Problem set 1
Prediciendo el ingreso

María Camila Arias Mario Velásquez Martín Velásquez Daniela Vlasak

I. Introducción

La predicción del salario real de los ciudadanos es de gran importancia para los gobiernos. Por un lado, esto permite que se pueda hacer mejor vigilancia del recaudo fiscal, y por el otro, contribuye a una mejor focalización de los programas sociales. En Colombia esto es relevante porque la evasión del impuesto de renta por parte de personas naturales alcanzó un equivalente al 3,6 % del PIB en 2021 (DIAN, 2022), por lo que no sólo es importante detectar a los potenciales evasores, sino además poder asignar los programas sociales a quiénes verdaderamente lo necesitan (y no a quienes reportan menos ingresos de los que realmente perciben). En este sentido, este trabajo se enfoca en desarrollar modelos de predicción para el salario real por hora en Colombia.

Para esto se utilizarán los datos de la Gran Encuesta Integrado de Hogares (GEIH) 2018, los cuales brindan una riqueza de observaciones representativas a nivel de las 13 áreas metropolitanas y capitales departamentales del país. Dentro de las principales variables a considerar se encuentran características laborales, la edad y el sexo. Esta última variable es especialmente relevante, teniendo en cuenta que múltiples autores han encontrado una brecha salarial de género en Colombia (Galvis-Aponte, 2010; Badel and Peña, 2010; Sabogal, 2012; Barreto Nieto et al., 2020). Ahora bien, es importante mencionar que aunque la investigación económica se ha centrado en estudiar las brechas salariales de género, no se encontraron trabajos empíricos que intentaran mejorar la predicción del salario como tal en Colombia.

Particularmene, los modelos de este trabajo se enfocan en predecir el salario para personas mayores de edad y ocupadas; teniendo como principales variable predictoras a la edad y el sexo. Dentro de los modelos analizados se encuentra que la edad es un buen predictor del ingreso, y que no mantiene una relación monótona creciente. Alrededor de los 50 años se encuentra un pico de máximo salario, y a partir de ahí el salario decrece. Por otro lado, el sexo también es un buen predictor del salario horario, y los modelos sugieren que existe una brecha salarial de género en Colombia de alrededor del 9,47 %, y que el pico de máximo salario por edad difiere entre hombres y mujeres, siendo estas útlimas quiénes alcanzan más jóvenes este punto.

El trabajo está estructurado de la siguiente manera. La segunda sección describe el proceso de obtención de los datos de la GEIH a través de *webscraping*, la depuración e imputación de los datos y se presenta unas estadísticas descriptivas alrededor de las variables de interés. El tercer capítulo analiza un modelo predictor basado completamente en la edad. El cuarto capítulo analiza un segundo modelo predictor que se centra en el sexo como variable predictora y después se refina usando otros predictores laborales y la edad. Finalmente, el quinto capítulo desarrolla modelos predictores que incorporen las variables predictoras estudiadas en los capítulos anteriores y permita mayor complejidad a partir de la interacción entre predictores, de manera que se llegue a minimizar el error cuadrático medio (ECM).

II. Datos

2.1. Descripción de los datos

Los datos empleados en este trabajo provienen de la Gran Encuesta Integrada de Hogares (GEIH), realizada en 2018. Este estudio usa la información correspondiente al "Reporte de Medición de Pobreza Monetaria y Desigualdad". Esta información ha sido usada desde 2009 para estimar indicadores directos de la pobreza, como el Índice de Necesidades Básicas Insatisfechas (NBI) y el nuevo Índice de Pobreza Multidimensional (IPM); e indicadores indirectos como la medición de la pobreza monetaria a partir de la línea de la pobreza (DANE, 2019). En particular, sólo usaremos las observaciones a nivel de persona para la ciudad de Bogotá. Los datos que proporciona la GEIH con ideales para explorar las variables que pueden predecir el ingreso, como lo es el nivel de educación alcanzado y situación laboral, entre otros.

2.2. Proceso de adquisición de los datos

Para acceder a los datos se realizó *webscraping* sobre el repositorio del Profesor Ignacio Sarmiento (`https://ignaciomsarmiento.github.io/GEIH2018_sample/`). En este se encontraba la información de la GEIH 2018 dividida en 10 bloques de información. Cada bloque tenía su link correspondiente. Sin embargo, no bastaba con usar el comando `html_table` porque cada link llevaba a una tabla que no estaba en formato HTML. Por lo anterior, para poder acceder a la información se inspeccionó un elemento de la tabla. En la ventana de inspección se revisó la pestaña "Network" y se actualizó la página para poder ver los detalles actualizados. De esta lista de detalles, nos dimos cuenta que, en particular, el elemento con el nombre "geih_page_1.html" era el que más tiempo había tomado en cargar, y como la tabla en la que estamos interesados es la última en cargar en la página web, supusimos que ese era el elemento que nos interesaba. Después, revisamos el URL que llamaba a ese elemento y concluimos que esta era la dirección que debíamos llamar en nuestro código de R para hacer *webscraping*.

Para obtener cada bloque de información que estaba contenido en una página web diferente hicimos un loop (de 1 a 10), modificando únicamente el número x que aparece al final de cada página "https://ignaciomsarmiento.github.io/GEIH2018_sample/pages/geih_page_x.html". Adentro del loop usamos el comando `html_table` para leer cada tabla, y las fuimos almacenando en una lista. Finalmente, usamos el comando `bind_rows` para transformarlo en un data frame. En total, la base reconstruida contiene 32.177 observaciones y 178 variables.

2.3. Proceso de limpieza

De la base original se extrajeron todas las observaciones que tuvieran menos de 18 años (*age<18*) y que estuvieran desempleados (*ocu==0*), reduciendo la muestra a 16.542 observaciones. Para optimizar la velocidad de procesamiento se dejaron sólo las variables que necesitábamos para este problem set, es decir, nos quedamos con 12[1] de las 178 variables que traía la base original.

Ahora bien, al revisar los valores, nos dimos cuenta que para el 40 % de los salarios no se tenía información y una observación no tenía información sobre el máximo nivel educativo alcanzado. Para este último caso optamos por eliminar la observación porque sólo representaba el 0,006 % de la muestra.

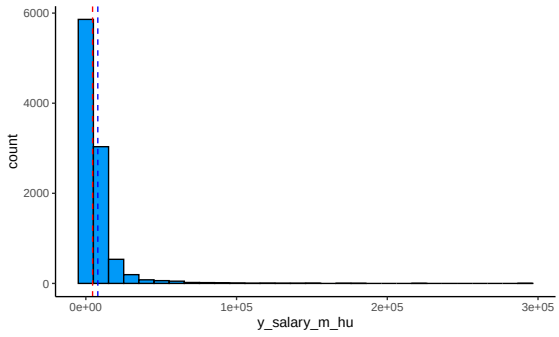
[1] Estas serán listadas en la sección de descriptivas.

Para imputar las faltantes de los salarios, primero revisamos su distribución agregada. En la distribución agregada de la Figura 1a se evidencia que los salarios reales por hora tienen una cola hacia la derecha, de ahí que su mediana (línea roja) esté corrida hacia la izquierda del promedio (línea azul). Como eliminar todas las observaciones faltantes no es una opción porque representan el 40 % de la muestra, buscamos en la literatura la variable más apta para realizar la imputación. De acuerdo con el DANE (2022a), en Colombia existe una alta correlación entre el estrato socioeconómico y el nivel de ingresos de la persona. La Figura 1b sugiere que para nuestra muestra esta relación se cumple, a mayor estrato se observa un salario horario promedio mayor. Asimismo, la Tabla 1 deja ver que las variables a usar en los modelos de predicción (edad y sexo) son prácticamente independientes al estrato socioeconómico. Esto se ve al analizar la magnitud de sus coeficientes, ambos muy cercanos a cero. Para pasar al siguiente estrato una persona tendría que cumplir 100 años, y sólo por ser hombre uno disminuiría 0.08 estratos; en ambos casos, son cambios irrisorios. Esto nos permite concluir que la edad y el sexo son independientes del estrato, por lo que si imputamos el salario por medio del estrato se espera que no afectemos los modelos de predicción.

Por lo anterior, decidimos imputar los valores de los salarios reales por hora faltantes a través del promedio condicional al estrato socioeconómico, de manera que se pudiera recuperar la mayor cantidad de información. Como primer intento consideramos hacer la imputación condicional a la edad de los individuo, pero nos dimos cuenta que no se capturarían muchas brechas importantes en ingresos que no están asociadas a la edad de las personas, consideramos que el estrato consigue agrupar muchas características subyacentes determinantes en los ingresos. Tras realizar la imputación, logramos completar la totalidad de los valores faltantes. En total, nuestra base limpia contiene 16.541 observaciones.

Figura 1: Distribución del salario horario.

