Profesor: Ignacio Sarmiento

Camila Beltrán - 202213110
Julian Pulido - 201814833
Alejandra Guevara - 202324570
Juan P. Rodríguez - 201716771

Problem Set 1

Big Data y Machine Learning para Economía Aplicada, 2024-I
Fecha: **3 Marzo 2024**.

Universidad de los Andes

Índice

Índice de cuadros

Índice de figuras

1. Introducción

La evasión de impuestos es un tema de interés de política pública para cualquier país, considerando el impacto que puede tener la recaudación fiscal en términos de estabilidad fiscal e inversión estatal. Por ejemplo, para Estados Unidos The Internal Revenue Service (2016) evidencia que hay una diferencia entre la estimación del recaudo y lo que realmente se recauda, lo cual se conoce como una brecha tributaria, del 15 %. Es decir, que por cada dólar que debería recaudar el Estado, solo 85 centavos son pagados de manera voluntaria y oportuna. En Colombia, Ávila y Cruz (2015) reflejan que la tasa de evasión del Impuesto al Valor Agregado (IVA) para 2012 fue de 23 %, mientras que el impuesto a la renta de personas jurídicas y naturales en 2006 se estimó en 32 %. Por su parte, Parra y Patiño (2010), estiman a través de un modelo de cointegración una tasa de evasión de 31,4 % entre el impuesto del IVA y el impuesto de renta para 2009.

En línea con esta problemática, Ávila y Cruz (2015) ofrecen tres explicaciones para la brecha tributaria existente en Colombia. En primer lugar, un sub-reporte en la declaración de renta de personas naturales (*under-reporting*), personas que no declaran (*non-filing*), o un menor pago en las obligaciones con el estado (*underpayment*). El primer elemento es un problema complejo de resolver, en especial porque se estima que el 55,5 % de la población ocupada tiene un empleo informal (DANE 2023), y las transacciones en efectivo permiten ocultar la información a la Dirección de Impuestos y Aduanas Nacionales (DIAN).

Este trabajo realiza un análisis predictivo del ingreso de las personas ocupadas (formales e informales) en Bogotá con una muestra de la Gran Encuesta Integrada de Hogares (GEIH), con el propósito de que los resultados puedan servir de insumo para los análisis posteriores de evasión fiscal en Colombia, en particular en la declaración de renta de las personas naturales. La GEIH es realizada por el Departamento Administrativo Nacional de Estadística (DANE) en Colombia. Esta encuesta aborda una variedad de temas a nivel de hogar, incluida información detallada sobre el mercado laboral, como ingresos, cargo, sector económico, entre otros. Además, la GEIH recopila datos sobre condiciones socioeconómicas, como tipo de vivienda, nivel educativo, acceso a servicios de salud y pensión, así como información demográfica, como el número de hijos en el hogar.

Es importante resaltar que los resultados de la predicción de ingresos están sujetos a la presencia de sesgos inherentes de las encuestas. En primera instancia, la veracidad de la información depende de la información que brinda el encuestado. Si bien algunas preguntas como el número de hijos o la ocupación pueden ser fácilmente verificadas, los ingresos laborales o propiedades pueden ser variables más difíciles de corroborar. En particular, se pueden observar problemas de *truncamiento*[1] o *censura*[2]. Adicionalmente, está presente la manipulación de la encuesta, por ejemplo, Camacho y Conover (2011) encuentran que las encuestas del Sistema de Identificación de Potenciales Beneficiarios de Programas Sociales (SISBEN) tenían una manipulación justo en el umbral de elegibilidad que determina el acceso a subsidios y ayudas estatales, una vez la población adquirió conocimiento sobre la metodología.

En relación al problema que abordamos, estos problemas error de medición pueden presentarse en el sub-reporte de ingresos o en la ausencia de datos de la muestra. Con el propósito de mejorar la robustez de los resultados, se implementan modelos de aprendizaje para predecir el ingreso del individuo, considerando características socio-económicas de hogar e individuo. Adicionalmente, se realiza una imputación de valores faltantes de la muestra, la cual se detalla en la siguiente sección. Finalmente, se presentan técnicas de remuestreo y validación cruzada para la elección del modelo con menor error de predicción.

Con el objetivo de que los resultados sean replicables, el trabajo cuenta con un repositorio en GitHub. El repositorio cuenta con cuatro carpetas. En la primera se encuentra el documento final. La segunda carpeta contiene el código de R, dividido en cinco scripts; en caso de correr el código, se recomienda usar el *00_main_script*, el cual es la base del código. La tercera carpeta cuenta con los resultados y las bases de datos con las cuales se trabajó. Finalmente, la cuarta carpeta contiene los gráficos y tablas usadas en el documento final.

[1] Al observar un subconjunto de la población, los individuos con ingresos muy bajos y muy altos podrían estar ausentes.
[2] Reportes de ingresos muy altos pueden ser modificados a un valor más bajo para proteger la identidad de los encuestados.

2. Datos

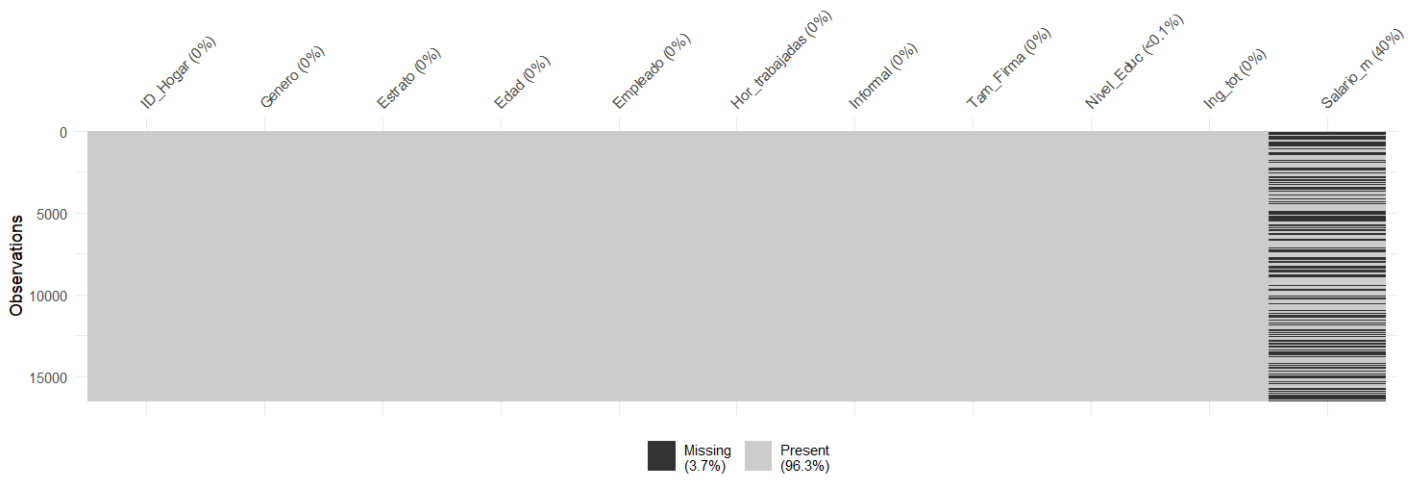
La Gran Encuesta Integrada de Hogares (GEIH) es una encuesta de periodicidad mensual recolectada por el DANE. Esta encuesta hace un seguimiento al mercado laboral en Colombia y, además, consulta información de características generales del hogar, fuentes de ingresos, entre otros. La GEIH tiene una cobertura a nivel nacional, presenta sus resultados a nivel de área urbana (cabeceras), área rural (resto) y 13 ciudades principales con sus áreas metropolitanas. Para este estudio vamos a utilizar la información de 2018 para Bogotá, extraída a través de *web scrapping*[3] del repositorio de GitHub del profesor Ignacio Sarmiento.

Para la adquisición de los datos, se utiliza *web scrapping* con el software R Studio, los datos se encuentran divididos en diez partes, cada una con sus respectivos enlaces. Adicionalmente, tienen una visualización de tablas que son cargadas desde vínculos externos. Se identifica que estos vínculos externos tienen una sintaxis similar, solo cambiando el número de página para cada parte de la muestra. Por lo tanto, se implementa una iteración que toma en cuenta el número de vínculos, en este caso diez, para que descargue cada una de las tablas y las una hacia abajo, creando una base final consolidada de 32.177 observaciones y 178 variables.

Considerando que el interés del estudio es predecir ingresos, se restringe a los ingresos laborales de las personas mayores de edad de Bogotá. La GEIH permite hacer un filtro por edad, las personas con una edad igual o mayor a 18 años representa la eliminación de 7.609 observaciones. Adicionalmente, entre las variables de mercado laboral de la encuesta, es posible identificar si una persona se encuentra ocupada. Al excluir las personas del hogar que no están ocupadas, se eliminan 8.933 observaciones. De esta manera, el número de observaciones utilizado en el análisis es de 16.542 (aproximadamente el 51 % de la muestra inicial).

Las variables que se utilizan en el estudio hacen referencia a características de los empleados que son relevantes a la hora de hacer el análisis de ingresos. Entre ellas se encuentran: variable indicadora del hogar, individuo y lugar; variables propias del individuo como el género y edad; variables socio-económicas como el estrato, la educación máxima; su oficio, si es o no contribuyente del sistema de salud y si es jefe de hogar; y variables del mercado laboral como el número de horas trabajadas, el tamaño de la firma, los años de experiencia en la empresa o industria, si es autoempleado, si esta es informal o no, además de los ingresos salariales de los individuos.

Figura 1: Proporción de *missing values* de las variables más representativas



De las 21 variables de referencia se usan para este estudio, 5 de ellas tiene *missing values*. La figura 1 muestra las variables más representativas (entre la variable dependiente y controles) que fueron usadas a lo largo del presente trabajo. Tal y como se observa en la última columna del gráfico, la variable de interés en el modelo que representa el salario mensual, tiene un 40 % de observaciones faltantes.

Por otra parte, se agregaron dos variables nuevas a la base de datos. Como se menciona en la sección 3 el salario también guarda una relación cuadrática con la edad. De esta manera, se incorporó la edad al cuadrado. Además, se transformó el salario mensual mediante una función logarítmica, con el fin de acotar el rango y facilitar la

[3]Proceso de extracción de contenido y datos de páginas web.

interpretación.

2.1. Imputación de datos

Para la imputación de valores faltantes con respecto al ingreso, se realiza una imputación por el promedio del salario por oficio. La variable oficio tiene la ventaja de tener una desagregación específica, con 99 valores. Al realizar la imputación por el ingreso promedio por oficio, se evidencia que únicamente hay valores faltantes para dos observaciones de la muestra, con el oficio número 60 (administrador de explotación agropecuaria, cooperativas agropecuarias, mayordomo y/o capataz de finca). Estas dos observaciones se eliminan de la muestra.

