

Problem Set 1: Predicting Income

Big Data y Machine Learning para
Economía Aplicada, 2023-1



Integrantes: Jorge Leonardo Rodríguez Arenas¹, Iván Camilo Velásquez Abril², Santiago González Cepeda³, María José Colmenares Wolff⁴

1. Introducción

Tanto la complejidad del sistema tributario como los datos disponibles, son variables que hacen del estudio tributario algo no tan sencillo. De lo anterior, se deriva lo que se conoce como la brecha fiscal, lo cual es la diferencia entre las estimaciones de los ingresos fiscales frente a lo que realmente es el recaudo real que es pagado de manera oportuna y voluntaria. En el caso de Estados Unidos la brecha fiscal estimada fue de USD \$456 mil millones para los años 2014 – 2016, por otro lado la tasa de cumplimiento voluntario que mide el cumplimiento relativo entre el estimado y el real se ubicó en 85 % según el Servicio de Impuestos Internos (IRS por sus siglas en inglés) para los mismos años. Es posible, que este fenómeno se replique en Colombia, ya que ha sido caracterizada por tener una estructura tributaria antitécnica, compleja, donde hay una combinación de gravámenes relativamente elevados con un cúmulo de exenciones Clavijo (2011) y puede haber una caída para que se presente una subdeclaración de ingresos por parte de las personas, lo cual tendría un efecto negativo en el recaudo tributario.

Para tener una mejor aproximación de los ingresos reales de la población, se utilizarán diferentes modelos donde la edad y el género hacen parte de la forma como se quiere hacer una predicción de dichos ingresos. Esto tendrá diferentes beneficios que se derivan directamente de un mayor poder de tributación a ciertos individuos, lo que implica un mayor recaudo, el cual se utiliza para inversión en infraestructura, salud, educación, entre otros. En ese sentido, los controles escogidos para las estimaciones están sustentados en la ecuación minceriana de ingresos, toda vez que sugiere ciertas variables que pueden ayudar a predecir los ingresos laborales de las personas. Según Mincer (1974) existía una relación positiva entre la escolaridad de un individuo y sus ingresos posteriores, asimismo con la experiencia. Por lo cual, se considera utilizar esta teoría para así poder generar unos resultados razonables. Además, tener una mejor aproximación a los ingresos reales de la población, permitirá tener un mejor conocimiento de la distribución de ingresos y, al desagregarlo por variables como la edad y el género, pueden ser de gran utilidad para la construcción o ajuste de políticas públicas que tengan como estrategia la redistribución a ciertos grupos poblacionales.

De esta manera, la Gran Encuesta Integrada de Hogares (GEIH) es una herramienta muy útil ya que brinda información sobre las condiciones de empleo del hogar, es decir, si las personas trabajan o no, cuál es su trabajo, cuántos son sus ingresos, si tienen seguridad social y salud y si están en búsqueda de empleo. Además de unas características generales de la población como sexo, edad, nivel de educación,

¹Cod. Uniandes: 201715669

²Cod. Uniandes: 201114762

³Cod. Uniandes: 201719971

⁴Cod. Uniandes: 201811929

estado civil, etc. DANE (2022). Igualmente, este será un enfoque que solo tomará datos de Bogotá dejando de lado el resto de información a nivel nacional, tanto de cabeceras, como regional, departamental y las capitales de los departamentos.

En este estudio, únicamente se utilizará una derivación de la GEIH que fue utilizada por el Departamento Administrativo Nacional de Estadística para construir medidas de pobreza monetaria y desigualdad. Esta derivación agrupa las variables que se quieren estudiar en este problema en particular y descarta otras tantas variables que están incluidas en la GEIH que no son necesarias para el objetivo del ejercicio. Como grupo demográfico, solo se utilizarán los datos de individuos residentes en Bogotá. Como resultado, esta predicción de ingresos lo que busca es que se puedan conocer los fraudes que se presentan y de igual forma, poder identificar familias e individuos vulnerables que serían potenciales beneficiarios de políticas que focalicen el gasto.

A modo de conclusión, se encontraron las edades picos, tanto agregadas como desagregadas por género. Al desagregarla por género, se encuentra que los hombres alcanzan su edad pico a una edad más tardía que las mujeres, la cual se puede explicar porque los hombres, dado el sistema de pensiones actual, deben permanecer por un mayor tiempo en el mercado laboral que las mujeres, además que las mujeres suelen tener desventajas sistemáticas que, al mismo tiempo las empuja fuera del mercado laboral. Además, se encuentra que efectivamente existe una brecha significativa entre los salarios de hombres y mujeres, así se controle por diferentes factores o se aísle únicamente el efecto del género sobre el salario. Esto nos parece que es una conclusión a la que hay que prestarle atención en términos de políticas públicas, ya que, a 2018 en Colombia, aún se estaba pagando de manera diferente trabajos iguales únicamente porque la persona a la que se le paga es una mujer.

De igual forma, se puede concluir que la variable Oficio tiene un gran peso en las estimaciones del salario laboral ya que los modelos que no la incluyen no son tan efectivos como los modelos donde se tiene en cuenta. Para estimar con precisión el ingreso laboral es necesario construir modelos complejos, que incluyan la mayor cantidad posible de variables (respetando el sentido económico y teórico de los modelos), ya que los modelos más simples están muy lejos de estimar con precisión el salario laboral y su MSE es considerablemente superior.

