Big Data y Machine Learning

Problem set 1 - Modelo de salario individual por hora

Profesor:
Ignacio Sarmiento-Barbieri

Elaborado por:
Laura Sarif Riversa Sanabria;
Jorge Eliecer Viafara Morales;
Nicolas Jacome Velasco;
Zaira Alejandra Garcia Bernal.

Link Repositorio: https://github.com/GeorgeWton1986/T1_BDML

2025

Universidad de los Andes

1. Introducción

La Dirección de Impuestos y Aduanas Nacionales (DIAN) es la principal fuente de ingresos del Gobierno Central, según el Comité Autónomo de la Regla Fiscal (2024). No obstante, la economía colombiana enfrenta un déficit fiscal impulsado, en parte, por una recaudación tributaria inferior en 10,8 billones de pesos a lo proyectado en el Marco Fiscal de Mediano Plazo (MFMP) hasta noviembre de 2024. La gestión de la DIAN es clave en el control tributario y la lucha contra la evasión fiscal (González, 2018). Díaz y González (2024) destacan que una fiscalización efectiva requiere herramientas gerenciales y operativas que permitan detectar focos de evasión y contrabando.

Diversos estudios atribuyen la evasión fiscal a factores como el bajo nivel educativo, la desconfianza en el gobierno y la ineficiencia en la administración de recursos públicos (Pinedo, del Águila y Alvarado, 2022). En contraste, los contribuyentes con mayor cumplimiento presentan una cultura tributaria sólida y estabilidad económica. La evasión fiscal reduce la recaudación, afecta el financiamiento del gasto público y compromete programas esenciales como salud, educación e infraestructura. Además, genera inequidad y perjudica la competitividad empresarial (Yikona, 2011). Espitia y Suárez (2017), así como Rojas, Martínez, Álvarez y Farfán (2024), señalan que la evasión contribuye al déficit fiscal y puede provocar crisis financieras, especialmente en economías con alta informalidad laboral y menor confianza en las instituciones (Fergusson, Molina y Riaño, 2017).

Este estudio estima un modelo econométrico con datos de la Gran Encuesta Integrada de Hogares (2018) del DANE para Bogotá. Se analizarán las condiciones salariales de personas mayores de 18 años, empleando distintas metodologías para predecir ingresos. La estructura del trabajo incluye la limpieza de datos, análisis de variables, predicción de ingresos y conclusiones.

2. Datos

En la presente sección se realiza una descripción de los datos y el proceso de adquisición y limpieza obtenidos en la Gran Encuesta Integrada de Hogares (GEIH) para 2018, realizada por el Departamento Administrativo Nacional de Estadística para Bogotá del "Informe de Pobreza Monetaria y Desigualdad". Esta sección se enfoca en las personas empleadas mayores de dieciocho (18) años y tiene una muestra anual aproximada de 315,000 hogares a nivel nacional.

La GEIH proporciona información estadística del tamaño y estructura de la fuerza laboral, incluyendo datos sobre empleo, desempleo y población fuera de la fuerza de trabajo, los ingresos laborales y no laborales de los hogares, y la pobreza monetaria y extrema de la población residente en el país. Esta encuesta permite caracterizar a la población según sexo, edad, parentesco con el jefe del hogar, nivel educativo, afiliación al sistema de seguridad social en salud, grupos poblacionales y otras formas de trabajo, como producción de bienes y servicios para autoconsumo y trabajo en formación, entre otros.

Proceso de adquisición de los datos: Para llevar a cabo el análisis, utilizamos "pacman" para facilitar la carga e instalación de varios paquetes en R que permiten la recolección, manipulación, exploración y visualización de datos. Así mismo, utilizaremos "p_load( )" para instalar y cargar múltiples paquetes en una sola línea de código.

La base de datos se encuentra almacenada en una página web, por lo que se realizó scrapping del sitio web https://ignaciomsarmiento.github.io/GEIH2018_sample/. En primer lugar, se exploró la URL para identificar la información almacenada, evidenciando que se encontraba dividida en 10 data chunks. Se inspeccionó cada una de ellas y se realizó una iteración para consolidar la información de cada tabla y crear un data frame llamado df_GEIH con 32.177 observaciones con 177 variables. En el proceso de extracción, se emplea el paquete rvest para analizar el contenido de la página web y el paquete dplyr para combinar la lista de tablas en el data frame, adicionalmente, se aplica una primera limpieza de los datos con el condicional if para eliminar la columna innecesaria de índice.

Proceso de limpieza de los datos: En primer lugar, se filtraron las personas ocupadas mayores de 18 años utilizando las variables "ocu" y "age". El término de personas ocupadas hace referencia a aquellos individuos que están trabajando en alguna actividad económica, ya sea de manera formal o informal, y que se encuentran dentro de la edad laboral. Para este análisis, se consideró que la edad legal para trabajar en Colombia es de 18 años, a partir de la cual no se aplican las restricciones que rigen para los menores de edad.

