PS1 - Equipo 03

Catalina Leal Rojas, Lucas Daniel Carrillo Aguirre, Lucas Eduardo Veras Costa, Maria Paula Basto Lozano

8 de septiembre de 2025

El link del repositorio es: https://github.com/mbastol06/PS1_Equipo3

1. Introducción

Durante los últimos diez años, se ha expandido con ímpetu el uso de los métodos de ciencia de datos para contribuir a la toma de decisiones de política pública (Hossin, 2023, Berkeley D-Lab, 2024)., tomando como insumos grandes cantidades de datos de muestras poblacionales y transformándolas en modelos que permitan predecir comportamientos de variables fundamentales para la economía, como el mercado laboral, los niveles de precios y las condiciones socioeconómicas de los individuos. Entre estas necesidades está también la de predecir los niveles de ingresos precisamente para observar patrones en las condiciones particulares de los ciudadanos y formular decisiones fiscales y laborales más eficientes.

Durante el ejercicio se pudo observar, tras los ejercicios de depuración de datos, que los salarios por hora se concentran en los tramos bajos y presentan una cola derecha extensa. La comparación entre el salario por hora y su logaritmo confirma que la mediana es menor que la media y que buena parte de la variabilidad proviene de un grupo pequeño de observaciones de altos ingresos. En términos descriptivos, el salario promedio por hora bordea los $8,100 mil pesos, la jornada semanal típica ronda las 48 horas, y la distribución por sexo es equilibrada. En el componente categórico, predominan la educación terciaria, los puestos de trabajo en empresas particulares y medianas o grandes, y una presencia de una formalidad mucho mayor que en el promedio nacional.

El perfil edad–salario muestra una forma cóncava, de forma que los ingresos crecen con la edad a ritmos decrecientes y alcanzan un máximo hacia la mitad de la vida laboral. Las estimaciones obtenidas ubican la edad pico alrededor de los 45 o 50 años y los intervalos de confianza por bootstrap son estrechos, lo que implica que existe una mayor precisión en la estimación. Además, al estudiar la brecha género–salario, el coeficiente asociado a ser mujer es negativo y estadísticamente significativo en todos los modelos, lo que confirma

definitivamente una brecha en contra de las mujeres. La brecha es modesta cuando no se controlan covariables, pero se amplía al introducir educación, intensidad laboral y características del empleo; en la especificación más completa, el diferencial ronda el 10 % a favor de los hombres, manteniendo signo y significancia bajo distintas variantes y verificándose también mediante el teorema de Frisch–Waugh–Lovell. Esto puede implicar que existe una discriminación sistemática en el mercado laboral o que existen variables no identificadas en el estudio.

Desde la perspectiva predictiva, fueron cosntrastadas varias especificaciones que incorporan no linealidades e interacciones. Los modelos con controles bien elegidos y una flexibilidad moderada (polinomios de edad e interacciones parciales) tuvieron los mejores desempeños, con RMSE en torno a 0.45 y estabilidad bajo validación cruzada leave-one-out; en particular, el modelo cúbico con interacciones muestra el compromiso más robusto entre sesgo y varianza. El análisis de errores revela una distribución centrada en cero con colas asociadas a segmentos menos representados—trabajadores en microempresas, menor formalidad y menos horas—y unos pocos casos con errores muy altos pese a perfiles comunes, potencialmente útiles para ejercicios de gestión de riesgo y auditoría. En conjunto, los resultados sugieren que las heterogeneidades por educación, tipo de relación laboral y estructura de la firma explican gran parte de las diferencias salariales, y que incorporar dichas dimensiones mejora ostensiblemente la capacidad predictiva de los modelos.

2. Datos

- **Fuente de los datos:**

  La información utilizada para este estudio fue obtenida desde el *"Reporte de Medición de Pobreza Monetaria y Desigualdad"*, correspondiente a la Gran Encuesta Integrada de Hogares (GEIH) hecha por el Departamento Administrativo Nacional de Estadística (DANE) en el año 2018.