(a) Muestra total



(b) Por estrato

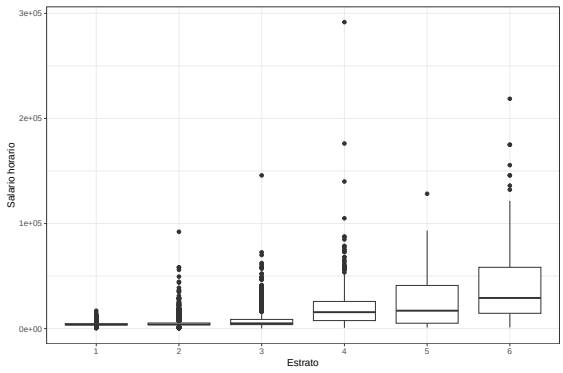


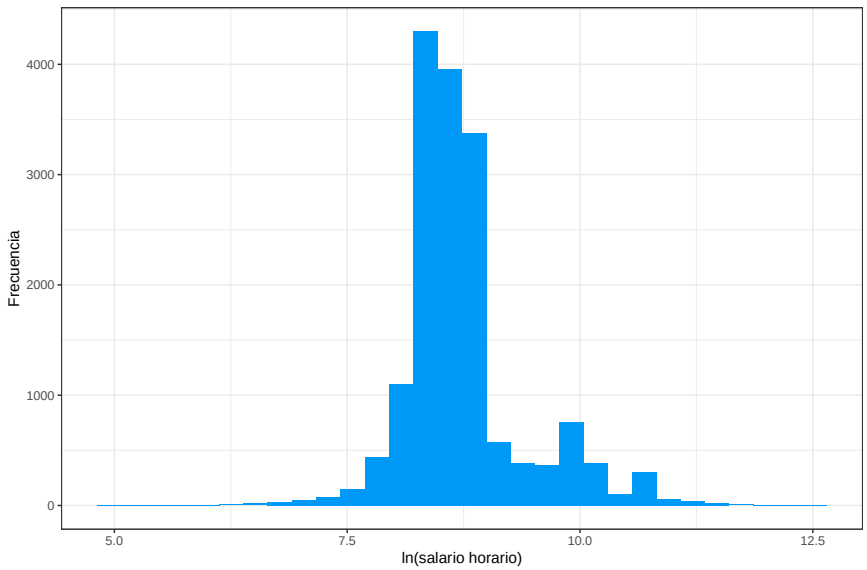
Tabla 1: Relación estrato vs edad y sexo

	Estrato socioeconómico
Edad	0.010
	(0.001)
Sexo (hombre=1)	−0.085
	(0.016)
Constante	2.194
	(0.025)
Observaciones	16,541
R^2	0.020
R^2 ajustado	0.020
ECM	1.001 (df = 16538)
F-Estadístico	169.676 (df = 2; 16538)

2.4. Estadísticas descriptivas

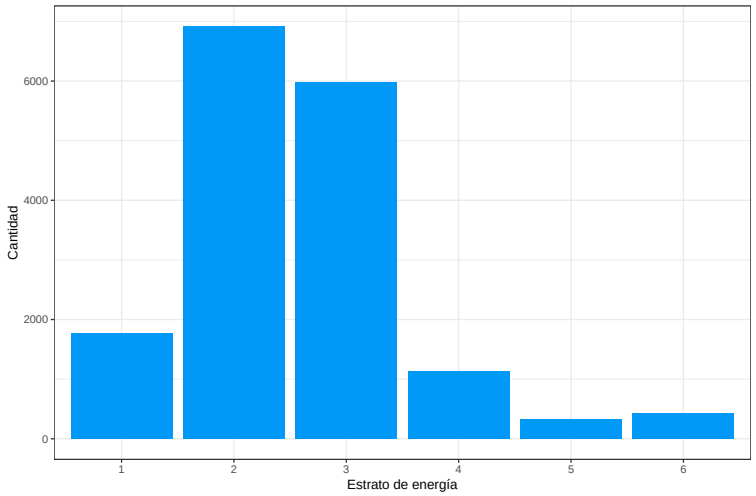
Tras la imputación de los datos, realizar un análisis exploratorio resulta crucial para comprender la distribución y características más importantes de la población de estudio. Presentar estadísticas descriptivas sobre el conjunto de datos permite un mejor entendimiento de los mismos, facilitando la interpretación adecuada de los resultados del análisis y la identificación de posibles limitaciones. Este enfoque contribuye a la pertinencia de las conclusiones, al reconocer las restricciones del conjunto de datos y su aplicabilidad en otros contextos, garantizando la calidad y fiabilidad de los resultados del estudio.

Figura 2: Logaritmo salario por hora



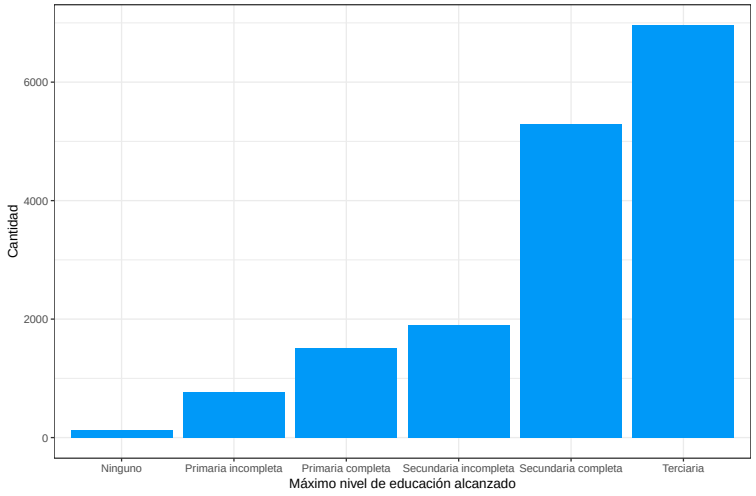
La Figura 2 presenta la distribución de la variable de resultado para el estudio, que es el logaritmo natural del salario real por hora de trabajadores colombianos. Esta transformación en logaritmo resulta conveniente por dos razones. La primera es que los datos se vuelven menos dispersos y la segunda es que la interpretación de los coeficientes de la regresión lineal se vuelve como diferencias en porcentaje. Es posible notar una importante concentración de los datos que cuentan con el logaritmo de su salario por hora entre 7.5 y 10, que equivalen a salarios reales de entre 2 mil y 22 mil pesos por hora. Ahora bien, la mediana se encuentra entre 8.25 y 8.5, siendo este el valor con mayor frecuencia. Estos salarios reales equivalen a entre 4 mil y 5 mil pesos por hora.

Figura 3: Estrato



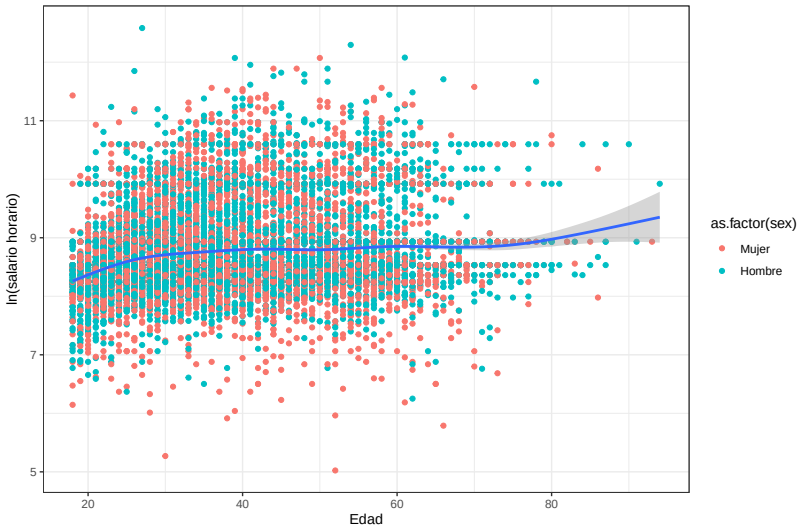
En términos socioeconómicos, la Figura 3 revela que la mayoría de observaciones son de estratos 1, 2 y 3. En este caso, la mediana es el estrato 2 y es posible apreciar que una fracción pequeña de la muestra pertenecen a los tres estratos más altos. Lo anterior deja ver que la mayoría de nuestra muestra se relaciona con una clase media-baja, lo que es consistente con la distribución de ingresos en el país.

Figura 4: Educación máxima



En contraste, la figura 4 deja ver que nuestra muestra tiene un número alto de observaciones que alcanzaron la educación terciaria. Esto se debe a que nuestra muestra es el resultado de filtrar la GEIH para las personas ocupadas, por lo que tiene sentido que haya una concentración de personas con el máximo nivel educativo. Entre más alto el nivel de educación alcanzado, más posibilidades de obtener un trabajo.

Figura 5: Distribución del salario por edad y sexo.