Lo anterior fue el resultado de la exploración de los datos, en la que se identificaron variables similares, como "ocu" y "dsi". Ambas son variables dicotómicas: "dsi" toma el valor de 1 si la persona está desempleada y 0 en caso contrario, mientras que "ocu" toma el valor de 1 si la persona está ocupada y 0 si no lo está. Dado que nuestro análisis se enfoca en las personas que están trabajando, se eligió la variable "ocu", ya que permite estudiar específicamente a las

personas ocupadas. De este filtro, se da como resultado una base de datos con 16.542 observaciones con 177 variables.

Elección de variables: Se construye un nuevo dataframe a partir de la Gran Encuesta Integrada de Hogares (GEIH), seleccionando las variables relevantes para el presente análisis. El objetivo es modelar y explicar el salario por hora de una persona en función de diversas características individuales y laborales. A continuación, se describen las variables elegidas y su pertinencia en la estimación del modelo, justificando su inclusión con base en la teoría económica y la literatura empírica sobre determinantes salariales.

- **Y_total_m_ha (Ingreso total mensual):**Variable dependiente que representa el salario mensual total del trabajador, eje central en estudios de salarios. En términos estadísticos, esta variable presenta 1.778 valores faltantes, lo que representa un 89,25% de los datos completos. La media del salario por hora es de 8.541 COP, con una desviación estándar de 13,86. Los valores oscilan entre un mínimo de 0,47 COP y un máximo de 350.583 COP. El primer cuartil es de 3.796 COP, la mediana se encuentra en 4.837 COP y el tercer cuartil en 7.899 COP. El histograma revela una asimetría positiva debido a la presencia de valores extremos elevados.
- **Age (Edad):** La edad influye en los salarios, ya que está vinculada a la experiencia laboral y las habilidades adquiridas a lo largo del tiempo. En general, la relación entre edad y salario es positiva hasta cierto punto, ya que la experiencia suele traducirse en mayores ingresos. Sin embargo, en etapas más avanzadas, esta relación puede estabilizarse o incluso disminuir. Según Castillo-Robayo y García-Estévez (2019), los jóvenes enfrentan tasas de desempleo más altas en el mercado laboral, mientras que las personas de mayor edad suelen experimentar períodos de desempleo más cortos, aunque de mayor duración. Estadísticamente, la edad no presenta valores faltantes y cuenta con el 100% de los datos completos. La media de la edad es de 39 años, con una desviación estándar de 13,48. El mínimo registrado es 18 años, el primer cuartil es 28 años, la mediana 38 años y el tercer cuartil 50 años. El valor máximo es de 94 años. A partir del histograma, se observa una mayor concentración de individuos en edades menores al segundo y tercer cuartil.
- **Sex (Género):** Variable dummy (1 = mujer, 0 = hombre) utilizada para analizar la brecha salarial de género. Según la teoría de discriminación laboral de Joan Robinson (1933), las mujeres pueden recibir menores salarios debido a una menor elasticidad en su oferta laboral (Cuervo Alvarado, 2022). En términos estadísticos, esta variable no presenta valores faltantes y cuenta con el 100% de los datos completos.
- **MaxEducLevel (Máximo nivel educativo):** Variable clave que afecta el salario, ya que un mayor nivel educativo se asocia con mayor capital humano y mejores oportunidades laborales, según la teoría de Becker (1964). En este estudio, la variable presenta solo un valor faltante, lo que representa un 99,99% de datos completos. Esta variable es categórica y se distribuye en niveles que van desde "Ninguno" hasta "Superior o universitaria". Según el histograma, la categoría más frecuente es "No sabe, no informa".
- **P6426 (Tiempo en la empresa):** El número de años que una persona ha trabajado en la misma organización, puede estar relacionado con el aumento salarial por antigüedad. El capital humano, como lo describen Becker y Lucas Jr., es un motor del desarrollo organizacional y un activo clave para las empresas (Cuervo Alvarado, 2022). Esta variable no presenta valores faltantes y cuenta con el 100% de los datos completos. En promedio, los trabajadores han estado en sus empresas por 63,76 meses (5,3 años), con una desviación estándar de 89,48 meses. El primer cuartil es de 0 meses, la mediana de 24 meses (2 años), el tercer cuartil de 84 meses (7 años) y el máximo de 720 meses (60 años). El histograma sugiere una distribución sesgada a la derecha, indicando que la mayoría de los trabajadores tienen menor antigüedad.
- **P6870 (Tamaño de la empresa):** El tamaño de la empresa es una variable clave en la determinación salarial. Según Cuervo Alvarado (2022), las pequeñas empresas (1-10 empleados) juegan un papel relevante en la industria, aunque las empresas grandes tienden a ofrecer mejores salarios y compensaciones. Esta variable no presenta valores faltantes y tiene el 100% de los datos completos. Se clasifica en categorías que van desde trabajadores independientes hasta empresas con más de 100 empleados. El análisis de la distribución sugiere que la mayoría de los trabajadores están empleados en empresas de gran tamaño.
- **Depto (Departamento):** El departamento donde vive un empleado puede influir en el salario, debido a diferencias en especialización, demanda de habilidades y tipo de empleo disponible. En este estudio, la variable no presenta valores faltantes y cuenta con el 100% de los datos completos. Se identificó que todas las observaciones corresponden a Bogotá.
- **Formal (Formalidad del empleo):** Distinción entre empleo formal e informal. Según Orlando (2000), los trabajadores formales gozan de mejores salarios y estabilidad laboral, mientras que los informales enfrentan menores ingresos y mayor precariedad. Finalmente, la variable no presenta valores faltantes y cuenta con el 100% de los datos completos. Se identificó que todas las observaciones corresponden a Bogotá.
- **P6100 (Seguridad social):** La OIT hace referencia al régimen al que un trabajador está afiliado dentro del sistema de salud. El acceso a la seguridad social, especialmente a través de los diferentes regímenes (contributivo, especial y subsidiado), está estrechamente vinculado con la estabilidad económica y las oportunidades de ingresos. La variable presenta 1.403 valores faltantes, lo que indica un 91,51% de datos