  Este reporte busca hacer una caracterización de la población sobre rubros como los ingresos, el estrato, la edad, el sexo, la cantidad de horas trabajadas, entre otros.

- **Web Scraping:**

  Los datos necesarios para el análisis están consignados en una página web (https://ignaciomsarmiento.github.io/GEIH2018_sample/)) que dirige a diez enlaces diferentes para cada una de las particiones de la base de datos original. Así, la forma más pertinente para extraer los datos resultó ser el método de raspado web o web scraping. Sin embargo, debido a que la página está construida en formato Java Script, las instrucciones revisadas a lo largo del curso para este método fueron insuficientes.

Para acceder a la fuente cruda de las tablas de cada uno de los diez chunks de datos fue necesario hacer una revisión de los elementos de la página por medio de la herramienta Ïnspección" integrada en el explorador Google Chrome. Allí, en el panel de elementos en la ventana de "Red" se identificó un apartado con la fuente HTML de la página, que dirigía al mismo enlace de arriba, con un "pages/geih_page_1.html" adicional para la primera página. Agregando esa terminación sobre cada una de las otras nueve páginas y se hizo entonces un web scraping común para un HTML con tablas. Este proceso de extracción se hizo iterativamente sobre cada uno de los enlaces y se compiló en una tabla combinada con 32,177 observaciones para 178 variables diferentes.

- **Limpieza de los datos:**

  Para hacer que los datos fuesen adecuados para hacer el análisis central de este trabajo, a saber, la modelación del salario por hora en esta muestra poblacional específica, fue necesario hacer un proceso de depuración e imputación de datos de la siguiente manera.

  Primero, se redujo la muestra poblacional a los individuos mayores a 18 años de edad, que son los que están en edad de trabajar y tienen datos de salario disponibles; así como también se eliminaron los individuos que trabajaran por cuenta propia toda vez que la contabilidad de sus ingresos es inexacta y difícil de rastrear. Esto también implicó la eliminación de observaciones que tuvieran salarios de cero o nulos.

  Segundo, se eliminaron, de la muestra inicial de 178, todas las variables que en cada una de las observaciones tuvieran consistentemente valores faltantes o valores idénticos. Además, se desestimaron todas las variables que tuvieran valores faltantes en más del 60 % de las observaciones. Estos dos pasos fueron necesarios para evitar utilizar observaciones y variables que podían no aportar nada a la base más que ruido.

  Después de estos dos pasos, se hizo la imputación de datos para cada uno de los valores faltantes en la base de datos resultante. Esto se hizo por medio de dos métodos. El primero implicó identificar todas las variables categóricas de la muestra, que eran todas aquellas que se correspondían con la indagación de salarios de los individuos (por ejemplo, preguntas de respuestas binarias de sí o no, o de respuestas con un rango del 1 al 3). Luego de que se hizo esto, se calculó la moda de las observaciones en esas variables para todos los individuos que compartieran el mismo estrato social, y los valores faltantes se imputaron con ella. Esto, entendiendo que es más probable que individuos que compartan condiciones socioeconómicas similares puedan tener características en sus ingresos comparables.

El segundo método implicó identificar las variables numéricas, que eran todas aquellas que tenían valores en magnitud monetaria, como los ingresos, los auxilios y las deducciones. A los valores faltantes de estas variables se les aplicó el método de K Vecinos más Cercanos (*o K-Nearest Neighbors*) que busca las similitudes entre las observaciones de la base e identifica, por medio de un algoritmo de aprendizaje supervisado, los k vecinos más cercanos para cada una (en este caso los 5 más cercanos). Así, con los valores de los vecinos, se imputaron las observaciones con valores faltantes en dichas variables.

Lo último que se hizo para la organización de la base de datos fue reemplazar con los valores del percentil 97.5 todos los valores de las observaciones de las variables de ingreso que se ubicaran más allá de ese umbral. Esto para evitar que esas variables, que suelen tener una dispersión importante en ese último segmento poblacional, tuviesen un impacto indeseado sobre la estadística descriptiva de la muestra.

