PROBLEM SET 1: PREDICTING INCOME

GITHUB REPOSITORY: https://github.com/dfosorio111/BD-ML---PS1.git

Los ingresos que recibe el sector público son de suma importancia para llevar a cabo los planes de gasto del gobierno. En Colombia, la Dirección de Impuestos y Aduanas Nacionales (DIAN) es la entidad encargada de recoger los impuestos, que representan el principal flujo positivo de las finanzas públicas. Sin embargo, la DIAN se enfrenta a problemas de evasión de distinta índole. Cálculos del gobierno estiman que la cifra ascienda a $80 billones.

Una de las prácticas más comunes para evadir impuestos es el subreporte del ingreso por parte de las personas naturales. Por ese motivo, el presente trabajo busca construir un modelo predictor de ingresos que permita a la DIAN identificar a aquellos individuos que subreporten sus ingresos.

Los datos a utilizar provienen de la Gran Encuesta Integrada de Hogares (GEIH). La GEIH es una encuesta realizada mensualmente para medir el nivel de empleo en la economía. En este caso se cuenta con la GEIH anualizada para Bogotá en 2018. Se accedió a los datos por medio de *web-scrapping*, lo que requirió un esfuerzo para descargar, juntar y organizar los datos. Para eso se desarrolló un código que itera sobre la lista de urls donde se ubican los datos. El proceso contó con la dificultad de que los enlaces visibles no contenían los datos en un formato de tabla fácilmente importable, por lo que fue necesario inspeccionar la página y obtener el enlace de las tablas anexas.

Se lograron juntar 32,177 observaciones, de las cuales solo fueron utilizadas 16,542 luego de restringir la muestra a los mayores de edad ocupados. La primera problemática consistió en elegir la variable de ingreso a utilizar, pues la base de datos cuenta con distintas clases de ingreso según su fuente. En este punto fue necesario apoyarse en la literatura económica existente, que ha estudiado ampliamente los determinantes del ingreso.

Luego de revisar las distintas variables contenidas en la base de datos, decidimos escoger la variable de ingresos laborales (y_total_m) como medida de ingreso. La variable se construye tomando el salario para los empleados y las ganancias para los independientes en periodicidad mensual.

Varios factores llevaron a la escogencia de dicha variable como medida de ingreso sobre las otras opciones y específicamente sobre los ingresos totales, que resultaba la otra posible medida de ingreso. En primer lugar, el grueso de la literatura se concentra sobre los ingresos laborales al utilizar variables explicativas como la edad, la experiencia y el género. Con el objetivo de apoyarse en la literatura existente, los estudios pasados sobre el tema influenciaron la decisión.

La literatura económica ha encontrado una relación indiscutible entre los ingresos laborales y la edad, el sexo, el nivel educativo, por mencionar algunos. Además, el mecanismo que explica la relación es claro teóricamente. Sin embargo, la evidencia frente a los determinantes de otro tipo de ingresos es menos robusta.

Si bien la evidencia empírica resulta importante, la razón clave para escoger ingresos laborales y no ingresos totales tiene que ver con los datos disponibles en la GEIH. Al ser una encuesta de mercado laboral, la GEIH captura distintas variables de interés que pueden afectar los ingresos devengados por la actividad laboral. En contraste, la GEIH no captura determinantes del ingreso por capitales y dividendos, por ejemplo. Debido a que esa porción de los ingresos no se podrá explicar con los datos disponibles, resulta mejor excluir las fuentes de ingreso distintas a la laboral.

Finalmente, recordemos que la muestra fue restringida a los ocupados. Para casi todos los ocupados, a excepción de los más ricos, los ingresos laborales significan la principal fuente de ingresos. En contraste, para los desocupados las fuentes de ingresos son más diversificadas y responden a determinantes que no se podrían analizar por medio de la GEIH. Al concentrarnos en los ocupados, el ingreso laboral resulta la opción más atractiva.

Se presentan dos problemas con nuestra variable de ingreso. En primer lugar, la variable de ingresos laborales no es observada para algunas de las observaciones. Mientras que la muestra está compuesta por 16,542 observaciones, la variable de ingresos laborales solo tiene 14,764 valores, dejando a 1,778 observaciones sin ningún valor. Para los ejercicios de inferencia causal se eliminan las observaciones sin valores asignados en la variable de ingresos laborales con el objetivo de mantener inalterados los datos. No se prevé que eliminarlas sea problemático, pues tan solo representan alrededor del 10% de la muestra y luego de eliminarlas aún se tiene un número elevado de observaciones. Sin embargo, para los ejercicios de predicción se vuelven a tomar esas observaciones y el valor faltante es reemplazado por el ingreso laboral promedio de quienes comparten el mismo directorio (llave de vivienda). Para aquellas observaciones que permanecieron vacías luego del reemplazo se procedió a hacer un segundo reemplazo, tomando los valores de ingreso laboral promedio para quienes comparten un mismo nivel educativo. El otro posible problema del ingreso laboral como variable dependiente tiene que ver con la capacidad predictiva del modelo sobre los individuos de altos ingresos. En el tramo superior de la distribución el rubro más importante de los ingresos no deviene de la actividad laboral, sino de otras fuentes, lo que puede derivar en estimaciones erróneas. Se abordará dicho problema más a fondo al realizar los ejercicios de predicción.

Con el objetivo de comprender a fondo los datos con los que se cuentan se procede a realizar distintos gráficos y estadísticas descriptivas. A continuación, se observan las estadísticas para las variables de ingreso laboral y edad.

Estadísticas Descriptivas

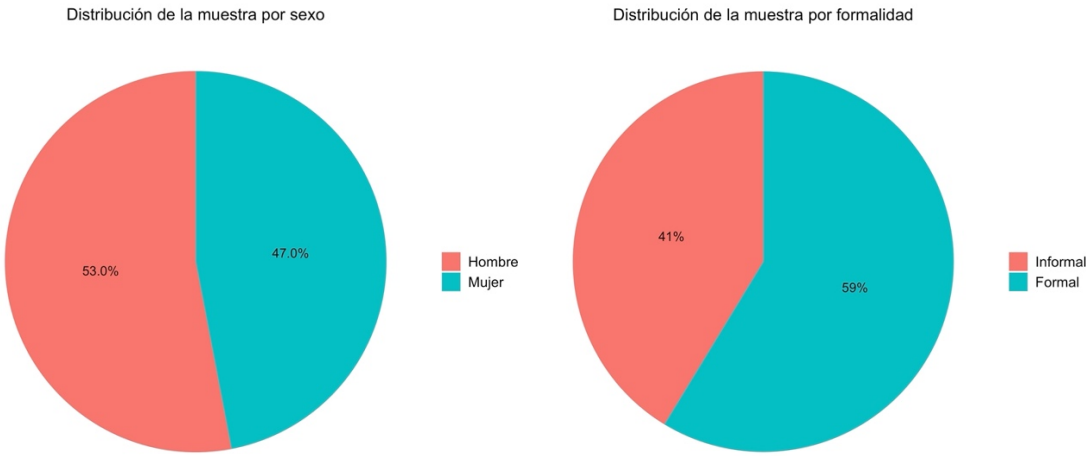
Estadístico	N	Promedio	Desv.Est.	Mín.	Máx.
Ingreso laboral	14,764	1,617,551	2,431,319	84	70,000,000
Edad	16,542	39.44	13.48	18	94

Como ya se mencionó, se tiene registro de ingreso laboral para 14,764 observaciones. El ingreso laboral mensual promedio es de $1,617,551 con una desviación que excede casi por un factor de dos al promedio. Lo anterior quiere decir que hay diferencias notables

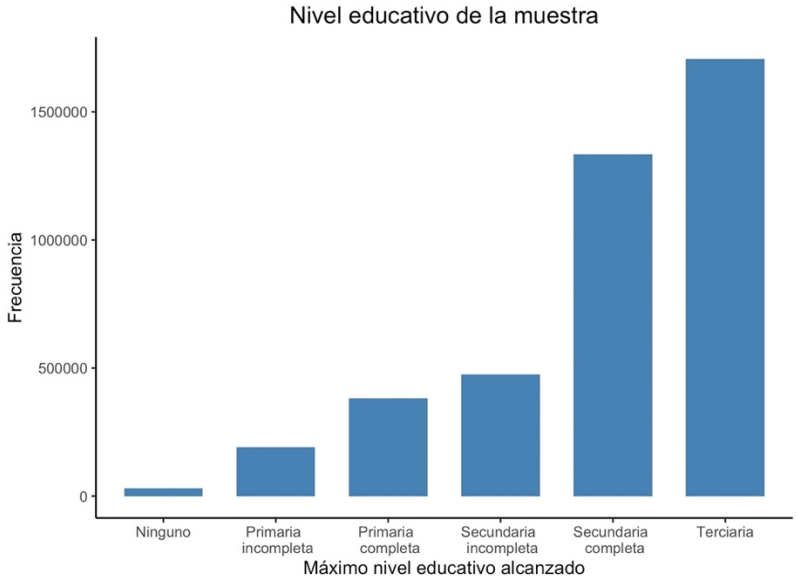
entre los salarios de la muestra. El individuo que menos ingreso laboral devenga recibe mensualmente $84, mientras que el que más dinero recibe gana $70 millones mensuales.

Para la variable de edad la muestra está completa, pues las 16,542 observaciones cuentan con un valor asignado. La edad promedio de la muestra es de 39.44 años, mientras que la desviación es de 13.48. Recordemos que la varianza de la edad se vio reducida de manera importante al ser uno de los filtros de la muestra ajustada que no incluye a quienes tienen menos de 18 años de edad, afectando la desviación estimada. Por ese motivo el individuo de menor edad en la muestra tiene 18 años, mientras que el mayor tiene 94 años.

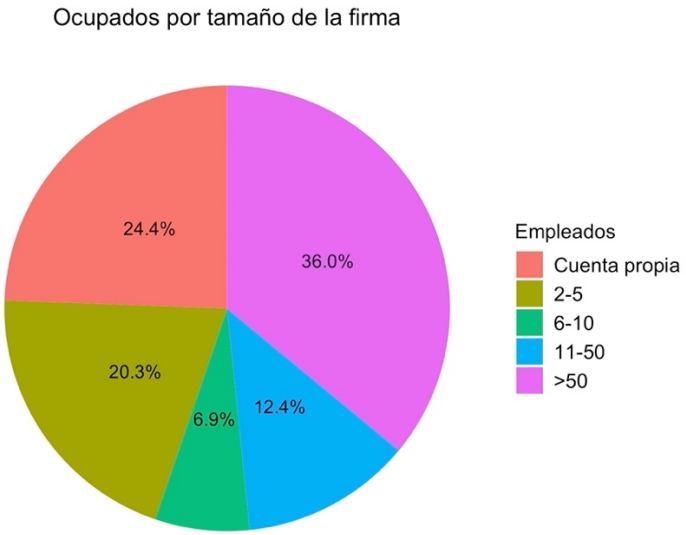
La muestra se encuentra relativamente balanceada tanto por sexo como por formalidad. El 53% de los individuos de la muestra son hombres, mientras que el 47% son mujeres. En cuanto al nivel de formalidad, 41% de los ocupados de la muestra trabajan de manera informal.



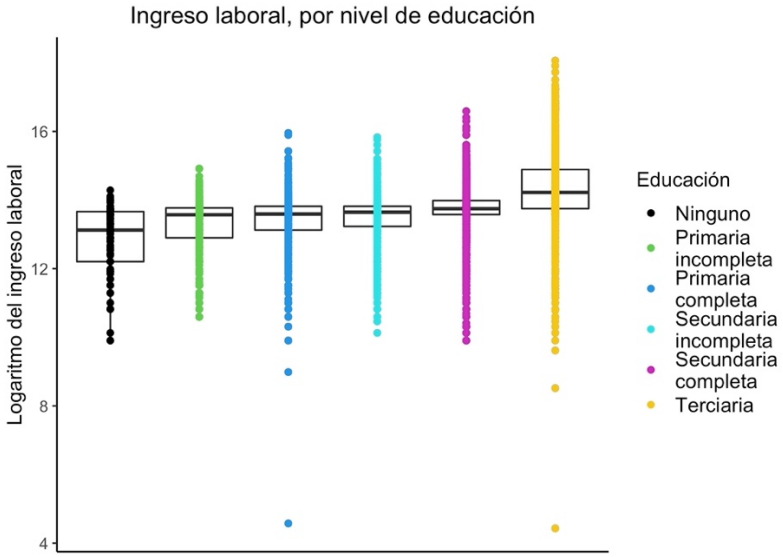
En cuanto al nivel educativo de los encuestados, el 42% de los individuos manifiesta haber completado estudios de educación terciaria. El segundo grupo más común lo componen quienes manifiestan haber completado la secundaria (31%). El grupo más reducido lo componen quienes no cuentan con ningún tipo de educación (0.7%).