completos. El análisis de distribución muestra que la mayoría de los trabajadores están afiliados al régimen contributivo, seguido del régimen especial.

Por último, la base de datos incluye los tipos de datos Numeric, con la variable continua *salario_hora*, e Integer, con las variables categóricas *age, ocu, sex, max_nivel_educ, tiempo_empresa, tamaño_empresa* y *depto*.

Valores faltantes: Se procedió a identificar los datos faltantes. Para algunas variables, se utilizó el promedio de los valores disponibles, lo que resultó en una nueva base de datos denominada "df_salario" con 16,542 observaciones y 10 variables. En cuanto a la imputación de los valores faltantes, se abordó de acuerdo con el tipo de variable.

Primero las variables categóricas, como educación y seguridad social, se ordenaron de manera descendente y se calculó el valor más común en cada categoría, que fue utilizado para reemplazar los datos faltantes. Luego, la variable continua, salario por hora, se analizó su distribución mediante un gráfico, observándose que presenta una cola hacia la derecha. Debido a esta distribución, se optó por imputar los valores faltantes con la mediana, ya que representa mejor el valor central de los datos. Finalmente, se verificó que no quedaran valores faltantes en la base de datos. Con el resumen de estadísticas de la nueva data frame creado, se identifica la cantidad de valores faltantes (n_missing) en las variables salario_hora, segu_social y max_nivel_educ. A continuación, se procede a validar la información e imputar los valores faltantes de acuerdo a las características de las variables para poder estimar los modelos más adelante.

GRÁFICO 1 - PORCENTAJE DE VALORES FALTANTES POR VARIABLES

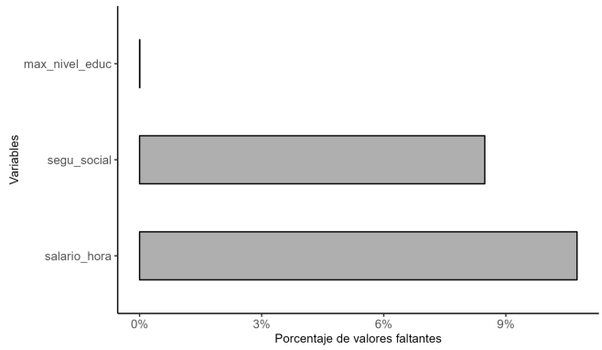
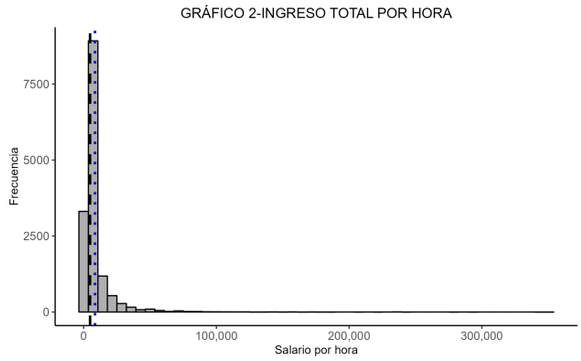


GRÁFICO 2-INGRESO TOTAL POR HORA



3. Perfil de edad - salario

En economía es usual formular modelos con variables que tengan una condición cuadrática, este tipo de modelos lo que buscan es captar los efectos marginales decrecientes o crecientes en la función. En este sentido, el modelo de estimación planteado de *edad-salario* está en función de una sola variable Age.