2.2. Estadística descriptiva

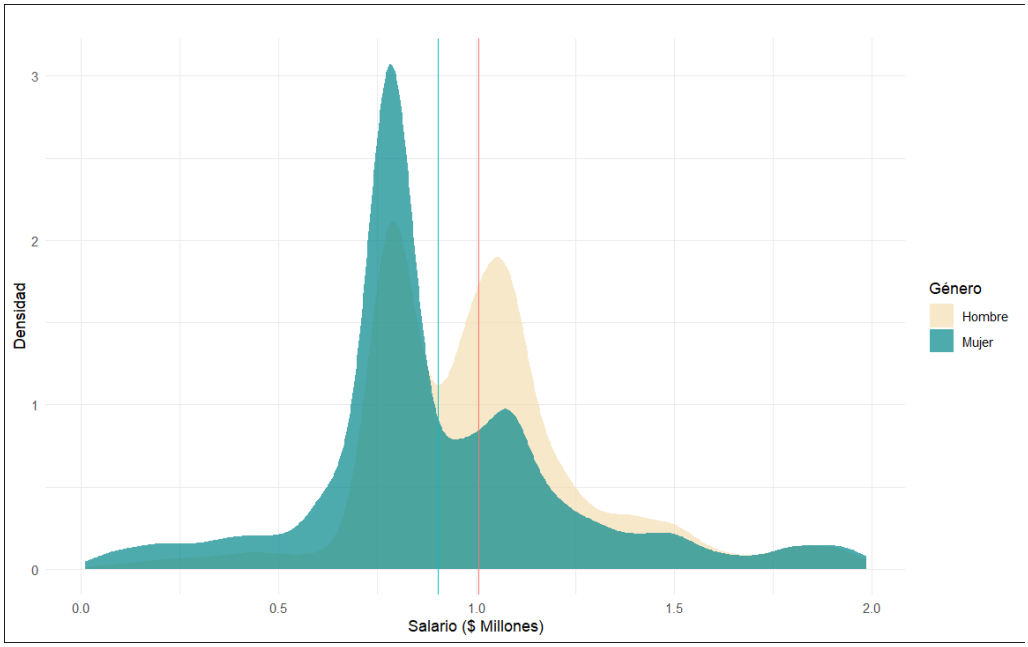
Tal y como se había mencionado en la introducción, los resultados de este documento se concentran en la población de Bogotá (2018) que labora y es mayor de edad. El resumen de las variables más importantes que fueron usadas en el estudio se presentan en la tabla 1.

Tabla 1: Estadísticas descriptivas de las variables cuantitativas más relevantes

Variable	N	Media	Des. Est.	Min	Max
Edad	16,539	39.4	13.5	18	94
Horas trabajadas	16,539	47.4	15.7	1	130
Salario mensual	16,539	1,583,275.0	1,834,617.0	10,000.0	34,000,000.0

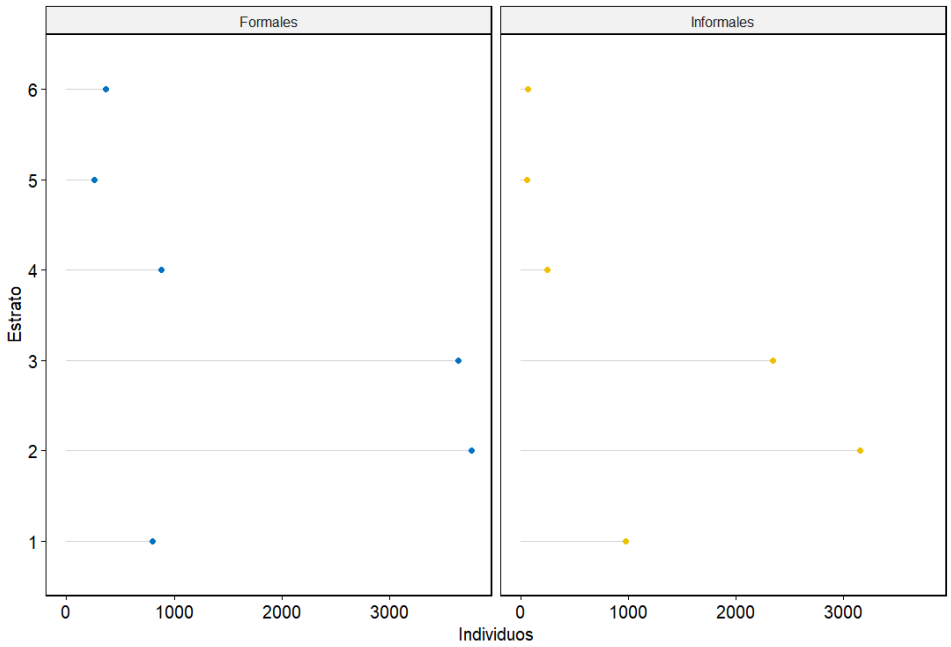
La edad es similar en los dos géneros. Mientras que una mujer promedio de la muestra tendrá 39 años, un hombre estará cerca a los 40 años. Esta equivalencia no se observa con todas las variables. Por ejemplo, al comparar las horas trabajadas (divididas por género) se observa que una mujer, en promedio, trabaja 5 horas menos que un hombre. Esto ofrece una evidencia sugestiva de que el salario de las mujeres podría ser menor.

Figura 2: Distribución de Salarios por Género



Con esta premisa, en la figura 2 se exhibe la distribución del salario mensual acotado hasta $2 millones de pesos -lo que representa cerca de 2.5 salarios mínimos en 2018-, abarcando a cerca del 80 % de la muestra. A primera vista, la media condicional por género presenta una diferencia favorable para el hombre; además, la densidad de las mujeres se concentra en salarios más bajos. Este problema se profundizará en la sección 4.

Figura 3: Frecuencia de individuos por estrato, segmentado por el grado de formalidad laboral



Por otra parte, los individuos de la muestra se concentran en los estratos 2 y 3, juntando 78 % de los hogares encuestados. La gráfica 3 muestra la frecuencia de la población de estudio dividido por estrato y la vinculación laboral. En este caso, en la capital el 41,3 % de los trabajadores mayores de 18 años son informales. La relación entre trabajadores formales e informales es similar hasta el estrato 3, sin embargo, luego del estrato 4 la proporción de trabajadores informales desaparece.

3. Teoría edad-salario

Para la modelación del ingreso de la población ocupada en Colombia, se hace un primer acercamiento considerando el perfil de ingreso - edad, en el cual se establece una relación cuadrática entre estas dos variables. Intuitivamente, refleja el comportamiento del ingreso cuando un individuo empieza a participar en el mercado laboral, y en la medida que aumenta su edad, se espera que al tener un mayor nivel de experiencia o de escolaridad, se vea reflejado en mayor salario. Posteriormente, alcanza su punto máximo de salario, el cual empieza a estabilizarse y decrecer, dada la pérdida de productividad del trabajador. Esta relación también ha sido estudiada con la relación entre el ingreso en función de la escolaridad, el nivel de experiencia y el nivel de experiencia cuadrático, también conocida como *la ecuación de Mincer*, Lemieux (2006) y Mincer (1974). Desde otro punto de vista y considerando la relación directa entre horas de trabajo e ingreso laboral, algunos modelos consideran que un individuo toma decisiones de optimización entre ocio y consumo de manera distinta a lo largo de su ciclo de vida. Esta literatura sugiere que al inicio y al final de la vida laboral el costo de oportunidad del ocio es bajo, dado que se perciben menores sueldos, mientras que en los años de edad productiva el costo de oportunidad del ocio es alto, por lo que el individuo decide trabajar más durante esos años (Borjas 2015).

Con el propósito de brindar evidencia sobre el perfil edad-ingreso para la muestra de interés, se estima la siguiente regresión por mínimos cuadrados ordinarios (MCO):

$$Ln(Salario_i) = \alpha + \beta_1 Edad_i + \beta_2 Edad_i^2 + X_i + u_i \tag{1}$$

En donde X_i agrupa variables de control tales como el género, una variable dicótoma de si el individuo es informal, el máximo nivel de educación, el oficio, las horas trabajadas, el estrato de su vivienda y el tamaño de la firma donde labora. Los resultados se muestran en la tabla 2, la primera columna refleja un modelo simple que presenta la relación entre la edad y el logaritmo del salario. La segunda columna de esta tabla muestra el modelo

con controles [4] en vista de que podrían existir variables omitidas que estén correlacionadas con el término de error, como lo es el caso de las horas trabajadas y su determinación por el ocio asociado a la edad.