Ahora bien, en lo que respecta a las principales variables predictoras de este trabajo, se observa que, por lo general, las personas de mayor edad tienen un mayor salario real por hora. La Figura 5 permite notar que entre los 20 y 30 años hay un aumento del salario por hora a medida que aumenta la edad. No obstante, entre los 30 y los 70 el promedio del logaritmo del salario real por hora parece mantenerse estable. Finalmente, la gráfica sugiere que a partir de los 70 años el salario real por hora comienza a aumentar, al igual que la variabilidad de los datos y el error estándar. Esto último no tiene mucho sentido ya que hacia la vejez es que los trabajadores comienzan a retirarse del trabajo, por lo que el aumento promedio del salario puede estar asociado a observaciones atípicas y a un aumento de la varianza en este rango de edad, y no a que verdaderamente aumente el salario real por hora promedio. Esto tiene sentido al ver las pocas observaciones que hay después de los 70 años. Por otro lado, la misma figura no permite concluir diferencias entre los salarios de las mujeres y los hombres.

Finalmente, la Tabla 2 expone las estadísticas descriptivas de otras variables predictoras que son utilizadas en el modelo. La muestra tiene en promedio 39 años. De estas estadísticas también se comprueba que los más jóvenes de la muestra cuentan con 18 años. Por otro lado, se observa que el 47 % de la muestra son mujeres. Esta cifra es un poco baja teniendo en cuenta que según el Censo de 2018 (DANE, 2018), las mujeres representaban el 51,2 % de la población colombiana; de manera que habría menos participación laboral femenina de la correspondiente con la proporción poblacional. En cuanto a otras variables laborales, se puede decir que en promedio las personas de la muestra trabajan 47 horas a la semana (lo cual tiene sentido porque en 2018 en Colombia la jornada laboral permitía trabajar un máximo de 48 horas a la semana). Por otro lado, se resalta que un individuo reportó trabajar 130 horas a la semana. Lo anterior implica que, en promedio, cada día de la semana trabajó 18 horas. Probablemente estas sean observaciones cuya influencia deba ser

estudiada en el modelo. Finalmente, un 58,7 % de la muestra es empleado formal[2]; un 30,9 % es cuenta propia, es decir, es empleado independiente y, en promedio, llevan trabajando 63,8 meses en el trabajo actual.

Tabla 2: Descriptivas de las variables explicatorias.

Predictora	N	Promedio	Desviación	Min	Máx
Edad	16,541	39.436	13.483	18	94
Mujer	16,541	0.470	0.499	0	1
Horas semanales trabajadas	16,541	47.008	15.543	1	130
Es empleado formal	16,541	0.587	0.492	0	1
Si es cuenta propia	16,541	0.309	0.462	0	1
Duración en el trabajo	16,541	63.761	89.489	0	720

[2]Se considera trabajador formal al que paga seguridad social.

III. Perfil edad-ingreso

Las estadísticas descriptivas muestran que la relación salario real por hora y edad no es obvia ni estrictamente lineal. Por esta razón, surge la necesidad de plantear un modelo que capture de forma más adecuada esta relación para entender si hay una relación estadísticamente significativa entre la edad y el salario.

Para estimar adecuadamente esta relación, los artículos más relevantes en esta materia han propuesto una relación cuadrática entre el salario, la edad del trabajador o la experiencia. Es decir, al comienzo de la vida laboral los trabajadores aumentan sus ingresos, pero llega un punto en el que comienzan a decrecer. Según Mincer (1974), el modelo adecuado para estimar los ingresos de los trabajadores debe tener como variables independientes la experiencia del trabajador como variable lineal y al cuadrado. Así, consistente con los postulados y la experiencia académica en el tema, para estimar el perfil edad ingreso se estimará un modelo como el incluido en la Ecuación 1. La forma funcional propuesta permitirá capturar la evolución no lineal del salario de los individuos por hora frente a la edad que tienen.

$$\ln salario\_hora_i = \beta_1 + \beta_2 edad_i + \beta_3 edad_i^2 + \varepsilon_i \tag{1}$$

Al realizar la estimación del modelo, se obtienen los resultados presentados en la Tabla 3. Esta, consistente con la ecuación de Mincer y la literatura en la materia, permite ver que hay una tendencia lineal creciente en el salario a medida que aumenta la edad, pero que decrece con el tiempo.

Tabla 3: Regresión de salario contra edad

	Logaritmo del salario real por hora
Edad	0.03308
	(0.00219)
Edad^2	−0.00029
	(0.00003)
Constante	7.92465
	(0.04403)
Observaciones	16,541
R^2	0.03536
R^2 ajustado	0.03525
Error estándar residual	0.65483 (df = 16538)
Estadístico F	303.15305 (df = 2; 16538)

Asimismo, la Tabla 3 muestra que un aumento de 1 año en la edad de una persona implica, en promedio, un aumento de 3,3 % en su salario real por hora. Además, dado que el coeficiente de edad al cuadrado es negativo se tiene que los beneficios sobre el salario real por hora del aumento de la edad son decrecientes. En otras palabras, habrá una "edad pico" a partir de la cual las personas empezarán a ganar cada vez menos de salario real por hora.

Para ver con claridad la relación entre edad y salario, es posible tomar la primera derivada del modelo con respecto a la edad, Ecuación 2. Debido a que el logaritmo es una transformación monotónica el punto máximo se conserva. Por lo tanto, al igualarla a cero, cómo en la Ecuación 3 se logra llegar al punto de edad con máximo salario real por hora promedio.

$$\frac{\partial \log SalarioHora_i}{\partial Edad_i} = \beta_2 + 2\beta_3 Edad_i = 0{,}03308 + 2 * (-0{,}00029) * Edad_i \tag{2}$$

En ese orden de ideas, usando la Ecuación 3 es posible encontrar la edad en la que se alcanza el salario real por hora máximo y, por lo tanto, el umbral de edad en el que se espera que los ingresos de los individuos comiencen a bajar. El despeje realizado se presenta en la Ecuación 4 y permite ver que, de acuerdo con las estimaciones, se espera que los individuos alcancen un salario máximo en una edad cercana a los 57 años.

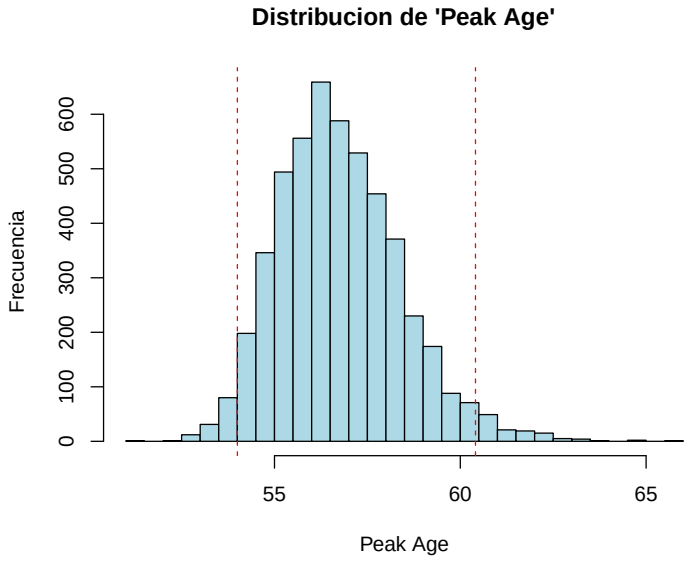
$$0 = 0{,}03308 - 0{,}00058 * Edad_i \tag{3}$$

$$0{,}00058 * Edad_i = 0{,}03308$$

$$Edad_i = \frac{0{,}03308}{0{,}00058} \approx 56{,}63 \tag{4}$$

Para complementar el anterior análisis y construir un intervalo de confianza sobre la edad a la que se estima que las personas alcanzan sus salarios máximos, se utilizó la metodología de *bootstrap*. Al realizar una iteración y estimar el parámetro con diferentes muestras, se obtiene que las personas alcanzan su salario máximo a los 56 años. Además, a partir de la distribución de betas estimados a partir de la iteración, se encuentra que, con un 95 % de confianza, el rango está entre los 54 años y los 61 años, como se evidencia en la Figura 6.

Figura 6: Distribución de la "edad pico" de salario por hora encontrada a través de la metodología *bootstrap*. Intervalos de confianza del 95 %.