Además, es importante recordar que el estimador β_1 no mide la variación en Log (w), porque, al realizar un análisis ceteris paribus sobre la variable Age^2, la variable Age será constante. De la estimación anterior, se tiene una aproximación luego de derivar la variable Log (w) entre la variable Age, así:

Ahora, se puede mencionar que, la relación Log (w) y ΔAge depende de la variable Age. Entonces, si se reemplaza el valor de 0 en la variable Age, se puede interpretar el coeficiente como el cambio marginal aproximado de pasar de 0 a 1. A continuación, estimaremos los coeficientes de la siguiente ecuación:

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

$$log(\hat{w}) = \hat{\beta}_0 + \hat{\beta}_1 Age + \hat{\beta}_2 Age^2 + u \quad (2)$$

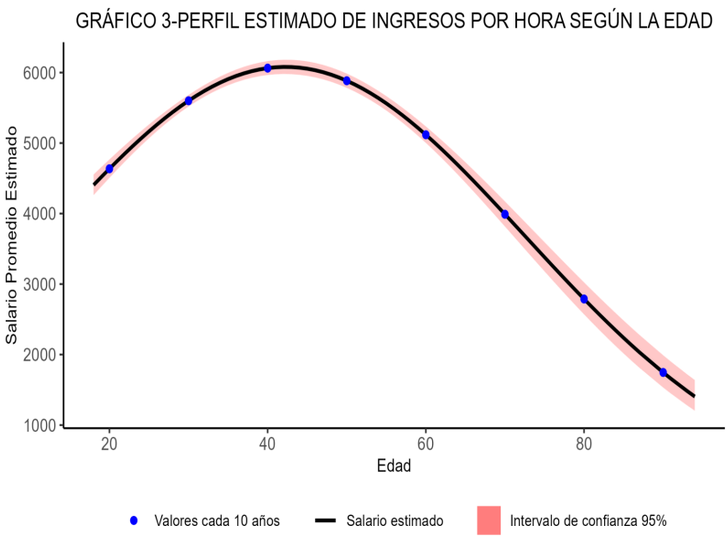
$$\frac{\Delta \log(\hat{w})}{\Delta Age} = \hat{\beta}_1 + 2\hat{\beta}_2 Age + u \quad (3)$$

TABLA 1-MODELO DE SALARIO

	Dependent variable:
	Log(Salario por Hora)
Edad	0.046***
	(0.003)
Edad al Cuadrado	-0.001***
	(0.00003)
Constant	7.735***
	(0.052)
Observations	16,542
R2	0.019
Adjusted R2	0.019
Residual Std. Error	0.781 (df = 16539)
F Statistic	163.894*** (df = 2; 16539)
Note:	*p<0.1; **p<0.05; ***p<0.01

Interpretación de los resultados de la regresión log (salario)

El resultado de la estimación indica que el coeficiente Age es positivo y Age^2 es negativo, esto predetermina que, valores pequeños en Age tiene un efecto positivo en log(w), por el contrario valores grandes en Age genera una condición de decrecimiento sobre el log(w). De esta manera, tomaremos los múltiplos de 18 años para realizar el análisis del cambio Δlog(w). Veamos, cuando la variable Age es igual a 18 años, Δlog(w) es igual a 2,66% Ahora, cambiamos el valor a 36 años, Δlog(w) es igual a 0,69% por último, Age sera 54 años y por lo tanto Δlog(w) es igual a -1,29%, lo cual responde a lo mencionado al principio de esta sección.



Con el análisis anterior, podemos observar que el comportamiento de la ecuación estimada es parabólico, es decir, en la "edad pico" se encuentra el punto de inflexión. Todas las edades menores a la "edad pico", la variable Δlog(w) será creciente y para edades mayores a la "edad pico" la variable Δlog(w) será decreciente. Así las cosas, es necesario conocer cómo se calcula el punto de inflexión.

$$\frac{\Delta \log(\hat{w})}{\Delta \text{Age}} = 0 \quad (4)$$

$$\frac{\Delta \log(\hat{w})}{\Delta \text{Age}} = \hat{\beta}_1 + 2\hat{\beta}_2 \cdot \text{Age} \quad (5)$$

$$0 = \hat{\beta}_1 + 2\hat{\beta}_2 Age \quad (6)$$

$$\text{Age}_{\text{pico}} = -\frac{\hat{\beta}_1}{2\hat{\beta}_2} \quad (7)$$

El punto de inflexión de la ecuación estimada donde está localizada la edad pico corresponde a 42.25 años. *Significancia:* Según los resultados obtenidos en la regresión todos los coeficientes obtenidos son estadísticamente significativos al 0.01 (99%). *Bondad de ajuste:* El R^2 es 0.0194 y R^2 ajustado da el valor de 0.01932 , esto indica que el modelo explica el salario en un 1.93%. Estos valores bajos indican que el modelo captura una parte de la relación y existen otros factores que afectan el logaritmo de salario no incluidos en el modelo.