- **Estadística descriptiva de los datos:**

  La base de datos final resultó en un total de 9,892 observaciones con 19 variables de interés. En el Cuadro 1 se presentan las estadísticas descriptivas tanto para las variables numéricas o binarias, como para las variables categóricas.

  En la muestra, el salario promedio de los individuos es de $1,610,063 pesos colombianos, o, lo que es lo mismo para el momento de la realización de la encuesta, 2.2 veces el salario mínimo legal vigente ($737,717). Por supuesto, este salario promedio contempla los auxilios de transporte y otras adiciones en los ingresos de los individuos. Esta estadística se corresponde con un salario por hora de $8,121 pesos en promedio y un promedio de 48.3 horas trabajadas por semana, lo que implicó una gran cercanía con el número de horas legales de trabajo para 2017, que fue de 48.

  Además, la muestra está balanceada en proporciones iguales entre hombres y mujeres, con un promedio de edad de 36.2 años, con un mínimo de 18 y un máximo de 86. Los trabajadores trabajan en microempresas un 23 % de las ocasiones, mientras que el 77 % de ellos tienen empleos formales. Esto contrasta con la información proveída por el DANE a finales de 2017 que indicaba que la proporción de ocupados informales en las 13 ciudades y áreas metropolitanas fue 47,0 % para el trimestre móvil diciembre 2017 - febrero 2018. Es decir, la muestra analizada en Bogotá para ese momento se ocupó ostensiblemente en mayor proporción de manera formal que el promedio nacional.

  Por último, el máximo nivel educativo alcanzado por los individuos de la muestra, en promedio, fue la educación terciaria, comprendiendo el 45 % del total. El tipo de

empleo más común, dando cuenta del 88.5 % del total, fue el empleo en empresa particular, además de que más de la mitad de los encuestados afirmaron trabajar en una empresa de más de 50 empleados; y el oficio más usual fue el de apoyo administrativo y logístico, con un 8.2 % de la muestra. Estos datos dan cuenta de una clara tendencia entre una contratación de tipo privado y un segmento poblacional con estudios superiores. Así como con el promedio de ingresos y el tipo de empleo, esto permite observar unas mejores condiciones laborales que el promedio nacional, explicado principalmente por una mayor cantidad de oportunidades académicas y profesionales.

Cuadro 1. Estadísticas descriptivas

(a) Variables numéricas

Variable	N	Media	Desv. Est.	Min	Max
Salario mensual	9,892	1,610,063	1,537,099	20,000	7,892,292
Salario por hora	9,892	8,121.29	8,099.11	326.67	41,470.34
ln(Salario)	9,892	8.71	0.69	5.79	10.63
Horas trabajadas	9,892	48.34	12.25	1.00	130.00
Edad	9,892	36.24	12.02	18.00	86.00
Mujer (1 = Sí)	9,892	0.50	0.50	0.00	1.00
Trabaja en microempresa (1 = Sí)	9,892	0.23	0.42	0.00	1.00
Trabajo formal (1 = Sí)	9,892	0.77	0.42	0.00	1.00

(b) Variables categóricas

Variable	Categoría	Descripción	Frecuencia	Porcentaje
Nivel educativo	7	Educación terciaria	4,478	45.27 %
Oficio	39	Apoyo administrativo y logístico	814	8.23 %
Tamaño de empresa	5	$>$ 50 empleados	5,107	51.63 %
Tipo de empleo	1	Empleado de empresa particular	8,757	88.53 %

