Big Data y Machine Learning: Problem Set 1

Ignacio Sarmiento

15 de septiembre de 2024

Integrantes:	Juliet Alejandra Molano Rizo	202226070
	Henry Nicolas Carvajal Cardenas	201718787
	Diego Fernando Cuesta Mora	202315672
	Jorge Ramirez	202116747

1. Introducción

La situación fiscal en Colombia se caracteriza por una preocupación significativa por la evasión de impuestos, que representa una fuga considerable de recursos públicos y evidencia deficiencias en la supervisión fiscal. La evasión tributaria no solo afecta la recaudación, limitando la capacidad del gobierno para financiar servicios públicos esenciales como educación y salud, sino que también refleja una actitud social que a menudo prioriza el beneficio personal sobre la contribución al bienestar colectivo (Delgadillo & Tavera, 2022).

Según la DIAN, los principales problemas para la recaudación fiscal incluyen la subdeclaración de ingresos, la inestabilidad y complejidad de la política tributaria, deficiencias en la administración de impuestos, y altas tasas de evasión y elusión fiscal. En este contexto, la subdeclaración de ingresos, donde individuos y empresas informales reportan menos ingresos de los que realmente perciben, plantea un desafío significativo para identificar con precisión los ingresos de los trabajadores en el país. Este fenómeno no solo afecta el equilibrio fiscal del país, sino que también limita la formalización del mercado laboral y la creación de un sistema de seguridad social inclusivo. (DIAN, s.f.)

En este contexto, desarrollar modelos de predicción de ingresos basados en datos reales, como los proporcionados por la Gran Encuesta Integrada de Hogares (GEIH), es crucial. Esta encuesta, realizada por el DANE, ofrece información detallada sobre el mercado laboral en Colombia, incluyendo la composición demográfica, las condiciones de empleo y los ingresos de los trabajadores. El objetivo del taller es construir un modelo predictivo del ingreso por hora de los individuos empleados en Bogotá, utilizando los datos del GEIH de 2018, que forman parte de la ”Medición de Pobreza Monetaria y Desigualdad”.

Según cifras del informe de Ocupación informal Trimestre abril - junio 2024 del DANE, en Colombia la informalidad laboral afecta a más del 56 % de la población trabajadora, siendo uno de los principales factores que contribuyen a la evasión fiscal. Esto significa que millones de trabajadores no reportan ingresos de manera adecuada, lo que afecta no solo al recaudo de impuestos, sino también al acceso a prestaciones sociales y a la sostenibilidad del sistema de salud y de pensiones. En ciudades como Bogotá, donde la informalidad es menor (33,7 %), la creación de un modelo que prediga con precisión los ingresos puede ayudar a identificar tanto a individuos en riesgo de evasión fiscal como a aquellos que podrían necesitar políticas de inclusión económica y formalización (Escobar, 2024).

Además, la predicción de ingresos permite a las autoridades fiscales como la DIAN desarrollar estrategias más eficientes para la detección de fraudes y subdeclaración de ingresos, y a los economistas obtener una mejor comprensión de los patrones salariales en la economía urbana. El análisis de estos datos ayuda a construir perfiles salariales en función de variables como la edad, el género, y la educación, y permite evaluar cómo estos factores influyen en la distribución de los ingresos. La capacidad de prever los ingresos por hora de los trabajadores no solo permite reducir la evasión fiscal, sino también diseñar políticas públicas más inclusivas y justas que promuevan el desarrollo económico y la equidad social en el país.

De esta manera, este taller se enmarca en un contexto nacional en el que la evasión fiscal y la informalidad laboral afectan de manera significativa el bienestar económico. El desarrollo de un modelo predictivo del ingreso por hora, basado en datos reales, representa una herramienta fundamental para mejorar la administración tributaria, reducir la desigualdad y avanzar hacia un sistema económico más formal e inclusivo.

Con el objetivo de que los resultados sean replicables, el trabajo cuenta con un repositorio de GitHub. El repositorio consta de cinco carpetas principales. En la primera carpeta denominada *Documento* se encuentra el documento final. La segunda carpeta denominada *Scripts* contiene los códigos con los que se replica el trabajo. La tercera carpeta *Base Datos* almacena la base de datos ajustada que es empleada para realizar el taller. La cuarta carpeta *Gráficas* almacena las visualizaciones que ayudaron al análisis de la información. Finalmente, *Latex* almacena los documento .tex con los resultados de las diferentes regresiones.

2. Datos

La Gran Encuesta Integrada de Hogares (GEIH) es una encuesta de periodicidad mensual recolectada por el DANE, cuyo principal objetivo es hacer un seguimiento detallado al mercado laboral en Colombia. Además de incluir información relacionada con el empleo, la encuesta también consulta características generales del hogar y de las diversas fuentes de ingresos. Con cobertura a nivel nacional, la GEIH presenta resultados tanto para áreas urbanas (cabeceras) como rurales, así como para 13 ciudades principales y sus áreas metropolitanas. En el contexto de este estudio, la información utilizada corresponde a la ciudad de Bogotá en el año 2018, la cual fue obtenida mediante técnicas de web scraping del repositorio de GitHub del profesor Ignacio Sarmiento.

Esta base de datos es esencial para el análisis propuesto, ya que proporciona una visión integral del comportamiento laboral y de ingresos en Bogotá, lo que permite modelar de manera precisa los factores que determinan el salario por hora de los individuos. La utilidad de esta información radica en su capacidad para capturar una amplia variedad de variables que pueden influir en los ingresos, como la edad, el nivel educativo, el género, el tipo de ocupación, formalidad, entre otras. Al aplicar esta base de datos en el problema planteado, podemos construir un modelo de predicción de ingresos que no solo nos permita entender las relaciones existentes entre estos factores, sino también identificar posibles anomalías o brechas en el reporte de ingresos, contribuyendo así a reducir la subdeclaración de salarios y, potencialmente, apoyar en el diseño de políticas públicas más efectivas, que por ejemplo, ayuden a reducir la brecha salarial entre hombres y mujeres.

El procedimiento detallado de cómo se realizó el web scraping para obtener esta información se describirá a continuación, asegurando que los lectores comprendan los pasos y herramientas utilizadas para acceder y procesar los datos.

2.1 Proceso de extracción de datos

Para el análisis propuesto, es necesario consolidar la información de la Gran Encuesta Integrada de Hogares (GEIH) de manera mensual. Esto es importante porque la consolidación de datos mensuales permite obtener una visión integral y continua del mercado laboral a lo largo del año.

Normalmente, los datos se almacenan en el directorio de microdatos anonimizados del DANE y son de acceso público para el libre análisis de información. Estos se encuentran divididos en archivos .zip, según las características requeridas para el análisis. Teniendo en cuenta el objetivo de este trabajo, sería necesario descargar los datos de cada mes del año 2018 y combinar al menos las bases de Ocupados, que contienen información sobre condiciones laborales y remuneración, y la base de características generales, que incluye datos sociodemográficos. Sin embargo, el enlace proporcionado en el taller recopila las principales variables de estas bases relacionadas, simplificando el proceso y evitando posibles problemas como actualizaciones, cambios en los datos o fallos en la página del DANE.

Para adquirir los datos de la Gran Encuesta Integrada de Hogares (GEIH), se empleó la técnica de web scraping utilizando el software R Studio. Este método puede tener implicaciones legales en relación con los términos de uso de los sitios web y el acceso a la información, por lo que es fundamental llevarlo a cabo conforme a las normativas vigentes, así como respetar el archivo robots.txt, que establece los permisos y restricciones de las páginas para este tipo de procesos.

Al consultar el enlace proporcionado para el taller, se observó que no existía un archivo robots.txt, lo que sugiere la ausencia de restricciones explícitas para el uso de web scraping. Con base en esta observación, se procedió con la extracción de los datos. La información proveniente del repositorio del profesor Ignacio Sarmiento estaba dividida en diez partes, cada una accesible a través de enlaces externos que contenían las tablas de datos.

El proceso comenzó con la definición de la URL base, desde la cual se extrajeron los enlaces de cada parte de la muestra utilizando el paquete `rvest` para leer el contenido HTML de la página. Para extraer la información, era necesario que el código detectara el enlace de cada una de las partes de la encuesta. Estos enlaces seguían una sintaxis uniforme, variando únicamente en el número de página que indicaba la parte específica del

conjunto de datos. Así, se implementó un proceso iterativo para acceder a cada enlace, ajustar la numeración de la página y localizar el elemento HTML correspondiente a la tabla.

Este enfoque permitió descargar y combinar cada tabla en un único conjunto de datos consolidado en R. De esta manera, se recorrieron los diez vínculos correspondientes y se unieron las tablas de manera vertical, generando una base de datos final con 32,177 observaciones y 178 variables. Finalmente, la base consolidada fue guardada en formato CSV y RDS para su posterior análisis.

2.2 Selección y limpieza de datos

Dado que el objetivo del estudio es predecir los ingresos por hora, se ha decidido centrar el análisis únicamente en los adultos residentes en Bogotá. La Gran Encuesta Integrada de Hogares (GEIH) permite aplicar un filtro por edad, lo que dio como resultado la eliminación de 7.609 observaciones correspondientes a personas menores de 18 años. Además, se identificó la ocupación laboral de los individuos y se excluyeron 8.933 registros de personas que no estaban empleadas. Como resultado, el análisis final incluyó 16.542 observaciones, lo que equivale aproximadamente al 51 % de la muestra original.

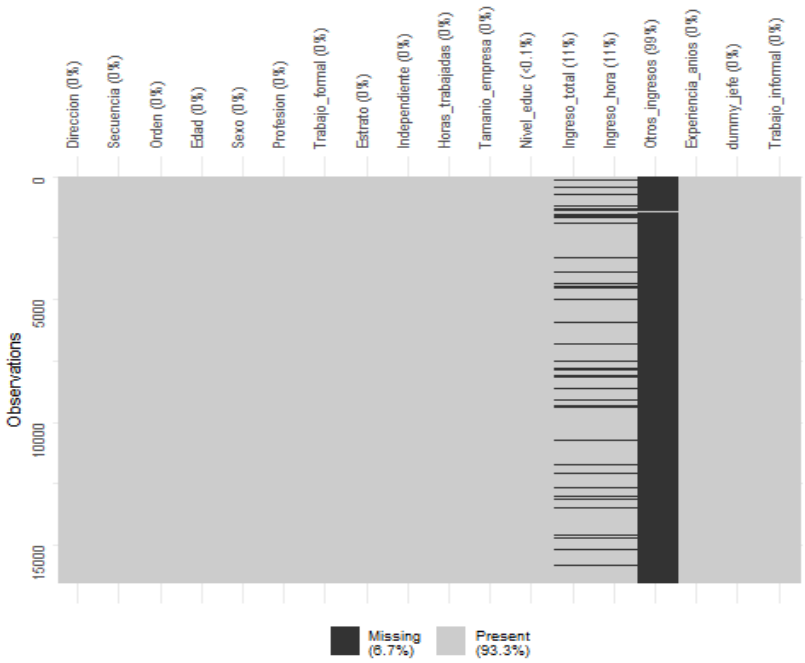
Para construir un modelo predictivo eficaz del salario por hora en Bogotá utilizando los datos de la GEIH, es esencial seleccionar variables que capturen adecuadamente los determinantes salariales y reflejen tanto las condiciones específicas del mercado laboral colombiano como los patrones observados a nivel internacional. Las variables seleccionadas para el modelo predictivo incluyen aquellas que permiten la identificación y referencia de individuos representativos en la muestra. Esto abarca las variables con características socio demográficas, como edad, género y nivel educativo, que proporcionan información sobre las características de los individuos. También se consideran variables laborales, que reflejan las condiciones de empleo, ocupación y tipo de trabajo. Finalmente, se incorporan variables relacionadas con los ingresos percibidos, que son cruciales para la predicción del salario. Estas variables buscan representar las siguientes características de los individuos:

- Edad y experiencia laboral: la intuición y el modelo teórico de Mincer (1974) sugiere que se da un aumento en los ingresos con la mayor experiencia del trabajador; sin embargo, este es un proceso decreciente en el cual la utilidad marginal de un año de experiencia disminuye con la acumulación de la misma.
- Nivel educativo: el nivel educativo, especialmente en países en desarrollo, es un determinante clave de los ingresos.
- Género: en Colombia, como en muchos países, existe una discriminación salarial por género que ocasiona que las mujeres tengan una remuneración salarial menor con la misma cantidad de experiencia y educación.
- Sector Económico (reflejado por la ocupación): las ocupaciones con mayores requisitos de habilidades, como profesionales o gerenciales, tienden a generar ingresos significativamente más altos que ocupaciones manuales o menos especializadas. Los estudios sugieren que sectores intensivos en capital humano, como la tecnología y las finanzas, ofrecen salarios superiores a sectores como la agricultura o el comercio minorista.
- Tamaño de la empresa:Diversos estudios han documentado la relación entre el tamaño de la empresa y sus ingresos. Según Oi e Idson (1999), los empleados de empresas más grandes tienden a recibir salarios más elevados, lo cual se atribuye a economías de escala y una mayor capacidad de negociación.
- Vinculación laboral (Independiente o asalariado):De acuerdo con un estudio del Banco Mundial (2013), los trabajadores independientes en Colombia enfrentan mayores vulnerabilidades económicas y tienden a reportar menores ingresos netos debido a la inestabilidad y falta de acceso a beneficios laborales.
- Formalidad: La formalidad laboral ha sido identificada como un factor importante en la determinación de los ingresos. Los trabajadores formales, que tienen acceso a seguridad social y contratos laborales, tienden a ganar más que sus contrapartes informales. Estudios de Perry et al. (2007) demuestran que la informalidad está asociada con menores ingresos y condiciones laborales más precarias, lo que perpetúa la desigualdad en el mercado laboral.
- Horas Trabajadas: la cantidad de tiempo dedicada al trabajo, en general, tiene implicaciones en el nivel de ingresos percibidos. Esto es especialmente significativo en trabajos a medio tiempo o intermitentes que se traducen en un menor salario. Sin embargo, este es un factor que disminuye su productividad en donde mayor tiempo no necesariamente se refleja en un mayor salario, especialmente en países en desarrollo.

- Estrato y posición en el hogar: Ser jefe de hogar puede implicar un nivel de ingresos más altos en la medida que este, por tradición, cumple un papel de proveedor. Además, el estrato socioeconómico influye significativamente en la capacidad de generar ingresos donde esté refleja riqueza y un mayor acceso a mejores oportunidades.

De las 18 variables escogidas se observa que tres de ellas contienen un porcentaje considerable de missing values. La figura 1 permite observar visualmente como se encuentran la distribución de estos missing values, en la cual los *ingresos totales* e *ingresos por hora* contienen un 11 % de observaciones faltantes, mientras que *otros ingresos* se encuentra sin datos casi en su totalidad (99 %), por lo cual se elimina de la base de datos.

Figura 1: Distribución de missing values



Por otro lado, dado que las demás variables son relevantes para el estudio, se realizó la imputación de los valores faltantes. En primer lugar, se calcula el ingreso por hora promedio para reemplazar estos valores, teniendo en cuenta una agrupación basada en la relación laboral (si el trabajador es independiente o no) y la profesión, ya que esta puede influir en el nivel de remuneración por el mismo tiempo de trabajo. Posteriormente, se imputa el nivel educativo utilizando la moda, es decir, el valor más frecuente de esta variable, sustituyendo los valores faltantes con este valor. Se eligió la moda porque es una medida adecuada para variables categóricas, ya que permite preservar la categoría más común y representativa en los datos.

Finalmente, dado que los valores atípicos extremos pueden distorsionar significativamente los resultados del análisis estadístico, afectando la precisión y la interpretación de los modelos, se realiza un proceso de *winsorización* para limitar estos valores al percentil 99. Esta técnica mejora la robustez del análisis y proporciona estimaciones más fiables. Además, se crean variables adicionales, como el logaritmo del ingreso por hora, factores para variables categóricas y nuevas variables derivadas, como la edad al cuadrado y la experiencia en años. Estas nuevas variables enriquecen el análisis al ofrecer perspectivas adicionales y permitir una evaluación más completa y detallada en el análisis estadístico.

2.3 Estadística descriptiva

La tabla 1 presenta las estadísticas descriptivas para el conjunto de 16,540 observaciones.

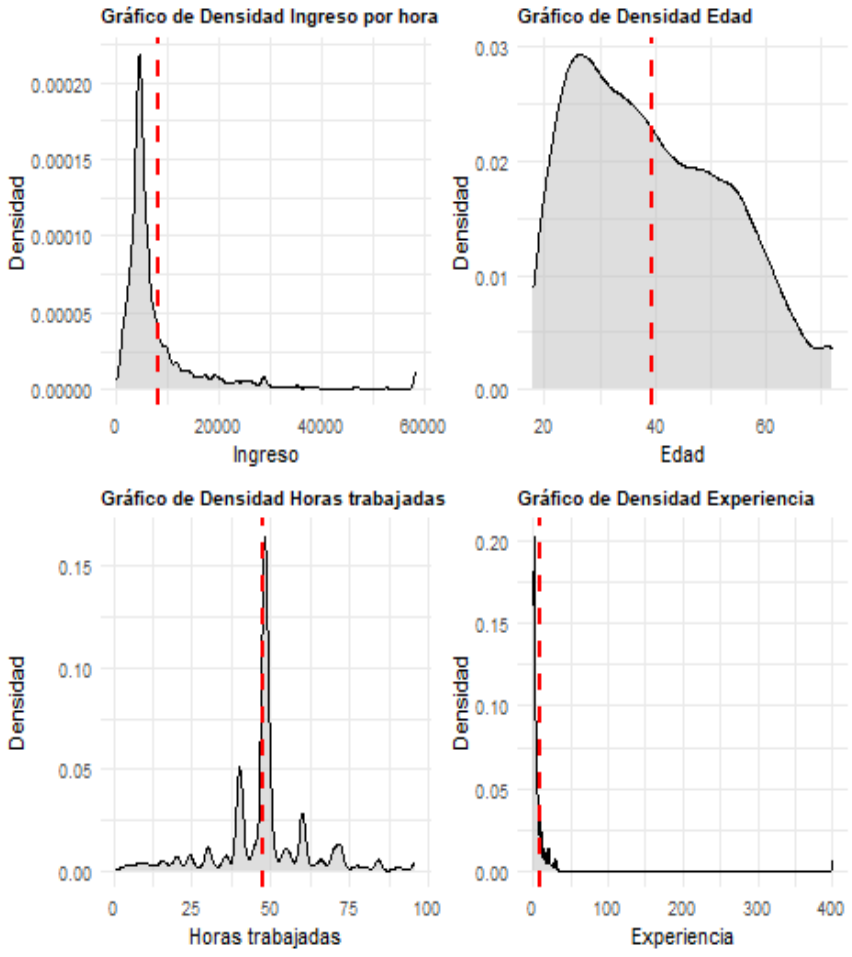
Tabla 1: Estadísticas Descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Ingreso por Hora (Winsorizado)	16,540	8,278.1	9,328.7	0.5	58,333.3
Edad (Winsorizada)	16,540	39.4	13.4	18	72
Sexo (Mujer)	16,540	0.5	0.5	0	1
Trabajo Informal	16,540	0.4	0.5	0	1
Horas Trabajadas (Winsorizado)	16,540	47.3	15.4	1	96
Experiencia (Winsorizada)	16,540	8.9	40.0	0.0	400.0
Independiente	16,540	0.3	0.5	0	1
Jefe de hogar	16,540	0.5	0.5	0	1

Se observa que la edad de los individuos varía en un rango de 18 a 72 años, con una media de aproximadamente 39 años. En cuanto al sexo, la distribución es bastante equilibrada, con una proporción cercana al 50 % de mujeres en la muestra. La variable de trabajo informal revela que el 40 % de los individuos están empleados en el sector informal, lo que refleja la dinámica del mercado laboral colombiano, que históricamente ha mantenido un alto grado de informalidad. Esta informalidad puede impactar en el ingreso, dado que los trabajadores informales a menudo enfrentan menos estabilidad y menores oportunidades para incrementar sus ingresos. En contraste, aproximadamente el 30 % de la muestra trabaja como independiente, lo que puede indicar una mayor flexibilidad laboral pero también un desafío adicional en términos de ingresos y cobertura de seguridad social, ya que los trabajadores independientes a menudo necesitan mayores ingresos para cubrir sus propios gastos de seguridad social y otros beneficios.

Ahora, respecto a la distribución general de los datos, la figura 2 nos permite observar el comportamiento de algunas de las variables más importantes para nuestro estudio:

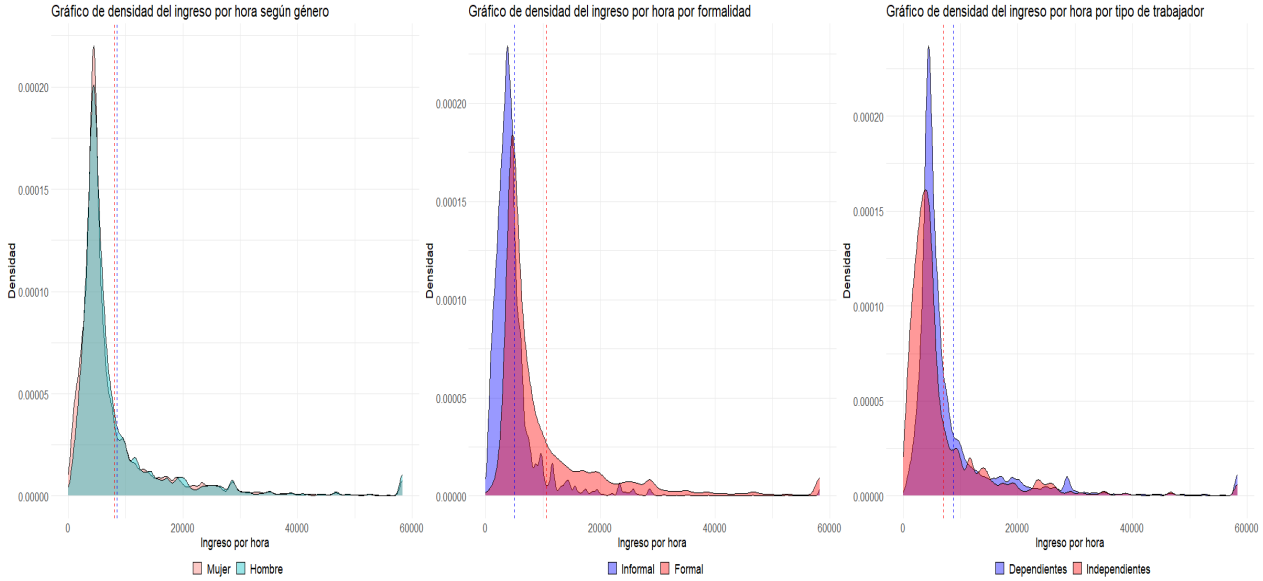
Figura 2: Densidad del promedio del ingreso



La gráfica de densidad para el ingreso por hora revela una distribución asimétrica, con un pico situado antes de la media. Esto indica que la mayoría de la población tiene ingresos por debajo del promedio, mientras que los ingresos significativamente altos de algunos individuos elevan el promedio general. En contraste, la gráfica de densidad para las horas trabajadas presenta una distribución más uniforme, con un ligero sesgo hacia la derecha, sugiriendo que algunos trabajadores laboran más horas que el promedio, aunque esta proporción no es significativa. Por último, la gráfica de densidad para la experiencia muestra una distribución con una cola extendida hacia la derecha, lo que indica que, a pesar de que la mayoría de los individuos tienen niveles de experiencia moderados, también hay casos con experiencia excepcionalmente alta.