Análisis de la "edad pico" con sus respectivos intervalos de confianza: Luego de estimar y analizar las ecuaciones y gráficas podemos sugerir que el comportamiento del salario en función de la edad en la población encuestada indica que los salarios son bajos a la edad de 18 años. Además, se nota que el salario mejora con el aumento de los años. No obstante, el ritmo de crecimiento del salario es más grande en comparación con edades medias, como por ejemplo, los 36 años, donde se registró un cambio de 0,69%. De otro lado, se registró la edad pico para este estudio en 42.25 años de edad, este punto es importante por dos razones. La primera, es el punto de inflexión entre el cambio positivo y el cambio negativo en el ingreso. La segunda, que en promedio se referencia la edad pico a los 50 años, esto puede sugerir que podemos tener otras variables que no fueron incluidas en la estimación que generen una disminución de la edad y obtener mejores salarios.

En conclusión, es de notar que esta variable no es la única que determina el valor, podríamos incluir, por ejemplo, la industria, la experiencia laboral, la cantidad de años de estudio, ranking institución educativa, etc. así como otras habilidades sociales que pueden ser difíciles de cuantificar, en cuyo casos podría ayudar a explicar el comportamiento del salario en función de la edad.

4. La brecha salarial de género

$$log(w) = \beta_1 + \beta_2 Female + u \quad (8)$$

TABLA 2-MODELO DE SALARIO SEGÚN GÉNERO

	Dependent variable:
	Log(Salario por Hora)
female	-0.079***
	(0.012)
Constant	8.646***
	(0.008)
Observations	16,542
R2	0.002
Adjusted R2	0.002
Residual Std. Error	0.787 (df = 16540)
F Statistic	41.438*** (df = 1; 16540)
Note:	*p<0.1; **p<0.05; ***p<0.01

Los responsables de las políticas se han preocupado durante mucho tiempo por la brecha salarial de género. Con la siguiente definición del modelo, se obtiene: El modelo es significativo al 0,01 (99%), es decir, la variable del modelo explica el salario por hora que el modelo explica el salario en un 0.2%. El *error estándar* para este modelo es de0.787 y esto indica la dispersión entre las 16542 observaciones y la estimación. La variable female es significativa al 0.01 (99%), el error estándar es de 0.012 y explica que, si es mujer, en promedio el salario disminuye en 7.9%

¿Salario igual para trabajos iguales?

En esta sección, se utiliza el modelo *Frisch-Waugh-Lovell* (FWL) para estimar la brecha salarial condicional, incorporando variables de control que reflejan las características semejantes de los trabajadores y los puestos de trabajo. De acuerdo con la tabla 3, el coeficiente de la variable "female" en ambos modelos sugiere que, en promedio, las mujeres ganan un 10.7% menos que los hombres en salario por hora, incluso después de controlar por variables como el nivel educativo, la antigüedad en la empresa y la edad. Ambas variables son estadísticamente significativas, y se observa que el error estándar del coeficiente en el modelo condicional es 0.011, mientras que en el modelo incondicional es ligeramente más alto, 0.012.

TABLA 3-COMPARACIÓN DE ESTIMACIONES DE GÉNERO EN SALARIO

	Dependent variable:	
	Log(Salario por Hora)	
	Modelo Condicional (1)	Modelo Incondicional (2)
Mujer	-0.107***	
	(0.011)	
Nivel Educativo	0.284***	
	(0.005)	
Tiempo en Empresa	0.001***	
	(0.0001)	
Edad	0.005***	
	(0.0005)	
Residuos de Mujer		-0.107***
		(0.012)
Constant	6.702***	8.609***
	(0.039)	(0.006)
Observations	16,542	16,542
R2	0.192	0.005
Adjusted R2	0.192	0.005
Residual Std. Error	0.709 (df = 16537)	0.786 (df = 16540)
F Statistic	983.030*** (df = 4; 16537)	75.872*** (df = 1; 16540)
Note:	*p<0.1; **p<0.05; ***p<0.01	
	Significancia: * p<0.1, ** p<0.05, *** p<0.01	

El R^2 de ambos modelos es relativamente bajo, lo cual indica que estos modelos no tienen un buen ajuste con la muestra. Generalmente, se espera que el R^2 se acerque al 40% - 80% para considerar los modelos más explicativos, pero en este caso ambos valores son menores (0.192) y (0.005), lo que sugiere que las variables elegidas no explican del todo las diferencias salariales. El estadístico F en ambos modelos es significativo, lo que indica que al menos una de las variables es relevante para predecir el salario por hora. Sin embargo, el modelo condicional tiene un ajuste superior, ya que se controla por variables clave que influyen en los salarios, lo que lo convierte en un modelo más robusto y confiable en comparación con el modelo incondicional.

Al comparar las estimaciones de los coeficientes para female en los dos modelos (uno con controles y otro con los residuos de female), se observa que las estimaciones son iguales cuando se utiliza el Bootstrap. El valor del error estándar de la estimación de female mediante el método FWL con bootstrap es 0.011, lo que es consistente con los resultados obtenidos en el modelo con controles. Además, se observa que la variabilidad en la estimación de female es más baja en el modelo con bootstrap, en comparación con el modelo FWL ajustado por los residuos, lo que sugiere una mayor precisión en las estimaciones obtenidas mediante el bootstrap.