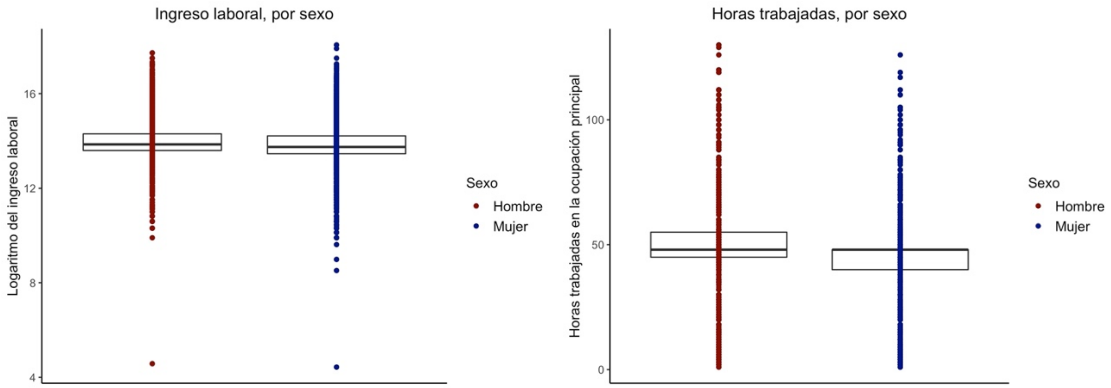
Los tamaños de las firmas en las que trabajan los individuos también dejan ver una muestra diversa. El 36% trabaja en una empresa de más de 50 empleados, mientras que el 24% trabaja por cuenta propia. Un poco más del 20% trabaja en firmas que tienen de 2 a 5 empleados.



Se observa que hay diferencias entre el ingreso laboral dependiendo del nivel de educación, lo que va en vía con la literatura existente. La mediana del ingreso laboral de quienes tienen educación terciaria es la más alta de todos los grupos. A medida que disminuye el nivel educativo alcanzado se observa que el ingreso laboral se reduce. Es importante resaltar la variabilidad de los datos, pues un gran número de observaciones se ubican por fuera de los dos cuartiles en el medio de la distribución.



Al desagregar las cifras de ingreso laboral por sexo se observa que la mediana del ingreso laboral de los hombres es superior a la mediana del ingreso laboral de las mujeres. En el gráfico también se observa una variación elevada en los datos. Es posible que la brecha de género en el ingreso laboral se deba al número de horas trabajadas. Como se observa en el gráfico de la derecha la mediana de horas trabajadas para los hombres es superior a la mediana de horas trabajadas para las mujeres. Es posible que las horas trabajadas expliquen, al menos en parte, la brecha de género en el ingreso laboral. Lo anterior va en vía con la literatura, que encuentra que las mujeres dedican más tiempo al cuidado del hogar y los hijos. La brecha de género en el ingreso laboral será explorada más adelante por medio de estimación econométricas.



Habiendo analizado los datos con los que cuenta la base, podemos adentrarnos en el análisis econométrico de la muestra, que nos permitirá hacer inferencia causal y predicciones sobre el ingreso laboral.

Uno de los determinantes clásicos del ingreso es la edad. La literatura económica ha encontrado en diversos estudios una curva en forma de U invertida para la trayectoria de los ingresos, compuesta por un primer tramo ascendente en donde a medida que aumenta la edad de un individuo aumenta su nivel de ingreso para llegar a un máximo, que es sujeto de debate y constantes estimaciones en la economía, y finalmente decrecer al final de la vida laboral.

El mecanismo que explica la curva es relativamente directo. Al entrar al mercado laboral los jóvenes devengan un ingreso reducido por su falta de experiencia. A medida que se van adquiriendo nuevos aprendizajes y más practica el ingreso va aumentado. Sin embargo, la relación no se mantiene lineal, pues en cierto punto el conocimiento de quienes llevan más tiempo en el mercado laboral se va haciendo en cierta medida obsoleto. A pesar de la experiencia, en cierto punto de la curva las empresas prefieren contratar jóvenes que adultos mayores. Lo anterior lleva a que al final del tramo de la curva los ingresos se vean reducidos.

El fenómeno descrito ha sido ampliamente estudiado en la literatura económica, que identifica la relación cuadrática prevista. Por ejemplo, Borjas (2016) [1] se menciona que los salarios de las personas tienden a ser bajos cuando las personas son más jóvenes, crecen a medida que las personas envejecen y llegan a un nivel máximo cerca a los 50 años, para luego mantenerse estables o decrecer. De manera similar, Murphy y Welch (1990) [2] encuentran una relación cuadrática del salario con los años de experiencia, con un pico salarial alrededor de los 30 años de experiencia.

Si bien la forma de la curva está relativamente establecida, un punto de debate particular es el máximo de la curva, que resulta diferente según el tiempo y el lugar de recopilación de los datos. Por ese motivo, es particularmente valioso el ejercicio de estimar la curva de ingresos para Bogotá.

En la Figura 1 se muestra un gráfico construido con los datos de la Gran Encuesta Integrada de Hogares (GEIH) del DANE para Bogotá en el año 2018. Siguiendo la literatura antes mencionada se muestra en el eje X la edad como proxy de la experiencia laboral (Carvajal, Peeples, Popovici y Rabionet, 2021) y en el eje Y el logaritmo del ingreso laboral (Murphy y Welch, 1990).

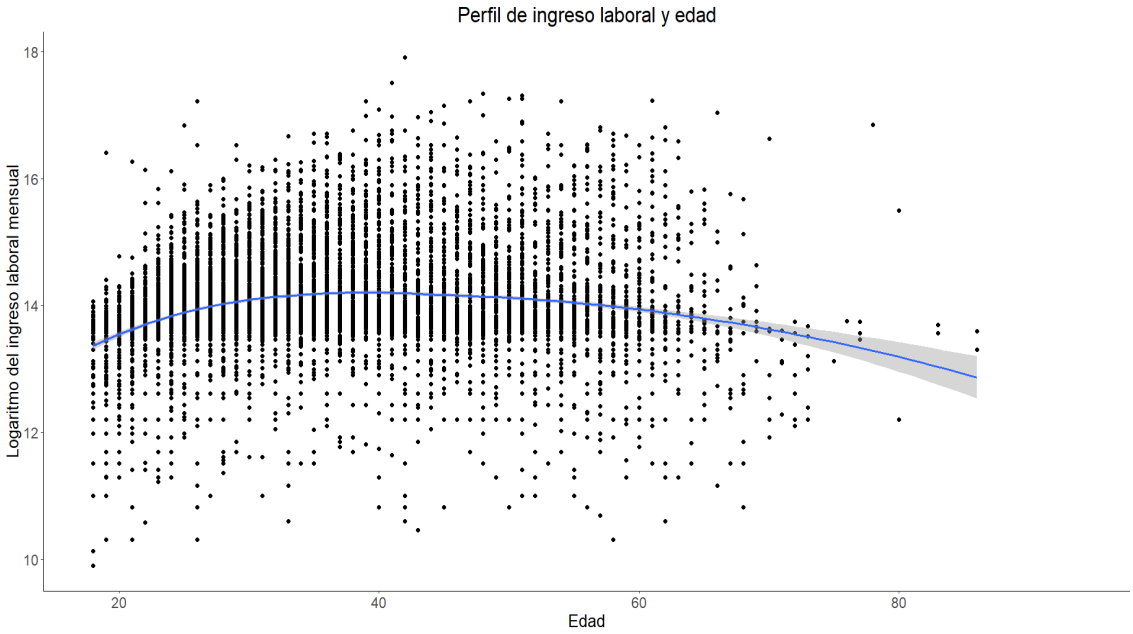


Figura 1: Relación entre el logaritmo del ingreso laboral y la edad a partir de los datos obtenidos de la GEIH.

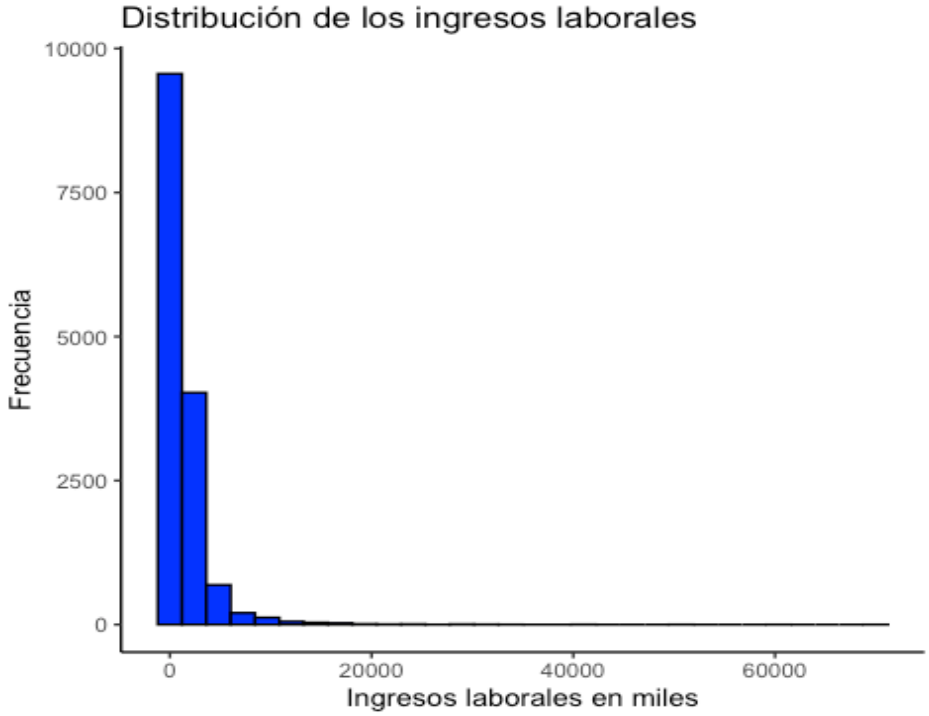
Se puede observar la relación cuadrática esperada entre la edad de las personas y su salario laboral con una forma de curva cóncava. Sin embargo, al menos gráficamente parece que

en Bogotá la edad en la cual se obtiene el pico de salario es menor a los 50 años de vida o 30 años de experiencia.

Teniendo en cuenta la relación cuadrática se plantea la siguiente ecuación, que será la primera estimación empírica del presente trabajo. La intención es esclarecer cuál es la edad en la que el ingreso laboral del individuo se maximiza.

$$\ln\left(Ingreso\ laboral\right)_i = \beta_1 + \beta_2 Edad_i + \beta_3 Edad_i^2 + \mu_i \qquad (1)$$

El modelo semi-logarítmico cuenta con dos ventajas que ameritan el ajuste en la variable de ingresos laborales. Primero, la variable dependiente tiene una distribución que se aleja de la distribución normal al acumular un gran número de observaciones en la cola izquierda, como se observa en el siguiente gráfico.



La transformación logarítmica permite normalizar la distribución de la variable en cuestión. Segundo, al aplicarle logaritmo a la variable dependiente cambia la interpretación de los coeficientes obtenidos en la estimación. Al ser un modelo *log-lin*, el estimador corresponde a una semi-elasticidad, que brinda la posibilidad de analizar el cambio porcentual del ingreso ante un año más de edad.

El modelo fue estimado por medio de la metodología de Mínimos Cuadrados Ordinarios (MCO) utilizando *bootstrap* para calcular los errores.

Tabla: Ingresos frente a la edad

	Variable dependiente:
	log(Ingresos laborales)
Edad	0.087^{***}

	(0.003)
$Edad^2$	-0.001***
	(0.00004)
Constante	12.288***
	(0.073)
Observaciones	14,764
R^2	0.052
Nota:	*p<0.1; **p<0.05; ***p<0.01

El intercepto arroja un valor de 12.29 y teóricamente ese sería el ingreso de una persona con 0 años de edad. Los signos resultantes para los parámetros de edad van en vía con lo predicho con la literatura, indicando una curva en forma de U invertida para los ingresos laborales frente a la edad. Debido a que para el modelo solo se realizó la transformación logarítmica en la variable dependiente, los parámetros no indican un efecto marginal o una elasticidad, sino una semi-elasticidad, que se puede obtener por medio de la derivación de la ecuación (1) y el reemplazo de los parámetros.