Ahora, la figura 3 muestra la distribución del ingreso por hora para trabajadores formales e informales, segmentado por sexo.

Figura 3: Densidad del ingreso por sexo, formalidad y tipo de trabajador



En primer lugar, al comparar los ingresos por sexo, se observa que los hombres, en promedio, reciben mayores ingresos por hora que las mujeres. Así mismo, las mujeres presentan una mayor concentración en los niveles bajos de ingreso, mientras que los hombres tienden a ubicarse en niveles de ingreso un poco más elevados.

En cuanto a la diferenciación por formalidad laboral, muestra que los trabajadores formales tienen un nivel de ingresos superior en comparación con los trabajadores informales. Los primeros se concentran en niveles más altos de ingresos, mientras que los informales predominan en los niveles más bajos y con una concentración mayor. Esto viene explicado por las desventajas económicas que enfrenta la población laboral que trabaja fuera de la formal y se dedican a actividades que comparativamente son menos productivas.

Finalmente, al observar el tipo de trabajador, los dependientes presentan ingresos promedio más altos en comparación con los trabajadores independientes, quienes tienen una mayor densidad en niveles más bajos de ingreso por hora. Este patrón indica una desventaja relativa para los trabajadores que operan de manera independiente. Esto puede deberse a que existe una mayor estabilidad en los ingresos de los trabajadores dependientes y la facilidad en el acceso a beneficios laborales que suelen tener los trabajadores dependientes.

3. Modelo Perfil Edad-Ingreso

Para la modelación de los ingresos de la población ocupada en Bogotá, como primer ejercicio es fundamental considerar el perfil de ingreso en función de la edad, ya que refleja una relación cuadrática bien documentada en la literatura económica. Esta relación describe cómo, al inicio de la vida laboral, los ingresos de los individuos tienden a aumentar debido a la acumulación de experiencia y educación. Sin embargo, tras alcanzar un punto máximo, los ingresos se estabilizan o incluso disminuyen debido a la disminución en la productividad. Este fenómeno ha sido estudiado en diversas economías y ha llevado al desarrollo de modelos empíricos para su cuantificación.

Uno de los marcos teóricos más utilizados para este análisis es la ecuación de Mincer (1974), que vincula los ingresos de los trabajadores con la educación y la experiencia laboral. Esta ecuación ha sido un referente en los estudios sobre el mercado laboral y ha sido complementada por investigaciones posteriores, como la de Lemieux (2006), que profundiza en las variaciones de los retornos a la experiencia y la educación en distintos contextos. De acuerdo con estos estudios, la relación entre edad e ingreso sigue una forma de Ü invertida", donde los ingresos aumentan inicialmente, pero luego decaen debido a factores como la pérdida de productividad y la proximidad a la jubilación.

Además, desde una perspectiva más amplia, autores como Becker (1964) en su teoría del capital humano, explican cómo la inversión en educación y formación a lo largo del ciclo de vida laboral de los individuos impacta directamente en sus ingresos. La teoría del capital humano sostiene que los ingresos aumentan a medida que los trabajadores invierten en su propia formación y habilidades, lo que justifica el aumento de ingresos en las primeras etapas de la vida laboral. Card (1999) también contribuye a este debate, destacando cómo los retornos a la educación y a la experiencia pueden variar significativamente en función de factores como las políticas laborales y las características del mercado.

Además de la ecuación de Mincer, otra vertiente importante para la comprensión de esta relación es la que aborda las decisiones de optimización entre ocio y trabajo, particularmente a lo largo del ciclo de vida laboral. Borjas (2015) analiza cómo los trabajadores equilibran las horas de trabajo y el ocio en función de su edad, argumentando que al inicio y al final de la carrera laboral, el costo de oportunidad del ocio es bajo, lo que da como resultado menores ingresos laborales. Sin embargo, durante los años de mayor productividad, este costo aumenta, llevando a los trabajadores a dedicar más tiempo al trabajo y, por ende, generar mayores ingresos.

Finalmente, estudios más recientes, como los de Heckman, Lochner y Todd (2006), amplían la comprensión sobre la relación entre la acumulación de capital humano y los ingresos a lo largo del ciclo de vida. Estos autores destacan la importancia de las inversiones tempranas en habilidades y educación como determinantes clave del perfil de ingresos a lo largo del tiempo, y cómo estas inversiones pueden amplificar o mitigar las desigualdades salariales en una economía.

De esta manera, 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):

$$\ln(\text{Ingreso por hora}_i) = \alpha + \beta_1 \text{Edad}_i + \beta_2 \text{Edad}_i^2 + X_i'\beta + u_i$$

Donde ln(Ingreso por hora$_i$) corresponde a los ingresos por hora del individuo i. Se realiza la transformación logarítmica con el fin de facilitar la interpretación de los coeficientes de la regresión. La Edad y la Edad^2 corresponden a la edad y el cuadrado de la edad para cada individuo i. La inclusión del término cuadrático permite modelar la relación en u invertida entre salarios y la edad y el término u_i corresponde al término error idiosincrático, que representa las variables que no están en nuestro modelo y que explican los salarios.

No obstante, en un modelo que intenta predecir los ingresos de los individuos basándose exclusivamente en la edad y el cuadrado de la edad, se pueden presentar problemas significativos debido a la omisión de variables importantes. Factores críticos como el nivel educativo, la experiencia laboral específica, las horas trabajadas, entre otros, también influencian significativamente los ingresos. La exclusión de estas variables puede llevar a sesgos en las estimaciones y a una comprensión incompleta de las dinámicas del ingreso. Esto, a su vez, podría dar como resultado predicciones inexactas o en políticas mal orientadas basadas en dichas estimaciones. Por lo tanto, es crucial incorporar un conjunto más amplio de variables para capturar de manera más completa los factores que afectan los ingresos de los individuos.

Por esta razón, en la ecuación anterior, X_i incluye un conjunto de variables de control como el género del individuo, una variable dicotómica que determina si el individuo trabaja en el sector informal, otra que indica si es jefe de hogar, el máximo nivel educativo alcanzado, experiencia[1], oficio, las horas laboradas, una variable dummy que toma el valor de 1 si es trabajador independiente y 0 si es asalariado, el estrato de la vivienda y el

[1]Se usa la variable p6426 "¿cuánto tiempo lleva ... Trabajando en esta empresa, negocio, industria, oficina?çomo proxy de experiencia laboral. Sin embargo, esta aproximación puede presentar limitaciones. Dado que la variable mide exclusivamente la duración en el empleo actual, podría dar como resultado una subestimación de la experiencia laboral total del individuo si este ha ocupado puestos previos en otras empresas. Además, no refleja la movilidad laboral entre diferentes empresas, un aspecto que podría indicar una experiencia laboral más diversa y enriquecedora. Asimismo, no diferencia entre la progresión en roles o los cambios de puesto dentro de la misma empresa, factores que pueden influir significativamente en los ingresos.

tamaño de la empresa en la que trabaja.

En la Tabla 2 se presentan los resultados de los análisis realizados de los modelos. La primera columna muestra los resultados del modelo sin variables de control. La segunda columna incluye los resultados del modelo con las variables de control[2] previamente mencionadas, pero excluyendo valores atípicos del modelo.

Tabla 2: Modelos de ingreso - edad

	Log(Ingresos por hora)	
	Sin Controles	Con controles Sin valores atípicos
	(1)	(2)
Edad	0.056***	0.037***
	(0.003)	(0.002)
Edad al cuadrado	-0.001***	-0.0004***
	(0.00003)	(0.00002)
Observaciones	16,540	15,662
R^2	0.022	0.686
R^2 Ajustado	0.022	0.684
Error Estándar Residual	0.799 (df = 16537)	0.402 (df = 15562)
Estadístico F	188.155*** (df = 2; 16537)	343.765*** (df = 99; 15562)
Notas:		*p<0.1; **p<0.05; ***p<0.01
		Errores estándar en paréntesis

Antes de profundizar en el análisis de los modelos, es importante aclarar que la selección de este último modelo (columna 2) se basó en un análisis exhaustivo de las observaciones influyentes y los valores atípicos. Este proceso comparativo incluyó tres versiones del modelo: una versión inicial con todas las observaciones, una ajustada que excluye las observaciones influyentes[3], y una tercera que elimina los valores atípicos[4]. Los resultados de este análisis se presentan en la Tabla 2, proporcionando una justificación sólida para la elección del modelo y permitiendo un análisis más profundo de los resultados.

Ahora bien, como se observa en la tabla 3, la estabilidad en los niveles de significancia y las pequeñas variaciones en la magnitud de los coeficientes, y en el R^2 Ajustado entre el modelo original (16.540 observaciones) y el modelo sin observaciones influyentes (15.898 observaciones) confirman la robustez de las estimaciones. Sin embargo, al remover los valores atípicos (15.662 observaciones), como se muestra en la tercera columna, se destaca un incremento considerable en el R^2 ajustado, de 0.572 a 0.689, pero no se evidencia cambios significativos en los coeficientes. Este incremento subraya la mejora en la capacidad explicativa del modelo, gracias a la reducción del ruido estadístico y a una mejor representación de las relaciones entre las variables. Esto conduce a un ajuste más efectivo y preciso de los datos. Así, el modelo de la columna (2) de la tabla 2 se identifica como el modelo con controles pero sin valores atípicos, demostrando ser el más representativo de las variables estudiadas.

[2]Los coeficientes de las variables de control, que incluyen género, informalidad, jefatura de hogar, máximo nivel educativo, experiencia, oficio, horas laboradas, tipo de trabajador (independiente o asalariado), estrato y tamaño de la empresa, fueron ocultados en la presentación de los resultados. Esta decisión se tomó para evitar sobrecargar la tabla con múltiples variables, lo que podría complicar la claridad y comprensión de los resultados en una tabla excesivamente extensa.