IV. Ganancias por género

4.1. Estimación de la brecha salarial no condicional

Históricamente, una de las variables más significativas que ha influido en los salarios de las personas es su sexo[3]. Esta realidad plantea preocupaciones importantes para quienes diseñan políticas públicas y es un tema central en el análisis del mercado laboral. Con los datos disponibles, podemos construir un modelo para investigar si existen diferencias salariales entre sexo. En concreto, el modelo que proponemos es la Ecuación 5.

$$\log\left(SalarioHora_i\right) = \beta_1 + \beta_2 * Mujer_i + u_i \tag{5}$$

En la ecuación, la variable *Mujer* se refiere al sexo femenino, que está definido como 1 cuando la persona es de sexo femenino y 0 cuando es de sexo masculino. Por otra parte, la variable *SalarioHora* indica el salario real por hora del individuo. Ahora bien, antes de presentar los resultados de la estimación, el hecho de que el parámetro esté asociado al sexo de la persona en la ecuación no necesariamente implica que el sexo de las personas determinen el salario que estas tienen. Es decir, la relación no es necesariamente causal. Esto porque en el modelo de regresión simple planteado no necesariamente se está garantizando el supuesto de exogeneidad, por lo que puede haber otras variables omitidas como la cultura de roles de género en que crecieron las personas, que termina afectando diferenciadamente a las mujeres de los hombres, y también el salario que terminan ganando. Por esta razón, el parámetro puede resultar indicativo, mas no tiene implicaciones causales. Entendiendo esto, en la Tabla 4 se presentan los resultados de estimar el modelo con los datos que se tienen. En esta, el primer modelo se refiere al estimado con los datos imputados, mientras el segundo muestra los resultados cuando los datos no son imputados, sino que los faltantes no son usados.

Tabla 4: Estimación no condicional de la brecha salarial entre hombres y mujeres

	Logaritmo del salario real por hora	
	(Imputados)	(Sin imputar)
Mujer	−0.038	−0.045
	(0.010)	(0.015)
Constante	8.740	8.641
	(0.007)	(0.010)
Observaciones	16,541	9,641
R^2	0.001	0.001
R^2 Ajustado	0.001	0.001
ECM	0.666 (df = 16539)	0.721 (df = 9889)
Estadístico F	13.482 (df = 1; 16539)	9.592 (df = 1; 9889)

[3]En esta sección se va usar el término sexo dado que según la metodología actualizada del DANE se hace una distinción entre sexo y género y los datos del 2018 son referentes al sexo.

En ambos casos, parece existir una correlación negativa entre el hecho de ser mujer y el salario real por hora promedio. En promedio, el hecho de ser mujer se relaciona con un salario real por hora inferior en 3,8 % (para el modelo con datos imputados) o en 4,5 % (para el modelo con datos sin imputar) con respecto al de los hombres. Este valor está por debajo al encontrado en la literatura. Badel and Peña (2010) reportan una brecha salarial de género del 14 % para Colombia, usando la misma GEIH en 2006, mientras que el DANE (2022b) encuentra una brecha salarial del 6,3 % en 2021. Ahora bien, es importante resaltar que a diferencias de estos autores, nuestra primera estimación no controla por otras covariables como los años de educación, la edad y el tipo de oficio. Lo anterior resalta la importancia de considerar otro tipo de variables. La debilidad de este primer modelo se ve reflejada en el bajo R^2, lo que indica también que no es muy bueno prediciendo el salario real por hora, y que logra explicar muy poco de la variación de esta misma variable.

No obstante, la predicción de la brecha salarial de género aún puede mejorar. La tabla Tabla 5 presenta una prueba t de las diferencias entre características de mujeres y hombres. Teniendo en cuenta estas diferencias, y que la mayoría de variables presenta diferencias estadísticamente significativas entre hombres y mujeres, se realizará la estimación de la brecha salarial controlando por estas características. Es importante notar que, a pesar de que el nivel de educación máximo es categórica, una prueba t resulta conveniente en este caso debido a que su interpretación es lineal: entre más alto es el valor de la variable, mayor es el nivel educativo de las personas.

Tabla 5: Prueba t para características entre hombres y mujeres.

Variable	P-Valor
Salario real por hora	0.28
Edad	0.10
Estrado	0.00
Tipo de oficio	0.00
Trabajo formal	0.59
Nivel máximo de educación	0.00
Cuenta Propia	0.00
Horas semanales trabajadas	0.00

4.2. Estimación de la brecha salarial condicional

A partir del análisis de diferencia de medias presentado en la anterior sección, se propone un modelo de regresión como el presentado en la Ecuación 6, de manera que se analice si al incluir variables que controlen por características de los trabajadores y de los oficios, se disipa la brecha salarial encontrada en la Tabla 4.

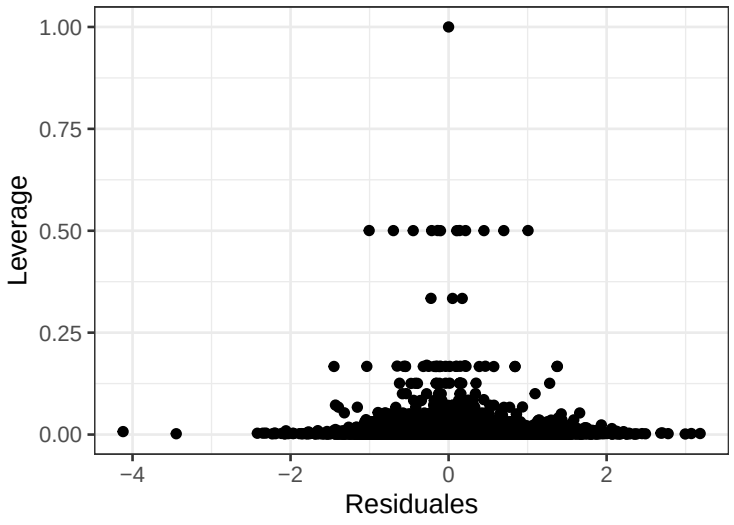
$$\begin{aligned} log(SalarioHora_i) =& \beta_0 + \beta_1 * Mujer_i + \beta_2 * Edad_i + \beta_3 * Edad_i^2 \\ &+ \beta_4 * Mujer_i * Edad_i + \beta_5 * Mujer_i * Edad_i^2 + \beta_6 * CuentaPropia_i \\ &+ \beta_7 * Formal_i + \beta_8 * HorasSemana_i \\ &+ \sum_{j=1}^{J} \gamma_j NivelEducativo_{i,j} + \sum_{f=1}^{F} \phi_f Oficio_{i,f} + u_i \end{aligned} \tag{6}$$

En la ecuación, las variables continuas o dummies tienen un único parámetro denotado por

β_i. Por su parte, las variables categóricas $NivelEducativo$ y $Oficio$ tienen su propio conjunto de parámetros asociados γ_j y ϕ_f, respectivamente.

Ahora bien, para la estimación del modelo se utilizarán cuatro formas diferentes. La primera, es a partir de regresar la ecuación presentada con todas las variables consideradas en la Ecuación 6. Los resultados de esta estimación se presentan en la primera columna de la Tabla 6. Posteriormente, se realizó un análisis del *leverage* de los datos; es decir, se analizó la posibilidad de que las estimaciones de los parámetros pudieran estar afectadas debido a la existencia de valores extremos. Los resultados de este análisis de presentan en la Figura 7. En esta, es posible ver que existen algunos valores extremos con mucha influencia sobre los parámetros estimados. Esto se ve porque entre más cercano es el *leverage* a 1, más pesa en las estimaciones. Por esta razón, se realizó una segunda estimación del modelo en el que no se utilizan las observaciones con un *leverage* tres veces mayor que la media. Los resultados de esta regresión se presentan en la segunda columna de la Tabla 6. Al comparar estos dos primeros modelos, no se observan grandes diferencias sobre los coeficientes de ninguna de las variables de interés. Contrario a lo esperado, al incluir las características laborales, no se redujo la brecha, sino que se aumentó. La brecha aumentó con respecto a la estimación no condicional presentada en la Tabla 4 en más de 2 puntos porcentuales, es decir, más cerca del valor que reporta la literatura (Badel and Peña, 2010; DANE, 2022b). En promedio, el hecho de ser mujer se correlaciona con un salario horario inferior al de los hombres en 9,39-9,47 %. Esto sugiere que para las mujeres empleadas de Colombia no se cumple que para cargos similares reciban un pago similar.