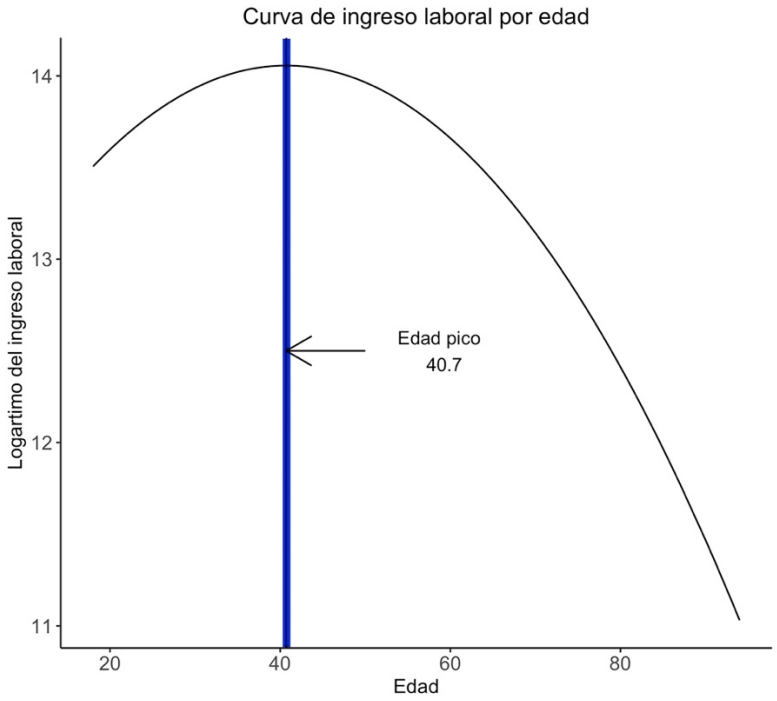
$$\frac{\partial \log (Ingresos\ laborales)}{\partial Edad} \frac{1}{\log (\ Ingresos\ laborales)} : \beta_2 + 2\beta_3 Edad = 0$$

Al mantener una relación cuadrática, la semi-elasticidad va a depender del tramo de la curva de edad sobre el cual se encuentre el individuo. Se realizó la estimación para un individuo de 39 años, correspondiente a una persona de edad promedio en la muestra, obteniendo una semi-elasticidad de 0.009. Lo anterior quiere decir que el modelo predice un incremento en el ingreso laboral del individuo entre los 39 y 40 años de 0.9% manteniendo lo demás constante.

El modelo se realizó con 14,764 observaciones, correspondiente al 73.4% de la muestra. La pérdida de observaciones debido a los valores faltantes para algunos individuos no será problemática para la estimación pues aún se cuenta con un número elevado de observaciones.

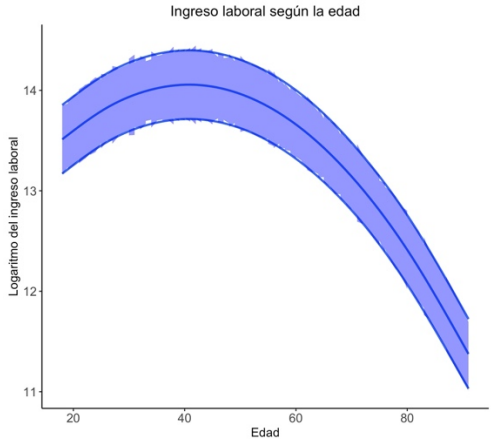
El ajuste del modelo dentro de la muestra es relativamente bajo. Como se observa en la tabla el R^2 obtiene un valor de tan solo 0.052, indicando que 5.2% de la variabilidad en salarios se explica por la edad. El bajo nivel de ajuste indica que hay variables importantes que no están incluidas en el modelo y determinan en gran medida los ingresos laborales de un individuo, sin dejar de ser relevante la edad. Algunos de los factores que podrían estar afectando la variabilidad de los ingresos laborales pueden ser el nivel educativo, la relación laboral que mantiene con el empleador, la formalidad, entre otras.

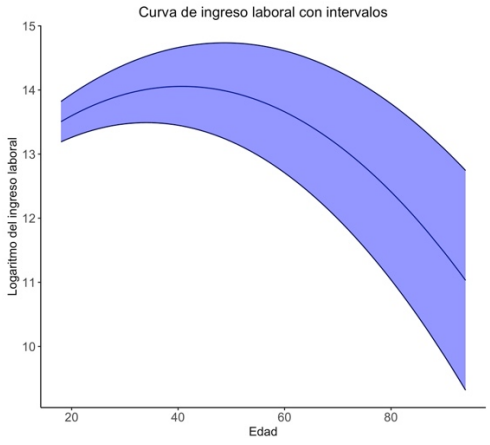
Con la estimación del modelo es posible construir los valores de ingreso laboral predicho para cada individuo y de esa manera dibujar la curva etaria de ingresos laborales para la muestra de nuestro modelo. A continuación, se resaltan los resultados del procedimiento.



Los resultado de la curva van en vía con la literatura estudiada sobre la determinación de ingresos laborales. En un primer tramo el ingreso laboral asciende a medida que un individuo envejece. La experiencia es el mecanismo más directo por el cual se da la relación positiva. El ingreso laboral va aumentando en menor medida por cada año para llegar a un punto de inflexión en los 40.7 años, indicando que esa es la edad que maximiza el ingreso laboral de un individuo. La zona resaltada en el gráfico indica el intervalo de confianza de esa edad máxima con una significancia del 5%. Se puede decir con un nivel de confianza del 95% que el ingreso laboral máximo de un individuo se da cuando este tiene entre 40.2 y 41.3 años. Luego de ese punto se observa un decrecimiento sostenido en los ingresos laborales, que puede deberse a la obsolescencia de los conocimientos con el tiempo y la entrada de nueva competencia con capacidades más demandas al mercado laboral. En el gráfico la pendiente se torna negativa y el decrecimiento en el ingreso laboral se va acelerando a medida que aumenta la edad.

Se debe resaltar que la estimación es sujeta a un término de error que captura variaciones en el ingreso que no dependen de la edad del individuo. Por ese motivo se realizaron dos gráficos que resaltan los intervalos de confianza dentro de los cuales, con el 95% de probabilidad, se ubica el ingreso de un individuo para una edad dada.





El gráfico de la izquierda se realizó al obtener los intervalos de confianza del ingreso para cada individuo, mientras que para dibujar el gráfico de la derecha fue necesario encontrar los errores de cada estimador para predecir un valor mínimo y máximo de ingreso para todas las edades. En el primer gráfico se observa una mayor varianza al inicio de la distribución, siendo el tramo de 20 a 40 año el más impredecible de la muestra. En el segundo gráfico se percibe una mayor varianza a medida que aumenta la edad.

Otro de los temas de suma relevancia en la literatura del mercado laboral es la igualdad de salarios para trabajos iguales. Existe una cantidad importante de literatura que discute las brechas salariales que existen entre hombres y mujeres, la cual muestra una discriminación del mercado laboral en contra de las mujeres, bajo la premisa de que personas igual de capaces y de preparadas, que desempeñan las mismas labores tienen salarios distintos sencillamente por su sexo.

En primer lugar, se presenta que, para el caso de la ciudad de Bogotá, de acuerdo con la GEIH del año 2018, en promedio las mujeres cuentan con ingresos menores que los hombres como se puede notar en la tabla de abajo. Para determinar lo anterior, se elaboró el siguiente modelo de regresión lineal simple que evalúa de forma incondicional la diferencia de salarios laborales promedio entre hombres y mujeres.

Tabla: ingresos y brecha de genero incondicional[1]

	Variable dependiente:
	Logaritmo del ingreso laboral
β_2	-0.240***
	(0.015)
β_1	13.098***
	(0.010)
Obs	14,764
R^2	0.018
F estadístico	270.200*** (df = 1; 14762)
Nota:	*p<0.1; **p<0.05; ***p<0.01

$$\ln(y_i) = \beta_1 + \beta_2 F_i + \mu_i$$

La tabla anterior muestra un ajuste del modelo a la muestra de 0,018 medido por el R^2 a la vez que el F estadístico confirma que al menos uno de los estimadores es distinto de 0 con un nivel de confianza del 99%.

[1] Errores estándar se encuentran entre paréntesis.

En la expresión anterior, y_i representa el ingreso laboral de la persona i y F_i hace referencia a si la persona i es una mujer. Por su parte, β_1 retorna el valor promedio del logaritmo del salario de los hombres, mientras que β_2 representa la diferencia porcentual entre el salario promedio de los hombres y el salario promedio de las mujeres.

Como se puede notar, el promedio del logaritmo del salario de los hombres es igual a 13,098, el cual es estadísticamente distinto de 0 bajo una significancia del 1%. Por su parte, las mujeres tienen un salario promedio 24% inferior al de los hombres, lo cual es estadísticamente significativo bajo un nivel de confianza del 99%.

La muestra antes evaluada se conformó por aquellas personas ocupadas, mayores de 18 años. Se eliminaron las observaciones que no contaban con reporte de ingresos laborales para respetar los datos originales y no incluir algún sesgo o error de medición que pueda afectar la inferencia causal. En consecuencia, todos los resultados aquí reportados son válidos para aquellas personas dispuestas a reportar sus ingresos laborales.

Es importante aclarar que previo a la eliminación de las observaciones con datos no reportados, la distribución de la muestra contenía un 46,6% de mujeres, mientras que una vez se eliminan las observaciones con datos faltantes las mujeres representaron el 46,9%. Por lo anterior, nos permitimos considerar que la distribución de genero se ve inalterada luego de la eliminación de observaciones con datos faltantes.

En la muestra utilizada, el salario promedio es de $1'623.822 COP y el salario mediano es de $996.120 COP. Por consiguiente, es claro que la diferencia promedio entre salarios de hombres y mujeres cuenta con magnitudes superiores a los $200.000 COP. Teniendo en cuenta que en 2018 el auxilio de transporte era de $88.211 COP y que la línea de pobreza extrema nacional era de $117.605 COP[2] es claro que se trata de una brecha salarial relevante.

Ahora bien, para entender los determinantes de la brecha de género, es necesario ir más allá que la diferencia no condicional. En específico, es importante resaltar que las brechas de género no son siempre las mismas y que se debe comprender su comportamiento a lo largo del ciclo de vida (Kleven et al, 2019; Toczek et al, 2021). Para este fin, se plantea el siguiente modelo de regresión lineal:

$$\ln(y_i) = \beta_1 + \beta_2 F_i + \beta_3 edad_i + \beta_4 edad_i^2 + \beta_5 edad_i * F_i + \beta_6 edad_i^2 * F_i + \epsilon_i$$

De acuerdo con Kleven et al (2019) existe una penalidad por los hijos que genera el incremento de la brecha de género. Es decir, al inicio de la vida laboral los salarios de hombres y mujeres son considerablemente similares, condicional a factores relevantes. Sin embargo, una vez se tiene un hijo o hija, la vida laboral del padre se ve poco o nada alterada, mientras que la de la madre cambia sustancialmente, pues empieza buscar trabajos más flexibles que les permitan cumplir con el cuidado de los hijos (los cuales son peor pagos y tienen menores probabilidades de ascensos). En este sentido, los autores reclaman un cambio de paradigma bajo el cual las brechas salariales son explicadas cada vez menos (aunque aún importa) por pagos distintos a personas que realizan el mismo trabajo y cada vez más por una carga desnivelada en el cuidado de los hijos. Por lo que

[2] DANE

poco a poco se hace más relevante el lema de "paridad en el cuidado de los hijos y del hogar" sobre el lema de "igualdad de salarios para trabajos iguales".

El modelo antes descrito permite separarlo en una ecuación para los hombres y una ecuación para las mujeres, como se muestra a continuación:

Modelo hombres:

$$\ln(y_i) = \beta_1 + \beta_3 edad_i + \beta_4 edad_i^2 + \epsilon_i$$

Modelo mujeres:

$$\ln(y_i) = (\beta_1 + \beta_2) + (\beta_3 + \beta_5) edad_i + (\beta_4 + \beta_6) edad_i^2 + \epsilon_i$$

Tabla: brecha de género y perfil ingresos-edad[3]

β_3	0.090	(0.005)
β_4	-0.001	(0.0001)
β_5	0.003	(0.008)
β_6	-0.0002	(0.0001)

En la tabla de arriba no se incluyen los estimadores β_1 y β_2 ya que hacen referencia al valor promedio del logaritmo del salario de personas recién nacidas, lo cual no tiene sentido interpretativo en el presente ejercicio, aunque más adelante serán usados en la comparación de modelos condicionados para determinar la brecha de género. Por su parte, para el análisis de los estimadores se realiza el proceso descrito a continuación. Se debe tener en cuenta que consiste en un modelo log-lin, por lo que los estimadores adquieren una interpretación de semi-elasticidades.

Para el caso de los hombres, un aumento en un año cuando la edad es igual a x, es equivalente a un aumento en $(\beta_3 + 2\beta_4 x) * 100\%$ del salario, ceteris paribus. Por su parte, para el caso de las mujeres, cuando una mujer tiene x años, el aumento en un año de su edad le implicará un crecimiento del $\big((\beta_3 + \beta_5) + (\beta_4 + \beta_6)x\big) * 100\%$ en el salario, manteniendo todo lo demás constante.