2. Datos

Los datos de la encuesta “Medición de Pobreza Monetaria y Desigualdad 2018” creados a partir de la GEIH de 2018, es usada principalmente por el DANE para realizar las estimaciones del Índice de Pobreza Multidimensional (IPM), sin embargo, múltiples estudios usan estos datos para evaluar modelos de desarrollo rural (Serrano-Malaver et al. (2018)), evaluar los niveles de educación y género a partir de la descomposición del IPM (Cortes Sabogal et al. (2016)) y otros estudios relacionados con tributación y pobreza (García Ocampo (2018)). Usando los datos de esta encuesta se pretende predecir el nivel de ingreso (laboral) de los individuos de la ciudad de Bogotá ocupados mayores de 18 años, con el fin de contar con un modelo predictivo que permita bajo ciertos parámetros tener una medida que funcione para crear las alertas necesarias en los entes encargados de la recolección tributaria. No se hace uso de la variable de ingreso total, dado que, por definición en esta encuesta, dicha variable tiene información del valor aproximado de ingresos especies que en la practica puede generar ruido a la hora de tener modelos predictivos, ya que es muy arbitrario el valor que toman.

La obtención de los datos fue a partir de web scraping, donde se seleccionan 10 archivos de la web que componen toda la información de la encuesta para todos los meses del 2018 para la ciudad de Bogotá, sin embargo los datos originales se encuentran en la página oficial del DANE, dicha página no tiene

limitaciones para el uso de web scraping ⁵. La base originalmente cuenta con 42,069 observaciones para la ciudad de Bogotá de las cuales 34,460 son personas mayores de edad y 16,542 pertenecen a la población ocupada, de esta muestra de individuos se eliminan aquellas observaciones que presentan valores perdidos en la variable de salario nominal mensual (6,650 observaciones) dado que no se utilizara ninguna técnica de imputación dado el alto porcentaje de valores perdidos. Para el tratamiento de valores atípicos se realiza una técnica llamada “Winsorizar” que consiste en retener las observaciones atípicas, haciendo que estas tomen valores particulares dependiendo del percentil que se elija, para este caso los valores atípicos se internalizan en la muestra del tal manera que el 1 % de la muestra toman el valor del percentil 1, y el 1 % más alto de la distribución de los salarios toma el valor del percentil 99. Finalmente, la muestra objeto de estudio cuenta con 9,892 observaciones.

Cuadro 1: Estadísticas Descriptivas - Colombia 2018 por género

	Todos		Mujer		Hombre		Diff ¹
	Media	Desv. Estándar	Media	Desv. Estándar	Media	Desv. Estándar	
A. Individuo							
Edad	36,24	12,02	36,58	11,90	35,90	12,14	***
Prop. Amo(a) de casa	0,030	0,171	0,055	0,229	0,005	0,072	***
B. Hogar							
Prop. Hijos en el hogar	0,235	0,424	0,232	0,422	0,238	0,426	
Estrato	2,51	0,98	2,58	1,00	2,44	0,95	***
C. Educación²							
Estudiantes	0,010	0,101	0,011	0,103	0,010	0,099	
Primaria	0,005	0,067	0,005	0,074	0,004	0,060	
Secundaria	0,095	0,293	0,081	0,273	0,109	0,312	***
Media	0,346	0,476	0,320	0,467	0,371	0,483	***
Terciaria	0,453	0,498	0,503	0,500	0,403	0,491	***
D. Mercado Laboral³							
Salario mensual	1.682.266	1.876.130	1.606.083	1.837.522	1.757.622	1.910.751	***
Ingreso Total	1.872.592	2.509.096	1.790.266	2.309.982	1.954.025	2.689.353	***
Experiencia trab. Actual	49,73	73,20	49,45	73,71	50,01	72,71	
Horas trabajas en la semana	48,02	12,15	45,86	11,68	50,16	12,24	***
Prop. Informalidad	0,233	0,422	0,248	0,432	0,217	0,412	***
Observaciones	9.892		4.919		4.973		

Notas: Cálculos propios usando datos DANE Medición de Pobreza Monetaria y Desigualdad 2018. ¹ Indica el nivel de significancia estadística de la diferencia de medias entre hombres y mujeres bajo la $H_0: \mu(\text{Hombre}) - \mu(\text{Mujer}) = 0$. ² Los valores representados en esta sección corresponden a la proporción de personas que esta dentro de cada categoría. ³ La experiencia en el trabajo actual esta medida en meses. La unidad monetaria de las variables referentes al ingreso son pesos colombianos nominales. *** p<0.01, ** p<0.05, * p<0.1

En el cuadro 1, resume los estadísticos descriptivos de la muestra para el total y por género de los datos que se utilizara en la estimación de los modelos. En la muestra hay 4,919 mujeres y 4,973 hombres. La edad promedio de la población es de 36.2 años y las mujeres tienen una edad promedio mayor que los hombres; existe una mayor proporción de mujeres que son amas de casa que hombres. El nivel socioeconómico o estrato es en promedio de 2.5. En cuanto a los niveles educativos reportados se encuentra que existe una mayor proporción de mujeres con algún tipo de educación superior que hombres siendo 50.3 % y 40.3 % respectivamente. Un dato relevante es que para Bogotá la educación superior es la categoría predominantes en términos educativos. El salario mensual promedio es de 1,682,266 siendo este superior al salario mínimo del año 2018 (781,242), existe una diferencia estadísticamente significativa al 1 % entre hombres y mujeres en términos salariales, esta brecha a persistido a través de los años (2008-2019) en las diferentes ciudades tal como lo muestra Florez, Melo-Becerra, and Posada (2021). El ingreso total es mayor dado que incluye otras fuentes de ingreso no laborales. La experiencia promedio de los individuos es de aproximadamente 4 años en el trabajo que reportan actualmente, y las horas que trabajan a la semana son en promedio de 48.02, destacando que el numero

⁵Esto se verifica a través de https://microdatos.dane.gov.co/index.php/catalog/608/get_microdata/robots.txt

de horas trabajadas es menor en las mujeres. Finalmente, la proporción de informalidad es más alta en las mujeres que en los hombres. Para todas las variables reportadas, en la última columna del cuadro 1, se reporta si existen diferencias promedio estadísticamente significativas entre hombres y mujeres.