Tabla 2: Estimación Salario - Edad

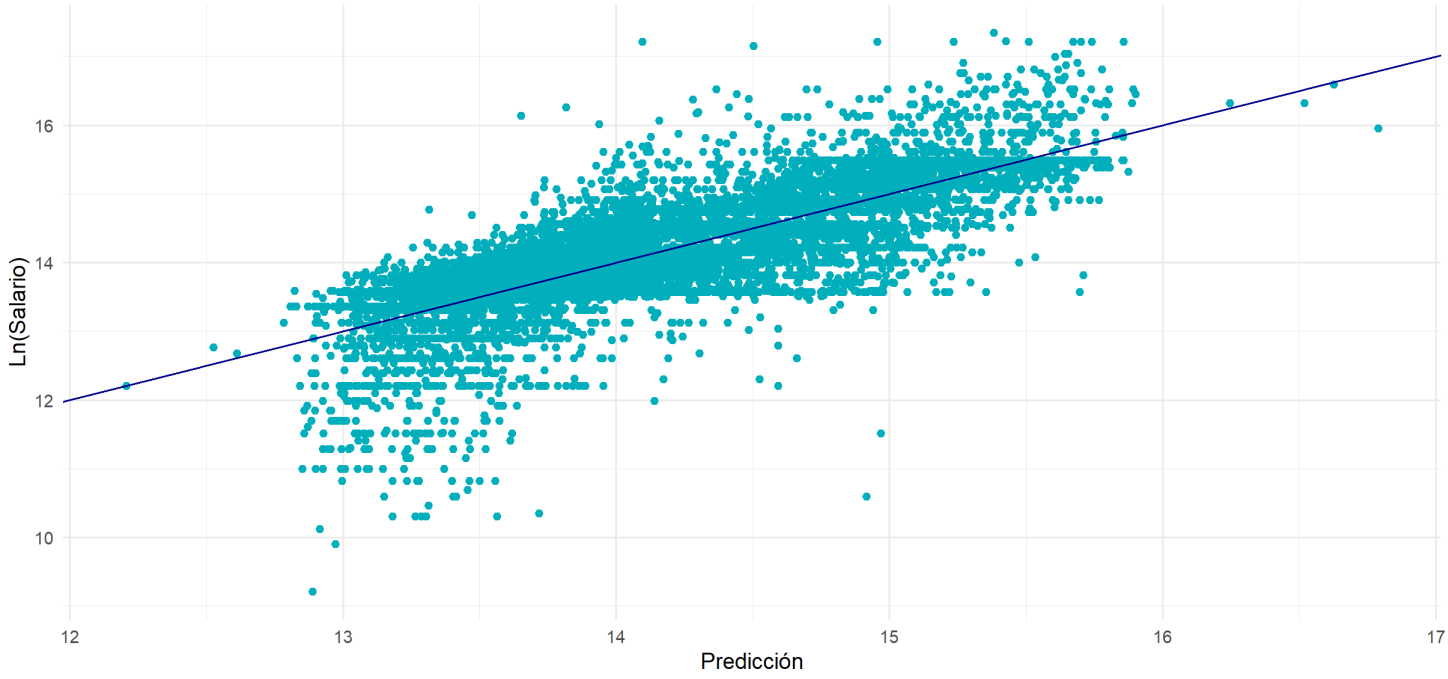
	Variable dependiente:	
	Ln(salario)	
	(1) sin Controles	(2) con Controles
Edad	0.046	0.024
	(0.002)	(0.001)
Edad^2	-0.0005	-0.0002
	(0.00003)	(0.00002)
Mujer		-0.073
		(0.008)
Informal		-0.084
		(0.010)
Horas Trabajadas		0.006
		(0.0002)
Constante	13.010	14.280
	(0.046)	(0.101)
Observaciones	16,540	16,539
R^2	0.029	0.644
R Ajustado^2	0.029	0.642
RMSE	0.683	0.414
Residual de Errores Std.	0.684 (df = 16537)	0.415 (df = 16441)
Estadístico-F	250.210 (df = 2; 16537)	306.540 (df = 97; 16441)

En primer lugar se observa que en ambos modelos el coeficiente de $edad$ es positivo, mientras que $edad^2$ es negativo, consistente con la teoría descrita al principio de esta sección. Sin embargo, la magnitud difiere en gran medida entre modelos, en particular, disminuyen en el modelo con controles. Adicionalmente, el signo de los controles también es congruente, pues el salario de una persona, en promedio, disminuye si es informal y es mujer, este último aspecto se discute con mayor profundidad en la sección 4. Finalmente, el salario medio es menor entre menor sea la educación del individuo y si vive en un menor estrato socio-económico.

El otro hecho que resalta en la tabla es que todas las variables son estadísticamente significativas a un nivel de confianza del 99 %, incluso la constante. Así mismo la significancia conjunta (estadístico F) establece que los coeficientes de las variables del modelo, al tiempo, son estadísticamente distintos de cero. Además, el segundo modelo explica mejor la variabilidad sobre el salario; en este mismo sentido, con su mayor complejidad, el segundo modelo tiene un mejor comportamiento (dentro de muestra) debido a que su RMSE es menor al del primer modelo, como se puede observar en la gráfica 4.

[4] Se ocultaron los coeficientes de las variables categóricas como el tamaño de la firma, el estrato, el máximo nivel de educación alcanzado y el oficio, debido a que estas variables cuentan con un gran número de categorías y se perdería claridad sobre los resultados en una tabla excesivamente larga.

Figura 4: Dispersión entre los valores observados del salario en su transformación logarítmica vs los valores predichos por el modelo con controles



Ahora bien, la edad tiene una relación cuadrática con el salario y el efecto parcial de la edad esta descrito por:

$$\frac{\partial y}{\partial Edad} = \beta_1 + 2\beta_2 Edad = 0 \tag{2}$$

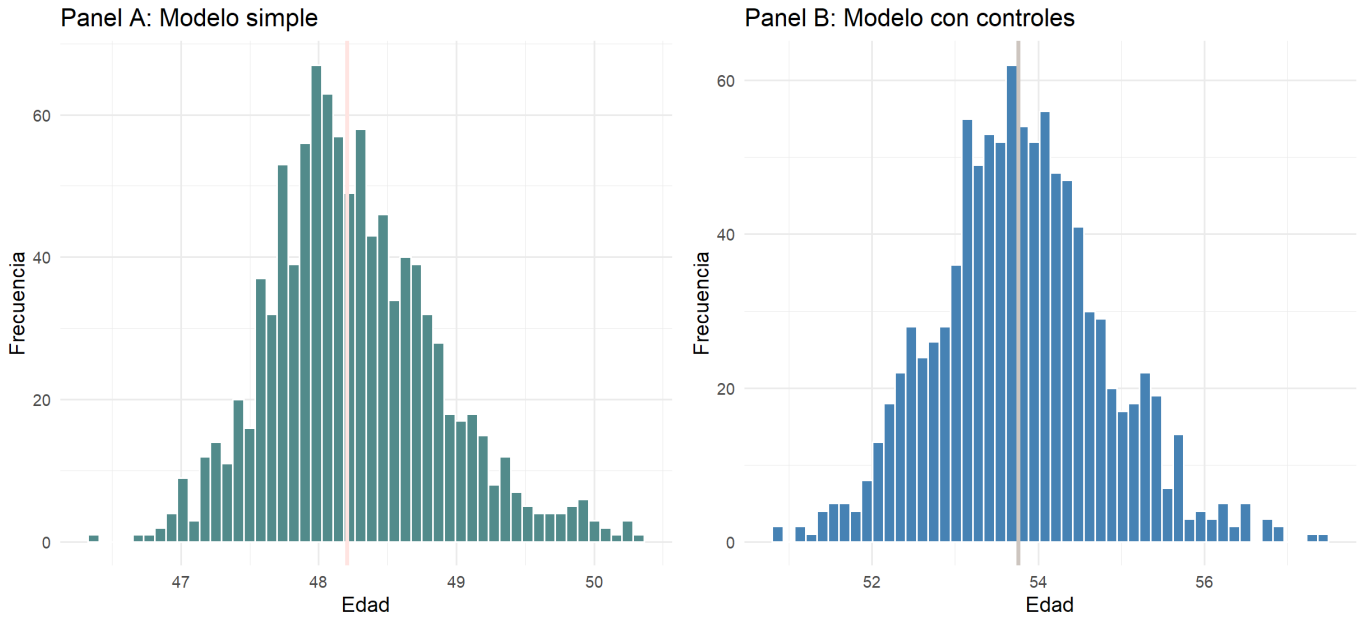
$$2\beta_2 Edad = -\beta_1 \tag{3}$$

$$Edad = -\frac{\beta_1}{2\beta_2} \tag{4}$$

De manera que, si reemplazamos los coeficientes de tabla 2 en la ecuación 4, vemos que el pico de edad promedio estaría en 48 años para el primer modelo (simple), mientras que para el segundo modelo (con controles) llegaría a los 54 años.

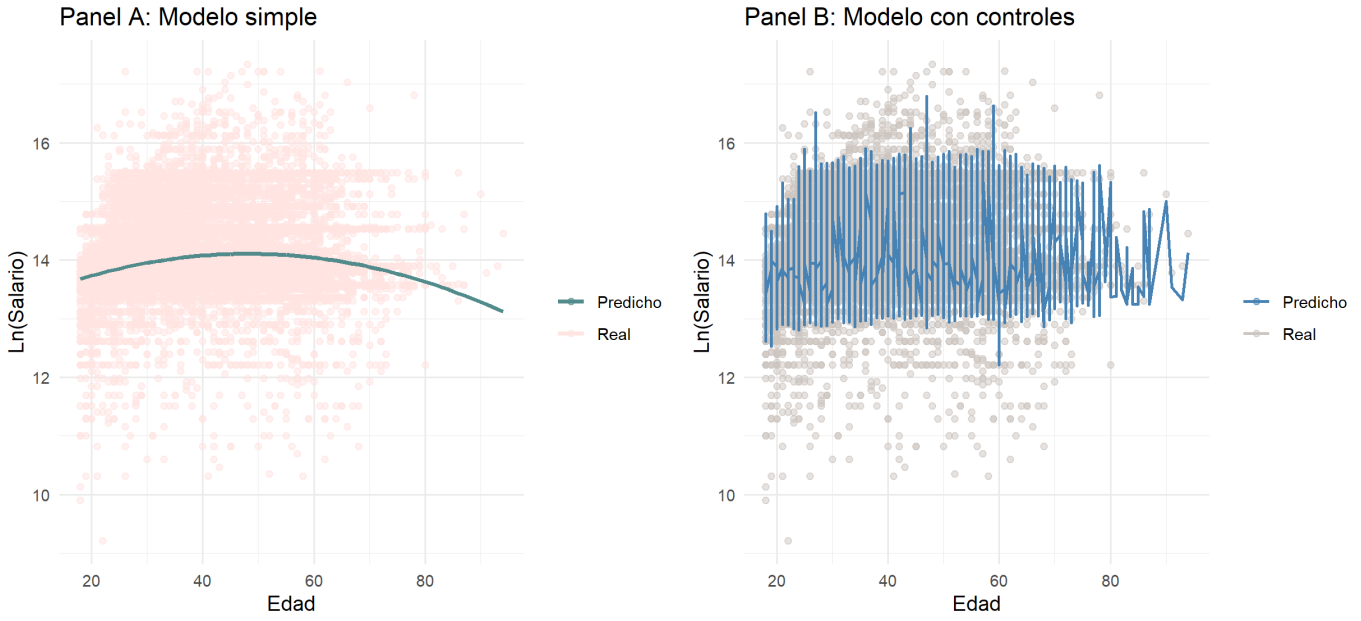
Sin embargo, estos dos modelos podrían ser susceptibles de heterocedasticidad, ya que la varianza de los errores podría variar de acuerdo con la edad. Por lo tanto, el calculo de los errores estándar podría ser erróneo. Con el fin de resolver este problema, se utiliza el método de *Bootstrap* (con 1.000) repeticiones para calcular los errores de una forma más acertada y así obtener el intervalo de confianza. La distribución de los resultados se presenta en la figura 5. Así, el modelo simple predice que dentro de los 47.17 y 49.70 años se encuentra el pico de ingresos real con 95 % de certeza, que es explicado por la edad. El modelo con controles propone un rango entre los 51.91 y los 55,92 años. Es decir, al controlar por otras características socieconómicas del individuo la edad en la cual se recibirán mayores ingresos laborales aumenta ligeramente.