La tabla de abajo reporta los resultados de un *Bootstrap* por lo que los errores estándar son robustos. De esta forma, a continuación, se muestran los intervalos de confianza de todos los estimadores del modelo.

Estimador	**2.5%**	**Promedio**	**97.5%**
β_1	12.0361	12.2255	12.4142
β_2	-0.3471	-0.0518	0.2441
β_3	0.0802	0.0904	0.1005
β_4	-0.0012	-0.0010	-0.0009
β_5	-0.0124	0.0034	0.0190
β_6	-0.0004	-0.0002	-0.0001

[3] Errores estándar entre paréntesis

En la tabla anterior es posible ver que los estimadores $\beta_1, \beta_3, \beta_4$ y β_6 son estadísticamente distintos de 0 bajo una significancia del 5%. Por su parte, β_2 y β_5 no son estadísticamente distintos de 0.

De igual forma, es posible calcular la edad para la cual hombres y mujeres obtienen su ingreso máximo. Para esto se optimiza cada uno de los modelos antes descritos, pues, como se puede comprobar con el valor de los estimadores, se trata de funciones estrictamente cóncavas.

Para los hombres:

$$\frac{\partial y}{\partial x}\frac{1}{y} = \beta_3 + 2\beta_4 edad = 0$$

Entonces, como por construcción el salario (y) es estrictamente mayor a 0:

$$edad^* = \frac{-\beta_3}{2\beta_4} = 43{,}53$$

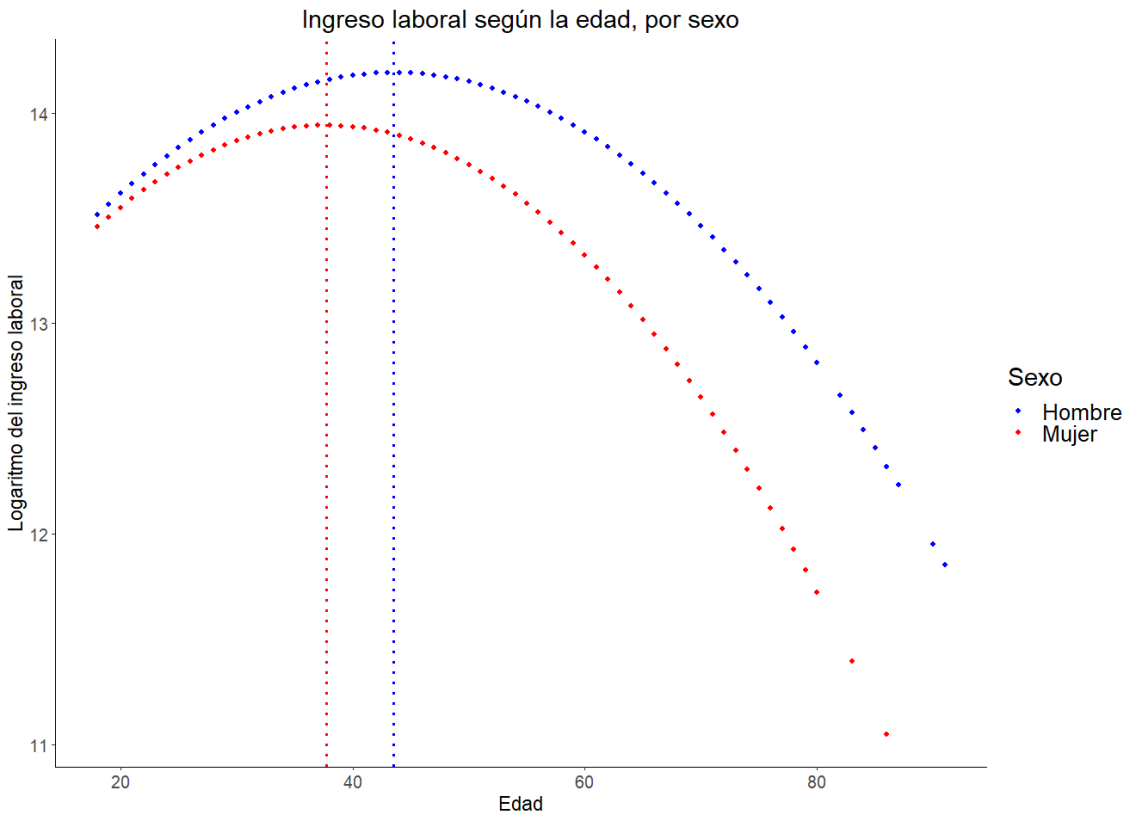
Mientras que, para las mujeres:

$$\frac{\partial y}{\partial x}\frac{1}{y} = (\beta_3 + \beta_5) + 2(\beta_4 + \beta_6) edad = 0$$

Entonces, como por construcción el salario (y) es estrictamente mayor a 0:

$$edad^* = \frac{-(\beta_3 + \beta_5)}{2(\beta_4 + \beta_6)} = 37{,}73$$

Al realizar el *Bootstrap* para calcular el intervalo de confianza de las edades en las que se encuentra el pico de salario por género se obtiene que para los hombres es de [42,77; 44,30] mientras que para las mujeres es de [36,94; 38,52]. Por lo anterior, bajo una significancia del 5%, las mujeres obtienen su salario máximo a una edad menor que los hombres.

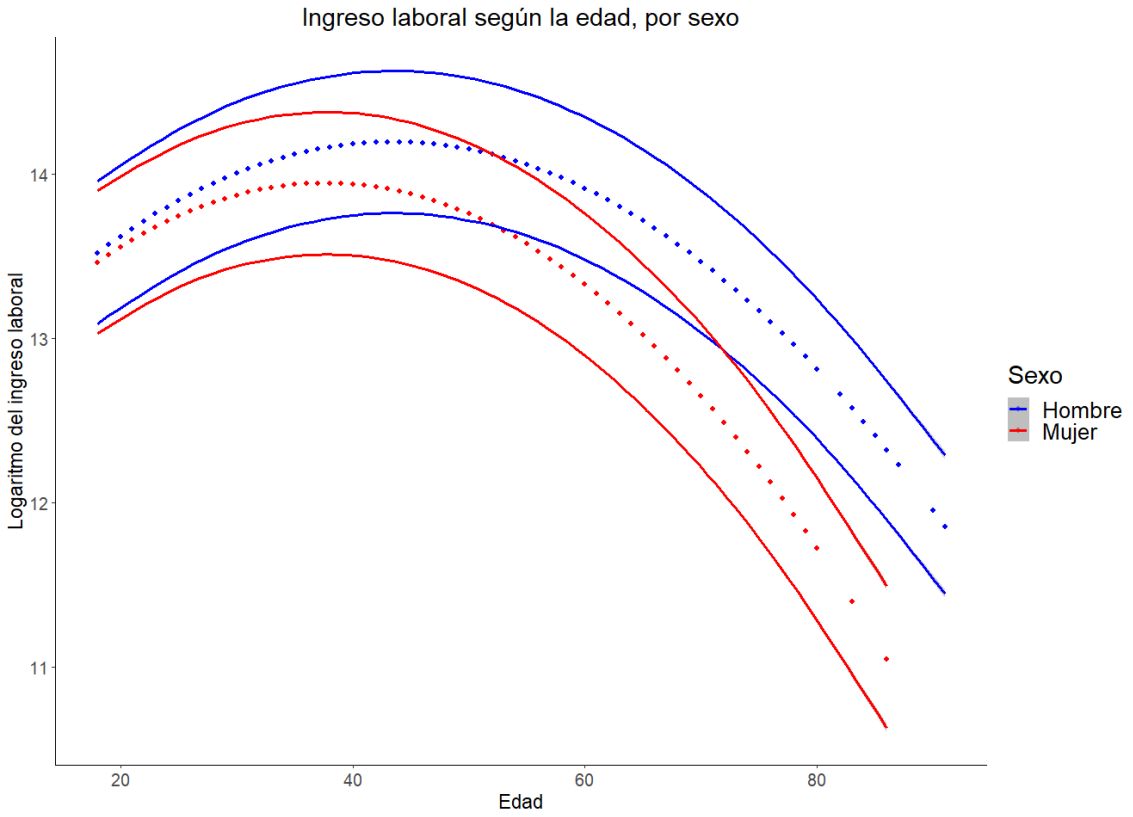


En el gráfico anterior se observa en rojo el valor predicho para el logaritmo del ingreso laboral de las mujeres y en color azul el de los hombres. De igual forma, en los mismos colores se traza una línea vertical que marca las edades en las cuales las personas de cada género obtienen el máximo valor en su ingreso laboral.

Con lo anterior, se puede notar que a medida que la edad de las personas aumenta, se hace más grande la brecha de género. Es importante mencionar que esto puede contemplar 2 efectos. En primer lugar, la ya mencionada penalidad por los hijos (Kleven et al, 2019) que profundiza las brechas salariales por las disparidades en los cuidados parentales. En segundo lugar, Kleven et al (2019) menciona que tiempo atrás, el peso de otros factores para explicar la brecha de género era mucho mayor, por lo que el lema “salarios iguales para trabajos iguales” puede que no tenga tanto peso para personas más jóvenes, pero sí un mayor impacto para las personas de mayor edad. Por consiguiente, se puede esperar que, al tratarse de una sección cruzada que evalúa en un mismo momento del tiempo personas de distintas edades y por tanto cosmovisiones y contextos diferentes, aquellas con mayor edad muestren una disparidad más marcada en los salarios entre hombres y mujeres.

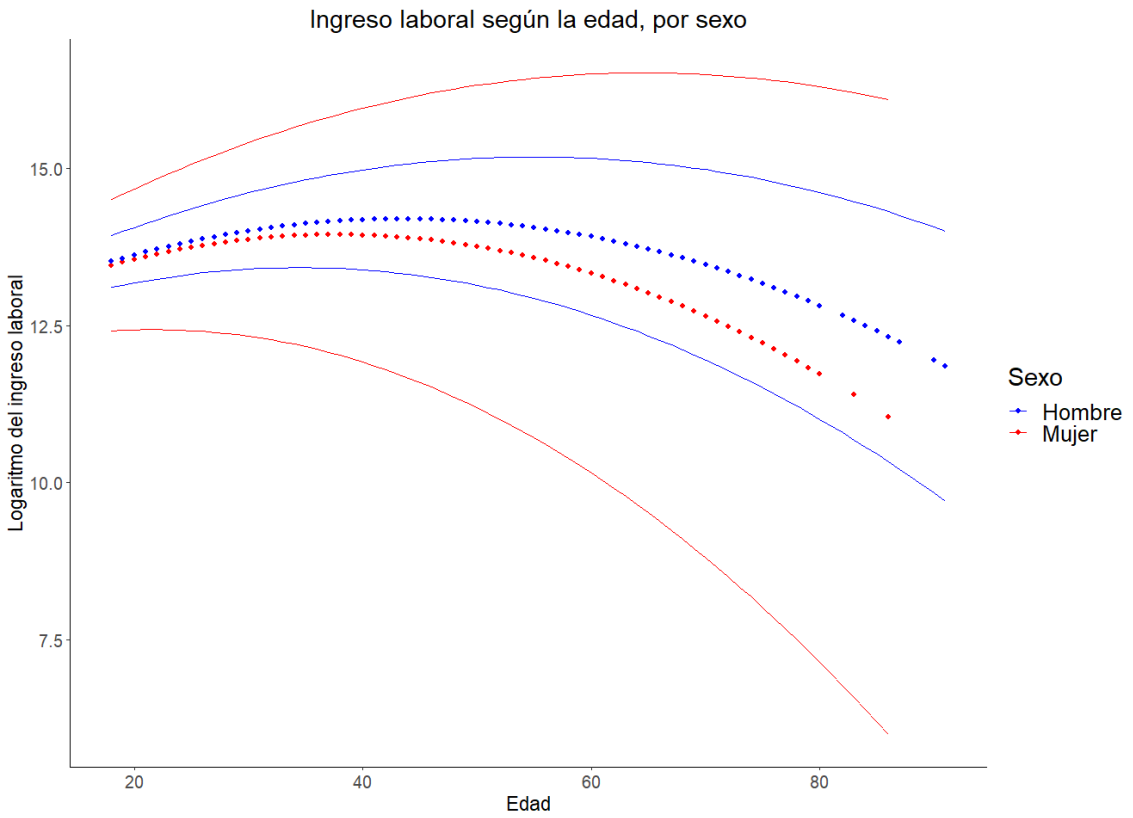
De hecho, de acuerdo con Toczek et al (2021) la brecha de género tiende a ser más alta para los empleados más viejos. Es decir, mientras que para los jóvenes las mayores desigualdades provienen de la penalidad de los hijos, para las personas mayores se contempla este mismo efecto y se suma un impacto grande de otros aspectos de discriminaciones culturales y sociales que implicaban dinámicas de brecha de género aún más profundas.