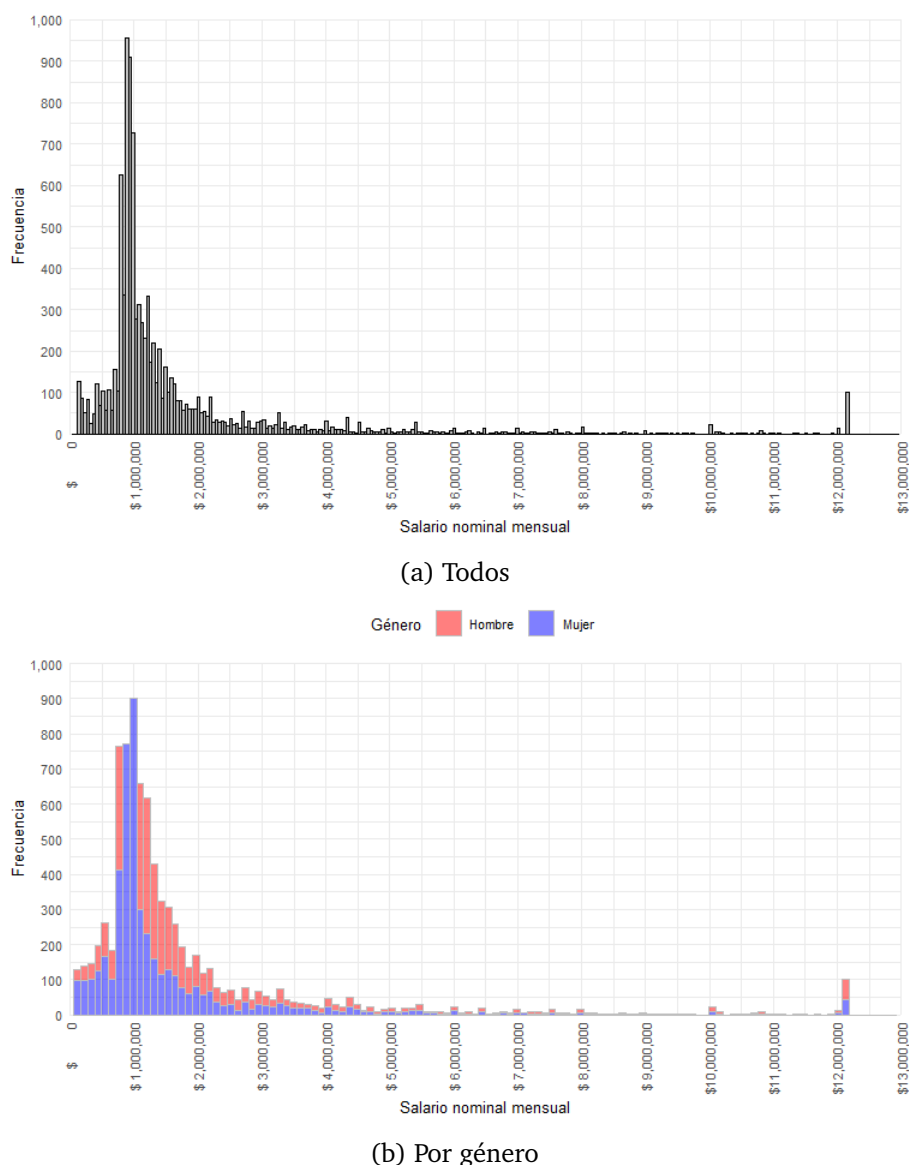


Figura 1: Distribución del salario nominal mensual

En la Figura 1, se muestra la distribución de los salarios nominales mensuales para la muestra, es evidente que la distribución está sesgada hacia la izquierda, y en su gran mayoría los salarios se centran entre 0 y 1,000,000. En cuanto a la distribución por sexo, los hombres tienen la distribución salarial más hacia la derecha.

3. Perfil de salario por edad

Lo primero para realizar el ejercicio de hacer la predicción de los salarios, es empezar por estudiar cómo diferentes variables tienen influencia en este. Para esta sección, evaluaremos la evolución del salario a través del tiempo, más específicamente en el tiempo de vida de las personas. Esto es relevante para entender la forma en que diferentes grupos de edad reciben ingresos, y uno de los objetivos de esta

sección es encontrar la edad en donde, en promedio, una persona alcanzaría sus ingresos máximos.

Para esto, se realizó un modelo de regresión sencillo, que sigue la siguiente ecuación:

$$\ln(salario) = \beta_0 + \beta_1 Edad + \beta_2 Edad^2 + u$$

Cuadro 2: Estimación Salario - Edad

<i>Dependent variable:</i>	
log_salario_m	
Edad	0.087*** (0.004)
Edad ²	-0.001*** (0.00004)
Constante	12.335*** (0.068)
Observations	9,892
R ²	0.060
Adjusted R ²	0.060
Residual Std. Error	0.711
F Statistic	315.885***

Note: *p<0.1; **p<0.05; ***p<0.01

En la tabla 3 se presentan los resultados de esta estimación. En esta, se observa que todos los coeficientes, incluyendo el de la constante son significativos a un nivel de confianza del 99 %. Para la variable de la edad, la estimación indica que por cada año adicional en la edad, en promedio el salario subiría alrededor del 8.7 %, la intuición de esto es que mientras las personas van adquiriendo experiencia laboral, van a poder ir mejorando su situación económica, al conseguir puestos que requieran mayores conocimientos y que tenga una mayor retribución dentro del mercado laboral.