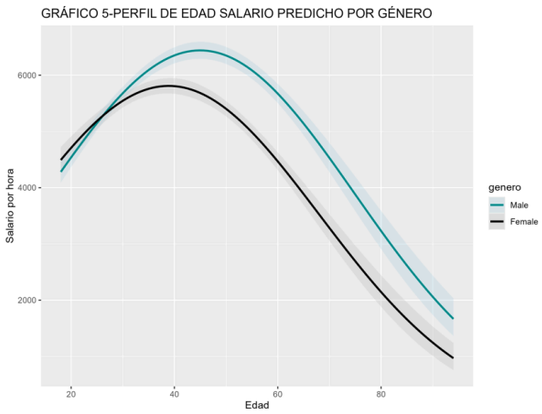
TABLA 4-COMPARACIÓN BETAS FEMALE

	Beta_female
female	-0.107
residuos_female	-0.107

TABLA 5-BOOTSTRAP ORDINARIO NO PARAMÉTRICO

	Estadístico	Valor
1	Estimación Original	-0.107
2	Sesgo	-0.0002
3	Error Estándar	0.011

Se puede concluir que la brecha salarial de género sigue siendo evidente. El coeficiente negativo para female en ambos modelos indica que las mujeres ganan, en promedio, menos que los hombres, lo que confirma la persistencia de la brecha salarial.

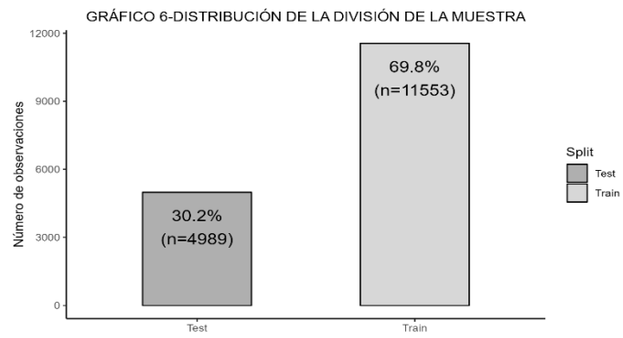


El gráfico muestra la relación entre la edad y el salario por hora predicho para hombres y mujeres, junto con los intervalos de confianza. Se observa una relación cuadrática, donde los salarios aumentan hasta un punto máximo alrededor de los 40-50 años y luego disminuyen, lo que concuerda con la teoría del capital humano. Se evidencia una brecha salarial de género: los hombres tienen salarios más altos que las mujeres en casi todas las edades, con una diferencia más pronunciada en la mitad de la vida laboral. En edades avanzadas, ambos grupos experimentan una caída en el salario, posiblemente debido a menor productividad, reducción de jornada o la pensión. También se observa indicando mayor incertidumbre en las predicciones para los trabajadores más jóvenes y mayores, mientras que en el rango de 30-60 años las estimaciones son más precisas. Este análisis sugiere la existencia de una desigualdad de género en los salarios, que podría estar relacionada con diferencias en oportunidades, segregación ocupacional o interrupciones en la carrera laboral de las mujeres. Un análisis adicional con variables como educación, ocupación o sector podría ayudar a entender mejor esta brecha.

Para concluir este análisis, retomamos lo que señala Sabogal (2012), quien destaca que, en Colombia, las mujeres perciben salarios más bajos que los hombres, a pesar del aumento en su participación laboral, el mayor número de horas trabajadas y ciertas características observables, como el nivel educativo, durante las últimas tres décadas.

5. Predicción del salario

El siguiente gráfico muestra cómo se distribuyen las observaciones entre el conjunto de entrenamiento y el de prueba.



Para realizar una comparación significativa de los rendimientos predictivos de los modelos anteriores, se proporcionaron cinco modelos adicionales que permiten identificar el mejor rendimiento incluyendo modelos no lineales y con mayor complejidad respecto a los estimados previamente. De este modo, se estudiaron un total 8 especificaciones:

$$log(w) = \beta_0 + \beta_1 Female + \beta_2 educ + \beta_3 tiempoempresa + \beta_4 age + u \quad (9)$$

$$log(female) = \beta_0 + \beta_1 educ + \beta_2 tiempoempresa + \beta_3 age + u \quad (10)$$

$$log(w) = \beta_0 + \beta_1 residuos female + u \quad (11) \qquad log(w) = \beta_0 + \beta_1 Female + \beta_2 tamañoempresa + u \quad (12)$$

$$log(w) = \beta_0 + \beta_1 Female + \beta_2 age + \beta_3 female * tiempoempresa + \beta_4 tiempoempresa + u \quad (13)$$

$$log(w) = \beta_0 + \beta_1 Female + \beta_2 tamañoempresa + \beta_3 tamañoempresa^2 + u \quad (14)$$

$$log(w) = \beta_0 + \beta_1 Female + \beta_2 educ + \beta_3 tiempoempresa + \beta_4 age + \beta_5 formal + \beta_6 segusocial + \beta_7 female * segusocial + \beta_8 age^2 + u \quad (15)$$

$$log(w) = \beta_0 + \beta_1 Female + \beta_2 educ + \beta_3 female * educ + \beta_4 age + \beta_5 age^2 + u \quad (16)$$

Luego de definir los modelos, se procede a realizar las predicciones, comparación de los resultados del RMSE e identificación del modelo con el mejor rendimiento predictivo.