Note que para un hombre de 23 años[4], al momento de cumplir 24 años se espera que su ingreso laboral incremente en 4.26%, ceteris paribus, mientras que para las mujeres se espera un incremento del 3,66%. Por su parte, a los 38 años (cuando las mujeres alcanzan su nivel máximo de salario), los hombres esperan en promedio un aumento de su salario del 1,15% al aumentar un año de su edad, manteniendo todo lo demás constante, mientras que las mujeres contemplan una reducción esperada del 0,07%. Como se puede notar, los niveles de salario entre hombres y mujeres comienzan de formas considerablemente iguales, pero con el paso del tiempo los incrementos de salario son menores para las mujeres, llegan a su salario máximo a una edad más temprana y comienzan a decrecer sus ingresos laborales con una pendiente más pronunciada.



El gráfico anterior se construyó por medio de realizar un Bootstrap y guardar las predicciones para los distintos remuestreos, para así hallar la distribución de las predicciones y construir los intervalos de confianza que se muestran en la imagen. Como se observa, bajo un nivel de confianza del 90% no hay diferencia estadística hasta poco después de los 70 años en el nivel de ingreso según la edad para cada género.

[4] Para el DANE los jóvenes se ubican entre los 15 y los 28 años. Se escoge arbitrariamente 23 años como un promedio entre 18 (el mínimo de esta muestra) y 28.
https://www.dane.gov.co/files/investigaciones/boletines/ech/ech/pres_ext_empleo_jul_22.pdf



Por su parte, en este gráfico se calcularon como valores máximos aquellos con el límite superior de todos los estimadores y los valores mínimos con el límite inferior, bajo un nivel de confianza del 95%. Similar al caso anterior, no se presentan diferencias estadísticamente significativas para los ingresos de hombres y mujeres a lo largo del ciclo de vida. Empero, se debe resaltar que la literatura presentada y la significancia económica analizada en el presente documento permiten ir más allá de la conocida "búsqueda de estrellas" y plantear debates relevantes en torno a las brechas de género en el ingreso laboral y sus determinantes.

Anteriormente se añadió la edad para entender el comportamiento de la brecha de género a lo largo del ciclo de vida. No obstante, es necesario controlar por otros factores que puedan causar endogeneidad por su relación con la edad y el sexo de los individuos, a la vez que explican el ingreso laboral.

Por un lado, se incluye el nivel educativo, pues es conocido en la literatura que los ingresos laborales dependen del nivel de formación de los individuos, a la vez que existen diferencias sistemáticas entre el nivel educativo de los hombres y las mujeres (Kleven, 2019; Toczek et al, 2021; Popovici et al, 2021).

Por otra parte, el oficio que desempeñan las personas y la relación laboral con la que cuentan también varía entre los distintos sexos. Existen profesiones mayoritariamente femeninas, otras profesiones mayoritariamente masculinas y también hay una mayor cantidad de hombres en posiciones de mando. Por este motivo, resulta relevante incluir como controles variables que indiquen el estatus ocupacional, el tipo de trabajo desempeñado y el tamaño de la firma (Toczek et al, 2021; Popovici et al, 2021).

Empero, Kleven et al (2019) discuten que para poder capturar el efecto de la penalidad de los hijos no se debe incluir en los controles determinantes del ingreso que respondan directamente al nacimiento de un hijo, como lo son el tipo de oficio o el tipo de firma. Adicionalmente, un mayor número de horas trabajadas revela un mayor costo de oportunidad del ocio y por consiguiente un mayor salario (Borjas, 2016; Popovici et al, 2021). De igual forma, por las disparidades en los cuidados del hogar y de los hijos es esperable que las mujeres trabajen en promedio un menor número de horas, al buscar mayor flexibilidad (Kleven, 2019).

Por último, se incluyen controles adicionales que resultan relevantes para el contexto colombiano como lo son: la informalidad, medida por el acceso a la seguridad social[5], donde los trabajadores formales suelen contar en promedio con un mayor salario, así como la informalidad es más prevalente entre las mujeres y está asociada al trabajo por cuenta propia[6]; si el trabajador realiza una segunda actividad que le signifique ingresos laborales.

Regresión condicional

	Variable dependiente:		
	Logaritmo del ingreso laboral		
	Sin controles	Regresiones condicionales	
	(1)	(2)	(3)
Mujer	-0.051	-0.142	-0.184*
	(0.128)	(0.090)	(0.106)
Edad	0.090***	0.050***	0.072***
	(0.004)	(0.003)	(0.003)
Edad al cuadrado	-0.001***	-0.001***	-0.001***
	(0.00005)	(0.00003)	(0.00004)
Interacción Edad*Mujer	0.003	0.002	0.002
	(0.006)	(0.005)	(0.005)
Interacción Edad^2*Mujer	-0.0002***	-0.0001	-0.0001*
	(0.0001)	(0.0001)	(0.0001)
Controles de nivel educativo	No	Sí	Sí
Controles de tipo de oficio, relación laboral, tamaño de la firma y horas trabajadas	No	Sí	No
Controles de formalidad y segunda actividad	No	Sí	Sí
Observaciones	14,763	14,763	14,763
R^2	0.084	0.561	0.378
Note:			*p<0.1; **p<0.05; ***p<0.01

[5] Variable construida por Manuel Fernández.
[6] DANE y ANIF.

En la tabla anterior se observan 3 modelos. El primer modelo es idéntico al observado previamente en este texto, pero luego de remover la observación con información faltante del nivel educativo. Como se puede notar, el modelo 1 no incluye ningún control. Por su parte, los modelos 2 y 3 incorporan los controles antes enunciados. La diferencia entre los 2 últimos modelos recae en aquellas variables relacionadas con el tipo de trabajo, el tipo de relación laboral, el tamaño de la firma o el número de horas trabajadas, teniendo en cuenta el trabajo de Kleven et al (2019).

En consecuencia, el modelo de regresión con todos los controles incorporados se presenta a continuación:

$$\begin{aligned}\ln(y_i) = \beta_1 + \ & \beta_2 F_i + \ \beta_3 edad_i + \beta_4 edad_i^2 + \beta_5 edad_i * F_i + \beta_6 edad_i^2 * F_i \\ & + \gamma(NivelEduc_i) + \phi(Oficio_i) + \omega(Relab_i) \\ & + \psi(HorasTrabajadas_i) + \varsigma(SegundaActividad_i) + \rho(Formal_i) \\ & + \eta(TamañoFirma_i)\epsilon_i\end{aligned}$$

Es relevante notar que el ajuste del modelo en la muestra (medido con el R^2) aumenta con la adición de controles, al pasar de 0,084 en el primer modelo, a 0,378 en aquel que incluye algunos controles y 0,561 en el que incluye todos los controles antes mencionados.

Para determinar la diferencia de ingreso entre hombres y mujeres se debe tener en cuenta el siguiente estimador:

$$(\beta_2 + \beta_5 edad_i + \beta_6 edad_i^2) * 100\%$$

Por su parte, el efecto marginal para los hombres de un aumento en la edad en un año se calcula de la siguiente forma:

$$(\beta_3 + 2\beta_4 edad_i) * 100\%$$

Mientras que el de las mujeres está determinado por la siguiente ecuación:

$$\big((\beta_3 + \beta_5) + (\beta_4 + \beta_6)edad_i\big) * 100\%$$

Como ejemplo, se evalúa a continuación para individuos de 23 años (para evaluar en las poblaciones jóvenes) y para individuos de 39 años (que corresponde a la edad promedio de la muestra).

En primer lugar, el modelo (1) menciona que, para individuos de 23 años, las mujeres tienen un ingreso 8,21% menor a los hombres, manteniendo todo lo demás constante. En el modelo (2) la diferencia es del 12,78% y en el modelo 3 del 19,86%. Por su parte, el mismo análisis para las personas de 39 años indica un menor salario de las mujeres frente al salario de los hombres en un 23,08% para el primer modelo; 16,40% en el segundo modelo y 28,32% en el tercer modelo.

Ahora bien, mientras un hombre de 23 años espera que su salario aumente en 4,26% cuando cumpla 24 años, ceteris paribus, para las mujeres este aumento esperado es del 3,66%, para el primer modelo. En el segundo modelo las cifra para los hombres es un

aumento del 2,68% y para las mujeres del 2,57%. Por último, para el tercer modelo se cuenta con aumentos de 3,75% y del 3,41% respectivamente.

Por su parte, un hombre de 39 años espera que su salario aumente en promedio un 0,94% cuando cumpla 40, manteniendo todo lo demás constante, mientras que para las mujeres se espera una disminución del 0,32%, para el primer modelo. En el segundo modelo las cifra para los hombres es un aumento del 0,10% y para las mujeres un incremento del 0,07%. Por último, para el tercer modelo se cuenta con aumentos de 1,32% y del 0,60% respectivamente.

Anteriormente en el presente documento se mencionaron los intervalos de confianza para hombres y mujeres del modelo sin controles. En el caso del modelo 2 el intervalo de confianza de la edad en la que los hombres alcanzan su salario máximo corresponde a [47,35; 51,17] mientras que el de las mujeres se encuentra entre [43,27; 46,76]. Por su parte, el modelo 3, que excluye los controles de tipo de oficio, relación laboral, tamaño de la firma y horas trabajadas tiene un intervalo de confianza para los hombres de [46,20; 49,13] y para las mujeres de [41,33; 43,47]. Es posible notar que los 3 modelos concluyen que, bajo una significancia del 5% las mujeres llegan a su salario máximo a una menor edad que los hombres, pero la distancia es mucho mayor para el modelo (3) que para el modelo (2).

Es claro que al incluir controles se modifica la magnitud las brechas de género que el modelo (1) mostraba. No obstante, aún existen estas brechas y la diferencia durante el ciclo de vida laboral es notoria. En el modelo (2), donde se incluyen todos los controles mencionados, se evidencia que las brechas son considerablemente menores a las de aquel modelo sin controles. No obstante, en el modelo (3) las brechas parecen en algunos casos incluso aún más pronunciadas.

De forma intuitiva, los datos anteriores pueden implicar que, acorde con Kleven et al (2019), los resultados cambien cuando se incluyen controles de tipo de oficio, relación laboral, tamaño de la firma y horas trabajadas, pues estos en muchos casos resultan de decisiones individuales de las familias por disparidades en los cuidados de los hijos y por tanto están asociados directamente con la penalidad de los hijos. Es relevante por tanto mencionar que la principal debilidad del presente modelo es no poder contar con una variable que indique si la persona tiene hijos, pues es claro que para las mujeres el hecho de tener hijos disminuye el ingreso laboral.

El peso de estos controles muestra que, para personas con el mismo oficio, tamaño de firma, horas de trabajo promedio y relación laboral (tipo de ocupación) las brechas entre hombres y mujeres son considerablemente menores, por lo que con el paso del tiempo se ha logrado cumplir con el reclamo de “salarios iguales para trabajos iguales” pero aún queda un terreno amplio por recorrer para poder cumplir con un nuevo lema que reclame “cargas de cuidado de los hijos equitativas para madres y padres”.

Para calcular los estimadores antes presentados se hizo uso del teorema de Frish-Waugh-Lovell (FWL). Para comprenderlo, separe las variables de interés (edad, sexo y sus interacciones) en una matriz X y agrupe todas las variables de control en una matriz Z, tal que resulte el siguiente modelo:

$$Y = X\beta + Z\gamma + \mu$$

Luego, con la matriz de proyección de los controles P_z es posible construir la matriz aniquiladora:

$$P_z = Z(Z'Z)^{-1}Z'$$
$$M_z = (I - P_Z)$$

Con lo anterior es posible reescribir el modelo de mínimos cuadrados ordinarios como:

$$M_z Y = M_z X\beta + \epsilon$$

Note que:

$$M_z Y = Y - Z\gamma$$

Es decir, hace referencia a los errores asociados de una regresión de las observaciones de Y contra las variables de control.

Por su parte:

$$M_z X\beta = X\beta - P_Z X\beta = (X - P_Z X)\beta$$

No es difícil comprobar que la expresión anterior se refiere a los errores asociados a una regresión linead de las variables de interés contra las variables de control. Por lo que una regresión de los residuales del modelo que regresa la variable Y contra los controles, contra los residuales del modelo que regresa las variables de interés sobre los controles permite encontrar resultados numéricamente idénticos para β.[7] Es decir:

$$resyz = (resxz)\beta + \epsilon$$

A continuación, se muestra la tabla con los resultados encontrados luego de haber aplicado la metodología de remuestreo de *Bootstrap* cuyos errores estándar son robustos.