Por otra parte, el estimador de la Edad al cuadrado, lo que está representando en la regresión es un efecto decreciente a medida que las personas van creciendo. Esto puede darse por diferentes razones, como un retraso en el aprendizaje de nuevas tecnologías que puede sacar a personas de mayor edad fuera del mercado laboral, también es posible que una persona de mayor edad tenga más dificultad a la hora de encontrar nuevos empleos que le permitan obtener salario más altos, al tener una gran competencia con personas más jóvenes, por último, también es necesario considerar que muchas personas esperar jubilarse después de haber cumplido las condiciones para hacerlo, y esto también puede ser una fuente de reducción de los ingresos para algunas personas. Es por esto que este tiene un coeficiente negativo y que, a pesar de que parezca una variable casi insignificante por la magnitud estimada, es clave para poder generar la estimación de la edad pico.

Por último, la estimación de la constante está indicando un nivel mínimo de salario que una persona de la edad más pequeña, que para este caso se refiere a una persona de 18 años, entraría en promedio a ganar dentro del mercado laboral. Como se indicó al inicio de esta sección, el objetivo principal era hallar la edad pico, es decir, la edad en la que, en promedio, una persona podría obtener los ingresos más altos.

En la figura 2, se puede observar lo que sería el perfil estimado de la relación entre la edad y los salarios. En la primera mitad de la gráfica, se puede observar que el efecto positivo de la estimación de la edad es

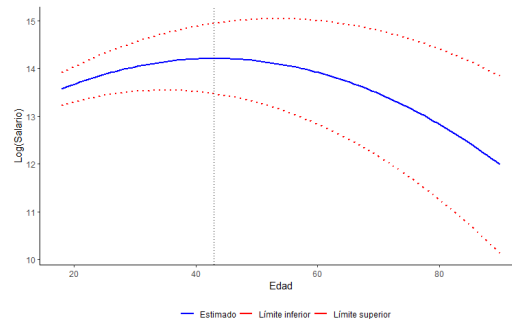


Figura 2: Estimación de la edad pico

más fuerte, por lo que se mantiene en una subida hasta llegar a la edad pico de los salarios, la cuál está representada con la línea vertical, que se encuentra, según las estimaciones, en los 43 años. Después de esta edad, el papel decreciente de la edad al cuadrado toma fuerza y se puede observar una caída en los salarios hasta llegar a niveles mucho más bajos a comparación de los que en promedio se estaría ganando una persona al comenzar su vida laboral. Como se explicó anteriormente, esto puede darse por diferentes razones, pero una de las principales es el cumplimiento de los requisitos de pensión, por lo que se deja de recibir ingresos como salario, y que es muy bajo el porcentaje de personas que después de los 65 años siguen en el mercado laboral.

La gráfica anterior también incluye los intervalos de confianza en los que cada edad se mueve. Estos fueron contruidos a través de los errores estándar del bootstrap, lo cuál genera una robustez adicional al modelo, ya que, al resamplear el modelo, las estimaciones iniciales de la regresión lineal siguen siendo validas, y caen dentro de los intervalos de confianza contruidos para niveles de confianza del 95 %. Es importante notar que los intervalos de confianza a medida que la edad aumenta, estos también crecen. Esto puede explicarse también por el tema de pensiones, ya que hay una gran proporción de personas que dejan de trabajar a edades mayores, pero también existe una cantidad de personas que siguen en grandes posiciones en donde puedan estar con un buen salario y esto tira el intervalo de confianza hacia los lados.

4. Brecha salarial de género

En Colombia, la brecha salarial general entre hombres y mujeres, según la media, es de 12,9 % para el año 2019 (sobre Salarios (2018)). Este fenómeno representa una de las mayores injusticias sociales que se viven actualmente, de allí el interés en estudiarlo en esta sección. Inicialmente, se evaluó un modelo simple (incondicional), donde los retornos salariales dependen única y exclusivamente si la persona es una mujer, con el fin de poder evidenciar si existe o no esta desigualdad antes mencionada.

Concretamente, el modelo de regresión que se realizó, sigue la siguiente ecuación:

$$\ln(salario) = \beta_1 + \beta_2 Mujer + u$$

Donde $\ln(salario)$ representa el salario nominal mensual que recibe el individuo, y $mujer$ es un indicador que toma el valor de uno si la persona se identifica como mujer y cero de lo contrario.

En el cuadro 3 se puede evidenciar como existe una brecha salarial casi que evidente, esto puede verse en como se comporta la semi elasticidad del ingreso cuando la persona es mujer. Es decir, que ser mujer en promedio afecta en un 32 % el salario mensual que se recibe, lo que refleja la desigualdad laboral

con respecto a los hombres.

Por otra parte, un eslogan muy común es “igual salario por igual trabajo”. Una forma de interpretar esto es que para los empleados con características laborales y laborales similares, no debería existir una brecha salarial de género. Es por esto, que en esta parte se estimó una brecha de ingresos condicional que incorporó variables de control tales como características similares del trabajador, la cuales fueron si el trabajador poseía o no un nivel de educación superior, las horas de trabajo usuales, el tipo de trabajo ya sea formal o informal.