Figura 7: Análisis Leverage-Residuales



Estos resultados fueron respaldados al hacer una tercera estimación del modelo que estimara los efectos fijos por oficio y niveles de educación a partir del teorema de Frisch–Waugh–Lovell (FWL). De acuerdo con el teorema de FWL, los resultados obtenidos a partir de la primera estimación son equivalentes a realizar el proceso en dos etapas. En la primera etapa, se estiman las siguientes tres regresiones:

$$
\begin{aligned}
Mujer_i =& \alpha_0 + \alpha_1 * Edad_i + \alpha_2 * Edad_i^2 \\
&+ \alpha_3 * CuentaPropia_i + \alpha_4 * Formal_i + \alpha_5 * HorasSemana_i \\
&+ \sum_{j=1}^{J} \rho_j NivelEducativo_{i,j} + \sum_{f=1}^{F} \delta_f Oficio_{i,f} + \mu_i
\end{aligned} \tag{7}
$$

$$
\begin{aligned}
Mujer_i * Edad_i =& \alpha_0^{'} + \alpha_1^{'} * Edad_i + \alpha_2^{'} * Edad_i^2 \\
&+ \alpha_3^{'} * CuentaPropia_i + \alpha_4^{'} * Formal_i + \alpha_5^{'} * HorasSemana_i \\
&+ \sum_{j=1}^{J} \rho_j^{'} NivelEducativo_{i,j} + \sum_{f=1}^{F} \delta_f^{'} Oficio_{i,f} + \epsilon_i
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
Mujer_i * Edad_i^2 =& \alpha_0^{''} + \alpha_1^{''} * Edad_i + \alpha_2^{''} * Edad_i^2 \\
&+ \alpha_3^{''} * CuentaPropia_i + \alpha_4^{''} * Formal_i + \alpha_5^{''} * HorasSemana_i \\
&+ \sum_{j=1}^{J} \rho_j^{''} NivelEducativo_{i,j} + \sum_{f=1}^{F} \delta_f^{''} Oficio_{i,f} + \varepsilon_i
\end{aligned} \tag{9}
$$

Es decir, en la primera etapa es necesario regresar las tres variables de interés frente a las demás variables de control. Una vez hecho esto, la segunda etapa consiste en regresar la variable de resultado, que en este caso es el salario por hora, frente a los mismos controles. De esta forma, la segunda etapa es la siguiente:

$$
\begin{aligned}
\log(SalarioHora_i) =& \beta_0^{'} + \beta_1^{'} * Edad_i + \beta_2^{'} * Edad_i^2 \\
&+ \beta_3^{'} * CuentaPropia_i + \beta_4^{'} * Formal_i + \beta_5^{'} * HorasSemana_i \\
&+ \sum_{j=1}^{J} \gamma_j^{'} NivelEducativo_{i,j} + \sum_{f=1}^{F} \phi_f^{'} Oficio_{i,f} + \nu_i
\end{aligned} \tag{10}
$$

Una vez realizadas ambas etapas, es posible estimar una regresión de los residuales de la segunda especificación frente a los de la primera. De esta manera, el parámetro asociado a las variables $Mujer$, $Mujer * Edad$ y $Mujer * Edad^2$ deben ser el mismo al obtenido por medio de la especificación. En la tercera columna de la Tabla 6 se muestra que, en ambos modelos, se estima una brecha salarial al inicio de la vida laboral del 9,47 % entre mujeres y hombres con cargos y características laborales similares. La brecha, según lo demuestran los resultados presentados en la Tabla 6, varía en el tiempo y no es siempre constante debido a que los salarios evolucionan de forma diferentes para hombres y mujeres. En la sección 4.3 de documento se estudiará con mayor detalle esta relación.

Tabla 6: Resultados de la estimación del logaritmo del salario por hora bajo diferentes modelos

	Variable dependiente:			
	Logaritmo del salario por hora			
	(1)	(2)	(3)	(4)
Mujer	−0.0947	−0.0939		
	(0.0667)	(0.0693)		
Edad	0.0244	0.0245		
	(0.0022)	(0.0023)		
$Edad^2$	−0.0002	−0.0002		
	(0.00003)	(0.00003)		
$Mujer * Edad$	−0.0002	−0.0003		
	(0.0033)	(0.0035)		
$Mujer * Edad^2$	0.00002	0.00003		
	(0.00004)	(0.00004)		
Otros controles	✓	✓		
Mujer FWL			−0.0947	−0.0947
			(0.0665)	(0.0649)
$Mujer * Edad$ FWL			−0.0002	−0.0002
			(0.0033)	(0.0033)
$Mujer * Edad^2$ FWL			0.00002	0.00002
			(0.00004)	(0.00004)
Constante	8.6110	8.4410	−0.0000	
	(0.1294)	(0.0729)	(0.0038)	
Observaciones	16,541	15,769	16,541	
R^2	0.4702	0.4690	0.0035	
R^2 Ajustado	0.4672	0.4670	0.0033	
ECM	0.4866	0.4866	0.4853	
Estadístico F	158.6782	239.2034	19.2099	

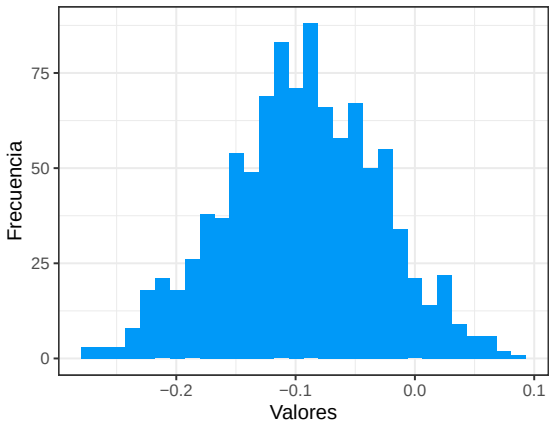
Nota: El modelo de la primera columna presenta los resultados de estimar la Ecuación 6 usando todos los datos obtenidos tras la limpieza. La segunda columna los resultados de la estimación de la Ecuación 6 sin incluir los que tienen un *leverage* tres veces superior a la media. La tercera columna presenta los resultados de la estimación usando el teorema FWL sobre toda la muestra. Finalmente, la última columna incluye la estimación por medio de FWL usando *bootstrap*.

Finalmente, se realizó una cuarta estimación de los parámetros asociados a la variables $Mujer$, $Mujer * Edad$ y $Mujer * Edad^2$ en las regresiones lineales usando el mismo teorema de FWL, pero estimando los errores estándar mediante el método de *bootstrap*. El *bootstrap* consiste en particionar arbitrariamente (permitiendo reemplazo) la muestra que se tiene en múltiples ocasiones y de esta manera, permite aumentar artificialmente el número de muestras disponibles para la estimación. Es, en últimas, llevar a cabo un proceso iterativo en el que se repite el método de FWL con muestras diferentes para aproximarse mejor a los errores estándar.

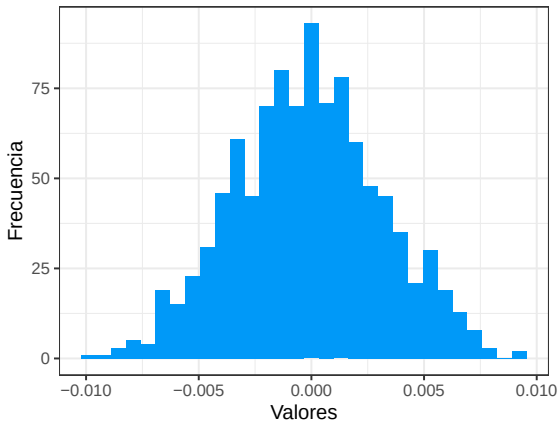
Cuando se lleva a cabo el proceso de *boostrap*, los resultados que se obtienen son similares a los de las anteriores tres estimaciones. En este caso, se obtiene una distribución de la estimación, ya que se tiene una diferente para cada repetición del proceso iterativo. En la Figura 8 se presenta una distribución de la estimación y la cuarta columna de la Tabla 6 muestra que la estimación es cercana a las anteriores.

Figura 8: Distribución de las estimaciones por *bootstrap*

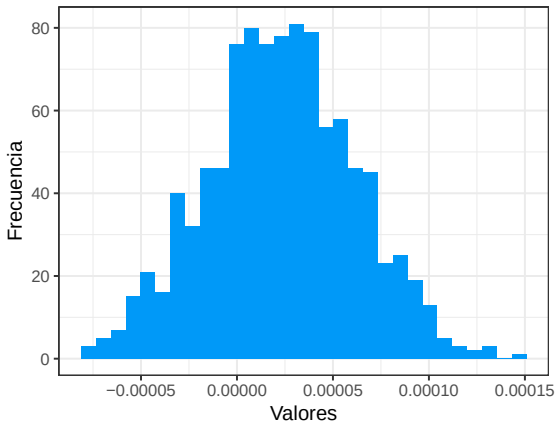
(a) Mujer



(b) Mujer por edad



(c) Mujer por edad al cuadrado



De esta forma, es posible ver al contemplar las características de los trabajadores y del oficio la brecha no disminuye, sino que aumenta con respecto a la brecha incondicional estimada. Y por ende, toma relevancia que las mujeres empleadas en Colombia no reciben"*equal pay for equal work*" . Este resultado además deja ver que la brecha encontrada en el modelo incondicional no era producto de una mala especificación del modelo producida por un problema de selección (variables omitidas),