Regresión condicional (Bootstrap)[8]

	Variable dependiente:		
	Logaritmo del ingreso laboral		
	Sin controles	Regresiones condicionales	
	(1)	(2)	(3)
Mujer	-0.051	-0.142	-0.184

[7] En el código de anexo es posible comprobar que la regresión lineal con la función lm() encuentra resultados idénticos al proceso de FWL hecho de forma manual y también idénticos a los resultados obtenidos con la función felm() que hace uso de esta metodología para absorber los efectos fijos y hacer el modelos más eficiente computacionalmente. Se hace manual únicamente para aquellos estimadores que contienen la variable female.

[8] En el código es posible comprobar que el FWL manual retorna el mismo valor para el Bootstrap. Todos los procesos realizados para el análisis de género utilizaron una semilla igual a 1000

	(0.151)	(0.099)	(0.127)
Edad	0.090	0.050	0.072
	(0.005)	(0.003)	(0.004)
Edad al cuadrado	-0.001	-0.001	-0.001
	(0.00006)	(0.00004)	(0.00005)
Interacción Edad*Mujer	0.003	0.002	0.002
	(0.008)	(0.005)	(0.007)
Interacción Edad^2*Mujer	-0.0002	-0.0001	-0.0001
	(0.0001)	(0.0001)	(0.0001)
Controles de nivel educativo	No	Sí	Sí
Controles de tipo de oficio, relación laboral, tamaño de la firma y horas trabajadas	No	Sí	No
Controles de formalidad y segunda actividad	No	Sí	Sí

En seguida se propuso abordar el problema de la estimación de ingresos laborales desde un punto de vista predictivo, esto es mediante la utilización de técnicas de Aprendizaje Supervisado y entrenamientos de diferentes modelos lineales para lograr predecir el comportamiento continuo de la variable continua de ingreso. En contraste con la metodología anteriormente desarrollada, esta aproximación propone que a partir de los parámetros dados para un modelo determinado se realiza el entrenamiento o ajuste de parámetros para entonces predecir el comportamiento de nuevas observaciones a partir del modelo previamente ajustado. El problema de ML en regresión lineal sugiere entonces una aplicación de optimización, en este caso del Error Cuadrático Medio entre los valores predichos por el modelo entrenado y las observaciones obtenidas.

Para desarrollar lo anterior, inicialmente se desarrollaron los modelos de regresión utilizando los parámetros anteriormente descritos para el perfil de ingresos-edad, así como también el perfil de ingresos considerando la brecha de género de descritos teniendo en cuenta los controles. En el código fuente desarrollado se puede observar que para construir modelos de aprendizaje supervisado se divide la base con las observaciones de la Gran Encuesta Integrada de Hogares (GEIH) en base de entrenamiento (train-set) y la base de testeo (test-set) con el 70% y 30% de observaciones, respectivamente.

Las métricas de evaluación reportadas en el análisis predictivo son $R^2, RMSE\ y\ MAPE$. La correlación R^2 permite medir el ajuste de las predicciones con el modelo, es en últimas la medida de la proporción de variabilidad de Y que se puede expresar por el vector de predictores X, dicho por James et al (2021) y por ende es una medida de variabilidad. Se utilizó esta métrica de evaluación debido a la naturaleza lineal del modelo, ya que desde el punto de vista predictivo construir un modelo de regresión lineal asume una relación lineal a priori entre los diferentes predictores de la base y la observación a predecir, en este caso nos permite medir el nivel de ajuste del modelo para predecir la observación de ingresos, a partir de los diferentes predictores, controles, interacciones, etc.

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

Donde se tiene que
$TSS = \sum(y_i - \bar{y})^2$ suma de los cuadrados
$RSS = \sum(y_i - \hat{y}_i)^2$ suma de los errores al cuadrado

Por otro lado, el Error Cuadrático Medio MSE es utilizado porque para modelos de ML, en particular los modelos de regresión lineal donde no intervienen otros factores como la función de activación de verosimilitud y diferentes modelos como en regresión logística, el problema de aprendizaje supervisado deriva en la optimización de la función del MSE tal que permita minimizar la función del error.

$$MSE = \frac{RSS}{n} = \frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2$$

Finalmente, la métrica de Error Porcentual Absoluto Medio (MAPE) mide el error en términos porcentuales de manera intuitiva y de interpretación directa, razón por la cual es reportada en el informe. En la sección de predicción presentada a continuación la observación y_i a predecir corresponde a los ingresos totales mensuales reportados por la base de datos de la GEIH. La variable para predecir es $y_i = \text{y\_total\_m}$.

Para la especificación 1 se tuvo en cuenta los parámetros de edad, $edad^2$, mujer y los controles tenidos en cuenta en el modelo (2) para el perfil de brecha salarial de género anteriormente descrita, la especificación 2 es el logaritmo de los ingresos totales mensuales Como se observa en la tabla los resultados para las métricas de evaluación de predicción para la especificación 1 en el test-set son $R^2 = 0.204607, RMSE = 1783646, \text{MAPE} = 1.059662$ mientras que para la especificación 2 son $R^2 = 0{,}427828, RMSE = 0{,}653602, \text{MAPE} = 0{,}034417$.

Función de especificación 1

$$\begin{aligned} y_i = \beta_1 + \beta_2 F_i + \beta_3 edad_i + \beta_4 edad_i^2 + \beta_5 edad_i * F_i + \beta_6 edad_i^2 * F_i \\ + \gamma(NivelEduc_i) + \omega(Relab_i) + \psi(HorasTrabajadas_i) \\ + \varsigma(SegundaActividad_i) + \rho(Formal_i) + \eta(TamañoFirma_i)\epsilon_i \end{aligned}$$

Función de especificación 2

$$\begin{aligned} \ln y_i = \beta_1 + \beta_2 F_i + \beta_3 edad_i + \beta_4 edad_i^2 + \beta_5 edad_i * F_i + \beta_6 edad_i^2 * F_i \\ + \gamma(NivelEduc_i) + \omega(Relab_i) + \psi(HorasTrabajadas_i) \\ + \varsigma(SegundaActividad_i) + \rho(Formal_i) + \eta(TamañoFirma_i)\epsilon_i \end{aligned}$$

Para la especificación 3 se tuvo en cuenta los parámetros de edad, $edad^2$, mujer y los controles tenidos en cuenta en el modelo (3), es decir sin contar los controles de la relación laboral, las horas trabajadas y el tamaño de la firma para contrastar los resultados de los modelos obtenidos en el análisis inferencial anteriormente reportado para el perfil de

brecha salarial de género anteriormente descrita. Como se reporta en los resultados las métricas de evaluación de predicción para la especificación 3 en el test-set son $R^2 = 0.204607, RMSE = 1783646, \text{MAPE} = 1.059662$ mientras que para la especificación 2 son $R^2 = 0{,}347942, RMSE = 0{,}697739, \text{MAPE} = 0{,}036327$, mientras que en la estimación de parámetros por inferencia desarrollado anterior es $R^2 = 0.378$. Se concluye que el modelo predictivo se acerca al performance obtenido, más sin embargo el ajuste es mejor mediante la estimación por los diferentes métodos presentados anteriormente.

Función de especificación 3

$$\ln y_i = \beta_1 + \beta_2 F_i + \beta_3 edad_i + \beta_4 edad_i^2 + \beta_5 edad_i * F_i + \beta_6 edad_i^2 * F_i + \gamma(NivelEduc_i) + \varsigma(SegundaActividad_i) + \rho(Formal_i)$$

Para la especificación 4 se tuvo en cuenta los parámetros de edad, $edad^2$ sin controles ya que se busca construir un modelo predictivo con los parámetros de edad y edad al cuadrado como es reportado por diversos autores entre otros Murphy y Welch (1990), es decir sin contar los controles de la relación laboral, las horas trabajadas y el tamaño de la firma. Como se reporta en los resultados las métricas de evaluación de predicción para la especificación 3 en el test-set son $R^2 = 0{,}033196, RMSE = 0{,}849609, \text{MAPE} = 0{,}043231$. Se observa que el modelo predictivo en comparación con las especificaciones anteriores presenta una disminución drástica en la medida de ajuste, así como también un aumento significativo en el Error Cuadrático Medio debido a la ausencia de controles sobre el modelo y las interacciones para asegurar semi-elasticidad del modelo $log - lin$

Función de especificación 4

$$\ln y_i = \beta_1 + \beta_2 Edad_i + \beta_3 Edad_i^2 + \epsilon_i$$

Para corregir lo anterior, la especificación 5 que modela las interacciones entre la variable que representa el género y la edad fue incluida para evaluar el desempeño del modelo para estas interacciones, en ausencia de controles. Los resultados obtenidos en desempeño son $R^2 = 0{,}057431, \text{RMSE} = 0{,}838893, \text{MAPE} = 0{,}042783$. Se puede observar que en relación a la especificación 4 que no contempla las interacciones que implica la semi elasticidad del modelo, se presenta un mejoramiento en el ajuste del modelo $0{,}033196 < R^2 < 0{,}057431$, mientras que el Error Cuadrático Medio desmejora marginalmente.

Función de especificación 5

$$\ln(y_i) = \beta_1 + \beta_2 F_i + \beta_3 edad_i + \beta_4 edad_i^2 + \beta_5 edad_i * F_i + \beta_6 edad_i^2 * F_i + \epsilon_i$$

Para concluir el contraste y comparación de los modelos estimados mediante las técnicas de Bootstraping, la aplicación del Teorema de Frish-Waugh-Lovell (FWL) y diferentes estrategias en el análisis inferencial con los modelos de predicción supervisado se realizó la especificación de control con un único estimador mujer (female) haciendo referencia

al sexo. Los resultados en desempeño obtenidos son $R^2 = 0{,}01633, \text{RMSE} = 0{,}856988$, $\text{MAPE} = 0{,}043364$. Es relación con las especificaciones reportadas en el documento es evidente que el ajuste obtenido por la regresión es cercano al 1.6% siendo el menos ajustado, del mismo modo el Error Cuadrático Medio es alto en relación con las especificaciones 2-10.

Función de especificación 6

$$\log y_i = \beta_1 + \beta_2 F_i$$

A continuación, se procedió a realizar el estudio de las relaciones entre los controles dispuestos en la GEIH, interacciones y no linealidades propuestas a partir de un análisis descriptivo de las variables propuestas.

Modelo	Muestra	R2_Score	RMSE	MAPE
Modelo 1	Train-set	0,173903	2267685	3,834311
	Test-set	**0,204607**	**1783646**	**1,059662**
Modelo 2	Train-set	**0,428339**	**0,675582**	**0,035431**
	Test-set	**0,427828**	**0,653602**	**0,034417**
Modelo 3	Train-set	0,34307	0,724217	0,037579
	Test-set	0,347942	0,697739	0,036327
Modelo 4	Train-set	0,040181	0,875394	0,044235
	Test-set	0,033196	0,849609	0,043231
Modelo 5	Train-set	0,063492	0,864699	0,043863
	Test-set	0,057431	0,838893	0,042783
Modelo 6	Train-set	0,011357	0,888441	0,044579
	Test-set	0,01633	0,856988	0,043364
Modelo 7	Train-set	0,534237	0,609806	0,031804
	Test-set	0,512997	0,602998	0,031578
Modelo 8	Train-set	0,535404	0,609041	0,03174
	Test-set	**0,515325**	**0,601555**	**0,031494**
Modelo 9	Train-set	0,53761	0,607594	0,03166
	Test-set	0,514097	0,602316	0,031543
Modelo 10	Train-set	0,542316	0,604494	0,031482
	Test-set	**0,518477**	**0,599596**	**0,031366**

Con base a la tabla anterior se observa que la medición de R^2 tiene un valor esperado de 0.3209909 y 0.314822 con desviaciones estándar de 0.226577 y 0.2163 en el train-set y el test-set para el método de validación por grupos (70%train-set y 30% test-set). Del mismo modo el promedio de MAPE (Error porcentual) sobre el base obtenido es de 0.41666 y 0.138577 en el train-set y test-set, con desviaciones estándar de 1.2008 y 0.32367 respectivamente.

Con base en la Tabla se obtuvo que la especificación 10 que contempla los predictores de base, los controle relacionados el tipo de trabajo, el tipo de relación laboral, el tamaño de la firma o el número de horas trabajadas, las no linealidades propuestas entre las horas trabajadas al cuadrado y al cubo en el modelo, los meses trabajados, los controles de estratos e impuestos por dividendos presenta el mejor desempeño con $R^2 = 0.518477, RMSE = 0.5995$. Lo anterior sugiere que los controles asociados al estrato socioeconómico, así como los ingreso por dividendos reportados en impuestos devienen en predictores que pueden aumentar significativamente el desempeño en la predicción del modelo. Cabe resaltar que, aunque los resultados de las métricas de evaluación para el ajuste R^2 obtenidas en simulación sean aceptables dentro del alcance del proyecto, los modelos de regresión lineal en el parto de aprendizaje automático supervisado devienen de fuertes supuestos y asunciones sobre la relación lineal entre los predictores y las observaciones ($y_i = ingresos\ mensuales\ totales$) que en la práctica puede que no se cumplan, por otro lado es posible que la base suministrada GEIH no contemple la totalidad de los predictores o variables de interés que puedan construir un modelo lineal más robusto. Más aun, debido a las implicaciones de independencia, comportamiento lineal y demás factores, es posible que el ruido o error en las mediciones no cumpla una distribución N requerida para las aproximaciones supuestas.