Cuadro 3: Estimación brecha de salario por género

Variable dependiente:	
Logaritmo del salario mensual	
Mujer	-0.139*** (0.015)
Constante	14.087*** (0.010)
Observaciones	9,892
R ²	0.009
R ² Ajustado	0.009
Error Estándar Residual	0.730 (df = 9890)
Estadístico F	90.051*** (df = 1; 9890)
Nota: *p<0.1; **p<0.05; ***p<0.01	

Concretamente, el modelo de regresión que se realizó, sigue la siguiente ecuación:

$$\ln(\text{salario}) = \beta_1 + \beta_2 \text{Mujer} + \beta_3 \text{Edad} + \beta_4 \text{Edad}^2 + \beta_5 \text{Superior} + \beta_6 \text{HorasTrabUsual} + \beta_7 \text{Informal} + u$$

Donde se incorporaron controles como los de edad cuyo propósito es integrar ese comportamiento que se observaba en la sección pasada, el cual es encontrar un pico de edad donde las personas maximicen sus ingresos. En ese sentido, también se puede ver que la variable indicativa *Superior* la cual toma el valor de uno si la persona cuenta con educación superior o cero de lo contrario, de igual forma con la variable *HorasTrabUsual*, pues esta mide cuantas horas son destinadas al trabajo a la semana en promedio y por ultimo la variable indicadora *Informal*, cuyo propósito es clasificar a las personas y ubicarlas según su tipo de trabajo, entendiendo que toma el valor de uno si se encuentra en un trabajo informal y cero de lo contrario.

En el cuadro 4 se muestra como añadiendo controles al modelo aún se presencia la brecha salarial, es decir, no importa si las características del trabajo son similares o no, sigue habiendo una brecha preexistente aún así controlando por esas características. Así mismo, también se observa una significancia estadística a un nivel de 99 % para todos los coeficientes incluida la constante, lo cual permite seguir realizando un análisis mas confiable. Por otro lado, cabe resaltar el papel de la *educación superior* dado que, según lo que muestran los datos, existe un efecto positivo en el retorno salarial, ya que en promedio subiría en 68.79 % si se cuenta con ese nivel educativo, manteniendo todo lo demás constante. Intuitivamente se puede pensar en que las personas que cuentan con una carrera profesional tienen tanto los conocimientos técnicos y analíticos que les permitiría desenvolverse mejor en su vida profesional en comparación con una persona que no obtuvo un grado de bachiller, actualmente se nota que el mercado laboral es mas competitivo y demanda conocimientos y habilidades informáticas que en su mayoría son enseñadas en las universidades (e. g. ciencia de datos, big data, machine learnig, IA, etc).

Sin embargo, realizar una actividad económica informal tiene grandes consecuencias en los salarios mensuales de las personas. Según lo arrojado por la estimación, pertenecer al mercado informal reduce

Cuadro 4: Estimación de la brecha salarial por género condicional

	<i>Variable dependiente:</i>	
	Logaritmo del salario mensual	
	(1)	(2)
Mujer	-0.139*** (0.015)	-0.144*** (0.011)
Edad		0.054*** (0.003)
Edad ²		-0.001*** (0.00004)
Educación superior		0.626*** (0.012)
Horas trabajo usual		0.012*** (0.0005)
Informal		-0.566*** (0.014)
Constante	14.087*** (0.010)	12.183*** (0.058)
Observaciones	9,892	9,892
R ²	0.009	0.444
R ² Ajustado	0.009	0.443
<i>Nota:</i>	*p<0.1; **p<0.05; ***p<0.01	

en un 64.79 % el ingreso mensual en promedio de las personas, controlando por otra serie de variables. Este resultado es bastante preocupante y triste, pues la mayoría de las personas ubicadas en este sector presentan condiciones de vulnerabilidad, de manera que, la brecha en la desigualdad de los ingresos es todavía mas grande.

Cuadro 5: Estimación de la brecha con el teorema de FWL

	<i>Variable dependiente:</i>	
	Logaritmo del salario mensual	FWL
	(1)	(2)
Mujer	-0.144*** (0.011)	
x_residuales		-0.144*** (0.011)
Controles	Si	No
Observaciones	9,892	9,892
R ²	0.444	0.016
R ² Ajustado	0.443	0.016
<i>Nota:</i>	*p<0.1; **p<0.05; ***p<0.01	

En el cuadro 5 se utiliza el teorema Frisch-Waugh-Lovell pues este nos permite “limpiar” el efecto de las variables de control. Dicho de otro modo, los coeficientes estimados del modelo con un subconjunto de variables explicativas no fueron afectados por la presencia o ausencia de otras variables explicativas en

el modelo. Por lo cual, se puede seguir apelando a que existe una brecha salarial, ya que no importa que controles se utilicen ya sea de características similares del trabajo o del puesto siempre va a existir la presencia de que ser mujer tiene un efecto negativo en las ganancias salariales mensuales.

Cuadro 6: Parámetros estimados por bootstrap

Estadísticos del Bootstrap			
	Original	Sesgo	Error Estándar
β_0	0	0,0001	0,0053
Mujer	-0,0144	-0,0001	0,0164

Lo anterior también se sustenta con el cuadro 6 (bootstrap), donde se hace uso del Bootstrap junto con el teorema Frish-Waugh-Lovell (FWL) con el fin de poder estimar la incertidumbre de los parámetros estadísticos resultantes del modelo y asimismo la sensibilidad de los modelos y poder identificar la importancia de las variables. Por consiguiente, lo que se observa es que los errores estándar no cambian de una forma significativa, lo cual es un buen resultado ya que, no se está cayendo en errores de sobre estimación y la inferencia estadística es acertada, es válido decir que el modelo es robusto y preciso.

Cuadro 7: Edad Pico usando bootstrap

	Edad pico	Intervalo de confianza	
		L.I	L.S
Mujeres	47	45,32	49,6
Hombres	52	49,25	55,83

Finalmente, el cuadro 7 muestra las edades pico para cada género junto con los sus intervalos de confianza respectivos. De esta manera, se puede decir que, como los límites no se sobreponen, las diferencias en las edades pico son significativas entre hombres y mujeres a un nivel de significancia del 5%. Intuitivamente, se puede inferir que la edad de los hombres es más elevada en el caso colombiano, debido a que los hombres tienen una edad de pensión más alta comparados con las mujeres, por ende, deben permanecer más en el mercado laboral. De igual manera se presenta gráficamente las edades y su relación con el salario por género en la Figura 3.