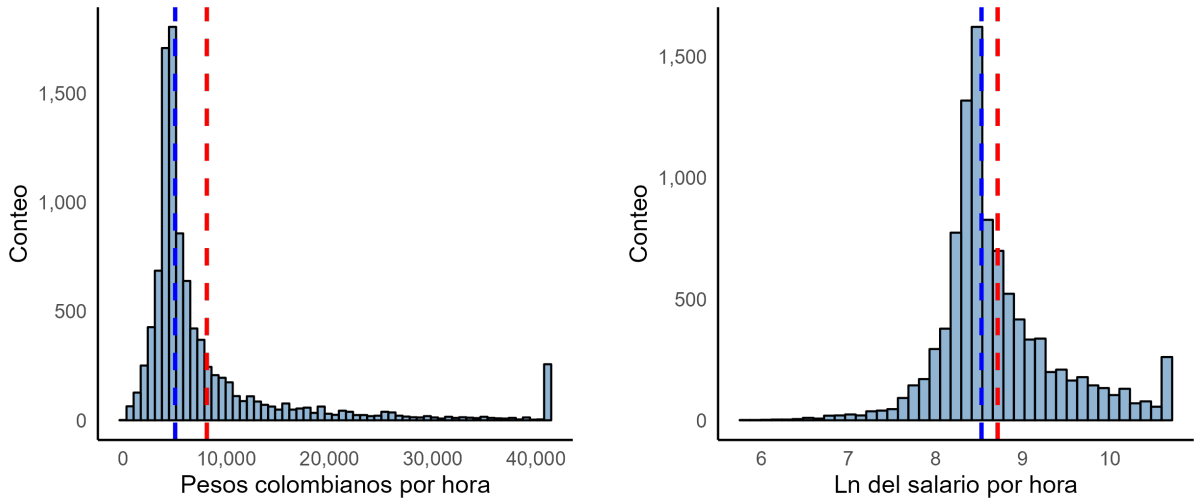
Nota: En el panel superior se presentan las estadísticas descriptivas de las variables numéricas o binarias. En el panel inferior se presentan las frecuencias para las variables categóricas.

En la Figura 1 se presentan las distribuciones del salario por hora (panel izquierdo) y su logaritmo natural (panel derecho) en la muestra analizada. Estos histogramas dan cuenta de dos condiciones clave: por un lado, los salarios se concentran en la parte más baja de la distribución, entre los $3,000 y los $9,000 pesos por hora.

Por otra parte, que, a pesar de que la distribución del logaritmo es más o menos simétrica, la diferencia entre la media y la mediana da cuenta de que los salarios más altos siguen empujando las estadísticas hacia arriba. A pesar de que, por cuenta del corte de ingresos en el percentil 97.5 las concentraciones en los ingresos más altos no son tan altas como podrían llegar a serlo en una muestra original, la muestra sigue

presentando una cola al final de casi 500 individuos (de un total de casi diez mil) que tienen ingresos de más de $40,000 pesos por hora. Este hecho es fundamental para entender cómo la muestra podría dar cuenta de una importante desigualdad salarial en una ciudad que, de por sí, contiene brechas socioeconómicas pronunciadas.

Figura 1. Distribuciones del salario por hora y su logaritmo natural



(a) Salario por hora

(b) Logaritmo natural del salario

Nota: En el panel izquierdo se encuentra la distribución del salario por hora en pesos. En el panel derecho se encuentra la distribución del logaritmo natural del salario por hora. La línea roja es la mediana y la azul es la media.

3. Perfil edad-salario

En esta sección analizaremos la relación entre el perfil edad-salario. En primer lugar, es importante discutir qué medida de ingreso utilizar. La base de datos de la GEIH incluye diversas medidas, como ingreso por primas monetarias, ingreso mensual, ingreso usual en el mes e ingreso efectivo en el mes, entre otras.

En los estudios que analizan el perfil salarial de edad, es común emplear el ingreso por hora, ya que permite aislar el efecto de las horas trabajadas en el mes, lo cual podría sesgar los resultados. Asimismo, suele utilizarse la variable de ingreso en logaritmos, lo que mejora la interpretabilidad del coeficiente estimado, que pasaría a representar cuánto impacta, en términos porcentuales, una unidad adicional de la variable independiente sobre el ingreso, en este caso un año adicional en la edad.