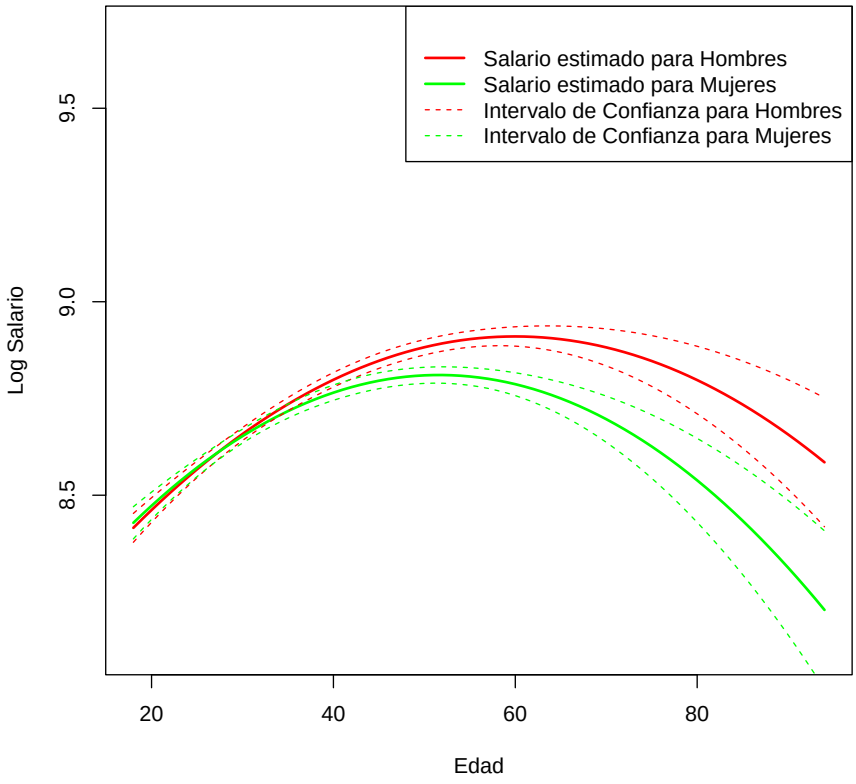
sino que en realidad la brecha persiste, aún controlando por características laborales. Lo anterior es consistente con otras estimaciones similares que se han realizado en Colombia (también usando la GEIH) controlando por factores laborales similares (Badel and Peña, 2010).

4.3. El perfil del salario-edad para hombres y mujeres

Buscando dar interpretabilidad a los hallazgos de la anterior sección, se analizará de forma gráfica la evolución de los salarios reales por hora para los hombres y mujeres de la muestra. Para eso, inicialmente, se realizará una estimación con la Ecuación 11 del salario real por hora. Este resulta simple, al tener menos variables independientes y, por lo tanto, se espera menor varianza en sus resultados y que estos tengan una mayor consistencia con el resto de la muestra. Se utilizó este modelo y no el de la Ecuación 6 en aras de lograr mayor interpretabilidad, pues en dicha ecuación tocaba fijar muchas variables de control lo que restringía el alcance general de la predicción. Más adelante se explorarán casos utilizando la Ecuación 6.

$$\begin{aligned}\log(\text{SalarioHora}_i) =& \beta_0 + \beta_1 * Mujer_i + \beta_2 * Edad_i \\ &+ \beta_3 * Edad_i^2 + \beta_4 * Mujer_i * Edad_i \\ &+ \beta_5 * Mujer_i * Edad_i^2 + \epsilon_i\end{aligned} \tag{11}$$

Figura 9: Perfil edad salario según sexo y características del individuo estimados bajo Ecuación 11



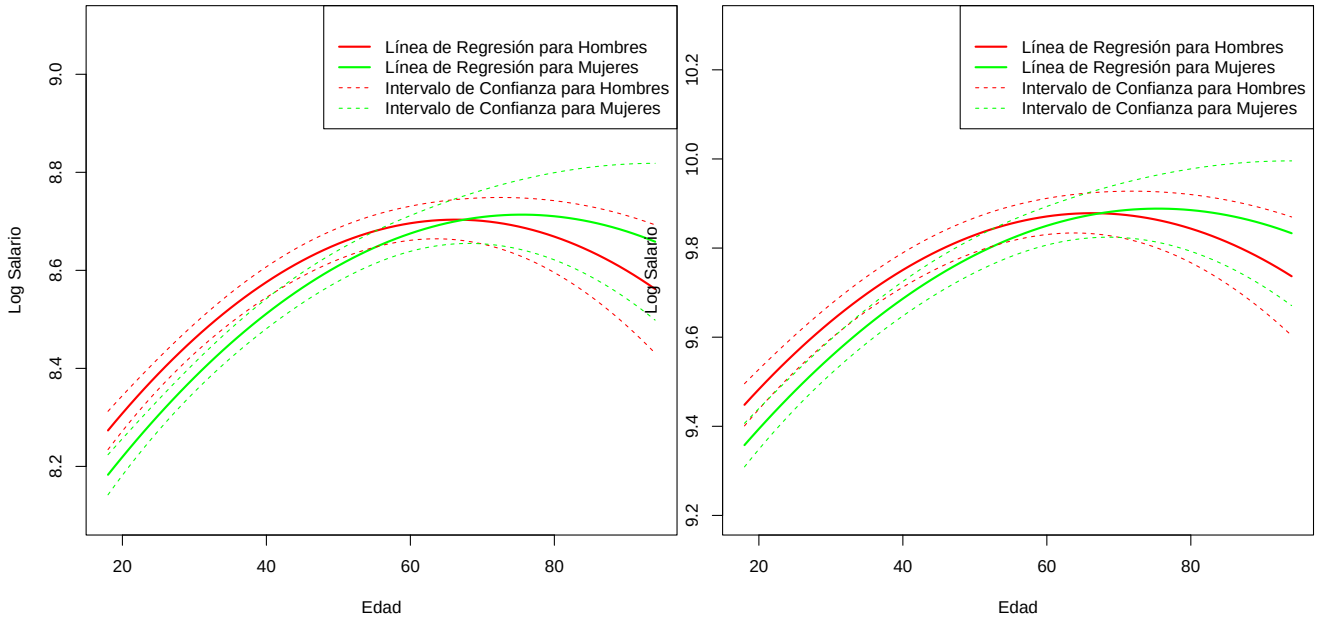
Como resultado de realizar la estimación del perfil por medio de una ecuación con menos variables explicativas, el perfil edad-salario estimado tiene menor varianza que un modelo más complejo, como el de la Ecuación 6. En la Figura 9, es posible notar el salario esperado para hombres y mujeres en las diferentes edades estimados bajo la Ecuación 11. De esta gráfica es importante notar dos cosas: su robustez y sus resultados consistentes con la literatura económica. En lo que respecta a la robustez, la gráfica permite notar que la diferencia entre el salario percibido por los hombres y las mujeres es estadísticamente significativa a lo largo de su vida laboral. Sobre sus conclusiones más consistentes con la literatura, la gráfica permite notar que el pico salarial era cercano a los 50 años para las mujeres y 60 para los hombres.

Según se esperaba, los resultados son consistentes con la literatura que ha estudiado esta relación en Colombia. Por ejemplo, Guataquí et al. (2009) encontraron que en Colombia los trabajadores alcanzan su salario máximo entre los 51 y 60 años cuando no se controla por tipo de empleo. Asimismo, encontraron que al controlar por tipo de empleo, la variable edad muestra efectos no lineales en ambos casos pero con tasa marginal creciente para quienes trabajan de manera independiente (cuenta propia) y decreciente en el caso contrario. Si bien los autores no buscaron diferencias entre sexos, uno de los hallazgos de los autores es similar y consistentes con los del análisis realizado hasta el momento: el salario por hora alcanza un máximo y no es siempre creciente en la edad. Estudios más recientes en el país, como el de Mora et al. (2023), han encontrado que existen brechas sistemáticas en el salario que reciben las mujeres, pero que sus retornos a la experiencia son mayores. En este caso, la conclusión de los autores únicamente aplican en las primeras etapas de la vida laboral cuando el salario de las mujeres es superior al de los hombres, pero luego la tendencia se revierte.

Figura 10: Perfil edad salario según sexo y características del individuo estimados bajo Ecuación 6

(a) Vendedores ambulantes

(b) Ejecutivos



Ahora bien, para comparar los resultados de la Figura 9 y la Ecuación 11 con la Ecuación 6 se realizaron predicciones que incluyen otros condicionales laborales. En la Figura 10a, se presenta el perfil edad salario asumiendo que una persona es vendedora ambulante, a domicilio, de lotería o de periódicos, que no trabaja por cuenta propia, que está empleada formalmente y trabaja 48 horas a la semana usualmente. El valor de las últimas 3 variables se seleccionó de forma consistente con que, de acuerdo con la muestra que se tiene, es es valor más común. Según se puede apreciar en la gráfica, se espera que los hombres de la mayoría de edades tengan un salario por hora mayor al de las mujeres. No obstante, debido a los numerosos parámetros, la diferencia no resulta ser significativa con un 95 % a partir de los 40 años.