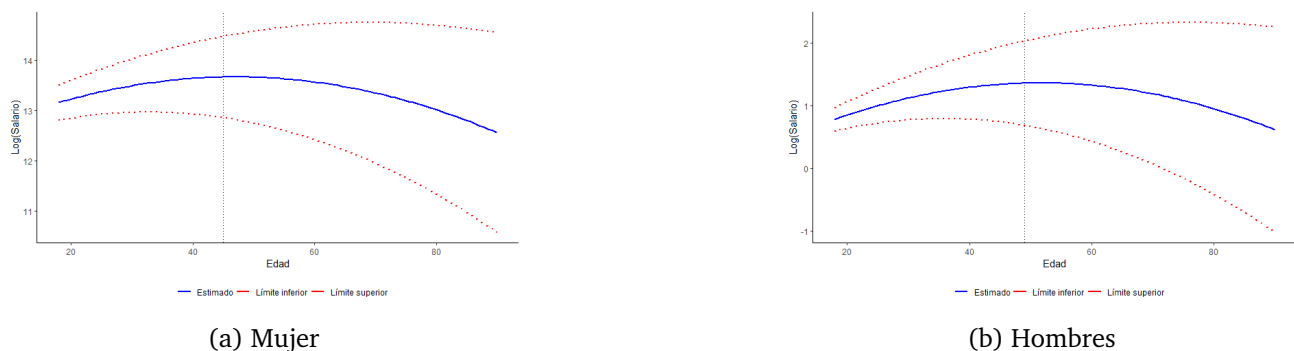


Figura 3: Relación salario-edad por género

5. Predicción de Salarios

Para evaluar el poder de predicción de los modelos anteriores y de los que se proponen en esta sección, primero se establece una semilla que permite la replicabilidad de los modelos estimados, y en este caso la muestra se divide aleatoriamente en dos partes, una parte correspondiente al 70 % y otra al 30 %, esto estableciendo que el 70 % es la muestra de entrenamiento y el 30 % es la muestra de testeo. Dado que los modelos usan la variable Oficio como predictor y que esta tiene grandes bondades al reconocer las diferencias salariales entre las distintas ocupaciones, el muestreo aleatorio se realizó de manera estratificada con el fin de garantizar que en las dos muestras existan individuos de cada ocupación y de esta manera poder estimar los parámetros de manera completa.

Los nuevos modelos estimados en esta sección son:

- Modelo 6:

$$\begin{aligned} \ln(\text{Salario}_i) = & \beta_0 + \beta_1 \text{Mujer}_i + \beta_2 \text{Mujer}_i \cdot \text{Edad}_i + \beta_3 \text{Mujer}_i \cdot \text{Edad}_i^2 + \beta_4 \text{Edad}_i + \beta_5 \text{Edad}_i^2 + \beta_6 \text{Superior}_i \\ & + \beta_7 \text{Horas_Trabajo}_i + \beta_8 \text{Informal}_i + \beta_9 \text{Media}_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{\text{Oficio}=w\}}] + \epsilon_i \quad (6) \end{aligned}$$

- Modelo 7:

$$\begin{aligned} \ln(\text{Salario}_i) = & \beta_0 + \beta_1 \text{Mujer}_i + \beta_2 \text{Mujer}_i \cdot \text{Edad}_i + \beta_3 \text{Mujer}_i \cdot \text{Edad}_i^2 + \beta_4 \text{Edad}_i + \beta_5 \text{Edad}_i^2 + \beta_6 \text{Superior}_i \\ & + \beta_7 \text{Horas_Trabajo}_i + \beta_8 \text{Informal}_i + \beta_9 \text{Media}_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{\text{Oficio}=w\}}] + \beta_{10} \text{Experiencia}_i + \epsilon_i \quad (7) \end{aligned}$$

- Modelo 8:

$$\begin{aligned} \ln(\text{Salario}_i) = & \beta_0 + \beta_1 \text{Mujer}_i + \beta_2 \text{Mujer}_i \cdot \text{Edad}_i + \beta_3 \text{Mujer}_i \cdot \text{Edad}_i^2 + \beta_4 \text{Edad}_i + \beta_5 \text{Edad}_i^2 + \beta_6 \text{Superior}_i \\ & + \beta_7 \text{Horas_Trabajo}_i + \beta_8 \text{Informal}_i + \beta_9 \text{Media}_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{\text{Oficio}=w\}}] + \beta_{10} \text{Experiencia}_i \\ & + \sum_{r=1}^6 \alpha_r [\mathbb{1}_{\{\text{Estrato}=r\}}] + \epsilon_i \quad (8) \end{aligned}$$

- Modelo 9:

$$\begin{aligned} \ln(\text{Salario}_i) = & \beta_0 + \beta_1 \text{Mujer}_i + \beta_2 \text{Mujer}_i \cdot \text{Edad}_i + \beta_3 \text{Mujer}_i \cdot \text{Edad}_i^2 + \beta_4 \text{Edad}_i + \beta_5 \text{Edad}_i^2 + \beta_6 \text{Superior}_i \\ & + \beta_7 \text{Horas_Trabajo}_i + \beta_8 \text{Informal}_i + \beta_9 \text{Media}_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{\text{Oficio}=w\}}] + \beta_{10} \text{Experiencia}_i \\ & + \sum_{r=1}^6 \alpha_r [\mathbb{1}_{\{\text{Estrato}=r\}}] + \beta_{11} \text{Experiencia}_i^2 + \epsilon_i \quad (9) \end{aligned}$$