[3]Observaciones influyentes son aquellas con un apalancamiento 2 o 3 veces superior al apalancamiento medio

[4]Un valor atípico se define como una observación con un residuo studentizado superior a 2, según el enfoque de William Greene.

Tabla 3: Análisis de observaciones influyentes y valores atípicos para el modelo con controles.

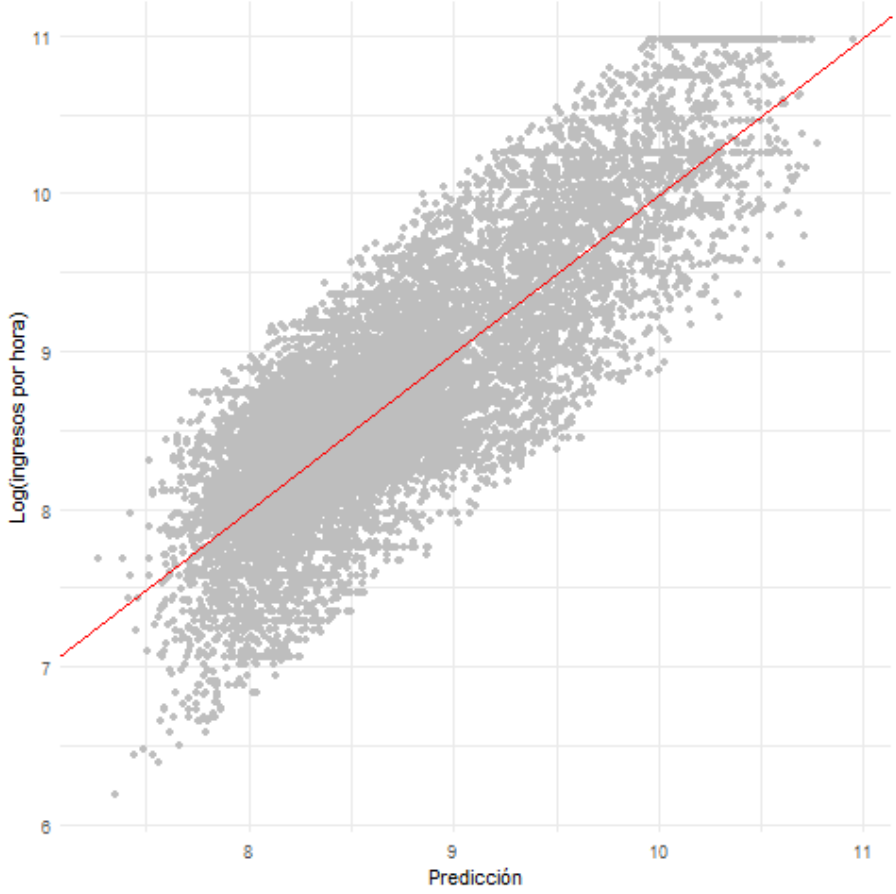
	Log(Ingresos por hora)		
	Con controles	Sin obs. influyentes	Sin Valores atípicos
	(1)	(2)	(3)
Edad	0.039***	0.040***	0.037***
	(0.002)	(0.002)	(0.002)
Edad al cuadrado	-0.0004***	-0.0004***	-0.0004***
	(0.00002)	(0.00002)	(0.00002)
Observaciones	16,540	15,898	15,662
R^2	0.571	0.571	0.686
R^2 Ajustado	0.569	0.569	0.684
Error Estándar Residual	0.531 (df = 16440)	0.529 (df = 15827)	0.402 (df = 15562)
Estadístico F	221.214*** (df = 99; 16440)	300.390*** (df = 70; 15827)	343.765*** (df = 99; 15562)
Notas:		*p<0.1; **p<0.05; ***p<0.01	
		Errores estándar en paréntesis	

Haciendo la anterior salvedad y retomando el análisis de los modelos de regresión mostrados en la Tabla 2, se ofrece una visión detallada de cómo la inclusión de variables adicionales mejora la estimación del logaritmo del ingreso por hora en función de la edad y su cuadrado. El modelo en la primera columna, que solamente emplea estas dos variables, arroja un coeficiente de 0.056 para la edad y -0.001 para la edad al cuadrado, ambos con significancia estadística al nivel del 1 %. Estos coeficientes sugieren que, manteniendo constantes los demás factores, un aumento de un año en la edad está vinculado con un incremento aproximado del 5.6 % en los ingresos por hora. Por otro lado, el término cuadrático indica una disminución del 0.1 % en la tasa de crecimiento de estos ingresos por cada año adicional. Esto refleja la típica relación cuadrática entre la edad y los ingresos, donde los ingresos se incrementan hasta un máximo y luego disminuyen a medida que avanza la edad. No obstante, este modelo básico solo explica el 2.2 % de la variación en los ingresos por hora, como lo muestra su R^2 ajustado de 0.022.

Al introducir controles adicionales en la segunda columna—tales como género, informalidad, jefatura de hogar, máximo nivel educativo, experiencia, oficio, horas laboradas, tipo de trabajador (independiente o asalariado), estrato y tamaño de la empresa—los coeficientes de la edad se ajustan a 0.037 y -0.0004 para la edad al cuadrado, con una precisión aumentada manifestada por errores estándar reducidos a 0.002 y 0.00002, respectivamente. Este ajuste sugiere que las variaciones en los ingresos previamente atribuidas a la edad son en parte explicadas por estos factores adicionales. Esto indica que, con los nuevos controles, cada año adicional de edad está asociado con un aumento del 3.7 % en el ingreso por hora, mientras que el impacto decreciente de cada año adicional se reduce a 0.04 %. Este modelo más complejo mejora dramáticamente en términos de ajuste, elevando el R^2 ajustado a 0.689, lo que significa que ahora explica el 68.9 % de la variabilidad en los ingresos.

Además, el Error Estándar Residual en el modelo mejorado es considerablemente más bajo, reduciéndose de 0.799 en la primera columna a 0.402 en la segunda. Esta reducción no solo indica que las predicciones del modelo están mucho más cerca de los valores reales observados, sino que también subraya la eficacia de incluir una variedad más amplia de variables, como se observa en la Figura 4. El Estadístico F aumenta significativamente de 188.155 a 353.765, reflejando una mejora en la significancia global del modelo gracias a los controles adicionales.

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



Estos resultados destacan que mientras el primer modelo proporciona un punto de partida útil para entender la relación entre edad e ingresos, es el segundo modelo el que ofrece una comprensión más profunda y precisa. La mejora en el R^2 ajustado, junto con una disminución en el Error Estándar Residual y un aumento en el Estadístico F, justifica claramente la inclusión de variables adicionales para captar de manera más efectiva los determinantes de los ingresos, proporcionando una base más sólida para políticas económicas y estrategias de gestión laboral basadas en una comprensión más completa de las dinámicas de ingreso.

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

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

$$-\beta_1 = 2\beta_2 \text{Edad}$$

$$\text{Edad} = -\frac{\beta_1}{2\beta_2}$$

De manera que, si remplazamos los coeficientes de tabla 3 en la ecuación anterior, se observa que el pico de edad promedio estaría en 42.64 años para el primer modelo (simple), mientras que para el segundo modelo (con controles) llegaría a 46.36 los años.

En el análisis de regresión de los modelos presentados, surge una preocupación significativa relacionada con la heterocedasticidad, un fenómeno que puede comprometer la validez de las inferencias estadísticas derivadas de los modelos. La heterocedasticidad ocurre cuando la varianza de los errores del modelo no es constante, lo cual es particularmente relevante en nuestro contexto, donde la varianza de los errores podría variar con la edad. Esta variabilidad en los errores puede ser resultado de diferencias en la dispersión de los ingresos a lo largo de la vida laboral de los individuos, ya que factores como cambios en la estabilidad laboral, la acumulación de experiencia o diferencias en la recompensa económica a lo largo de la carrera pueden influir significativamente.

La presencia de heterocedasticidad en los datos puede llevar a estimaciones de los errores estándar que son erróneas, poniendo en riesgo la confiabilidad de las pruebas estadísticas, como las t-tests para la significancia de los coeficientes, y por ende, la interpretación de los resultados del modelo. Esto ocurre porque los métodos convencionales para calcular los errores estándar asumen homocedasticidad, es decir, que todos los residuos tienen la misma varianza, independientemente del valor de las variables independientes como la edad.

Para abordar esta problemática, se aplica el método de Bootstrap con 1,000 repeticiones en el análisis. Este método permite recalibrar los errores estándar de una manera más robusta y confiable. Mediante el re-muestreo de los datos y la estimación repetida del modelo, el Bootstrap genera una distribución empírica de los estimadores, permitiendo una evaluación más precisa de los errores estándar y los intervalos de confianza asociados.

Tabla 4: Errores estándar bootstrap

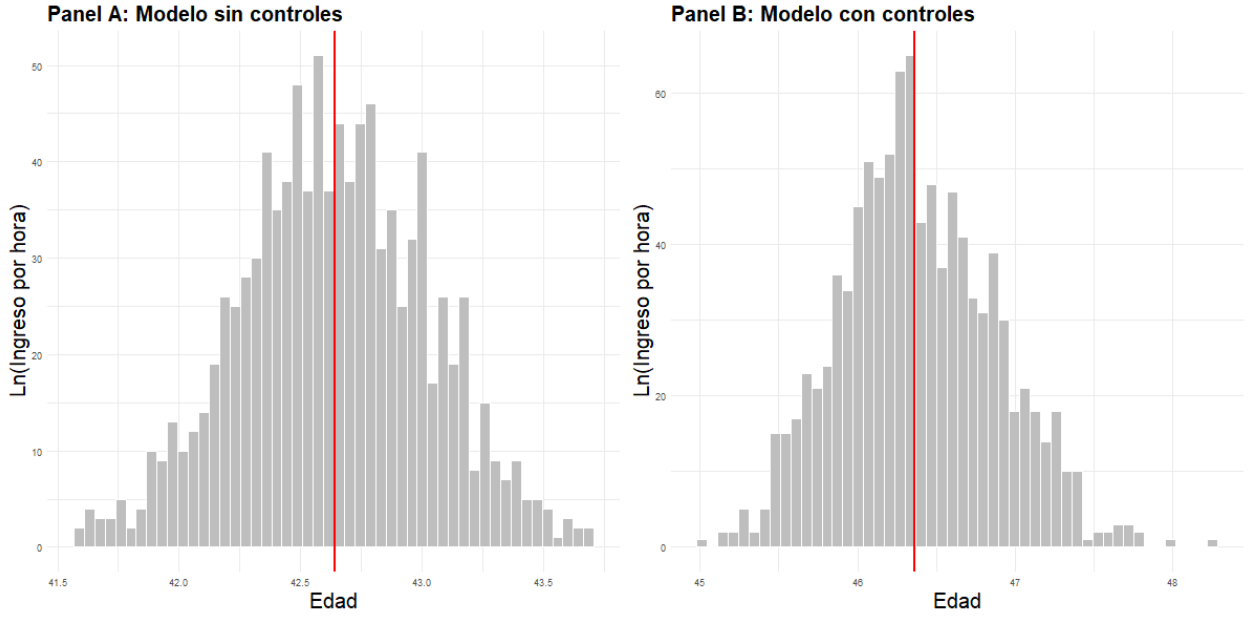
Modelo	Coeficiente	Sesgo	Error estándar
Sin controles	42.64	-0.0084	0.39
Con controles	46.36	0.0161	0.49

La tabla 4 proporciona resultados del bootstrap con 1.000 repeticiones para dos modelos distintos, uno sin variables de control y otro con ellas. Para el modelo sin controles, el coeficiente estimado es 42.64, con un sesgo casi nulo (-0.0084) y un error estándar de 0.39, lo que sugiere una estimación bastante precisa. En el modelo con controles, el coeficiente es 46.36, con un sesgo leve (0.0161) y un error estándar de 0.49, indicando que la inclusión de variables de control afecta tanto la magnitud del coeficiente como la precisión de la estimación, aunque introduce un pequeño sesgo.