Figura 5: Histograma del método de remuestreo (*Bootstrap*) del pico máximo de ingresos explicado por la edad



Finalmente, el gráfico 6 muestra la dispersión entre la edad de un individuo de la muestra y su salario (transformado mediante una función logarítmica). En el panel A se observa el modelo simple, el cual plantea una relación cuadrática con un menor ingreso -asimétrico- tanto al inicio como al final de la vida laboral de una persona promedio. El panel B muestra la misma relación, pero con un modelo más complejo que incluye controles. En este caso, la relación entre la edad y el ingreso tiende a ser menos clara, pues el modelo reduce el sesgo.

Figura 6: Relación entre la edad y el logaritmo del salario de los valores observados y predichos por un modelo simple y uno complejo (con controles)



En conclusión, si bien el modelo complejo permitiría una mejor predicción del salario, el modelo simple -descrito en la ecuación 1 (sin el vector **X**)- permite observar una relación no lineal entre la edad y el salario, tal como se comentó en el inicio de la sección. Factores como una menor productividad en la vejez, un cambio en las preferencias de ocio al final de la vida laboral producen que el ingreso laboral de una persona disminuya en sus últimos años laborales.

4. Brecha salarial por género

En economía laboral el análisis de la brecha salarial de género ha tomado mayor relevancia. De hecho, en 2023 se condecoró a la economista Claudia Goldin por estudiar el mercado laboral de las mujeres y, en especial, por entender las razones socioeconómicas que explican este fenómeno. Entre sus hallazgos identifica que para Estados Unidos, esta brecha se ha visto explicada históricamente por una diferenciación entre ocupación y educación. No obstante, encuentra que actualmente existen diferencias entre hombres y mujeres con la misma ocupación, la cual se relaciona con el nacimiento del primer hijo. Por otro lado, el avance en el cierre de la brecha ha tenido un ritmo lento, por ejemplo, para el 2022 una mujer promedio recibía el 84 % del salario de un hombre promedio en Estados Unidos, una cifra muy similar al año 2002, cuando se calculó en 80 % (Aragão 2023).

En esta sección se incorpora al análisis la diferenciación de ingresos entre hombres y mujeres. Se incluye porque, a pesar del aumento en la participación femenina en el mercado laboral, se siguen evidenciando brechas salariales.

4.1. Brecha salarial incondicional

Como primera aproximación, llevaremos a cabo una estimación que denominaremos: regresión *naive* de género. Este enfoque implica prescindir de la inclusión de controles adicionales y surge de la necesidad de explorar de manera preliminar la relación aparente entre el género y el logaritmo del salario, sin ajustes o correcciones por otras variables potencialmente relevantes. La adopción de esta perspectiva nos permite obtener una visión directa y simplificada de las posibles disparidades salariales entre géneros antes de incorporar factores adicionales que podrían influir en esta relación.

$$Ln(Salario) = \beta_0 + \beta_1 Mujer + e_i \tag{5}$$

La ecuación 5, tiene como variable dependiente el logaritmo del salario de las muestra de individuos en Bogotá y como variable explicativa, una dicótoma que identifica como 1 si el individuo i es mujer.

Si bien la regresión *naive* de genero actúa como un primer vistazo, proporcionando una estimación inicial de cómo el género podría estar asociado con las variaciones en el logaritmo del salario, es crucial reconocer que este enfoque conlleva un sesgo inherente al no considerar posibles factores de confusión o variables de control que podrían modular la relación.

4.2. Pago igual por el mismo trabajo

Una vez establecida la ecuación 5 y siguiendo la hipótesis inicial en la que el mercado laboral colombiano tiene un comportamiento similar al mercado norteamericano y en otros países, en el cual las mujeres tienen un ingreso promedio menor que hombres con características similares; es preciso realizar una regresión en donde se tenga presente características que tienen los trabajadores, puesto que, de no existir diferencias significativas, esto implicaría la no existencia de una brecha del salario por género.

A continuación, se planteará el modelo a estimar y la motivación de realizar el análisis con las características que se exponen en la ecuación 6:

$$Ln(Salario_i) = \beta_0 + \beta_1 Mujer_i + \beta_2 Edad_i + \beta_3 Educación_i + \beta_4 HrsTrabajo_i + \beta_5 Oficio_i + \beta_6 Jefehogar_i + e_i \tag{6}$$

La ecuación propuesta, encapsula un modelo que busca observar la existencia de una brecha salarial por género. Cada término de la ecuación representa una dimensión clave que podría influir en los salarios de las personas. La variable $Edad_i$ representa la edad del individuo. Se espera que la edad tenga una relación positiva con el salario, ya que el acumulo de experiencia a lo largo del tiempo suele estar asociado con salarios más altos.

La variable $Educación_i$ refleja el máximo nivel educativo alcanzado por la persona. Un mayor nivel educativo generalmente se asocia con salarios más altos debido a la adquisición de habilidades y calificaciones.

La variable $HrsTrabajo_i$ representa las horas de trabajo por semana. Se podría esperar que las horas de trabajo estén positivamente relacionadas con el salario, ya que un mayor tiempo dedicado al trabajo suele traducirse en mayores ingresos.

La variable $Oficio_i$ es una variable categórica de 1 a 99, que representa la diversidad de ocupaciones presentes en la muestra de datos. Diferentes ocupaciones pueden tener estructuras salariales distintas, y esta variable permite capturar esas variaciones.

La variable $Jefehogar_i$ es una variable dicótoma que indica si el individuo es el jefe de hogar (1) o no (0). Ser el jefe de hogar puede tener implicaciones en la distribución de ingresos y responsabilidades.

4.2.1. Estimación

Para estimar el modelo propuesto en la ecuación 6, se emplea el teorema de Frisch-Waugh-Lovell (FWL) como punto de partida. Este teorema es una técnica que permite analizar el impacto de una variable particular en un modelo de regresión múltiple, controlando o manteniendo constante el efecto de otras variables. La estimación se hace en dos etapas, y dado el teorema, el resultado del coeficiente debería ser el mismo que con la estimación tradicional de regresión múltiple.

En este sentido, se realiza la estimación en dos etapas:

- Etapa 1. Estimar la relación entre la variable explicada y los controles, y la relación entre la variable explicativa de interés (mujer) y los controles:

$$\ln(Salario_i) = \alpha_0 + \alpha_1 \text{Edad}_i + \alpha_2 \text{Educación}_i + \alpha_3 \text{HrsTrabajo}_i + \alpha_4 \text{Oficio}_i + \alpha_5 \text{Jefehogar}_i + e_i$$

$$\text{Mujer}_i = \pi_0 + \pi_1 \text{Edad}_i + \pi_2 \text{Educación}_i + \pi_3 \text{HrsTrabajo}_i + \pi_4 \text{Oficio}_i + \pi_5 \text{Jefehogar}_i + u_i$$

  Donde e_i y u_i son el término de error de las regresiones auxiliares.

- Etapa 2. Utilizar los residuales las regresiones auxiliares para estimar la relación entre $ln(Salario)_i$ y Mujer_i.

  Calcular los residuales:

$$\hat{e}_i = \ln(Salario_i) - \hat{\alpha}_0 - \hat{\alpha_1} \text{Edad}_i - \hat{\alpha}_2 \text{Educación}_i - \hat{\alpha}_3 \text{HrsTrabajo}_i - \hat{\alpha}_4 \text{Oficio}_i - \hat{\alpha}_5 \text{JefeHogar}_i$$

$$\hat{u}_i = \text{Mujer}_i - \hat{\pi}_0 - \hat{\pi}_1 \text{Edad}_i - \hat{\pi}_2 \text{Educación}_i - \hat{\pi}_3 \text{HrsTrabajo}_i - \hat{\pi}_4 \text{Oficio}_i - \hat{\pi}_5 \text{JefeHogar}_i$$

Reestimar a partir de los residuales:

$$\hat{e}_i = \phi_1 \hat{u}_i$$

Donde ϕ es el nuevo coeficiente estimado, y dado el teorema FWL, $\hat{\phi} = \hat{\beta}_1$ de la ecuación 6.

La tabla 3 exhibe los resultados de la estimación de la ecuación 6 con el teorema FWL (primera columna) y de la ecuación 5 con MCO (segunda columna). En el modelo sin controles, el coeficiente asociado a la variable de interés, es de -0.134 con un error estándar de 0.011. Este modelo sugiere una disminución promedio del 13.4 % en el salario de las mujeres, en comparación con los hombres. Además, el modelo tiene un error estándar residual de 0.691, que mide la dispersión de los datos alrededor de la línea de regresión.

Por otra parte, en el modelo con controles se observa un coeficiente es de -0.53, con un error estándar de 0.008. Por tanto, los resultados sugieren que existe una disminución promedio del 5.3 % en el salario de las mujeres, en comparación con los hombres. Además, el error estándar del modelo es menor, con un valor de 0.442, indicando una menor dispersión de los datos alrededor de la línea de regresión. Es importante resaltar que el coeficiente de los dos modelos estimados son estadísticamente significativos con un nivel de confianza de 99 %.

Tabla 3: Estimación de la brecha salarial por género

	Variable dependiente:	
	Ln(Salario)	
	Con controles	Sin controles
Mujer	−0.053	−0.134
	(0.008)	(0.011)
Observaciones	16,539	16,539
R^2	0.004	0.009
R^2 ajustado	0.004	0.009
Error estándar residual	0.439	0.691
Estadístico F	40.163	155.998

Nota: errores estándar en paréntesis.

Con el propósito de dar evidencia a la validación de los resultados, se realiza la estimación del modelo con controles por bootstrap con 1.000 repeticiones. La tabla 4 muestra los resultados y se observa que no hay diferencia entre el coeficiente estimado por FWL con la muestra original. Adicionalmente, la estimación por bootstrap nos genera un menor error estándar, lo cual hace que el coeficiente estimado sea más consistente.

Tabla 4: Estimación por bootstrap de la brecha salarial por género con controles

Coeficiente	Sesgo	Error Estándar
Edad	0.048	0.049

En la comparación entre ambos modelos se destaca el que tiene controles, puesto que proporciona una estimación más precisa del efecto de ser mujer en el salario. Además, tiene un menor error estándar, lo cual sugiere un mejor ajuste del modelo a los datos en comparación con el modelo de regresión *naive* de género.