- Modelo 10:

$$\begin{aligned}
Ln(Salario_i) = & \beta_0 + \beta_1 Mujer_i + \beta_2 Mujer_i \cdot Edad_i + \beta_3 Mujer_i \cdot Edad_i^2 + \beta_4 Edad_i + \beta_5 Edad_i^2 + \beta_6 Superior_i \\
& + \beta_7 Horas_Trabajo_i + \beta_8 Informal_i + \beta_9 Media_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{Oficio=w\}}] + \beta_{10} Experiencia_i \\
& + \sum_{r=1}^6 \alpha_r [\mathbb{1}_{\{Estrato=r\}}] + \beta_{11} Experiencia_i^2 + \beta_{12} Horas_Trabajo_i^2 + \epsilon_i \quad (10)
\end{aligned}$$

En este caso los nuevos modelos reconocen las no linealidades entre las distintas variables y su relación con el salario, como se muestra en el anexo en el cuadro 10, la experiencia en el actual empleo reportado y las horas trabajadas a la semana representan rendimientos marginales decrecientes frente al salario, algo que va muy acorde con la teoría económica.

En el cuadro 8, se presentan los MSE de los 10 modelos estimados, la columna uno muestra el MSE en la muestra de entrenamiento y en la columna dos se muestra el MSE en la muestra de testeo. Se evidencia que el MSE aumenta muy poco en la muestra de testeo con respecto a la de entrenamiento. El MSE va bajando a medida que aumentamos la complejidad del modelo, sin embargo, no se está realizando ningún sobre ajuste al aumentar la complejidad, ya que se tiene en cuenta la teoría económica a la hora de incluir predictores. El mejor modelo es el número 10, puesto que presenta el menor MSE tanto en entrenamiento como en testeo. Los parámetros de este modelo están contenidos en el anexo en el cuadro 10 en la columna 10, y los valores que toman tienen sentido económico puesto que reconoce que las relaciones cuadráticas de las variables que se eligieron para el ajuste presentan rendimientos marginales decrecientes.

Cuadro 8: MSE Entrenamiento y testeo

Modelo	MSE	
	Entrenamiento	Testeo
1	0,5349	0,5445
2	0,5051	0,5070
3	0,5300	0,5396
4	0,2213	0,2362
5	0,2213	0,2361
6	0,2197	0,2348
7	0,2128	0,2245
8	0,1873	0,1963
9	0,1869	0,1956
10	0,1770	0,1848

Notas: Cálculos propios usando datos DANE Medición de Pobreza Monetaria y Desigualdad 2018. La muestra de entrenamiento es el 70 % de la muestra total elegida aleatoriamente estratificando por la variable oficio.

En la Figura 4, se muestra la distribución de los valores observados del logaritmo del salario y los valores predichos del modelo 10, las líneas punteadas representan el percentil 5, 50 y 90 de izquierda a derecha respectivamente, el modelo es un buen predictor del logaritmo del salario en los percentiles inferiores de la distribución, salvo en el percentil uno que presenta una de las diferencias mas grandes entre el observado y el predicho (ver tabla XX), ya que la distribución del salario y su alta concentración en la

cercanía a la mediana el modelo esta estimando de manera correcta, sin embargo en la cola derecha de la distribución el modelo se queda corto a la hora de predecir, puesto que 6 de las 10 diferencias más grandes entre valores predichos y observados se encuentran ubicadas en percentiles mayores al 90 (ver tabla 11 del anexo). Dado que en las colas de la distribución el modelo presenta grandes diferencias con los datos observados, este modelo no funciona para predecir sobre los valores extremos de la distribución.

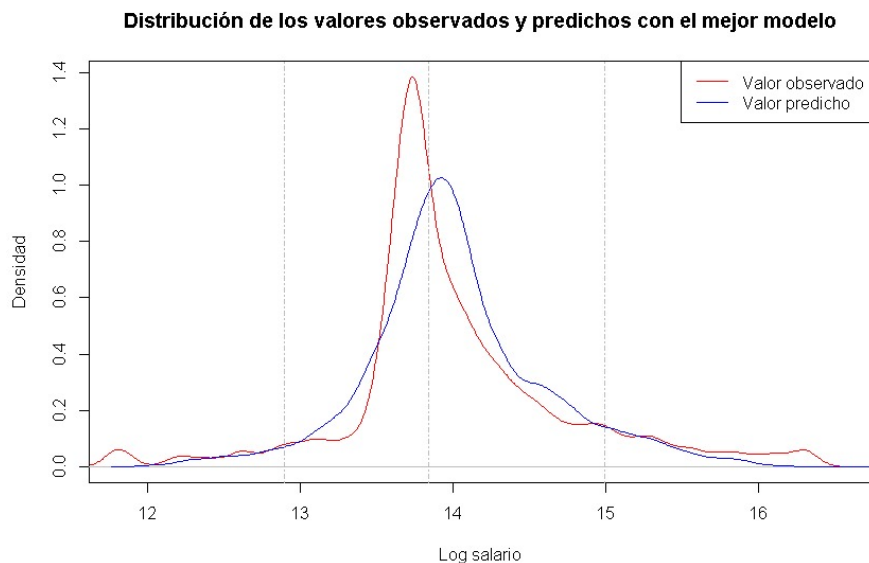


Figura 4: Distribución del logaritmo salario nominal mensual y sus valores predichos

Al estimar los dos mejores modelos con el algoritmo de validación LOOCV arroja un error referente a la variable oficio, dado que para algunas categorías de oficio se presentan individuos únicos. Por lo tanto, los modelos 9 y 10 se ajustan de tal manera de que la variable oficio no se incluya en la estimación. En la tabla 5 se presenta el MSE para los modelos 9 y 10 ajustados bajo la metodología de Cross-validation y LOOCV. Al comprar el MSE bajo las dos metodologías se encuentra que el modelo 10 sigue siendo el que mejor predice los datos; los MSE bajo las dos metodologías son muy parecidos. Así mismo, se observa que el MSE de ambos modelos aumento en comparación con la tabla 9 , lo que quiere decir que la variable oficio es una variable de gran importancia a la hora de estimar el salario.

