + \beta_2 \text{p6426} + \beta_3 \text{sex} \\ &+ \beta_4 \text{formal} + \beta_5 \text{age}^2 + \beta_6 \text{p6426}^2 \\ &+ \varepsilon\end{aligned}$$

5. **Modelo 7:**

$$\begin{aligned}\text{Log\_salario} = \beta_0 &+ \beta_1 \text{age} + \beta_2 \text{totalHoursWorked} + \beta_3 \text{formal} \\ &+ \beta_4 \text{estrato1\_3} + \beta_5 \text{estrato1\_4} + \beta_6 \text{estrato1\_5} \\ &+ \beta_7 \text{estrato1\_6} + \beta_8 \text{p6210\_6} + \beta_9 \text{relab\_2} \\ &+ \beta_{10} \text{relab\_5} + \beta_{11} \text{p6426} + \beta_{12} \text{age}^2 \\ &+ \beta_{13} \text{totalHoursWorked}^2 + \beta_{14} (\text{age} \times \text{formal}) \\ &+ \beta_{15} (\text{totalHoursWorked}^2 \times \text{estrato1\_3}) \\ &+ \beta_{16} (\text{p6426} \times \text{age}^2) + \varepsilon\end{aligned}$$

Después de entrenar cada uno de los modelos y probar en el subconjunto de la muestra para el test, calculamos los siguientes rmse para los modelos y ordenamos para ver cuales tuvieron el mejor desempeño:

Nombre del Modelo	**RMSE**
Modelo 0	0.546178
Modelo 7	0.555257
Modelo 1	0.617017
Modelo 2	0.700284
Modelo 3	0.703800

Table 6: Resultados de los Modelos y sus RMSE

Es importante recalcar que el objetivo principal de este análisis fue evaluar el desempeño predictivo de los diferentes modelos mediante el **RMSE**. Luego de entrenar y evaluar los cinco mejores modelos, podemos evidenciar que el **Modelo 0** y el **Modelo 7** son los que presentan los menores RMSE. Esto demuestra que estos dos modelos son los más adecuados para predecir los ingresos en nuestra muestra, dado su ajuste a la especificación de las ecuaciones. Los modelos con mayores no linealidades y más interacciones

lograron capturar mejor las complejidades de la relación entre las variables predictoras y los ingresos. Por otro lado, los modelos que parecían ser más simples no tuvieron el mismo desempeño predictivo, siendo este menor. Así lo reflejan los Modelos 1, 2 y 3, mostrando que la capacidad de predecir con precisión se ve influenciada por la falta de interacciones o variables. Se puede evidenciar que el **Modelo 0** es el que presenta mayor precisión predictiva de los estudiados, con un **RMSE** de 0.5462. Lo que explica que este modelo sea el mejor es la combinación de variables sociodemográficas, características del trabajador y, sobre todo, la incorporación de no linealidades al incluir interacciones de las variables al cuadrado y el uso de variables categóricas de alta importancia, como el estrato socioeconómico. El bajo **RMSE** de este modelo sugiere que la inclusión de los elementos mencionados previamente representa un aumento en la complejidad del modelo, lo que resulta en una mayor precisión al momento de predecir los ingresos.

Ahora bien, en la imagen se evidencia la distribución de los errores, aquí se observa que la mayoría de los errores se concentran cerca de 0, lo cual refuerza la capacidad predictiva del modelo. A continuación, se presenta la gráfica:

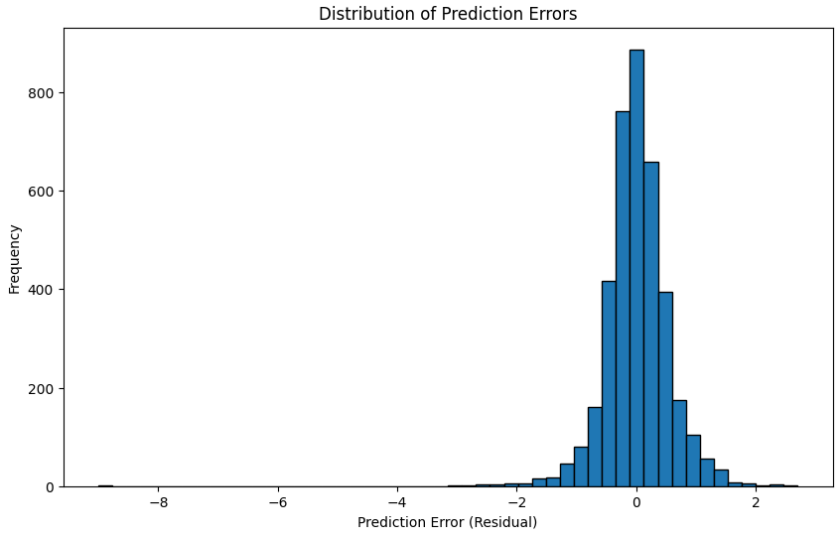


Figure 8: Distribución de los errores de predicción

En la cola de la distribución, tanto del lado negativo como del lado positivo, se encunetran las observaciones con errores más grandes, lo que muestra que algunas predicciones se desvían ampliamente de lo observado. De hecho, hay observaciones donde los errores son considerablemente negativos, lo que puede indicar una sobrepredicción de los ingresos en estos casos.

Estos casos extremos pueden generar preocupación y son dignos de revisión por parte de las autoridades, como la DIAN. Sin embargo, es posible que algunos de estos errores se expliquen por ingresos atípicamente bajos, lo cual podría estar fuera del alcance del modelo. Esto, sumado a los errores de predicción inherentes a cualquier modelo, sugiere que una investigación enfocada en los casos extremos podría ser prudente para determinar si existen anomalías que afecten significativamente la distribución.

Partimos de los dos modelos con menor RMSE para calcular el error con la aproxima-

cion de LOOCV intentando encontrar el modelo con menor error. Al minimizar el RMSE se busca construir un modelo de predicción de ingresos que sea lo más preciso posible, capturando las complejidades de los datos

Table 7: Comparación de RMSE entre Validación y LOOCV

Modelo	**RMSE (Validación)**	**RMSE (LOOCV)**
Modelo 0	0.546178	0.5347041
Modelo 7	0.555257	0.5459413

La tabla compara el error cuadrático medio (RMSE) de los dos modelos bajo dos enfoques: validación y LOOCV. Para el Modelo 0, el RMSE utilizando LOOCV es ligeramente menor al que se obtiene a travez del metodo de validacion. Esto indica que el Modelo 0 tiene un buen rendimiento predictivo y es relativamente robusto, ya que el RMSE de LOOCV, que es una evaluación más confiable, muestra una pequeña mejora. El Modelo 7 también muestra un patrón similar sin embargo su RMSE es marginalmente mayor.

En ambos modelos, los valores de RMSE son menores en el enfoque de LOOCV, lo que sugiere que estos modelos pueden generalizar bien a nuevos datos y no dependen excesivamente de una sola partición específica del conjunto de datos. La ligera reducción en el RMSE con LOOCV implica que ambos modelos son relativamente estables y no están significativamente afectados por la presencia o ausencia de observaciones individuales. Esto refleja que el Modelo 0, con un RMSE más bajo en ambos enfoques, es marginalmente mejor en términos de rendimiento predictivo que el Modelo 7

5.1 Conclusion

El análisis realizado proporciona una visión integral de los factores que influyen en los ingresos laborales en Colombia, destacando cómo la edad, el género, el nivel educativo y las características del empleo afectan la distribución salarial. El modelo predictivo revela que, además de las variables tradicionales como la experiencia y la educación, existen disparidades de género y diferencias asociadas al contexto socioeconómico que contribuyen a la desigualdad salarial. Estos hallazgos no solo ofrecen una herramienta valiosa para predecir ingresos individuales, sino que también aportan información para el diseño de políticas orientadas a reducir la brecha salarial y promover una mayor equidad en el mercado laboral colombiano.