Finalmente, se propone abordar el problema de estimar los ingresos totales y el estudio a profundidad de las brechas salarial de género, sobre todo en las no linealidades exploradas en el documento, mediante la construcción de modelos no lineales y de mayor complejidad como redes neuronales o SVM (funciones de kernel), arboles de decisión, que propongan un mayor desempeño en predicción y por ende mejoramiento en las métricas de evaluación, a costo computacional. Por ende, es fundamental concluir que los modelos de regresión lineal parten un fuerte comportamiento de regresión lineal a priori entre las variables y sus predictores.

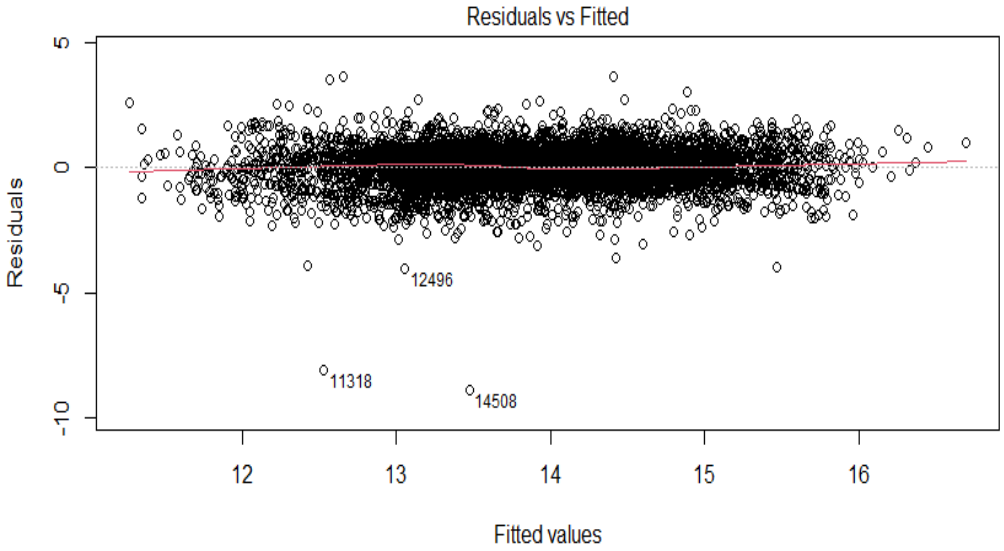
En seguida se propuso identificar los “outliers” y observaciones de alto apalancamiento a partir de la especificación 10 que tiene el mejor desempeño en métricas de evaluación, con el mayor valor de ajuste R^2 y Error Cuadrático Medio más bajo. Para esto se calculó la influencia estadística de las observaciones sobre la muestra y se muestra su distribución estadística, como se muestra en la Figura. Los outliers son observaciones de la muestra alejadas de los valores dados por la función de predicción o la función ajustada en el valor predicho $\hat{y}$, mientras que los puntos de alto apalancamiento son valores que para un valor de $\hat{y}_i$ dado presentan un valor inusual de x_i .

$$\hat{u}_i = y_i - \hat{y}_i$$

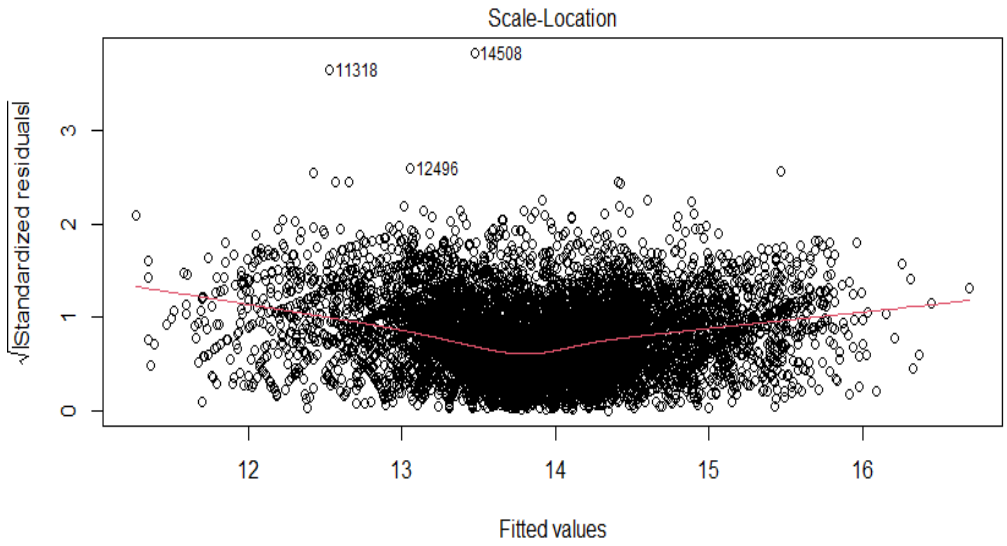
En seguida se encontró el vector diagonal de pesos de las observaciones $\hat{h}$, a partir de la influencia del modelo lineal de cada observación. Finalmente, el vector α se calculó mediante la ecuación:

$$\alpha = \frac{\hat{u}}{1 - \hat{h}}$$

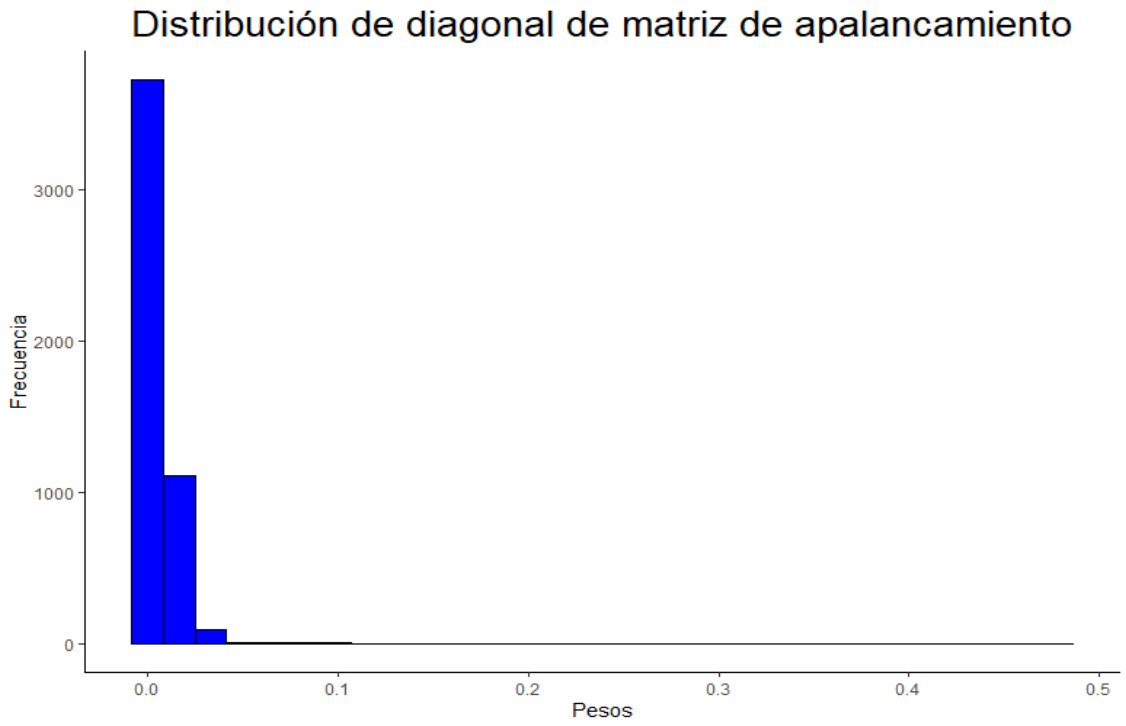
En la Figura se muestra el gráfico de residuos u^ para la base test-base construida para la identificación. A partir del gráfico podemos observar algunas observaciones11318, 12496 y 14508 se identifican como potenciales outliers a primera vista, debido a que la magnitud de la diferencia entre el valor predicho por el modelo y el observado $\hat{u} = y_i - \hat{y}_i$ es mayor a |4|. Se observa que las demás observaciones representadas presentan un residuo en el rango de $-4 < \hat{u} < 4$ en su mayoría, y la mayoría se encuentran sombre la línea del 0, es decir se espera que la distribución de la variable u^ tenga una distribución Normal con valor esperado $E(\hat{u}) = 0$



Algunos autores, entre otros James et al (2021) sugieren el gráfico de los residuos normalizados por el valor de su respectivo error estándar estimado. En la gráfica # se muestran los resultados. Como se puede observar la curva de tendencia para los residuales presenta un mínimo local en 13.6 aproximadamente, la mayoría de las observaciones en la distribución normalizada se encuentran entre 0 y 2, entonces aquellos valores que exceden el residuo estandarizado de 2 pueden ser considerados potenciales outliers que desmejorarían el modelo predictivo construido para la especificación 10.

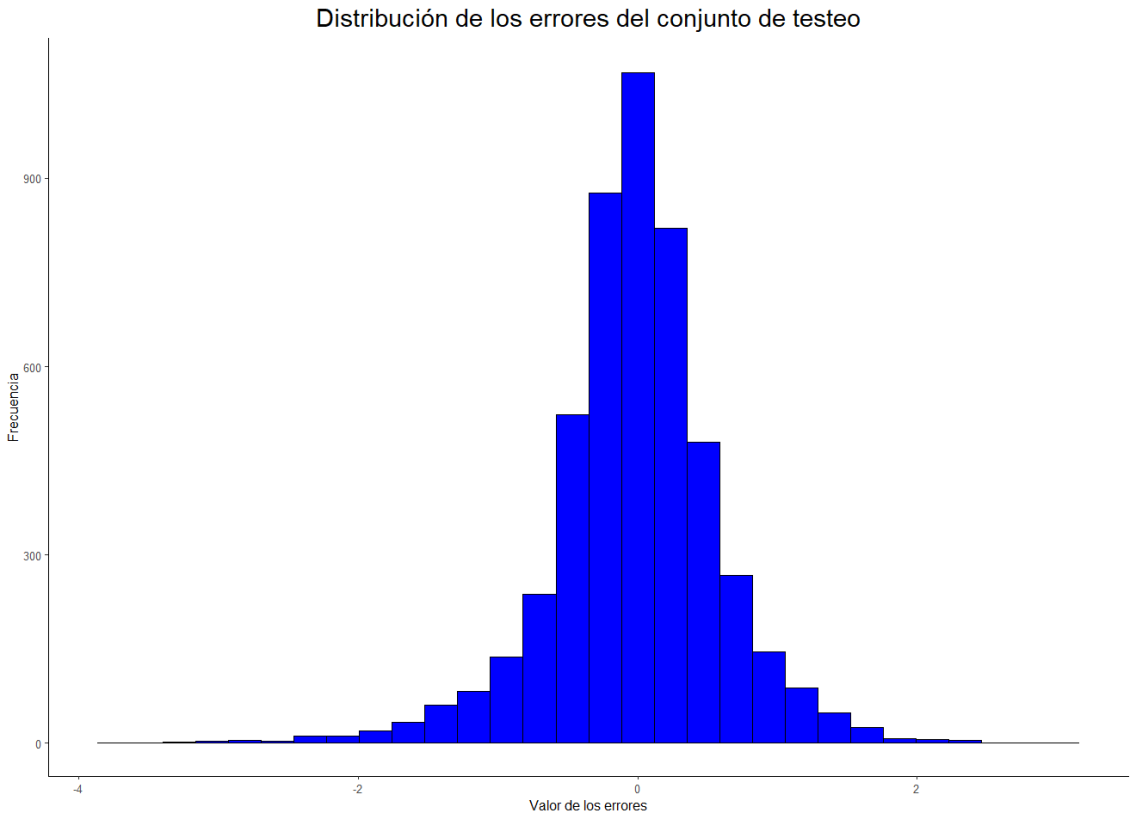


En seguida se procedió a computar la distribución de la función de pesos de la diagonal de la matriz de apalancamiento en el test sample, que define el peso que tiene cada observación sobre el modelo predictivo. Como se puede observar la mayor frecuencia de las observaciones cercanas a 3600 tienen un peso sobre el modelo que cercano a 0, cerca de 1000 observaciones tienen un peso entre 0.01 y 0.02 , mientras que un mínimo de observaciones tiene un peso superior a 0.1. La media de la distribución es 0.007455 y la desviación estándar 0.01378, entonces concluimos que la distribución de la influencia estadística $\hat{h} \sim N(0.007455, 0.01378)$, lo que corrobora los resultados esperados.