Cuadro 9: MSE Entrenamiento, testeo y LOOCV

Modelo	MSE		
	Entrenamiento	Testeo	LOOCV
9-Ajustado	0,227	0,231	0,219
10-Ajustado	0,216	0,221	0,218

Notas: En el código el modelo 11 es equivalente al modelo 9 ajustado (sin la variable oficio) y el modelo 12 es equivalente al modelo 10 ajustado bajo la misma especificación.

6. GitHub

Mediante este link :
namedDarkOrchid=1 GitHub grupo 9

Referencias

- CLAVIJO, S. (2011): «Estructura fiscal de Colombia y ajustes requeridos (2010-2020),» .
- CORTES SABOGAL, M. Á. ET AL. (2016): «Nivel de educación y género: análisis de su impacto en la pobreza a partir de la descomposición del índice de pobreza monetaria en Colombia 2010-2015,» .
- DANE (2022): «Gran Encuesta Integrada de Hogares,» Tech. rep., Departamento Administrativo Nacional de Estadística - DANE.
- FLOREZ, L. A., L. A. MELO-BECERRA, AND C. E. POSADA (2021): «Estimating the reservation wage across city groups in Colombia: A stochastic frontier approach,» *Borradores de Economía*; No. 1163.
- GARCÍA OCAMPO, TANIA LORENA Y CASTELLANOS SABOGAL, Y. T. (2018): «Tributación y pobreza en Colombia: un análisis desde la evolución del impuesto de renta y el índice de pobreza monetaria,» *Revista Activos*, 16, 79–98.
- MINCER, J. (1974): «Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2.» .
- SERRANO-MALAYER, M. A. ET AL. (2018): «Modelos de desarrollo rural e incidencia en medición de pobreza monetaria: Una aproximación econométrica desde el modelo Logit para el sector rural año 2017,» .
- SOBRE SALARIOS, O. I. M. (2018): «2019, Que hay detrás de la brecha salarial de género,» *Ginebra: Organización Mundial del Trabajo*.

A. Anexo

Cuadro 10: Modelos de predicción - Entrenamiento

	Variable dependiente: $\text{Log}(\text{Salario})$				
	Modelo (6)	Modelo (7)	Modelo (8)	Modelo (9)	Modelo (10)
Mujer	-0.032 (0.109)	-0.073 (0.107)	-0.074 (0.101)	-0.055 (0.101)	0.003 (0.098)
Edad	0.035*** (0.004)	0.034*** (0.004)	0.036*** (0.004)	0.033*** (0.004)	0.033*** (0.004)
Edad ²	-0.0003*** (0.0001)	-0.0003*** (0.00005)	-0.0004*** (0.00005)	-0.0003*** (0.00005)	-0.0004*** (0.00005)
Educación superior	0.392*** (0.021)	0.375*** (0.020)	0.308*** (0.020)	0.309*** (0.020)	0.306*** (0.019)
Horas trab. a la semana	0.013*** (0.001)	0.012*** (0.0005)	0.012*** (0.0005)	0.012*** (0.0005)	0.041*** (0.002)
Informal	-0.456*** (0.016)	-0.426*** (0.016)	-0.418*** (0.015)	-0.414*** (0.015)	-0.335*** (0.015)
Educación media	0.123*** (0.018)	0.110*** (0.017)	0.096*** (0.016)	0.097*** (0.016)	0.087*** (0.016)
Experiencia trab. actual		0.001*** (0.0001)	0.001*** (0.0001)	0.002*** (0.0002)	0.002*** (0.0002)
Experiencia trab. actual ²				-0.00000*** (0.00000)	-0.00000*** (0.00000)
Horas trab. a la semana ²					-0.0003*** (0.00001)
Mujer·Edad	-0.002 (0.006)	0.001 (0.006)	0.002 (0.005)	0.001 (0.005)	-0.003 (0.005)
Mujer·Edad ²	0.00001 (0.0001)	-0.00004 (0.0001)	-0.0001 (0.0001)	-0.00004 (0.0001)	0.00000 (0.0001)
Constante	13.614*** (0.188)	13.592*** (0.185)	13.351*** (0.174)	13.394*** (0.174)	12.730*** (0.173)
Efectos fijos de oficio	Si	Si	Si	Si	Si
Efectos fijos de estrato	No	No	Si	Si	Si
R ²	0.589	0.602	0.650	0.651	0.669
R ² - Ajustado	0.584	0.597	0.645	0.646	0.665
Observaciones	6,925	6,925	6,925	6,925	6,925

Notas: Cálculos propios usando datos DANE Medición de Pobreza Monetaria y Desigualdad 2018. *** p<0.01, ** p<0.05, * p<0.1

Cuadro 11: Top 10 de diferencias entre observado y predicho

Top	Log(Salario)	Valor predicho	Indice de Leverage	Diferencia absoluta	Percentil
1	11,81	14,04	0,04	2,23	1
2	13,65	15,65	0,03	2,00	21
3	16,32	14,37	0,01	1,94	100
4	16,32	14,39	0,01	1,93	100
5	16,32	14,53	0,01	1,79	100
6	15,88	14,16	0,01	1,71	97
7	13,10	14,71	0,04	1,61	7
8	15,42	13,85	0,03	1,57	95
9	12,21	13,76	0,05	1,56	2
10	15,80	14,25	0,01	1,55	97

Notas: Cálculos propios usando datos DANE Medición de Pobreza Monetaria y Desigualdad 2018. Estimaciones con base en el modelo número 10.