4.3. Perfil edad-ingreso por género

En esta sección se analiza el perfil edad-ingreso desagregado por género. Los resultados responden a la pregunta: ¿existe diferenciación en el comportamiento de los salarios para cada nivel de edad entre hombres y mujeres? La regresión a estimar está dada por:

$$Ln(Salario)_{i,g} = \alpha_{i,g} + \beta_1 Edad_{i,g} + \beta_2 Edad^2_{i,g} + u_{i,g} \tag{7}$$

En donde el subíndice *g* representa el género del individuo *i*, y asumimos que el término de error $u_{i,g}$ es independiente e idénticamente distribuido. Los resultados se muestran en la tabla 5; los coeficientes estimados son estadísticamente significativos con un nivel de confianza del 99 %, tanto para hombres como mujeres.

Los signos de los coeficientes se mantienen por género. Como se mencionó en la secciones anteriores, el primer coeficiente refleja que mayor edad se relaciona con un aumento en el salario, lo cual puede verse explicado por un incremento en experiencia. En este aspecto, las mujeres tienen una pendiente mayor que los hombres. El segundo coeficiente refleja la productividad del trabajador con rendimientos marginales decrecientes en relación a su edad, lo que implica una pendiente negativa, la pendiente caracterizada es menor para los hombres que para las mujeres.

Puesto que la diferencia entre los coeficientes es más relevante en el término cuadrático de la edad, y asumiendo que el salario representa la productividad del trabajador, los resultados indican que el mercado laboral percibe a las mujeres con un nivel más alto de pérdida de productividad al paso de los años, en comparación con los hombres.

Tabla 5: Estimación del perfil edad-ingreso por género

	Variable dependiente	
	Ln(Salario)	
	Hombres	Mujeres
Edad	0.048	0.049
	(0.003)	(0.004)
$Edad^2$	−0.0005	−0.001
	(0.00003)	(0.00004)
Observaciones	8,765	7,775
R^2	0.053	0.021
R^2 ajustado	0.053	0.021
Error estándar residual	0.627	0.731
Estadístico F	245.022	84.090

Nota: errores estándar en paréntesis.

Con el propósito de estimar el efecto marginal de la edad frente a los salarios, se realiza una predicción con las observaciones de la muestra a partir de las estimaciones del modelo (figura 7). Adicionalmente, la figura 8 compara las predicción entre hombres y mujeres; esto nos permite caracterizar el comportamiento de la brecha salarial para cada nivel de edad. Por un lado, se observa que para todos los niveles de edad existe una diferenciación salarial entre hombres y mujeres. En segundo lugar, esta brecha empieza siendo pequeña al inicio de la vida laboral. Luego, cerca de los 30 años comienza a ampliarse, manteniéndose constante a partir de los 50 años, aproximadamente.

Figura 7: Predicción del perfil edad-ingreso por género

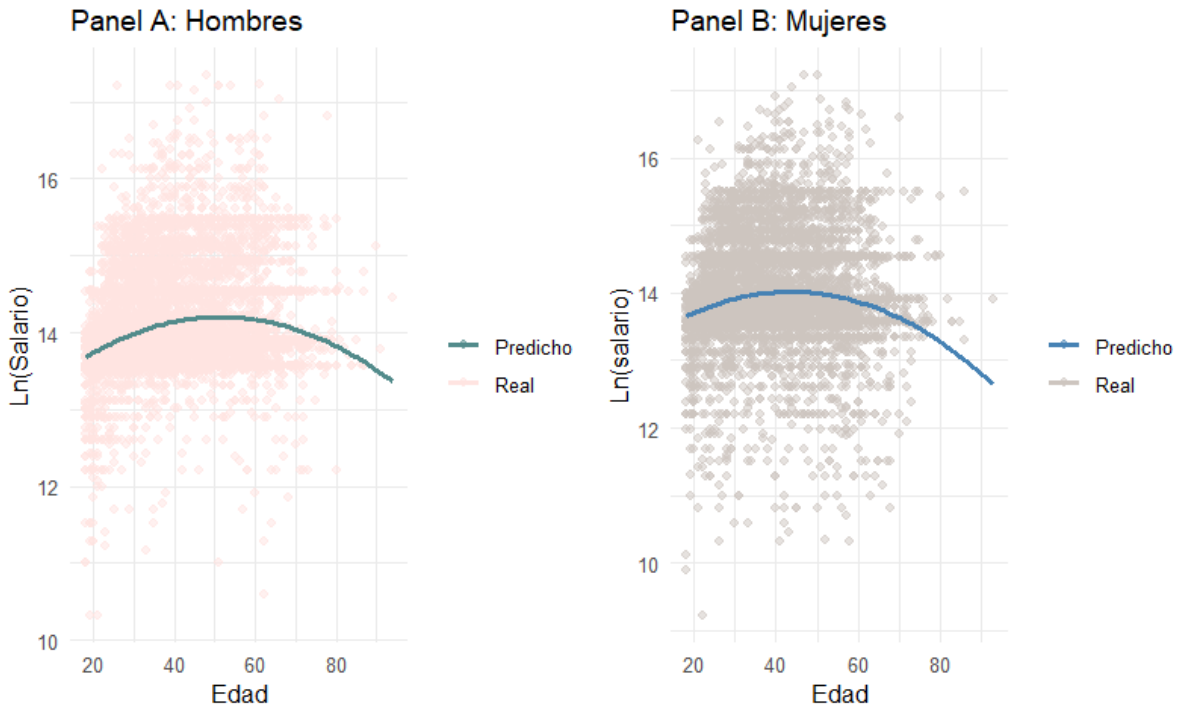
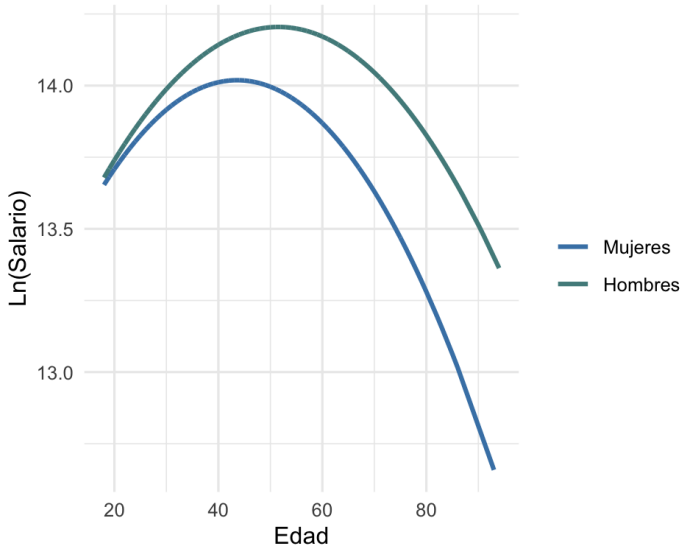
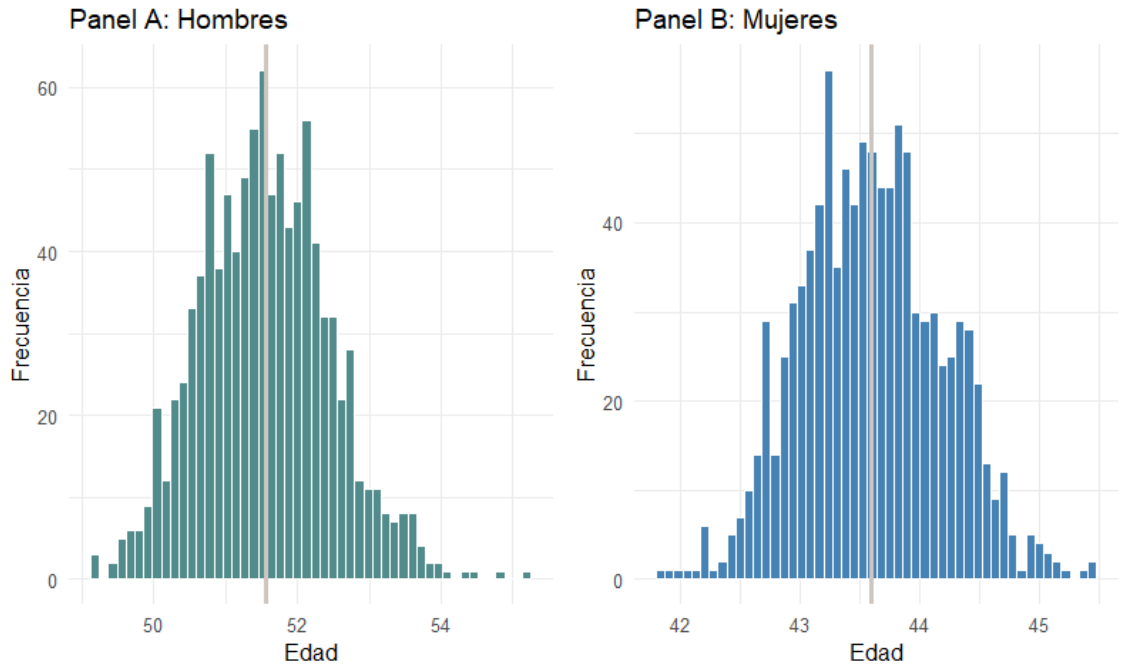


Figura 8: Brecha salarial de género por edad



A partir de los resultados de la predicción se encuentra que los hombres perciben su máximo ingreso a los 52 años de edad, mientras que para los mujeres es a los 44 años de edad. Para estimar los intervalos de confianza del 95 % de la edad de máximo ingreso, se realiza la estimación con 1.000 repeticiones de *bootstrap*. La distribución de los resultados se presenta en la figura 9, en donde, para los hombres, se estima que el rango de edad en que se percibe el ingreso más alto se encuentra entre 49.9 y 53.4 años, mientras que para las mujeres entre 42.5 y 44.7 años.

Figura 9: Distribución de la edad de máximo ingreso por género (estimación por *bootstrap*).



4.4. Discusión

La participación en el mercado laboral es una decisión individual y existe la posibilidad de que aquellos que eligen participar difieran sistemáticamente de quienes no lo hacen en términos de características no observadas que también podrían influir en el salario. En otras palabras, existe un sesgo de selección y eso puede llevar a que las estimaciones realizadas en este ejercicio se encuentren sesgadas, ya que solo se observan los datos de aquellos que participan en el mercado laboral.

Por lo anterior, nuestro modelo podría subestimar o sobreestimar la relación entre las variables explicativas y el salario si no se aborda el sesgo de selección. Para mitigar este riesgo, se podrían considerar métodos como el Modelo de Heckman o la descomposición de Oaxaca - Blinder. No obstante, los resultados encontrados en este ejercicio cobran fuerza al seguir la literatura que aborda el campo de investigación de brechas de salario e igualdad de género.