TABLA 6-COMPARACIÓN DE RMSE ENTRE MODELOS

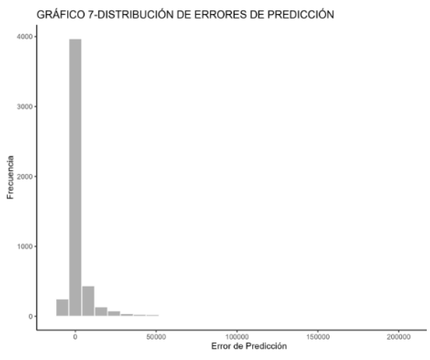
	Modelo	MSE
1	m1	11,121.560
2	m2	13,941.480
3	m3	11,696.090
4	m4	11,442.210
5	m5	11,665.160
6	m6	11,440.190
7	m7	10,940.350
8	m8	11,159.340

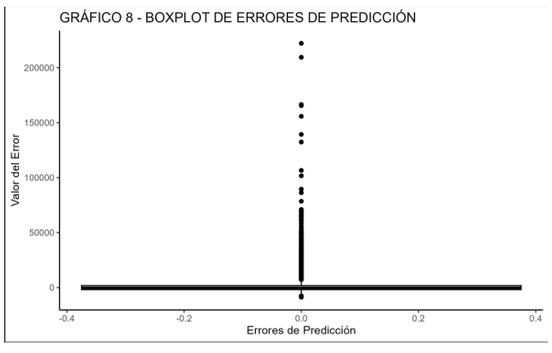
Se evaluaron ocho especificaciones diferentes usando Validation Set Approach para medir el rendimiento predictivo de los modelos con la métrica RMSE, la cual indica un mejor desempeño predictivo entre más bajo sea su nivel. El resultado hace evidente que el Modelo 7 tiene el menor valor de RMSE, con un valor de 13,448.67, por lo tanto, es el mejor en términos de RMSE. Esto sugiere que incluir controles adicionales como si la persona cuenta con seguridad social, si está en un trabajo formal, entre otros, aporta información relevante para predecir el cambio en los salarios.

El Modelo 7 tiene el mejor desempeño predictivo con un RMSE de 10,940.350. Esto es consistente con su mayor nivel de complejidad, ya que incorpora no solo los controles básicos, sino también términos adicionales como la variable formal, segu_social, la interacción female:segu_social y un término cuadrático age^2. El Modelo 1 sigue siendo un modelo competitivo, obteniendo el segundo mejor RMSE (11,121.560). Aunque es más simple que el Modelo 7, sigue capturando buena parte de la variabilidad del salario en función de female, max_nivel_educ, tiempo_empresa y age. Los modelos con no linealidades e interacciones presentan mejoras en algunos casos, como el Modelo 8, que incorpora la interacción female:max_nivel_educ y logra un RMSE similar al Modelo 1 (11,121.560).

En términos generales, los modelos que incluyen mayores niveles de complejidad tienden a mejorar el desempeño predictivo, pero no todos los ajustes mejoran la precisión. La inclusión de interacciones y términos no lineales es útil solo si se justifica teóricamente.

Especificación del modelo con error de predicción: Modelo 7





El histograma de errores de predicción muestra una distribución sesgada a la derecha, lo que indica que la mayoría de los errores son pequeños, pero existen algunas observaciones con diferencias extremadamente grandes entre el salario real y el predicho.

El boxplot confirma la presencia de valores atípicos con errores de predicción elevados. Las líneas rojas en el gráfico representan los percentiles 2.5% y 97.5%, y cualquier punto fuera de este rango es considerado un outlier. Al analizar las características de los outliers, encontramos casos llamativos, como un individuo con un salario de 75000, que es significativamente mayor que el promedio. Otros outliers presentan salarios bajos que tampoco fueron bien predichos por el modelo. Estas diferencias pueden deberse a errores en la recolección de datos, condiciones laborales particulares o factores relevantes que no fueron incluidos en la regresión.

Desde el punto de vista de la DIAN, estos valores atípicos pueden indicar posibles irregularidades. Los individuos con ingresos extremadamente altos podrían estar no declarando información, lo que podría justificar una revisión adicional. Otra posibilidad es que el modelo no haya capturado algunas características clave del mercado laboral, lo que explicaría por qué ciertas observaciones presentan errores tan grandes.