La aplicación de este método revela que, según el modelo simple, el pico de ingresos reales se estima con un 95 % de confianza que se encuentra entre los 41.87 y 43.39 años. En cambio, el modelo con controles adicionales, que incluyen características socioeconómicas como la educación y el tamaño de la empresa, entre otros, sugiere que este pico ocurre entre los 45.48 y los 47.35 años. Este desplazamiento hacia edades más avanzadas en el modelo ajustado por controles indica que, al tomar en cuenta factores socioeconómicos adicionales, la edad en la cual se reciben los mayores ingresos laborales tiende a aumentar. Esto podría reflejar el impacto positivo de la acumulación de capital humano y otros activos socioeconómicos a lo largo de la carrera de un individuo.

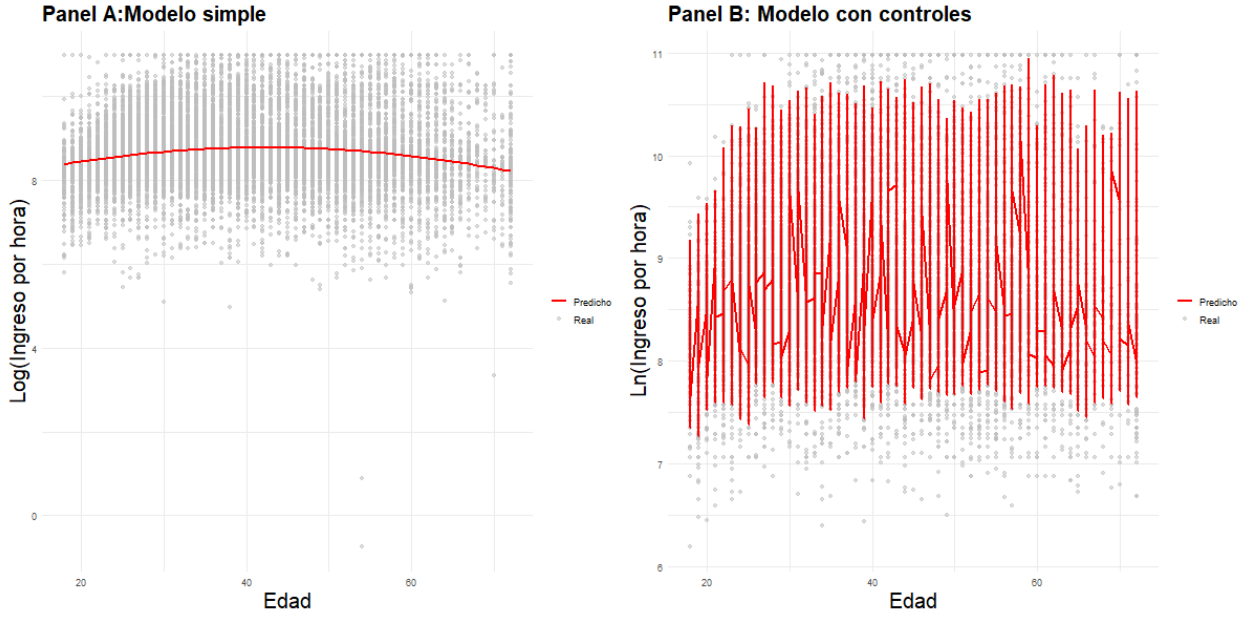
Estos hallazgos son presentados en la Figura 5, que ilustra la distribución de los resultados del análisis de Bootstrap, ofreciendo una visión gráfica del impacto de la heterocedasticidad y la eficacia de los métodos utilizados para mitigar sus efectos, asegurando así la integridad y la precisión de las conclusiones extraídas de los modelos de ingresos por edad.

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



La Figura 6 ilustra la dispersión entre la edad de los individuos de la muestra y sus ingresos, donde el ingreso por hora ha sido transformado usando una función logarítmica. En el Panel A, se presenta el modelo simple que sugiere una relación cuadrática entre la edad y el salario. Este modelo muestra que los ingresos son más bajos tanto al principio como al final de la vida laboral de una persona promedio, reflejando una curva asimétrica. Es decir, las personas suelen tener menores ingresos al inicio de su carrera, los cuales aumentan con la experiencia, pero luego disminuyen a medida que envejecen. En el Panel B, se presenta la misma relación, pero en un modelo más complejo que incorpora controles adicionales. En este caso, la relación entre la edad y los ingresos es menos evidente, ya que el modelo ajustado reduce el sesgo inicial que existía en la estimación simple. Este ajuste refleja cómo la inclusión de otros factores relevantes atenúa la claridad de la relación entre edad e ingresos, sugiriendo que otras variables además de la edad influyen en los salarios.

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



En conclusión, aunque el modelo complejo mejora la capacidad de predicción de los ingresos por hora al incluir

un mayor número de variables, el modelo simple (sin el vector de controles X) permite observar claramente una relación no lineal entre la edad y el ingreso, como se indicó al inicio de la sección. Esta relación muestra que los ingresos tienden a disminuir tanto en las primeras etapas como en las últimas etapas de la vida laboral. Factores como la reducción de la productividad en la vejez y el cambio en las preferencias hacia el ocio en los últimos años laborales explican parte de esta disminución de ingresos.

No obstante, es crucial tener en cuenta que los ingresos no dependen únicamente de la edad. Existen otros factores clave que también influyen, como el nivel educativo, el estrato socioeconómico, si la persona es jefe de hogar, el tipo de ocupación, las horas trabajadas, la condición de ser trabajador independiente o asalariado, el tamaño de la empresa en la que se desempeña y si se encuentra en la informalidad. Estos elementos, entre otros, interactúan con la edad y pueden alterar su impacto sobre los ingresos. Por esta razón, los modelos más complejos, que incorporan estos factores, ofrecen una interpretación más precisa y detallada de la relación entre edad e ingresos.

4. Brecha salarial por género

La brecha salarial por género, entendida como la diferencia en los ingresos laborales entre hombres y mujeres, ha sido un tema central de estudio de la economía laboral. En el caso colombiano, se ha identificado que las mujeres tienen salarios más bajos que los hombres, pese al aumento en su participación laboral, el mayor número de horas trabajadas y a la igualación de características observables, tales como la educación y los años de experiencia (Sabogal, A., 2012). La comprensión de los determinantes de esta brecha es fundamental para el diseño de políticas de desarrollo que la reduzcan y promuevan la equidad laboral. En este contexto, a lo largo del presente punto se estudiará la brecha salarial existente los hombres y las mujeres mayores de edad ocupados de Bogotá.

4.1. Brecha salarial por género incondicional

A continuación se estima a través de mínimos cuadrados ordinarios (MCO) la relación básica entre el logaritmo del salario por hora y el género. Este enfoque simple proporciona una intuición inicial de las diferencias en el ingreso entre mujeres y hombres. Sin embargo, es importante reconocer que esta aproximación básica está sujeta a problemas de sesgo, una vez que hay otros factores no considerados que de acuerdo con la literatura económica son determinantes del nivel del ingreso salarial, tales como el nivel de educación, los años de experiencia o la habilidad innata de las personas. En particular, se estima la siguiente regresión 4.1:

$$\ln(\text{Salario por hora}_i) = \beta_0 + \beta_1 \text{Mujer}_i + u_i \tag{4.1}$$

donde *Salario por hora*$_i$ representa el ingreso salarial por hora recibido por cada individuo *i*, *Mujer*$_i$ es una variable dicótoma que toma el valor de 1 si el individuo *i* es mujer y u_i representa el error asociado al individuo *i*. Los resultados de la presente estimación se encuentran en la tabla 4.1.

4.2. Brecha salarial por género condicional: ¿Equal pay for equal work?

Como se mencionó previamente, para la identificación correcta de la brecha salarial se deben tener en cuenta, además de la distinción por género de los individuos, otras características observables que de acuerdo con la literatura económica inciden en el valor del salario. Considerar este conjunto de nuevas variables permitiría obtener una estimación más plausible de la verdadera brecha salarial, una vez que se controlaría, en la medida de lo posible, por otras fuentes de variación. Si con base en esta estimación con controles se obtienen diferencias significativas en el salario entre hombres y mujeres, se podría argumentar la existencia de la brecha salarial asociada al género. Con este objetivo se estima la regresión 4.2:

$$\begin{aligned}\ln(\text{Salario por hora}_i) = \beta_0 + \beta_1 \text{Mujer}_i + \beta_2 \text{Edad}_i + \beta_3 \text{Edad}_i^2 + \beta_4 \text{Tamaño de la empresa}_i \\ + \beta_5 \text{Trabajo informal}_i + \beta_6 \text{Independiente}_i + \beta_7 \text{Oficio}_i + \beta_8 \text{Nivel de educacion}_i \\ + \beta_9 \text{Horas trabajadas}_i + \beta_{10} \text{Experiencia}_i + \beta_{11} \text{Estrato} + \beta_{12} \text{Jefe} + u_i\end{aligned} \tag{4.2}$$

donde *Salario por hora*$_i$ representa el ingreso salarial por hora recibido por cada individuo *i*, *Mujer*$_i$ es una variable dicótoma que toma el valor de 1 si el individuo *i* es mujer y u_i representa el error asociado al individuo *i*. Adicionalmente, se incluyeron los siguientes controles:

- *Edad*$_i$: Indica la edad del individuo *i*. Esta variable se espera que tenga una relación positiva, ya que a medida que aumenta la edad de los individuos puede aumentar la experiencia y sus capacidades laborales.

- *Edad*$^2{}_i$: Indica la edad del individuo *i*. Esta variable se incluye ya que la relacion entre la edad y el salario a menudo no es lineal, una vez que el salario aumenta hasta cierto punto (a medida que se asciende en jerarquia por ejemplo), pero después puede estabiliarse o reducirse (a medida de que disminuyen las responsabildiades o se alanza la jubilación).

- *Tamaño de la empresa*$_i$: Variable categórica que representa el tamaño de la empresa donde trabaja el individuo *i*. Esta variable recoge el hecho de que las empresas más grandes pueden ofrecer salarios más altos debido a los mayores recursos disponibles y quizás mayor nivel de eficiencia.
- *Trabajo informal*$_i$: Variable dicótoma que toma el valor de 1 si el *i* posse un trabajo informal desde el punto de vista del no acceso a seguridad social y 0 en otro caso. Esta variable se espera que tenga una relación negativa con el ingreso laboral, ya que el trabajo informal no debe cumplir con requisitos de seguridad social, salud, salario mínimo y por tanto pueden tener prestaciones salariales más bajas.