Una de las posibles explicaciones al comportamiento diferencial en los ingresos de hombres y mujeres por edad es la permanencia de roles de género en los hogares, los cuales limitan la participación laboral de las mujeres. En el caso de Colombia, Tovar y Urdinola (2019) encontraron que existe una brecha persistente entre la dedicación de tiempo a labores del hogar entre hombres y mujeres, en especial con el cuidado de los niños. Así mismo, se evidenció que cuando las mujeres tienen menores niveles de educación, estas diferencias son más relevantes.

En este sentido, es plausible considerar que estas actividades no remuneradas pueden limitar las horas disponibles al trabajo remunerado, explicando así que la diferencia salarial por género sea pequeña al inicio de la vida laboral, y aumente durante la edad fértil de las mujeres, reduciendo su participación en el mercado laboral al tener hijos y dedicar más tiempo a su cuidado y crianza. En el caso de nuestra estimación ese punto de inflexión se encuentra cerca de los 30 años.

Otra evidencia que soporta esta hipótesis es el estudio de Jiménez (2019), el cual encuentra que el perfil de las mujeres desempleadas en Costa Rica difiere de los hombres: mientras la población femenina desempleada se concentra en la edad más productiva y relacionada a la época de fertilidad, en los hombres se observa este fenómeno en una edad más temprana, lo que podría explicarse por posponer la entrada al mercado laboral por unos años adicionales de educación. De tal manera, las mujeres que se retiran del mercado laboral durante los años de fertilidad, lo hacen en sus años más productivos, por lo que retornar al mercado laboral puede generar penalidades en sus salarios vía productividad o experiencia.

Por otro lado, Fernández (2006) considera otros factores que aportan a la explicación de esta brecha salarial. Entre ellos se encuentran modelos de discriminación por gusto, en donde un empleador toma decisiones a partir de juicios de valor, e incluso se encuentra dispuesto a asumir mayores costos de producción o a una reducción en sus ingresos con el propósito de no contratar a una persona que haga parte de este grupo discriminatorio, en este caso, una mujer. Otro tipo de discriminación se da por asimetría en la información que percibe el empleador, considerada como discriminación estadística; en este escenario entra la percepción de la productividad del empleado. Si la percepción es una en la cual las mujeres puedan tener una menor productividad que los hombres, esto se vería reflejado en sus salarios o en su participación en el mercado laboral.

En relación a esta problemática, en Colombia se han implementado mecanismos institucionales en el sector público y privado para evitar problemas de discriminación hacia la mujer en el mercado laboral. Por ejemplo, la Ley 581 de 2000 conocida como la *"Ley de cuotas"* establece una participación mínima del 30 % en cargos públicos de poder para las mujeres. Así mismo, la Ley 1496 de 2011 establece sanciones a empresas del sector público o privado que no garanticen la igualdad salarial entre hombres y mujeres que desempeñen el mismo cargo. Por consiguiente, uno de los retos es la identificación de la efectividad de este tipo de políticas para tener certeza sobre si se mantiene la existencia de algún tipo de discriminación por sexo.

5. Predicción de ganancias

Teniendo en cuenta los modelos estimados previamente, en esta sección vamos a evaluar el poder predictivo de estas especificaciones y algunas adicionales con mayor complejidad y relaciones no lineales. Para ello, primero vamos a comparar a través de distintos enfoques el entrenamiento de todas las especificaciones y los resultados de predicción. Luego se discutirá acerca de los errores de predicción del modelo y, finalmente, validaremos a través de la metodología *Leave one out cross validation* (LOOCV) los dos modelos con menor error de predicción medio, y se hará una comparación con la primera validación hecha en todos los modelos.

5.1. Evaluación de modelos mediante métodos de remuestreo

En esta sección se discutirá el paso a paso de todos los enfoques utilizados para evaluar el poder predictivo de las siguientes especificaciones:

Modelos previamente estimados:

Recordemos primero todas las especificaciones que hemos hecho a lo largo del presente trabajo.

Modelo 1:

$$Ln(Salario_i) = \alpha + \beta_1 Edad_i + \beta_2 Edad_i^2 + u_i \tag{8}$$

Modelo 2:

$$\begin{aligned} Ln(Salario_i) =& \alpha + \beta_1 Edad_i + \beta_2 Edad_i^2 + \beta_3 Mujer_i + \beta_4 Informal_i + \beta_5 Oficio_i + \\ & \beta_6 Educación_i + \beta_7 HrsTrabajo_i + \beta_8 Estrato_i + \beta_9 Tam\_Firma_i + u_i \end{aligned} \tag{9}$$

Modelo 3:

$$Ln(Salario_i) = \alpha + \beta_1 Mujer_i + u_i \tag{10}$$

Modelo 4:

$$\begin{aligned} Ln(Salario_i) =& \alpha + \beta_1 Mujer_i + \beta_2 Edad_i + \beta_3 Educación_i + \\ & \beta_4 HrsTrabajo_i + \beta_5 Oficio_i + \beta_6 JefeHogar + u_i \end{aligned} \tag{11}$$

Modelos adicionales más complejos:

Así mismo, también agregamos 5 especificaciones adicionales que nos permitieran aumentar la complejidad del modelo con formas, no necesariamente lineales, en las que consideramos adicionar variables distintas a las que hemos visto a lo largo del presente trabajo.

Consideramos importante observar el efecto que tenía en el poder predictivo añadir variables como: $Experiencia$ —que mide los meses que lleva la persona trabajando para una empresa o industria— , $Autoempleado$ —una variable dicótoma que indica si una persona trabaja como independiente— , e interacciones entre los años de experiencia de las mujeres, las mujeres autoempleadas y las mujeres jefes de hogar.

En el mercado laboral colombiano la experiencia en una industria es muy valiosa en cuanto a la remuneración de los empleados, así mismo, los individuos autoempleados también ven afectado su ingreso, pues no cuentan con la estabilidad laboral que brinda un empleo. Por ejemplo, puede existir el caso en el que un individuo tenga una microempresa con ingresos fluctuantes. Por otra parte, como se discutió en la sección 4, el mercado laboral colombiano es bastante distinto para las mujeres, sobretodo aquellas que son cabeza de hogar o autoempleadas.

Modelo 5:

$$\begin{aligned} Ln(Salario_i) =& \alpha + \beta_1 Edad_i + \beta_2 Edad_i^2 + \beta_3 Mujer_i + \beta_4 Informal_i + \beta_5 Oficio_i + \\ & \beta_6 Educación_i + \beta_7 HrsTrabajo_i + \beta_8 Estrato_i + \beta_9 Tam\_Firma_i + \\ & \beta_{10} Experiencia_i + \beta_{11} HrsTrabajo_i^2 + \beta_{12}(Mujer_i \times JefeHogar_i) + \\ & \beta_{13}(Mujer_i \times Experiencia_i) + \beta_{14}(Mujer_i \times Autoempleado_i) + u_i \end{aligned} \tag{12}$$

Modelo 6:

$$\begin{aligned} Ln(Salario_i) =& \alpha + \beta_1 Edad_i + \beta_2 Edad_i^2 + \beta_3 Mujer_i + \beta_4 Oficio_i + \\ & \beta_5 Educación_i + \beta_6 HrsTrabajo_i + \beta_7 Estrato_i + \\ & \beta_8 Experiencia_i + \beta_9 HrsTrabajo_i^2 + \beta_{10}(Mujer_i \times JefeHogar_i) + \\ & \beta_{11}(Mujer_i \times Autoempleado_i) + u_i \end{aligned} \tag{13}$$

Modelo 7:

$$\begin{aligned} Ln(Salario_i) =& \alpha + \beta_1 Edad_i + \beta_2 Edad_i^2 + \beta_3 Mujer_i + \beta_4 Informal_i + \beta_5 Oficio_i + \\ & \beta_6 Educación_i + \beta_7 HrsTrabajo_i + \beta_8 Estrato_i + \beta_9 Tam\_Firma_i + \\ & \beta_{10} Experiencia_i + \beta_{11} Experiencia_i^2 + \beta_{12} Experiencia_i^3 + \\ & \beta_{13} HrsTrabajo_i^2 + \beta_{14} Ln(Edad_i) + \beta_{15} JefeHogar_i + u_i \end{aligned} \tag{14}$$

Modelo 8:

$$\begin{aligned} Ln(Salario_i) =& \alpha + \beta_1 Edad_i + \beta_2 Edad_i^2 + \beta_3 Mujer_i + \beta_4 Informal_i + \beta_5 Oficio_i + \\ & \beta_6 Educación_i + \beta_7 HrsTrabajo_i + \beta_8 Estrato_i + \beta_9 Tam\_Firma_i + \\ & \beta_{10} Experiencia_i^2 + \beta_{11} Ln(HrsTrabajo_i) + u_i \end{aligned} \tag{15}$$

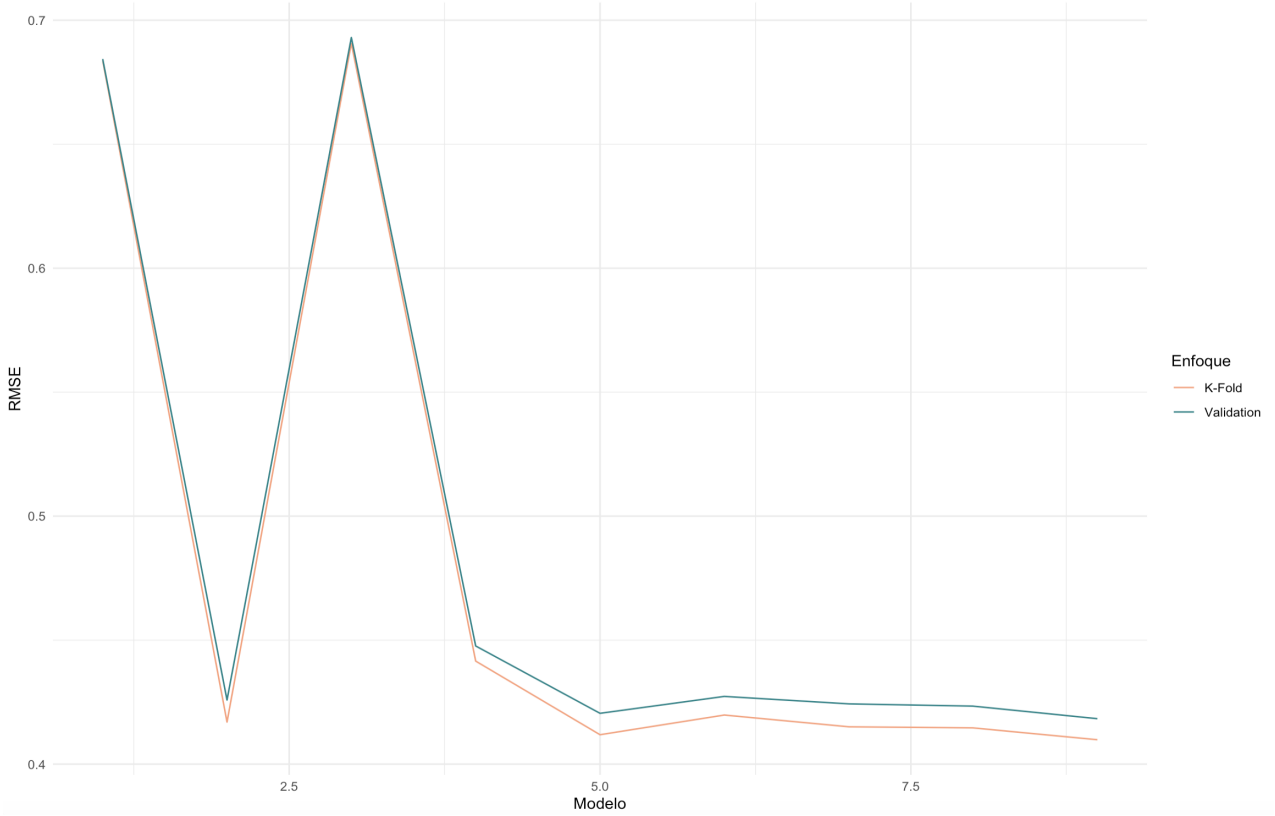
Modelo 9:

$$\begin{aligned} Ln(Salario_i) =& \alpha + \beta_1 Edad_i + \beta_2 Edad_i^2 + \beta_3 Oficio_i + \beta_4 Educación_i + \beta_5 Estrato_i + \\ & \beta_6 Tam\_Firma_i + \beta_7 Ln(HrsTrabajo_i) + \beta_8 Experiencia_i^3 + \\ & \beta_9 Ln(Edad_i) + \beta_{10}(Mujer_i \times JefeHogar_i) + \\ & \beta_{11} Autoempleado_i + u_i \end{aligned} \tag{16}$$

Una vez construimos todas las especificaciones que queríamos evaluar, decidimos aplicar 2 enfoques que nos permitieran comparar el poder predictivo de cada uno de los modelos. Primero, utilizamos el enfoque del conjunto de validación, para lo cual la muestra se dividió aleatoriamente en un 70 % destinada al entrenamiento y un 30 % destinada a testear cada uno los modelos. Luego, utilizamos el enfoque de *K-fold Cross validation*, para el cual dividimos el conjunto de datos en $k = 5$ subconjuntos de aproximadamente el mismo tamaño. De esta manera, cada modelo se entrenó y se evaluó 5 veces utilizando distintas combinaciones de subconjuntos de prueba. Este fue el proceso utilizado para capturar la raíz del Error Cuadrático Medio (RMSE) de cada modelo en ambos enfoques.

En la Figura 10 se muestra el comportamiento del RMSE en cada modelo para el enfoque de validación y el de *K-fold Cross Validation*. Como podemos observar, los resultados en ambos enfoques son bastante similares, sin embargo, el segundo es ligeramente menor que el primero. Además podemos apreciar que los modelos con mayor error de predicción son el 1 y el 3, que contaban con pocas variables predictoras, mientras que los otros modelos más complejos presentan un error de predicción mucho más bajo.

Figura 10: RMSE de todas las especificaciones con *K-fold* y Validación



Dado que los valores del RMSE en la gráfica son bastante parecidos para los modelos 4 en adelante, incluímos la Tabla 6 donde se muestran los RMSE calculados para cada modelo en cada enfoque.

Tabla 6: RMSE para todos los modelos

Modelo	RMSE_vsa	RMSE_kfold
1	0.684	0.684
2	0.426	0.417
3	0.693	0.691
4	0.448	0.442
5	0.421	0.412
6	0.427	0.420
7	0.424	0.415
8	0.423	0.415
9	0.418	0.410

Nota: vsa hace referencia al enfoque de conjunto de validación y kfold hace referencia al enfoque de *k-fold cross validation*.

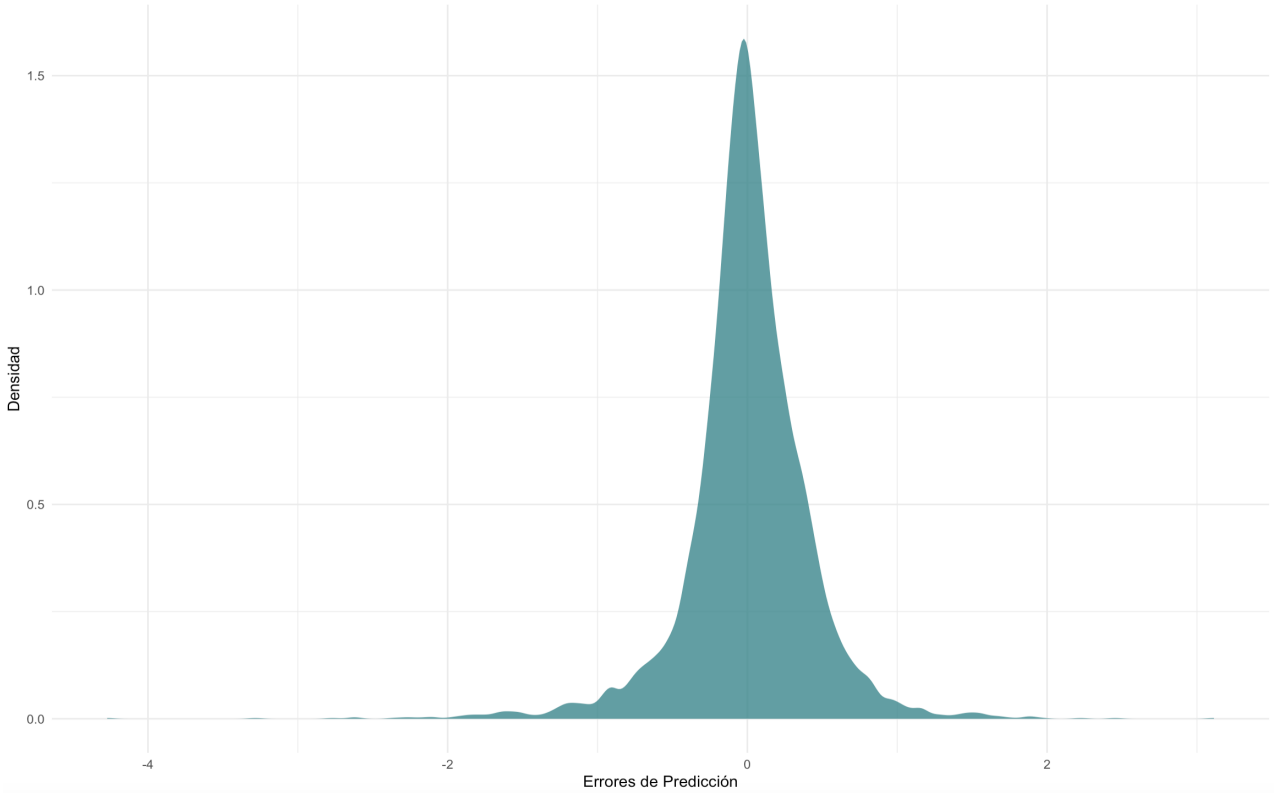
5.2. Discusión de *outliers*

Como podemos observar en la Tabla 6, el modelo con menor error de predicción es el modelo 9 -mejor balance de complejidad-, pues no regresa tantas variables como es el caso del modelo 7. El siguiente modelo con menor error de predicción es el modelo 5, el cual a pesar de ser más complejo que el modelo 9, contiene más variables explicativas relacionadas a las distintas condiciones que afectan el salario de las mujeres.

Ahora bien, si tomamos modelo con el menor error de predicción (modelo 9 - ecuación 16) se explorará aquellas observaciones en las cuales la predicción difiere sustancialmente de la observación. En la Figura 11 se muestra la distribución de estos errores de predicción. como se observa, hay una media centrada en cero y la curtosis no es muy amplia, sin embargo, podemos ver que existen algunos valores extremos, sobretodo en la cola negativa. Esto implica que existen varias personas reportando un ingreso mucho menor al que predecirían las características

propias del individuo.

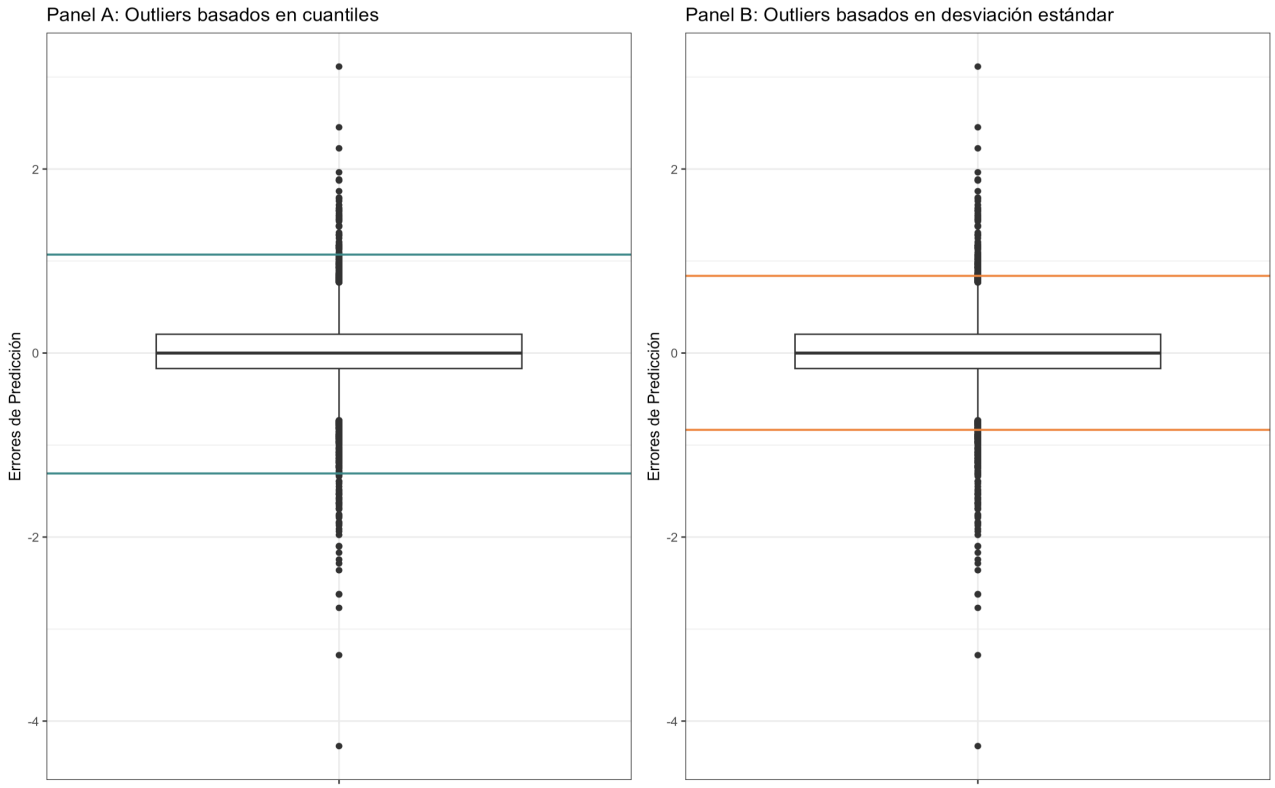
Figura 11: Distribución de los errores de predicción



Para identificar a estos *outliers* tenemos dos posibles definiciones: primero, aquellas observaciones que estén por debajo del percentil 1 % y por encima del percentil 99 %. Segundo, aquellas observaciones que se encuentren 2 desviaciones estándar más allá de la media. En la Figura 12 se muestran los diagramas de caja de la distribución de los errores de predicción; en el Panel A se encuentra la gráfica para la primera definición y en el Panel B para la segunda definición.