Se estimó el siguiente modelo para identificar el perfil edad-salario de las observaciones de la muestra.

$$\ln w_i = \beta_1 + \beta_2 Age_i + \beta_3 Age_i^2 + u_i \tag{1}$$

Donde w_i es el ingreso salarial por hora del individuo i y Age_i es su edad. Se realizó la estimación del modelo con las diferentes medidas de ingreso salarial real y nominal por hora:

Cuadro 2. Perfil de salario-edad

	Dependent variable:	
	log(y_salary_m_hu)	log(y_total_m_ha)
	(1)	(2)
age	0.059***	0.066***
	(0.003)	(0.003)
I(ageˆ2)	−0.001***	−0.001***
	(0.00004)	(0.00004)
Constant	7.407***	7.393***
	(0.064)	(0.065)
Observations	9,892	9,892
R^2	0.040	0.045
Adjusted R^2	0.040	0.045
Residual Std. Error (df = 9889)	0.671	0.677
F Statistic (df = 2; 9889)	206.012***	235.041***
Note:		*p<0.1; **p<0.05; ***p<0.01

- **Interpretación de los coeficientes:**

  De acuerdo con los resultados de las dos regresiones, el perfil salarial según la edad tienen la misma forma funcional, es decir una curva cóncava. En concreto, el término lineal de la edad muestra el efecto directo de "sumar un año", mientras que el término cuadrático (edad^2) captura cómo que existen efectos diferenciados sobre el salario en edades diferentes de la vida, es decir que la relación entre la edad y el salario no es lineal.

  El sumar un año de edad incrementa 5.9 % y 6.6 % el salario (el efecto directo) tomando como el salario real y nominal por hora respectivamente. Ambos coeficientes estimados son significativos estadísticamente y positivos lo que indica en primera instancia un efecto directo positivo de la edad sobre el salario. El coeficiente relacionado con la edad al cuadrado captura esa relación no lineal. Tanto para el modelo con el salario real como el nominal el coeficiente de la edad cuadrática es negativo lo que indica la concavidad de la relación, es decir que hasta un punto la relación entre edad y salario o es positiva y luego es negativa, en ambos modelos el término es negativo

y estadísticamente significativo.

Para estimar la semileasticidad de la edad sobre el logaritmo del salario solucionamos la condición de primer orden. Luego un año adicional de edad aumenta porcentualmente el salario en:

$$100 \times (\beta_2 + 2 \times \beta_3 \times edad)\,\%$$

La semieleasticidad de para las edades de 18 , 25, 45 y 50 años se presentan en el Cuadro 3. Como se observa, entre más joven más aporta un año más de edad en el incremento porcentual del salario, va decreciendo e incluso se vuelve negativo el aporte desde edad más avanzada lo que se denomina edad pico.

Cuadro 3. Semielasticidad del salario sobre el ingreso

Edad	Salario Real	Salario Nominal
18	3.588	3.957
25	2.676	2.927
35	1.372	1.456
45	0.069	-0.015
50	-0.582	-0.750

El coeficiente constante para ambos modelos es positivo y significativo, y representa el valor del logaritmo del salario cuando la edad es cero. Aunque este valor no tiene una interpretación económica directa, no se trabaja con 0 años, por lo que es el término que ajusta el modelo para la relación entre edad y salario dentro del rango de edades observado en la muestra.

- **Ajuste de la muestra:**

Para analizar el ajuste de los modelos dentro de la muestra, se presenta los siguientes estadísticos como el error estándar (RSE), el error cuadrático medio (RMSE), el R^2 ajustado, estadístico F y criterios de información Akaike (AIC) y bayeciano (BIC).