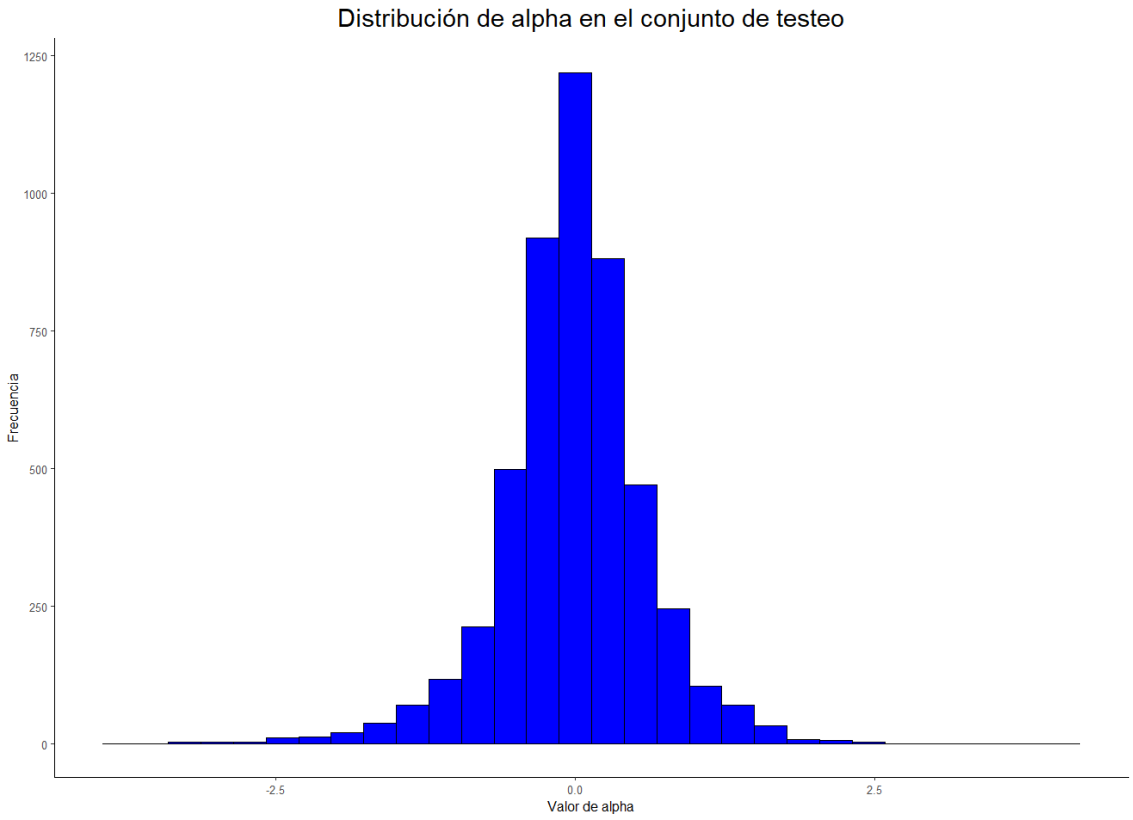


Con base en el gráfico anterior, se sabe que las observaciones en la loca de la distribución pueden ser outliers o puntos de alto apalancamiento que pueden tener un efecto significativo sobre el modelo. James et al (2021) sugieren que los outliers pueden generarse por errores en la recolección de datos, que para el caso del proyecto de investigación podría generarse debido a una variable aleatoria en la toma de muestras, protocolos de las mediciones realizadas por el DANE, pérdidas de datos o valores sobre las bases de datos construidas que podrían estar afectando el desempeño del modelo. Sin embargo, es posible que las bases de datos proveídas por la GEIH a partir de las cuales se realizó la construcción del modelo no contemplen predictores significativos para la estimación, variables con distribuciones aleatorias, fenómenos sociales, contextuales, entre otros lo cual genera una deficiencia en el modelo.

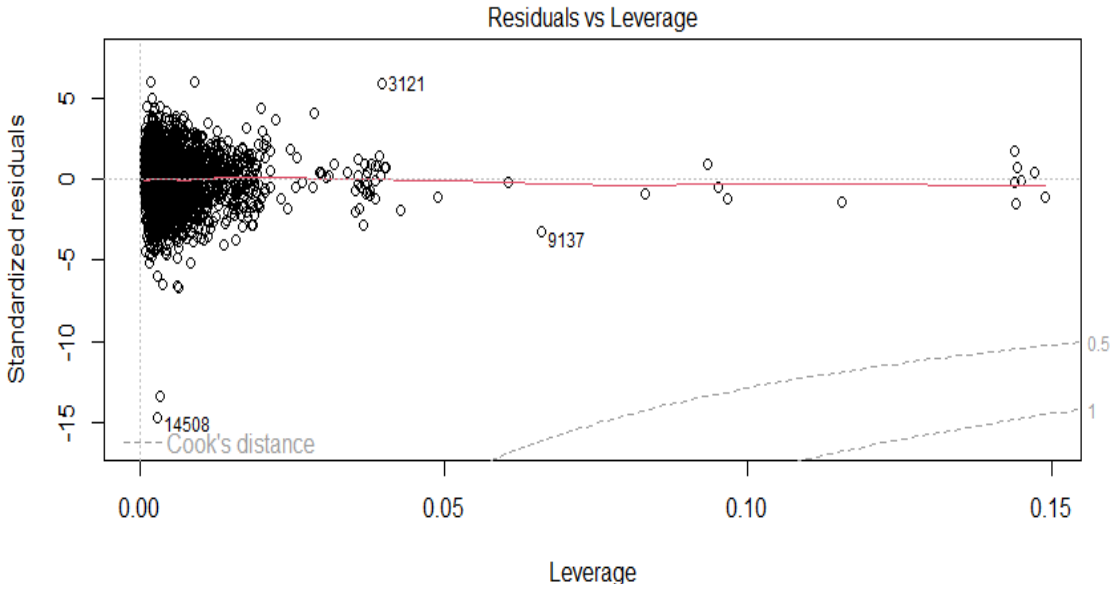
En la Figura se observa la distribución del vector de residuos para cada observación. Se tiene un valor medio -0.02309, desviación estándar 0.599211, medina -0.01742 y un valor mínimo y máximo de $-3.80633\ y\ 4.07172$ respectivamente siendo estas observaciones que la DIAN debe indagar en la recopilación de datos del hogar para potencial comportamiento irregular de ingresos totales mensuales para individuos ubicados en diferente escenarios, como lo son temprana edad, bajo nivel educativo, bajo estrato socio-económico actividades informales y condiciones de trabajo irregular que reporten altos ingresos, o individuos con alto nivel educativo, avanzada edad, formalidad laboral, años de experiencia laboral, alto nivel socio-económico que reporten ante el Estado ingresos marginales.



En la Figura se observa la distribución del vector α descrito anteriormente que es el residuo normalizado al peso correspondiente de cada observación sobre la muestra. Se tiene un valor medio -0.02339, desviación estándar 0.6073806, medina -0.01742 y un valor mínimo y máximo de -3.80633 y 4.07172, lo cual es consecuente con los resultados ya que el vector α sigue al vector de los residuos de las observaciones, teniendo en cuenta que la mayor frecuencia de observaciones tiene un peso que tiende a cero, excepto para los puntos 'outliers' o de alto apalancamiento sobre el modelo.



Finalmente, en la Figura se muestra el gráfico del vector de residuos $u^{\wedge}$ en relación con su respectivo apalancamiento, para aquellas observaciones con medición de los predictores x_i inusual dada una predicción. Como se puede observar la mayoría de las muestras presentan un respectivo residuo estandarizado en el rango entre $[-5,5]$ mientras que la medida de apalancamiento tiende a ser entre $[0 - 0.025]$. Aquellas observaciones de y_i predicho que superen el umbral de 0.025 tienen un apalancamiento mayor sobre el modelo, entonces pueden ser una potencial fuente de error en la toma de las observaciones o desajuste del modelo.



Finalmente se procedió a cambiar la estrategia de entrenamiento de los modelos. En las especificaciones reportadas anteriormente se utilizó el acercamiento de validación por grupos. Sin embargo, para concluir esta investigación se indagó por la robustez, costo computacional y desempeño de las métricas de evaluación para modelos a través de validación cruzada. A diferencia de ésta que divide la base en train-set, para realizar el entrenamiento y ajusto de los parámetros, test-set para validar su desempeño y realizar las mediciones en el error de predicción, la validación LOOCV divide las observaciones en n partes del tamaño de la muestra en el número de observaciones totales, realiza la el cálculo de los Errores Cuadráticos Medios para las n observaciones y realiza el promedio de MSE sobre las $n-1$ observaciones, para finalmente realizarla validación en predicción sobre la $n-ésima$ observación excluida, permitiendo así reducir la

Modelo	Muestra	R2_Score	RMSE	MAPE
Modelo 8	Train-set	0,535404	0,609041	0,03174
	Test-set	**0,515325**	**0,601555**	**0,031494**
Modelo 10	Train-set	0,542316	0,604494	0,031482
	Test-set	**0,518477**	**0,599596**	**0,031366**

Aplicando el método anteriormente de LOOCV descrito se obtuvo un resultado para las métricas de $R^2 = 0.5364, RMSE = 0.429358\ y\ MAE = 0.4293588$, y $R^2 = 0.5305, RMSE = 0.4323066\ y\ MAE = 0.4323066$ para la especificación descrita por el modelo 10 y el modelo 8 respectivamente

Conclusiones

En el presente trabajo se encuentra que, para Bogotá en 2018, de acuerdo con los datos de la GEIH existe una relación cuadrática de la edad y el ingreso laboral, donde las personas más jóvenes y las de mayor edad tienen menos ingresos que aquellas entre los 40 y 50 años (lo que forma una figura de u invertida).
Además, las brechas de género también varían con la edad y representan menores ingresos relativos para las mujeres en comparación con los hombres a medida que transcurre el ciclo de vida. Los factores relacionados con la flexibilidad laboral, el tamaño de la firma, el tipo de oficio y la relación laboral explican una proporción importante de la variabilidad de los salarios. El hecho de que los determinantes anteriores se relacionen estrechamente con decisiones posteriores al nacimiento del primer hijo (kleven et al, 2019) hacen que además del lema "salarios iguales para trabajos iguales" sea necesario resaltar la importancia de la equidad de género en el cuidado de los hijos y el hogar. En especial, en el documento se mostró que las mujeres encuentran en promedio su salario máximo en el ciclo de vida a una edad significativamente menor que los hombres.

Estos resultados se contrastan con el análisis inferencial del modelo (2) con los controles del modelo determinados donde se observa un aumento del coeficiente de correlación (R^2) a la línea ajustada de $R^2 = 0.204607$ para la especificación sin logaritmo en predicción, $R^2 = 0{,}427828$ con el logaritmo de los ingresos a $R^2 = 0.561$ en el modelo inferencial. Entonces se concluye que la inferencia de los parámetros de estimación β_i para el modelo (2) con las variables de control obtienen mayor correlación al modelo lineal ajustado que el modelo predictivo obtenido.

Las simulaciones obtenidas para los modelos predictivos arrojaron un valor máximo de las métricas de evaluación de $R^2 = 0.5364, RMSE = 0.429358\ y\ MAE = 0.4293588$, mientras que las distribuciones de los vectores de residuos y matriz diagonal que describe el comportamiento de los pesos para las observaciones sigue una distribución normal con valor medio que tiende a 0.

Bibliografía

Borjas, G. (2016). *Labor Economics.* Mc Graw Hill. Séptima Edición

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021) *An Introduction to Statistical Learning: with Applications in R.* Segunda edición.

Kleven, H., Landais, C., & Egholt Sogaard, J. (2019). *Children and Gender Inequality: Evidence from Denmark.* American Economic Journal.

Murphy, K., & Welch, F. (1990). *Empirical age-earnings profiles.* The University of Chicago. JSTOR.

Popovici, I., Carvajal, M., Peeples, P., & Rabionet, S. (2021). *Disparities in the Wage-and-Salary Earnings, Determinants and Distribution of Health Economics, Outcomes Research, and Market Access Professionals: An exploratory Study.* NCBI. Springer.

Toczek, L., Bosma, H., & Peter, R. (2021). *The Gender Pay Gap: Income Inequality Over Life Course – A Multilevel Analysis.* Frontiers in sociology. Volumen 6. Artículo 815376.