En conclusión, la DIAN podría enfocarse en analizar más a fondo los outliers para determinar si reflejan ingresos no declarados, errores en los datos o simplemente limitaciones del modelo. Para mejorar la precisión de las predicciones, sería recomendable incluir nuevas variables que expliquen mejor los salarios atípicamente altos o bajos. También sería útil realizar un análisis más detallado sobre la naturaleza de estos empleos y sus características tributarias, para evaluar si requieren una revisión adicional.

Cálculo de error predictivo con LOOCV

TABLA 7-COMPARACIÓN DE RMSE ENTRE MODELO 1 Y MODELO 7

	Modelo	RMSE_VSA	RMSE_LOOCV	RMSE_Leverage
1	M1	11,121.560	12,896.130	12,894.710
2	M7	10,940.350	12,723.800	12,721.740

Se compararon los errores predictivos de los modelos M1 y M7 utilizando dos métodos de validación: el Validation Set Approach (VSA) y el Leave-One-Out Cross-Validation (LOOCV). Los resultados muestran que el RMSE obtenido con LOOCV sigue siendo mayor que el obtenido con VSA para ambos modelos. Esto indica que la evaluación con LOOCV puede estar capturando mayor variabilidad en la predicción, en comparación con el VSA, que depende de una única partición de los datos. En ambos métodos, M7 mantiene un RMSE menor que M1, lo que refuerza la conclusión de que M7 tiene un mejor desempeño predictivo en términos de error cuadrático medio.

Los valores de RMSE calculados con LOOCV utilizando leverage muestran que las diferencias con el LOOCV tradicional son mínimas (por ejemplo, para M7, pasa de 12723.80 a 12721.74). Esto sugiere que las observaciones con alto Leverage no están afectando significativamente la estabilidad del modelo en términos de error predictivo. En términos prácticos, esto implica que aunque algunas observaciones pueden tener un alto Leverage, su impacto sobre la predicción global del modelo no es significativo. Si hubiera diferencias marcadas entre LOOCV normal y LOOCV con leverage, eso indicaría que ciertos puntos influyentes distorsionan la estimación de los coeficientes y, por ende, del error predictivo.

En conclusión, el modelo M7 sigue siendo el mejor en términos de error predictivo, independientemente del método de validación utilizado. Además, la inclusión de Leverage en los cálculos de LOOCV confirma que no hay observaciones individuales que estén afectando de manera desproporcionada la estabilidad del modelo. Esto da confianza en la robustez de los resultados obtenidos.

6. Referencias

1. Pinedo, W. C., del Aguila, W. C., & Alvarado, G. D. P. P. (2022). Un análisis de la evasión tributaria.Ciencia Latina Revista Científica Multidisciplinar,6(2), 3224-3241.
2. Pedro A. Cabra-Acela, 2021. “Premiar a los buenos contribuyentes, ¿un mecanismo efectivo? ” Documentos CEDE19419, Universidad de los Andes, Facultad de Economía, CEDE.
3. Leopoldo Fergusson & Carlos Molina & Juan Felipe Riaño, 2017. "Evado impuestos, ¿y qué? Una nueva base de datos y evidencia de Colombia",Documentos CEDE15444, Universidad de los Andes, Facultad de Economía, CEDE.
4. Díaz, D. & González, J. (2024).Controles tributarios y la evasión fiscal en Colombia. [Proyecto aplicado]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/64465
5. Bloom Monterroza, C. C. y Villalba Ayazo, D. S. (2024). Causas y consecuencias de la evasión tributaria [Tesis de pregrado, Universidad Cooperativa de Colombia]. Repositorio Institucional Universidad Cooperativa de Colombia https://hdl.handle.net/20.500.12494/57925
6. El déficit fiscal de 2024 superaría la meta del Marco Fiscal de Mediano Plazo y el ajuste requerido para 2025 es mayor que el contemplado en el decreto de aplazamiento, (CARF, 2024)
7. Rocha Combita, J. (2024) Elusion y evasion fiscal en Colombia. https://bibliotecadigital.iue.edu.co/jspui/handle/20.500.12717/3192
8. Cuervo Alvarado, M. (2022). Análisis de los determinantes de las asignaciones salariales entre hombres y mujeres en la industria manufacturera en Bogotá 2016-201
9. Castillo-Robayo, Cristian Darío, & García-Estévez, Javier. (2019). Desempleo juvenil en Colombia: ¿la educación importa?. Revista Finanzas y Política Económica, 11(1), 101-127. Epub October 10, 2020.https://doi.org/10.14718/revfinanzpolitecon.2019.11.1.7
10. Sabogal, A. (2012). Brecha salarial entre hombres y mujeres y ciclo económico en Colombia.
11. Becker, G. (1964). Human capital: a theoretical and empirical analysis, with special reference to education, Columbia University Press for the National Bureau of Economic Research.
12. Sabogal, A. (2012). Brecha salarial entre hombres y mujeres y ciclo económico en Colombia.