- *Independiente*$_i$: Variable dicótoma que toma el valor de 1 si el *i* trabaja por cuenta propia y 0 en otro caso. La relacion de esta variable con el salario puede ir en ambas direcciones. Por un lado, personas independientes pueden tener una mayor remuneración una vez que todos los ingresos del negocio los obtiene directamente. Sin embargo, la remuneración puede ser menor, una vez que el trabajo puede no ser igual de estable comparativamente a trabajar con un empleador.

- *Oficio*$_i$: Variable categórica de 1 a 99 que representa la diversidad de ocupaciones en la muestra de datos.Esta variable recoge las particularidades salariales de cada tipo de ocuapación.

- *Nivel de educacion*$_i$: Variable categórica que refleja el máximo nivel educativo alcanzado por la persona *i*. Un mayor nivel educativo normalmente se asocio con un nivel salarial más alto, debido al mayor nivel de preparación.

- *Horas trabajadas*$_i$: Variable que indica las horas trabajadas en la última semana por el individuo *i*. Se espera que las horas de trabajo estén positivamente relacionadas con el salario, ya que un mayor tiempo dedicado al trabajo suele traducirse en mayores ingresos.

- *Experiencia*$_i$: Variable indica el tiempo de experiencia laboral del individuo*i* en el sitio de trabajo. Se espera que la relación sea positiva una vez que mayor experiencia implica la mayor adquisición de habilidades o el ascenso jerárquico, lo que se traduce generalmente en mayores salarios. Esta variable se complementa con la edad, una vez que incorpora una dimensión diferente. En particular, el tiempo efetivo de experiencia puede ser muy menor comparativamente al número de años, por ejemplo por la entrada tardía en el mercado laboral.

- *Estrato*$_i$: Variable categórica del estrato socio-económico al que pertenece el individuo *i*. Esta variable recoge las particularidades salariales de cada tipo de estrato, relacionadas por ejemplo a posible discriminación en el proceso de contratación.

- *Jefe*$_i$: Variable dicótoma que toma el valor de 1 si el individuo *i* es jefe de hogar y 0 si no. Ser el jefe de hogar puede tener implicaciones en cuanto al rendimiento laboral.

4.2.1 Estimación: Brecha salarial por género condicional

Para llevar a cabo la estimación propuesta en la ecuacion 4.2, se aplicará el teormea Frisch-Waugh-Lovell (FWL). Este teorema facilita el análisis del impacto de una variable particular en un modelo de regresión multiple, mientras se mantiene constante el efecto de las otras variablers de control. En particular, este teorema es útil con grandes volúmenes de datos al mejorar la eficiencia en tiempo de la estimación. La estimación consta de dos etapas, y de acuerdo con el teorema, el resultado del coeficiente es equivalente al obtenido con la estimación tradicional de regresión multiple en una etapa, a veces más costosa computacionalmente dependiendo del volumen de información. De acuerdo con lo anterior, se llevará a cabo la estimación en dos etapas siguiente dicho teorema:

Etapa 1 (FWL): Se estiman mediante MCO dos regresiones auxiliares. La primera relaciona la variable explicada (logaritmo del salario por hora) con el conjunto de controles, pero excluyendo la variable de interés *Mujer*.

$$\begin{aligned}\ln(\text{Salario por hora}_i) = \alpha_0 + \alpha_1 \text{Edad}_i + \alpha_2 \text{Edad}_i^2 + \alpha_3 \text{Tamaño de la empresa}_i + \alpha_4 \text{Trabajo informal}_i \\ + \alpha_5 \text{Independiente}_i + \alpha_6 \text{Oficio}_i + \alpha_7 \text{Nivel de educacion}_i + \alpha_8 \text{Horas trabajadas}_i \\ + \alpha_9 \text{Experiencia}_i + \alpha_{10} \text{Estrato} + \alpha_{11} \text{Jefe} + e_i \quad (4.2.1.1)\end{aligned}$$

donde e_i corresponde al término de error.

$$\ln(\text{Mujer}_i) = \gamma_0 + \gamma_1 \text{Edad}_i + \gamma_2 \text{Edad}_i^2 + \gamma_3 \text{Tamaño de la empresa}_i + \gamma_4 \text{Trabajo informal}_i + \gamma_5 \text{Independiente}_i + \gamma_6 \text{Oficio}_i + \gamma_7 \text{Nivel de educacion}_i + \gamma_8 \text{Horas trabajadas}_i + \gamma_9 \text{Experiencia}_i + \gamma_{10} \text{Estrato} + \gamma_{11} \text{Jefe} + v_i \quad (4.2.1.2)$$

donde v_i corresponde al término de error.

Etapa 2 (FWL): Se utilizan los residuales las regresiones auxiliares para estimar la relación específica entre *ln(Salario por hora)*$_i$ y *Mujer*$_i$. De esta forma:

$$\hat{e_i} = \ln(\text{Salario por hora}_i) - \hat{\alpha_0} - \hat{\alpha_1} \text{Edad}_i - \hat{\alpha_2} \text{Edad}_i^2 - \hat{\alpha_3} \text{Tamaño de la empresa}_i - \hat{\alpha_4} \text{Trabajo informal}_i - \hat{\alpha_5} \text{Independiente}_i - \hat{\alpha_6} \text{Oficio}_i - \hat{\alpha_7} \text{Nivel de educacion}_i - \hat{\alpha_8} \text{Horas trabajadas}_i - \hat{\alpha_9} \text{Experiencia}_i - \hat{\alpha_{10}} \text{Estrato}_i - \hat{\alpha_{11}} \text{Jefe}_i \quad (4.2.1.3)$$

$$\hat{v_i} = \ln(\text{Mujer}_i) - \hat{\gamma_0} - \hat{\gamma_1} \text{Edad}_i - \hat{\gamma_2} \text{Edad}_i^2 - \hat{\gamma_3} \text{Tamaño de la empresa}_i - \hat{\gamma_4} \text{Trabajo informal}_i - \hat{\gamma_5} \text{Independiente}_i - \hat{\gamma_6} \text{Oficio}_i - \hat{\gamma_7} \text{Nivel de educacion}_i - \hat{\gamma_8} \text{Horas trabajadas}_i - \hat{\gamma_9} \text{Experiencia}_i - \hat{\gamma_{10}} \text{Estrato}_i - \hat{\gamma_{11}} \text{Jefe}_i \quad (4.2.1.4)$$

Con base en estos resultados, se obtiene el coeficiente de interés θ_1 :

$$\hat{e_i} = \theta_0 + \theta_1 \hat{v_i} \quad (4.2.1.5)$$

Finalmente, sabemos por cuenta del teorema FWL que:

$$\hat{\theta_1} = \hat{\beta_1} \quad (4.2.1.6)$$

4.3. Resultados

En la tabla 4 se muestran los resultados de la estimación de la brecha salarial por género incondicional (ecuación 4.1) en la primera columna y condicional (ecuación 4.2) en la segunda columna. En el primer caso, la estimación se llevó a cabo sin incluir controles y por medio de MCO. En el segundo caso, se estimó con controles y empleando MCO bajo el teorema FWL.

En cuanto a la aproximación sin controles, el coeficiente asociado a la variable de interés es -0.083 y es significativo a un nivel de confianza del 99 %. Este valor indica que en promedio, ser mujer se asocia con un menor salario por hora en 8.3 % en comparación con los hombres. No obstante, como se mencionó en la sección 4.1, esta estimación esta sesgada, dado que hay problemas de variable omitida.

En cuanto a la aproximación con controles, el coeficiente de interés estimado fue mayor en magnitud (es decir, en términos absolutos), y también significativo al 99 %. Especificamente, el valor de -0.122 indica que en promedio ser mujer se asocia con un menor salario por hora en 12.2 % en comparación con los hombres. Este último resultado recoge de una forma mas precisa el valor de la brecha de género, una vez que se está controlando por un conjunto de varibles, que de acuerdo con la literatura económica, son relevantes en la determinación de salario. Es decir, en esta aproximación se redujo el sesgo de la primera estimación, aunque se reconoce que aún existen otros factores inobservables que impiden eliminarlo completamente. En particular, las características inhatas de cada uno de los individuos podrían influir en el salario y por disponibilidad de datos en la GEIH no se están incluyendo. Lo mismo sucede por ejemplo con la motivación personal de los individuos.

Con esta advertencia, es posible concluir que si existe una brecha salarial de género desfavorable por las mujeres, pero que se debe ser cuidadoso con la interpretación de su magnitud. Finalmente, en cuanto al poder explicativo del modelo, la aproximación con controles tiene un mejor desempeño de acuerdo con el R^2 ajustado, aunque se mantiene en niveles cercanos a cero.

Tabla 5: Brecha salarial por género

	Variable dependiente: Ln(Salario por hora)	
	Sin controles	Con controles
	(1)	(2)
Female	-0.083^{***}	-0.122^{***}
	(0.013)	(0.010)
Observaciones	16,540	16,540
R^2 Ajustado	0.003	0.009
Notas:		*p<0.1; **p<0.05; ***p<0.01
		Errores estándar en paréntesis

Adicional a los resultados previos, se llevó a cabo la estimación de la especificación con controles empleando el teorema FWL para 1000 repeticiones de la muestra a través de Boopstrap. Los resultados de este ejercicio muestran que la magnitud del estimador de la brecha salarial por género de la tabla 5 de -0.122 se asemeja a la obtenida mediante Boopstrap. En particular, en la tabla 6 se observa que el sesgo de la estimación con la muestra original es bajo y de -0.0004. Por su parte, el error estándar del coeficiente de interés se elevó levemente desde 0.010 hasta 0.015, lo que indica que la medida de resampleo a través de boopstrap encontró una mayor variabilidad del coeficiente de interés. No obstante, a pesar de este aumento, el coeficiente identificado continúa siendo significativo a un nivel de confianza del 99 %. Lo anterior se debe a que, el intervalo de confianza a dicho nivel (-0.158,-0.081), se mantiene ampliamente lejos de cero.

Tabla 6: Brecha salarial por género: Boopstrap con controles

Variable	**Sesgo**	**Desviación Estándar**
Mujer	-0.0004	0.015

4.4. Perfil edad - salario por género Adicional a la brecha salarial encontrada por género en la sección anterior, es de interés conocer si el perfil de edad-ingreso también difiere dependiendo del género. Lo anterior, una vez que permite contrastar las edades implícitas en las que se considera que cada trabajador alcanza su mayor rendimiento. Esto es importante porque, si el pico del salario se alcanza a una menor edad para parte de las mujeres comparativamente a los hombres, en los años laborales posteriores la brecha del salario aumentará. Con el propósito de calcular esta relación, se estimará la siguiente regresión:

$$\ln(\text{Salario por hora}_{ig}) = \alpha_0 + \alpha_1 \text{Edad}_{ig} + \alpha_2 \text{Edad}_{ig}^2 + e_{ig} \tag{4.4.1}$$

donde *i* hace referencia a cada individuo, *g* indica su género y *e* representa el término de error.