Cuadro 4. Medidas de ajuste para los modelos de salario por edad

Modelo	RSE	RMSE	R^2*	F	AIC	BIC
Salario real	0.671	0.671	0.040	206.012	20186.076	20214.874
Salario nominal	0.677	0.677	0.045	235.041	20360.186	20388.984

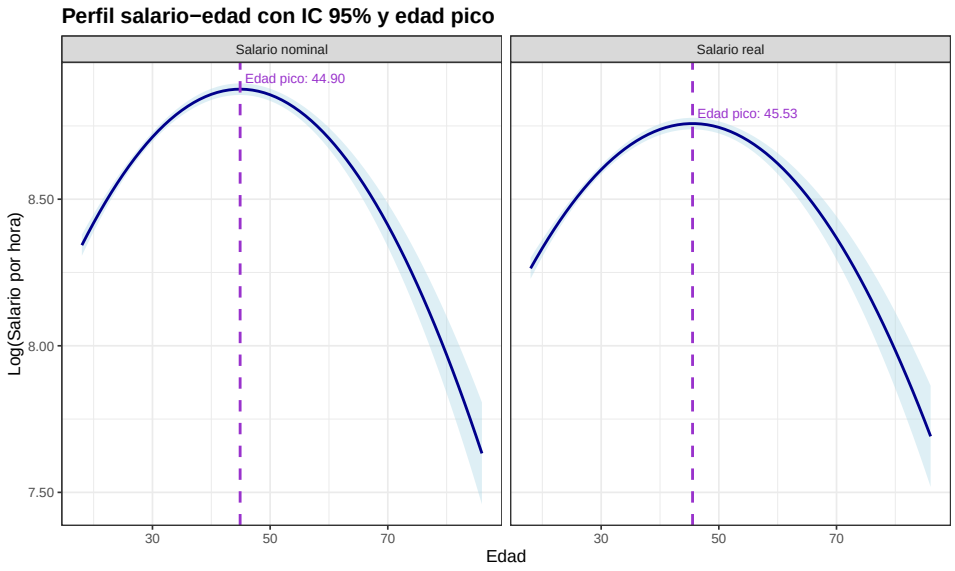
Los resultados muestran que ambos modelos ofrecen un desempeño explicativo limitado pero estadísticamente significativo. El coeficiente de determinación ajustado R^{2*} es de 0.040 en el modelo del salario real y de 0.045 en el del salario nominal, lo que indica que aproximadamente el 4–4,5 % de la variación en el logaritmo del salario se explica únicamente por la edad y su cuadrado. El error estándar residual (RSE), que estima la desviación promedio de los residuos ajustando por los grados de libertad,

es de 0,671 y 0,677 respectivamente, mientras que el RMSE, que se calcula sobre el total de observaciones, coincide numéricamente con esos valores debido al gran tamaño muestral. Ambos modelos presentan un F estadístico alto y significativo (206.0 y 235.0), que el modelo es globalmente estadísticamente significativo. Finalmente, los criterios de información AIC y BIC son menores en el modelo con salario nominal, lo que sugiere que este ofrece un ajuste más eficiente penalizando la complejidad del modelo.

- **Gráfico de ingreso salarial, edad pico y su intervalo de confianza:**

  A continuación se presenta el gráfico de los valores predichos con cada modelo:

Figura 2. Gráfico del logaritmo de salarios predichos, intervalo de confianza y edad pico



La edad pico se refiere a la edad en la cual el logaritmo del salario alcanza su máximo. Y se obtiene derivando la función del logaritmo del salario e igualando a cero , es decir que la edad pico se calcula como:

$$edad\_pico = -\frac{\beta_2}{2\beta_3}$$

Luego, para cada modelo la edad pico es:

- **Modelo salario real**: 45.53
- **Modelo salario nominal**: 44.89

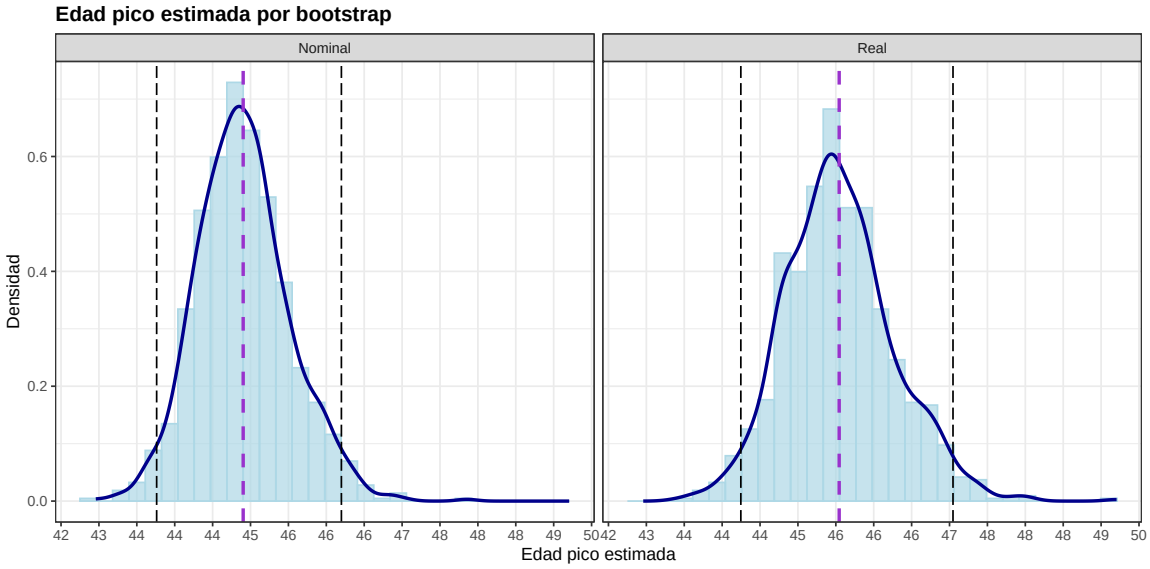
Luego por bootstrap no paramétrico se estimaron las edades pico para cada modelo con mil remuestreos con reemplazo (R). En cada réplica se re estimó cada modelo

obteniendo los coeficientes y luego calculando la edad pico para las R repeticiones. A partir de la distribución empírica de las R edades pico, se calcularon los intervalos de confianza al 95 % con el método de percentil, es decir que los límites corresponden a los cuartiles 2.25 % y 97.5 % de la distribución bootstrap del la edad pico. Los respectivos intervalos de confianza son:

- **Modelo salario real**: (43.74 , 46.21)
- **Modelo salario nominal**: (44.25 , 47.05)

La figura muestra, para cada modelo (nominal y real), la distribución bootstrap de la edad pico: el histograma en azul claro representa la frecuencia de las edades pico re-muestreadas y la curva en azul oscuro es la densidad kernel asociada. Además se agregó: la media de la edad pico por bootstrap para cada modelo, que es representado con la línea vertical morada ; y, los intervalos de confianza para cada modelo en la línea punteada gris. En conjunto, la gráfica permite comparar visualmente la localización y la dispersión de las edades pico entre ambos modelos, dado las densidades se solapan y los intervalos percentiles son similares, se la evidencia indica que las edades pico no difieren de forma sustantiva.

Figura 3. Distribución estimada e intervalos de confianza edad pico



4. La brecha género-salario

En esta sección analizaremos la relación entre ingresos y género femenino. En primer lugar, es importante discutir qué medida de ingreso utilizar. La base de datos de la GEIH incluye diversas medidas, como ingreso por primas monetarias, ingreso mensual, ingreso usual en el mes e ingreso efectivo en el mes, entre otras.

En los estudios que analizan la brecha salarial entre hombres y mujeres, es común emplear el ingreso por hora, ya que permite aislar el efecto de las horas trabajadas en el mes, lo cual podría sesgar los resultados. Asimismo, suele utilizarse la variable de ingreso en logaritmos, lo que mejora la interpretabilidad del coeficiente estimado, que pasaría a representar cuánto impacta, en términos porcentuales, una unidad adicional de la variable independiente sobre el ingreso.

También es importante señalar que nos interesa comparar únicamente a hombres y mujeres que trabajan. Por lo tanto, el análisis se centra en los individuos ocupados de la muestra.