Es importante resaltar que, bajo ambas definiciones, el punto de corte por encima de la media es muy parecido, sin embargo, para el punto de corte por debajo de la media sí hay un cambio importante, en el Panel A este punto es $-1{,}308635$, mientras que en el Panel B este punto es $-0{,}8347712$, esto implica que hay más personas que reportan menor salario que aquellas que reportan un mayor salario que la predicción de nuestro modelo.

Figura 12: Diagramas de caja de la distribución de los errores de predicción



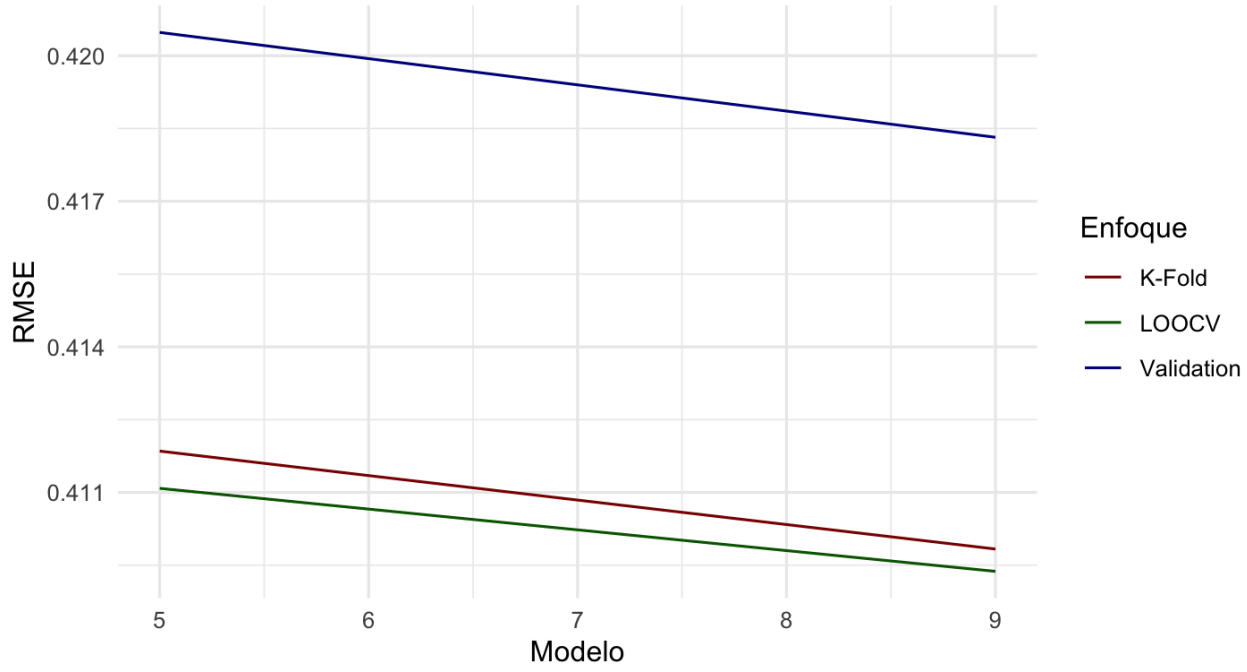
Ahora bien, queremos centrar nuestra discusión en estos sub-reportantes, pues son estos valores atípicos los cuales deberían ser investigados por la DIAN. Sin embargo, entendemos que nuestro modelo de predicción no es perfecto y tiene errores, por lo que no todos los valores atípicos serían potenciales evasores fiscales, sino simplemente hacen parte de los defectos del modelo. Teniendo en cuenta que la gran mayoría de los errores de predicción de nuestro modelo son cercanos a cero, tenemos suficiente información para afirmar que tiende a predecir muy bien el salario de las personas. En este sentido, es pertinente elegir la primera definición de *outliers*: aquellas observaciones que estén por debajo del percentil 1 %, son los valores atípicos que la DIAN debería investigar, pues no tiene sentido que exista una diferencia tan importante entre la predicción y el reporte del salario que hacen estas personas. De esta manera, existirían 50 personas que son potenciales evasores fiscales.

5.3. Comparación LOOCV

En esta sección vamos a comparar los modelos 9 y 5, que son los que menor error de predicción tienen según la metodología de validación y *k-fold cross validation* que vimos en la sección 5.1. Esta vez vamos a añadir una tercera comparación utilizando la metodología *Leave-one-out-cross-validation*(LOOCV) para ver las diferencias respecto a los dos enfoques anteriores. En este último enfoque el modelo se ajusta n veces, donde n es el número de observaciones en el conjunto de datos. Cada vez, se deja fuera del conjunto de entrenamiento una observación diferente y se utiliza como conjunto de prueba. De esta manera, se promedia el RMSE de todas las iteraciones para obtener una medida única.

En la Figura 13 se muestra el comportamiento del RMSE de los modelos 5 y 9 para los enfoques de validación, *K-fold Cross Validation* y *Leave-one-out-cross-validation*. Como podemos observar, los resultados en el enfoque de validación son mucho mayores que los otros dos enfoques, sin embargo, estos últimos se comportan muy parecido, siendo el LOOCV ligeramente menor.

Figura 13: RMSE de modelos 5 y 9 con *LOOCV*, *K-fold* y Validación



Dado que los valores del RMSE en la gráfica son bastante parecidos, incluimos la Tabla 7 donde se muestran los RMSE calculados para los dos modelos en cada enfoque. Acá podemos observar que las diferencias están en pequeños decimales.

Tabla 7: RMSE para modelos 5 y 9

Modelo	RMSE_vsa	RMSE_kfold	RMSE_LOOCV
5	0.421	0.412	0.411
9	0.418	0.410	0.409

Nota: vsa hace referencia al enfoque de conjunto de validación,
kfold hace referencia al enfoque de *k-fold cross validation*
LOOCV hace referencia al enfoque de *Leave-one-out-cross-validation*.

A partir de los resultados presentados en la Figura 13 podemos explorar el potencial enlace con la estadística de influencia, es decir, las métricas para evaluar la influencia de cada observación individual en la estimación de los parámetros del modelo. El enfoque LOOCV nos permite ver que cambio hay en el poder predictivo de los modelos al quitar una observación a la vez. Teniendo esto en cuenta, podemos ver que es un modelo bastante robusto, donde las observaciones con mayor influencia no afectan significativamente a las predicciones en general, dado que se observa que la diferencia entre el primer enfoque de validación y este último enfoque de LOOCV es de $0{,}009$ para el modelo con menor error de predicción. Por lo tanto, podemos afirmar que es un modelo con una muy buena capacidad predictiva y los errores de predicción no están directamente ligados a los outliers de la muestra.

Referencias

Aragão, Carolina (2023). *Gender pay gap in U.S. hasn't changed much in two decades*. URL: https://www.pewresearch.org/short-reads/2023/03/01/gender-pay-gap-facts/.

Ávila, Javier y Ángela Cruz (2015). *Colombia: Estimación de la evasión del impuesto de renta de personas jurídicas 2007-2012*. URL: https://www.dian.gov.co/dian/cifras/cuadernos%20de%20trabajo/colombia.%20estimaci%C3%B3n%20de%20la%20evasi%C3%B3n%20del%20impuesto%20de%20renta%20perosnas%20jur%C3%ADdicas%202007-2012..pdf (visitado 01-03-2024).

Borjas, George J (2015). *Labor Economics*. McGraw-Hill/Irwin.

Camacho, Adriana y Emily Conover (2011). «Manipulation of Social Program Eligibility». En: *American Economic Journal: Economic policy*.

DANE, Departamento Administrativo Nacional de Estadística (2023). *Empleo informal y seguridad social*. URL: https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-informal-y-seguridad-social#:~:text=Para%20el%20total%20nacional%2C%20en,anterior%20(57%2C6%25). (visitado 01-03-2024).

Fernández, Maria del Pilar (2006). «Determinantes del diferencial salarial por género en Colombia, 1997-2003». En: *Universidad de los Andes, Facultad de Economía, CEDE*.

Jiménez, Pamela (2019). «Gender Gaps in Costa Rica: Analysis of Time Use and Labor Income by Education». En: *Time Use and Transfers in the Americas*. Springer Cham.

Lemieux, Thomas (2006). *The "Mincer Equation" Thirty Years After Schooling, Experience, and Earnings*. URL: https://link.springer.com/chapter/10.1007/0-387-29175-x_11.

Mincer, Jacob (1974). *Schooling, Experience and Earnings*. URL: https://econpapers.repec.org/bookchap/nbrnberbk/minc74-1.htm.

Parra, Orlando y Ruth Patiño (2010). «Evasión de impuestos nacionales en Colombia: años 2001-2009». En: *Revista Facultad de Ciencias Económicas: Investigación y Reflexión*.

The Internal Revenue Service (2016). *The Tax Gap*. URL: https://www.irs.gov/newsroom/the-tax-gap (visitado 01-03-2024).

Tovar, Jorge y Piedad Urdinola (2019). «Home and Market Production Time Use Differentials in Colombia». En: *Time Use and Transfers in the Americas*. Springer Cham.