Tabla 7: Perfil salario - edad por género

	Variable dependiente: Ln(Salario por hora)	
	Mujeres	Hombres
	(1)	(2)
Edad	0.057^{***}	0.059^{***}
	(0.004)	(0.004)
Edad2	-0.001^{***}	-0.001^{***}
	(0.0001)	(0.00004)
Observations	7,775	8,765
Adjusted R^2	0.025	0.029
Note:		*p<0.1; **p<0.05; ***p<0.01
		Errores estándar en paréntesis

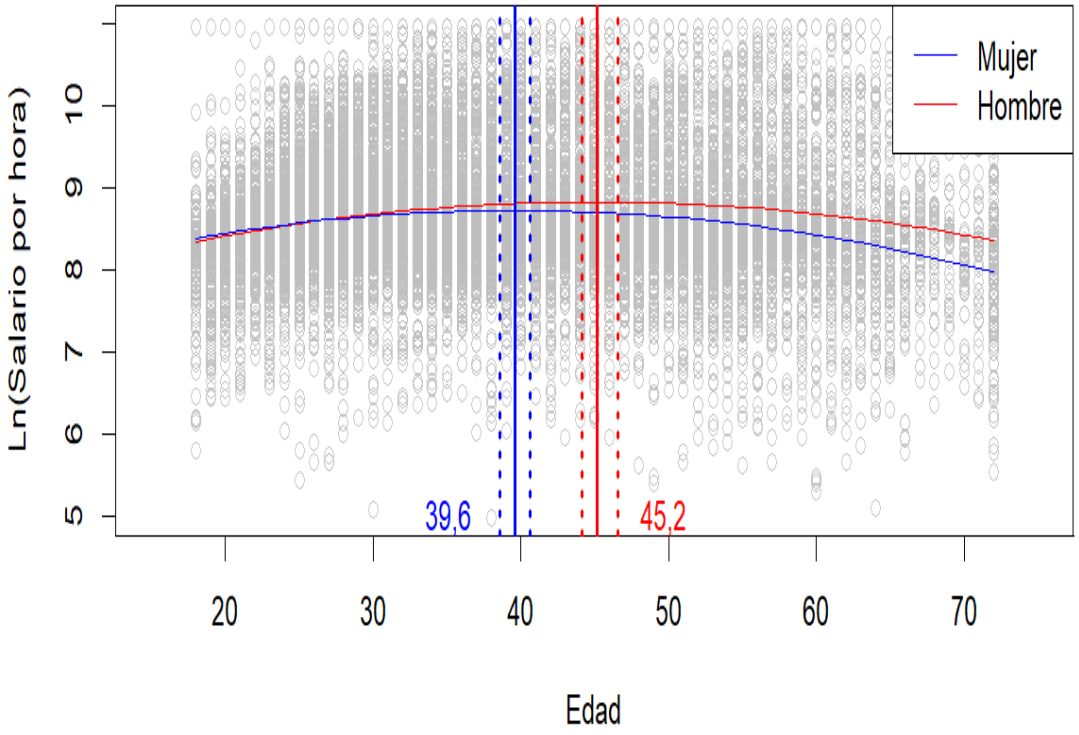
Los resultados de la tabla 7 indican para el primer coeficiente de ambas estimaciones que una mayor edad se relaciona con un mayor salario, asociado por ejemplo a un incremento en la experiencia. Por su parte, el segundo coeficiente refleja que la productividad del trabajador tiene rendimientos marginales decrecientes con la edad. Todos los coeficientes son estadísticamente significativos a un nivel de confianza del 99 %. Para facilitar la comparación entre los resultados de hombres y mujeres del perfil salario - edad e identificar las edades en las que cada género alcanza su mayor nivel de ingresos laborales, se calcula:

$$\text{Edad de máximo salario} = -\frac{\beta_1}{2\beta_2}$$

De acuerdo con esta estimación, las mujeres alcanzas su nivel máximo de salario por hora a los 39.6 años, mientras los hombres lo logran a los 45.2. Sin embargo, para conocer si estos resultados son diferentes estadísticamente, se debe conocer el intervalo de confianza de cada estimador. Para lograr lo anterior, se emplea la técnica de boopstrap con 1000 repeticiones, que recalcula dicha métrica para cada género. De acuerdo con los resultados de la figura 7, ambos niveles son estadísticamente diferentes a un nivel de confianza del 95 %.

Lo anterior se sabe debido a que los intervalos de confianza, representados con las líneas punteadas, no se superponen. Así, es posible concluir que las mujeres en Bogotá alcanzan su salario máximo de remuneración por hora a una edad más temprana que los hombres. Como se mencionó previamente, esto puede generar problemas, una vez que después de los 39.6 años, en promedio se amplía la brecha salarial. Sin embargo, este resultado se debe continuar analizando con cuidado, una vez que la estimación no contempló la inclusión de otras variables de control importantes, tal como se mencionó en la sección 3, puede tener problemas de sesgos asociados a factores no observables como la motivación y capacidad innata, como se resaltó a lo largo de la sección 4.2.

Figura 7: Perfil salario - edad por género



5. Prediciendo los ingresos

Considerando que el objetivo del trabajo es la predicción acertada del ingreso, pasamos a evaluar la capacidad de este para predecir correctamente este valor. Dado esto, se observa que los modelos con menor error de predicción fuera de muestra son los modelos 3 y 6, en estos modelos el número de predicciones que utilizamos es uno de 65 predictores y el modelo 6 explorando distintas combinaciones de no-linearidades. Estos modelos fueron ajustados tomando una proporción de los datos como set de entrañamientos, con un 70 % de la muestra, y el porcentaje restante se utilizó para la realización de pruebas, representando así un 30 % de los datos para la evaluación fuera de muestra.

En este punto es importante resaltar que el modelo con mayor poder de predicción fuera de muestra fue el modelo con mayor número distinto de variables añadidos en los puntos anteriores. Al aumentar la complejidad, incluyendo no linearidades con los modelos con pocas variables, no tiende a tener un mayor poder predictivo fuera de muestra, que es lo más relevante. Las fórmulas que representan estos modelos se pueden observar en la sección de Anexos.

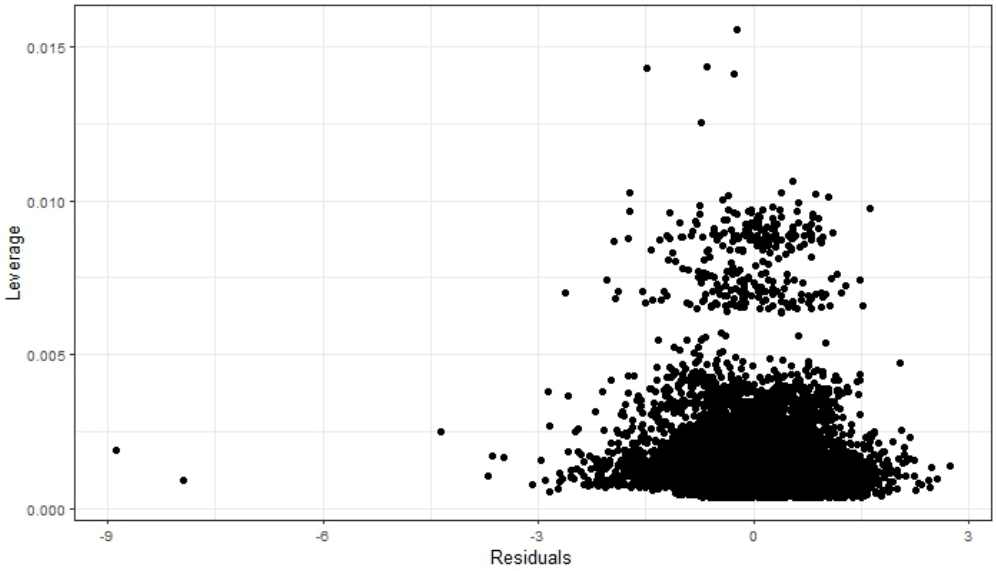
Nuestro *modelo 3*, que incluye variables como la experiencia del trabajador y variables categóricas (entre otras variables previamente winsorizadas), obtuvo el mejor desempeño de predicción fuera de muestra. Esto se debe a que estas variables adicionales aportan mayor información sobre los ingresos de los individuos. En la tabla 8 se puede observar los errores de predicción para cada uno de los modelos:

Tabla 8: RMSE Para distintos modelos probados

Modelo	RMSE	Número de Predictores
M1	0.79	2
M2	0.80	1
M3	0.79	3
M3.1	0.58	99
M4	0.80	17
M5	0.77	65
M6	0.76	73
M7	0.89	80
M8	0.80	82

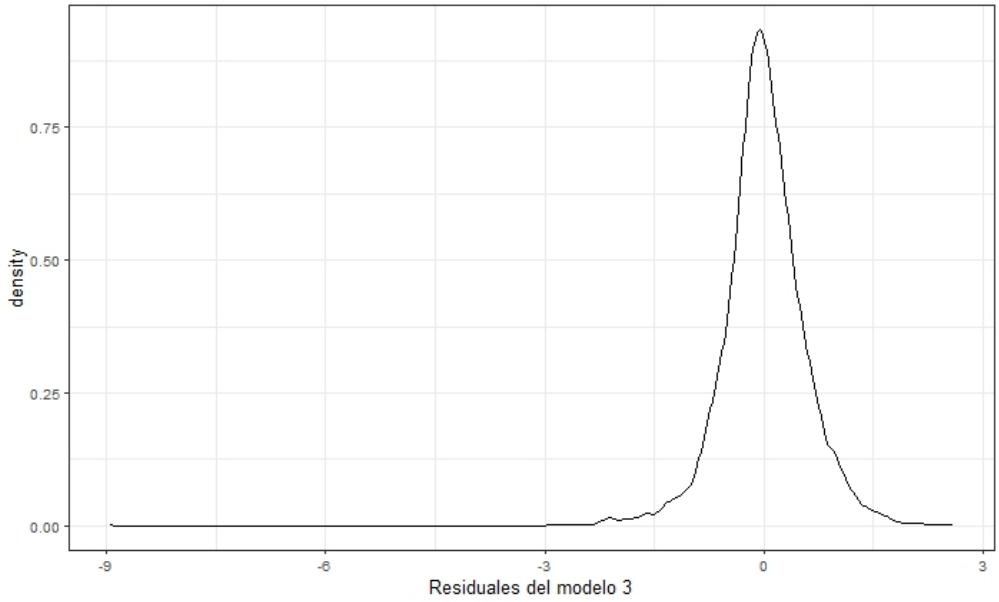
A continuación, y tomando el modelo tres con menor error de predicción, se procesó a analizar el error de estos datos respecto al leverage:

Figura 8: Leverage de los residuales de modelo con menor error de prediccion



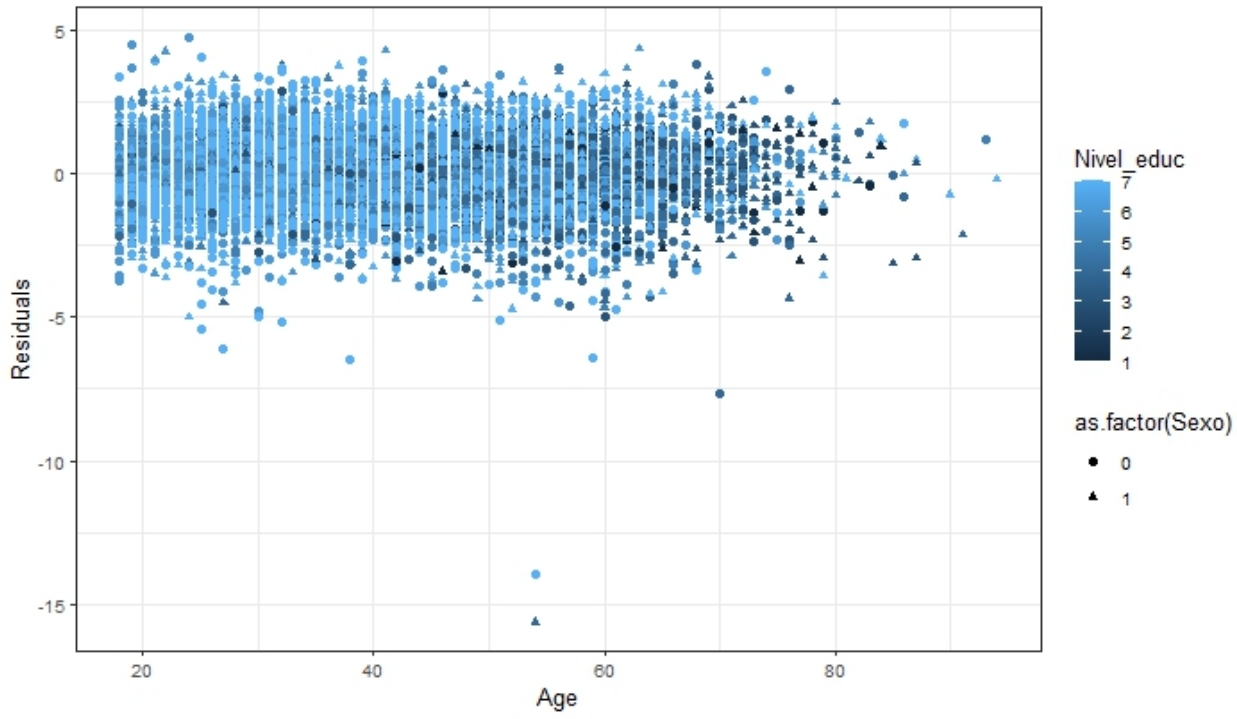
El gráfico de leverage en contra de los residuales en la figura 9 revela que, aunque la mayoría de las observaciones tienen residuales cercanos a cero y leverage bajo, hay un conjunto disperso de puntos con leverage alto, lo que indica que tienen un gran impacto en la estimación del modelo. Estos puntos influyentes podrían distorsionar los resultados si no son manejados adecuadamente, ya que un alto leverage sugiere que esas observaciones están alejadas del centro de los datos y podrían impactar sobre los parámetros del modelo de regresión.

Figura 9: Densidad de errores de predicción de modelo con menor error de predicción



Ahora, cuando revisamos la distribución de los errores, como se puede observar en la Figura 10, el gráfico de densidad de los errores de predicción del modelo 3 muestra que la mayoría de los residuales están concentrados cerca de cero, lo que indica un buen ajuste general del modelo y una alta precisión, sin subestimar o sobreestimar los valores observados.

Figura 10:



En el gráfico 10 podemos observar por medio de los errores studentizados, los residuos ajustados por la variación en la predicción, divididos por categorías de edad y nivel de educación, nos permite ver algunos resultados que pueden ser interesantes. Particularmente podemos observar a una persona que tiene más 45 años y un nivel alto de educación con un residuo studentizado menor a -3, al igual que múltiples personas por un nivel menor a -3, lo que parece sospechoso y valdría la pena que la DIAN analizara estos individuos, ya que dado que parece que el modelo sobrestimara el ingreso y podría darse el caso de que individuo reporte menores ingresos. Sin embargo, es necesario reconocer que el modelo puede incluir fallos en la predicción.

6. LOOCV

Por la estimación de LOOCV, podemos obtener un menor resultado para el RMSE en el modelo 6, como se observa en la tabla 9. La estimación por este método puede realizar una mejor estimación para el RMSE asociado a este modelo y efectivamente podría haber estimaciones con una alta influencia por datos específicos de la muestra u outliers que pueden afectar nuestra predicción. Este cambio en el RMSE puede deberse a que la selección inicial de la muestra para el modelo pudo no haber sido la óptima.

Tabla 9: RMSE para LOOCV y para una estimación de entrenamiento

Modelo	RMSE	LOOCV
M6	0.76	0.53
M3.1	0.58	0.52

Referencias

- Becker, G. (1964). Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education. Chicago: University of Chicago Press.
- Borjas, G. (2015). Labor Economics (7th ed.). New York: McGraw-Hill Education.
- Card, D. (1999). The Causal Effect of Education on Earnings. In O. Ashenfelter D. Card (Eds.), Handbook of Labor Economics, Vol. 3, Amsterdam: Elsevier Science.
- Delgadillo, D., Tavera, D. (2022). La evasión tributaria del impuesto de renta en Colombia. Caso 2020-2022, Universidad Libre. https://repository.unilibre.edu.co/bitstream/handle/10901/27665/Final%20-%20GT%20-%20Delgadillo%20-%20Tabera.pdf?sequence=3&isAllowed=y
- DIAN. (s.f.). El sistema tributario colombiano: Impacto sobre la eficiencia y la competitividad. https://www.dian.gov.co/dian/cifras/EstudiosExternos/Tributacion_y_competitividad.pdf
- Escobar, J. (2024). Dane dio a conocer cifra de informalidad en Colombia: hay más trabajadores que estarían en riesgo de no pensionarse. Infobae. https://www.infobae.com/colombia/2024/08/12/dane-dio-a-conocer-cifra-de-informalidad-en-colombia-cada-vez-son-mas-los-trabajadores-//en-riesgo-de-perder-la-pension/
- Lemieux, T. (2006). Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill? American Economic Review, 96(3), 461-498.
- Mincer, J. (1974). Schooling, Experience, and Earnings. New York: Columbia University Press for the National Bureau of Economic Research.
- Oi, W. Y., & Idson, T. L. (1999). Firm size and wages. In O. Ashenfelter & D. Card (Eds.), Handbook of labor economics (Vol. 3, pp. 2165-2214). Elsevier.
- Perry, G., Maloney, W., Arias, O., Fajnzylber, P., Mason, A., & Saavedra-Chanduvi, J. (2007). Informality: Exit and exclusion. World Bank.
- Heckman, J., Lochner, L., Todd, P. (2006). Earnings Functions, Rates of Return, and Treatment Effects: The Mincer Equation and Beyond. In E. A. Hanushek F. Welch (Eds.), Handbook of the Economics of Education, Vol. 1, Amsterdam: North-Holland.
- Sabogal, A. (2012). Brecha salarial entre hombres y mujeres y ciclo económico en Colombia. Coyuntura Económica: Investigación Económica y Social, 42(1), 53-91

7. Anexos

7.1. Modelos utilizados en la evalucaion de prediccion

Modelo 1:

$$m_1 : \log(\text{Ingreso por Hora}) = \beta_0 + \beta_1 \cdot \text{Edad} + \beta_2 \cdot \text{Edad}^2 + \varepsilon$$

Modelo 2:

$$m_2 : \log(\text{Ingreso por Hora}) = \beta_0 + \beta_1 \cdot \text{Sexo} + \varepsilon$$

Modelo 3:

$$m_3 : \log(\text{Ingreso por Hora}) = \beta_0 + \beta_1 \cdot \text{Sexo} + \beta_2 \cdot \text{Edad} + \beta_3 \cdot \text{Edad}^2 + \varepsilon$$

Modelo 3.1:

$$\begin{aligned} m_{3.1} : \log(\text{Ingreso por Hora}) = \beta_0 &+ \beta_1 \cdot \text{Edad} + \beta_2 \cdot \text{Edad}^2 + \beta_3 \cdot \text{Mujer} \\ &+ \beta_4 \cdot \text{Estrato factor} + \beta_5 \cdot \text{dummy\_jefe} + \beta_6 \cdot \text{edu\_factor} \\ &+ \beta_7 \cdot \text{Trabajo\_informal} + \beta_8 \cdot \text{Independiente} + \beta_9 \cdot \text{Horas\_trabajadas\_win} \\ &+ \beta_{10} \cdot \text{Experiencia\_win} + \varepsilon \end{aligned}$$

Modelo 4:

$$m_4 : \log(\text{Ingreso por Hora}) = \beta_0 + \left(\sum_{i=1}^{8} \beta_i \cdot \text{Edad}^i \right) \cdot \text{Sexo} + \varepsilon$$

Modelo 5:

$$\begin{aligned} m_5 : \log(\text{Ingreso por Hora}) = \beta_0 &+ \left(\sum_{i=1}^{8} \beta_i \cdot \text{Edad}^i \right) : \left(\sum_{j=1}^{8} \beta_j \cdot \text{Experiencia}^j \right) \\ &+ \beta_9 \cdot \text{Sexo} + \varepsilon \end{aligned}$$

Modelo 6:

$$\begin{aligned} m_6 : \log(\text{Ingreso por Hora}) = \beta_0 &+ \left(\sum_{i=1}^{8} \beta_i \cdot \text{Edad}^i \right) : \left(\sum_{j=1}^{8} \beta_j \cdot \text{Experiencia}^j \right) \\ &+ \left(\sum_{k=1}^{8} \beta_k \cdot \text{Experiencia}^k \right) + \beta_9 \cdot \text{Nivel\_educ} + \varepsilon \end{aligned}$$

Modelo 7:

$$\begin{aligned} m_7 : \log(\text{Ingreso por Hora}) = \beta_0 &+ \left(\sum_{i=1}^{8} \beta_i \cdot \text{Edad}^i \right) : \left(\sum_{j=1}^{8} \beta_j \cdot \text{Experiencia}^j \right) \\ &+ \left(\sum_{k=1}^{8} \beta_k \cdot \text{Experiencia}^k \right) + \left(\sum_{l=1}^{8} \beta_l \cdot \text{Experiencia}^l \right) \\ &: \left(\sum_{m=1}^{8} \beta_m \cdot \text{Nivel\_educ}^m \right) + \varepsilon \end{aligned}$$

Modelo 8:

$$\begin{aligned} m_8 : \log(\text{Ingreso por Hora}) = \beta_0 &+ \left(\sum_{i=1}^{8} \beta_i \cdot \text{Edad}^i \right) : \left(\sum_{j=1}^{8} \beta_j \cdot \text{Experiencia}^j \right) \\ &+ \left(\sum_{k=1}^{8} \beta_k \cdot \text{Experiencia}^k \right) + \left(\sum_{l=1}^{8} \beta_l \cdot \text{Experiencia}^l \right) \\ &: \left(\sum_{m=1}^{8} \beta_m \cdot \text{Nivel\_educ}^m \right) + \beta_9 \cdot \text{Independiente} \\ &+ \beta_{10} \cdot \text{Trabajo\_informal} + \varepsilon \end{aligned}$$