En el caso de los vendedores ambulantes, el pico de edad estimado para lo hombres es cercano a los 60 años; mientras que para las mujeres esta cifra es de 74 años. En el caso de los ejecutivos, si bien los salarios de estos son superiores tanto para hombres como para mujeres, el patrón es similar: el pico de salario por hora es alcanzado por los hombres cerca a los 60 años mientras que las mujeres lo alcanzan cerca a los 75. Sin embargo, los resultados son poco confiables por sus grandes intervalos de confianza y su poca consistencia con la literatura. Por ejemplo, Guataquí et al. (2009) encuentraron que en Colombia los trabajadores alcanzan su salario máximo entre los 51 y 60 años. Si bien hay posibles explicaciones de por qué se llega a resultados como los presentados en la figura, como que la única manera por la que mayores de 60 años buscarían ocuparse es si su paga es superior a la que tenían antes de la edad de pensión, resulta más confiable e interpretable la estimación presentada en la Figura 9. Es decir, el modelo de la Ecuación 11 es mejor para analizar de manera gráfica los perfiles edad sexo de las personas en Colombia que el modelo de la Ecuación 6 dado que este varía mucho debido a su gran cantidad de variables de control, pues al fijar dichas variables se pierde la capacidad de generalizar dichos perfiles.

V. Prediciendo los ingresos

Hasta el momento, los modelos propuestos en el trabajo no tienen como objetivo principal la predicción, sino la inferencia. Esto se evidencia en sus bajos niveles de ajuste. En esta sección, se busca construir un modelo cuya meta principal sea maximizar su capacidad predictiva. Para lograr esto, se va a buscar un modelo que minimice el error cuadrático medio (ECM) al reducir el sesgo, pero teniendo cuidado de que la varianza del modelo no aumente tanto como para contrarrestar las ganancias en la reducción del sesgo, dado el 'textittrade-off evidenciado en los modelos predictivos entre estos dos componentes.

5.1. Comportamiento de los modelos planteados

Para cumplir con el objetivo anterior, se plantearon cinco especificaciones de modelos diferentes a los planteados en puntos anteriores, estos se incluyeron en el siguiente análisis. El **Modelo 1** (Ecuación 12) y el modelo 2 capturan las relaciones más básicas. Se partió del modelo del punto 4 (**Modelo 3** (Ecuación 14)). Al comparar el ECM de los modelos, es notable cómo entre el modelo 2 y el 3 hay una reducción notable al incluir la interacción con el sexo y otras variables de control. Buscando complejizar más el modelo, se plantea una especificación en la que también se incluye la permanencia en el trabajo, que es una posible *proxy* de la experiencia de un trabajador en el empleo, para el **Modelo 4** (Ecuación 15). Más adelante, en el **Modelo 5** (Ecuación 16), se explora una relación cúbica entre la edad y el logaritmo del salario real por hora, como se puede ver en su ECM. Esta relación cúbica vale la pena ponerla en contexto de que estamos buscando un modelo predictivo no un modelo explicativo, por ende queremos un grado de complejidad que se ajuste bien al modelo sin perder una varianza moderada y este modelo lo cumple. Los siguientes modelos propuestos van a mantener la edad cúbica.

De este modelo en adelante, se busca interactuar la variable de sexo con las diferentes variables explicativas propuestas. Entre el **Modelo 6** (Ecuación 17) y **Modelo 7** (Ecuación 18) se pueden ver reducciones notables en el ECM. Entre las variables que se interactúan están la cantidad de horas a la semana que se trabaja, si el trabajo es formal o informal, si el trabajo es independiente o no, el nivel educativo y el tiempo de duración en el trabajo actual. Hasta el **Modelo 7** (Ecuación 18) se consigue reducir el sesgo del modelo predictivo al tomar en cuenta la relación entrelazada que tienen estas variables y el sexo de las personas para determinar su salario. Volviendo a la teoría económica, tiene sentido pensar que, por ejemplo, un hombre con un alto nivel de educación y una mujer con el mismo nivel no son necesariamente iguales en la forma como son remunerados. O por tomar otro ejemplo, una mujer que trabaja 40 horas a la semana posiblemente carga con muchas más responsabilidades por fuera del trabajo, mientras que un hombre en condiciones parecidas, en promedio, tiene menos responsabilidades no remuneradas por fuera del trabajo (Tribín et al., 2021). Sin embargo, para el modelo 8, al interactuar el sexo con la variable oficio, se ve un incremento pronunciado en el ECM. Esto indica que el *trade-off* entre sesgo y varianza ha sido superado por la varianza. Esto puede tener sentido dado que la variable oficio es una variable categórica, con una alta cantidad de categorías, más de ochenta, y por ende, tiene mucha variación lo que contribuye a una alta varianza en el modelo, y más aún con la interacción con sexo.

De este proceso resulta el **Modelo 7** (Ecuación 18), como el modelo con el menor error y mayor nivel de complejidad. Se destaca, como ya se ha mencionado antes, el rol clave de las interacciones de la variable de sexo con las otras variables explicativas. La especificación a la que se llega recalca el rol crucial del sexo al analizar los ingresos de las personas naturales y, por extensión, el posible recaudo de organismos como la DIAN. La recomendación más importante sería que al analizar a

personas naturales contribuyentes, es importante que instituciones de recaudo fiscal, como la DIAN, consideren lo determinante que puede ser el sexo y que al tratar de predecir si una persona está mintiendo en su declaración de renta es clave tener en cuenta si es hombre o mujer; de no ser así, cualquier conclusión estaría fuertemente sesgada.

Tabla 7: Valores del ECM para los modelos predictivos

Modelo	ECM
1	0.6524673
2	0.6650510
3	0.4957021
4	0.4904258
5	0.4902586
6	0.4901828
7	0.4896096
8	0.4907051

5.2. Especificación modelos

Modelo 1

$$\log(SalarioHora_i) = \beta_1 Edad_i + \beta_2 Edad_i^2 + \varepsilon_i \tag{12}$$

Modelo 2

$$\log(SalarioHora_i) = \alpha_1 Mujer_i + \epsilon_i \tag{13}$$

Modelo 3

$$\begin{aligned}\log(SalarioHora_i) =& \sigma_1 Mujer_i + \sigma_2 Edad_i + \sigma_3 Edad_i^3 + \sigma_4 Mujer_i * Edad_i \\ &+ \sigma_5 Mujer_i * Edad_i^2 + \sigma_6 HorasSemana_i + \sigma_7 Formal \\ &+ \sigma_8 CuentaPropia + \sum_{f=1}^{F} \phi_f Oficio_{i,f} + \sum_{j=1}^{J} \rho_j NivelEducativo_{i,j} + u_i\end{aligned} \tag{14}$$

Modelo 4

$$\begin{aligned}\log(SalarioHora_i) =& \mu_1 Mujer_i + \mu_2 Edad_i + \mu_3 Edad_i^2 + \mu_4 Mujer_i * Edad_i \\ &+ \mu_5 Mujer_i * Edad_i^2 + \mu_6 HorasSemana_i + \mu_7 Formal \\ &+ \mu_8 CuentaPropia + \mu_9 Permanencia_i + \sum_{f=1}^{F} \phi_f Oficio_{i,f} \\ &+ \sum_{j=1}^{J} \rho_j NivelEducativo_{i,j} + \nu_i\end{aligned} \tag{15}$$

Modelo 5

$$\begin{aligned}\log(SalarioHora_i) =&\pi_1 Mujer_i + \pi_2 Edad_i + \pi_3 Edad_i^2 + \pi_4 Edad_i^3 + \pi_5 Mujer_i * Edad_i \\ &+ \pi_6 Mujer_i * Edad_i^2 + \pi_7 Mujer_i * Edad_i^3 + \pi_8 HorasSemana_i \\ &+ \pi_9 Formal + \pi_{10} CuentaPropia + \pi_{11} Permanencia_i \\ &+ \sum_{f=1}^{F} \phi_f Oficio_{i,f} + \sum_{j=1}^{J} \rho_j NivelEducativo_{i,j} + \eta_i\end{aligned} \tag{16}$$

Modelo 6

$$\begin{aligned}\log(SalarioHora_i) =&\delta_1 Mujer_i + \delta_2 Edad_i + \delta_3 Edad_i^2 + \delta_4 Edad_i^3 + \delta_5 Mujer_i * Edad_i \\ &+ \delta_6 Mujer_i * Edad_i^2 + \delta_7 Mujer_i * Edad_i^3 +_8 HorasSemana_i \\ &+ \delta_9 Mujer_i * HorasSemana_i + \delta_{10} Formal + \delta_{11} CuentaPropia \\ &+ \delta_{12} Permanencia_i + \sum_{f=1}^{F} \phi_f Oficio_{i,f} + \sum_{j=1}^{J} \rho_j NivelEducativo_{i,j} + \iota_i\end{aligned} \tag{17}$$

Modelo 7

$$\begin{aligned}\log(SalarioHora_i) =&\theta_1 Mujer_i + \theta_2 Edad_i + \theta_3 Edad_i^2 + \theta_4 Edad_i^3 + \theta_5 Mujer_i * Edad_i \\ &+ \theta_6 Mujer_i * Edad_i^2 + \theta_7 Mujer_i * Edad_i^3 + \theta_8 HorasSemana_i \\ &+ \theta_9 Mujer_i * HorasSemana_i + \theta_{10} Mujer_i * Formal + \\ &\theta_{11} Mujer_i * CuentaPropia + \theta_{12} Mujer_i * Permanencia_i \\ &+ \sum_{f=1}^{F} \phi_f Oficio_{i,f} + \sum_{j=1}^{J} \rho_j Mujer_i * NivelEducativo_{i,j} + \upsilon_i\end{aligned} \tag{18}$$

Modelo 8

$$\begin{aligned}\log(SalarioHora_i) =&\tau_1 Mujer_i + \tau_2 Edad_i + \tau_3 Edad_i^2 + \tau_4 Edad_i^3 + \tau_5 Mujer_i * Edad_i \\ &+ \tau_6 Mujer_i * Edad_i^2 + \tau_7 Mujer_i * Edad_i^3 + \tau_8 HorasSemana_i \\ &+ \tau_9 Mujer_i * HorasSemana_i + \tau_1 0 Mujer_i * Formal \\ &+ \tau_{11} Mujer_i * CuentaPropia + \tau_{12} Mujer_i * Permanencia_i \\ &+ \sum_{f=1}^{F} \phi_f Mujer_i * Oficio_{i,f} + \sum_{j=1}^{J} \rho_j Mujer_i * NivelEducativo_{i,j} + \varsigma_i\end{aligned} \tag{19}$$

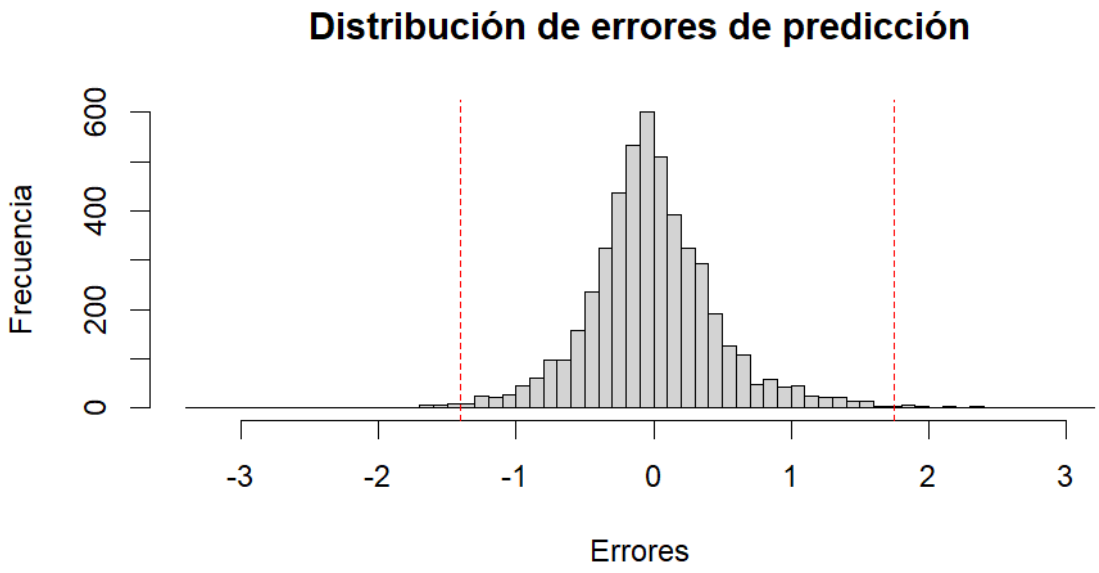
5.3. Distribución del modelo con mejor ECM

El modelo de predicción **Modelo 7** (Ecuación 18) fue el que obtuvo el menor *ECM*. En la Figura 11 se evidencia qué tan lejos está su predicción de los valores reales. Es evidente que los errores están centrados en cero y que la gran mayoría se encuentran entre los valores de -1 y 1.

Dado que es el modelo con menor *ECM*, se podría pensar que los *outliers* no se ajustan al modelo debido a que reportan datos erróneos. Es decir, podrían ser grupos de interés para la DIAN dado

que pueden ser evasores de impuestos. Sin embargo, hay que recordar que no hay un predictor perfecto, y que a medida que se aumenta la complejidad de un modelo su error de sesgo disminuye; pero, a medida que crece la misma complejidad el error producido por la varianza aumenta. Por lo anterior, es perfectamente válido que haya observaciones en las colas del histograma. Además, al ver la Figura 11 se nota que sigue una distribución normal y que no hay cúmulos en los extremos, es decir, no se evidencia un intento de manipulación de información u otra práctica deshonesta por parte de las personas.

Figura 11: Distribución de los errores de predicción del logaritmo del salario real por hora del modelo 6 con un intervalo de confianza del 99 %



5.4. LOOCV

Partiendo de los modelos que obtuvieron los dos ECM más bajos, se implementa el modelo de validación de *Leave One Out Cross-Validation* (LOOCV). Este fue un proceso largo y minucioso en el que se tomó cada observación de la base y se utilizó para probar el modelo predictivo de toda la muestra menos esa observación. Este es un proceso más completo para evaluar el poder predictivo del modelo, pero es un proceso costoso computacionalmente. Por esta razón, solo se realizó con los modelos **Modelo 6** (Ecuación 17) y **Modelo 7** (Ecuación 18) que tenían el menor ECM. Los resultados se pueden ver en Tabla 8, donde se observa que en el proceso de validación anterior se estaba sobreestimando el ECM de ambos modelos. Usando esta metodología, el ECM de ambos modelos es inferior y son más cercanos entre sí. Esto tiene sentido dado que incluyen un par de interacciones diferentes, pero de todas formas se mantiene que el **Modelo 7** (Ecuación 18) es el modelo donde el ECM más se acerca a cero.

A partir de este ejercicio se puede ver el poder que tiene un método de validación como LOOCV, sin embargo también queda claro los costos de tiempo y recursos que pueden tener, en este caso se podría argumentar que el proceso inicial de validación fue satisfactorio para indicar el modelo con mejor balance entre sesgo y varianza. Con respecto a la influencia de ciertas observaciones sobre el modelo, dado el análisis que se hizo en el punto anterior no se considera que haya un riesgo alto de

que haya observaciones que afecten el modelo, dado que no había un número notable de *outliers*. También, por construcción el modelo de LOOCV asegura darle menos peso a observaciones que potencialmente podrían estar contribuyendo desmedidamente al sesgo.

Tabla 8: Mediciones de ECM usando LOOCV

Modelo	**ECM con LOOCV**
6	0.4847495
7	0.484197

VI. Repositorio de GitHub

Link al repositorio. Para replicar los resultados descargar los paquetes y prerrequisitos especificados en el README. Además, utilizar RStudio, importar el repositorio por medio de git a RStudio y correr el código. Los archivos creados se suben directamente al repositorio, no se exportan a carpetas locales fuera del repositorio.

VII. Bibliografía

Badel, A. and X. Peña (2010): "Decomposing the Gender Wage Gap with Sample Selection Adjustment: Evidence from Colombia," *Economic Analysis Review*, 25, 169–191.

Barreto Nieto, C. A., A. Jimenez Ospina, D. F. Lemus Polanía, P. Montenegro Helfer, and D. P. Ramírez (2020): "Brecha Salarial De Género: Estudio De Caso De Los Contrastistas Independientes del Estado en Colombia," *Archivos de Economía. DNP. Dirección de Estudios Económicos*.

DANE (2018): "Censo Nacional de Población y Vivienda," .

——— (2019): "Documento Metodológico. Medición de Pobreza Monetaria y Desigualdad 2018." .

——— (2022a): "Análisis de las clases sociales en las 23 ciudades y áreas metropolitanas de Colombia," .

——— (2022b): "Breacha salarial de género en Colombia," .

DIAN (2022): "Informe Rendición de Cuentas 2022. Prospectiva 2023-2026." *Informe rendición de cuentas*.

Galvis-Aponte, L. A. (2010): "Diferenciales salariales por género y región en Colombia : una aproximación con regresión por cuantiles," Tech. rep.

Guataquí, J. C., A. F. García, and M. Rodríguez (2009): "Estimaciones de los determinantes de los ingresos laborales en Colombia con consideraciones diferenciales para asalariados y cuenta propia," .

Mincer, J. (1974): *Schooling, Experience, and Earnings*, Human behavior and social institutions, National Bureau of Economic Research.

Mora, J. J., D. Y. Herrera, J. F. Álvarez, and J. S. Arroyo (2023): "Returns to Human Capital in a Developing Country: A Pseudo-Panel Approach for Colombia," .

Sabogal, A. (2012): "Brecha salarial entre hombres y mujeres y ciclo económico en Colombia," *Coyuntura Económica: Investigación Económica y Social. Fedesarrollo*.

Tribín, A. M., A. D. Gómez-Barrera, and A. Pirela-Ríos (2021): "Distribución del cuidado, roles de género y poder de negociación en Colombia: análisis ENUT 2020-2021," *Quanta*.