Inicialmente estimamos el siguiente modelo:

$$\log(\omega) = \beta_0 + \beta_1 Female + u \tag{2}$$

Cuadro 5. Resultados del modelo incondicional

	Variable dependiente:				
	log(ysalarym)	log(ysalarymha)	log(yingLabm)	log(ytotalm)	log(ytotalmha)
	(1)	(2)	(3)	(4)	(5)
female	−0.149***	−0.045***	−0.147***	−0.238***	−0.090***
	(0.015)	(0.015)	(0.015)	(0.015)	(0.014)
Constant	13.977***	8.641***	14.088***	13.981***	8.667***
	(0.011)	(0.010)	(0.011)	(0.010)	(0.009)
Observaciones	9,892	9,892	9,892	14,764	14,764
R^2	0.010	0.001	0.009	0.018	0.003
R^2 ajustado	0.010	0.001	0.009	0.017	0.003
Error estándar residual	0.751 (df = 9890)	0.721 (df = 9890)	0.762 (df = 9890)	0.889 (df = 14762)	0.832 (df = 14762)
F Statistic	97.364*** (df = 1; 9890)	9.559*** (df = 1; 9890)	91.422*** (df = 1; 9890)	263.841*** (df = 1; 14762)	43.342*** (df = 1; 14762)
Nota:					*p<0.1; **p<0.05; ***p<0.01

El Cuadro 5 presenta los resultados de la estimación por mínimos cuadrados ordinarios utilizando diversas medidas de ingreso. *ysalarym* corresponde al ingreso nominal de la ocupación principal, *ysalarymha* al salario por hora de la ocupación principal, *yingLabm* al ingreso proveniente de todas las ocupaciones, *ytotalm* al ingreso total proveniente de ocupaciones e ingresos independientes, y *ytotalmha* al ingreso total de ocupaciones e independientes medido por hora.

Los resultados muestran evidencia de que las mujeres ganan menos que los hombres. Aunque existe cierta variación en los coeficientes de la variable $female$, todos son negativos y estadísticamente significativos, lo que indica pérdidas salariales para las mujeres.

Cabe señalar que en los resultados anteriores todos los coeficientes fueron estimados mediante mínimos cuadrados ordinarios. No obstante, es posible obtenerlos aplicando el Teorema de Frisch-Waugh-Lovell (FWL). Supongamos que queremos estimar el coeficiente de una variable X_1 en una regresión múltiple que también incluye una variable de control X_2, en el siguiente modelo:

$$y = \beta_1 X_1 + \beta_2 X_2 + u$$

Para estimar β_1 usando el Teorema de Frisch-Waugh-Lovell, se siguen los siguientes pasos:

1. Regrese y sobre X_2 y obtenga los residuos.
   Denotemos estos residuos como r_y:

$$r_y = y - \hat{y}_2 \quad \text{donde } \hat{y}_2 = X_2\hat{\beta}_2$$

2. Regrese X_1 sobre X_2 y obtenga los residuos.
   Denotemos estos residuos como r_{X_1}:

$$r_{X_1} = X_1 - \hat{X}_{1,2} \quad \text{donde } \hat{X}_{1,2} = X_2\hat{\gamma}$$

3. Regrese los residuos r_y sobre r_{X_1}.
   El coeficiente estimado será exactamente igual a $\hat{\beta}_1$ de la regresión original:

$$\hat{\beta}_1 = \frac{r'_{X_1} r_y}{r'_{X_1} r_{X_1}}$$

Este procedimiento permite estimar el efecto de X_1 sobre y, controlando por X_2, sin necesidad de realizar directamente la regresión múltiple completa. Además, resulta útil para reducir el costo computacional en aplicaciones con gran cantidad de variables.

Como ilustración de este procedimiento, estimamos el siguiente modelo mediante el método clásico de MCO y también aplicando el teorema de FWL:

$$\log(\omega) = \beta_1 + \beta_2 female + \beta_3 age + \beta_4 age^2 \tag{3}$$

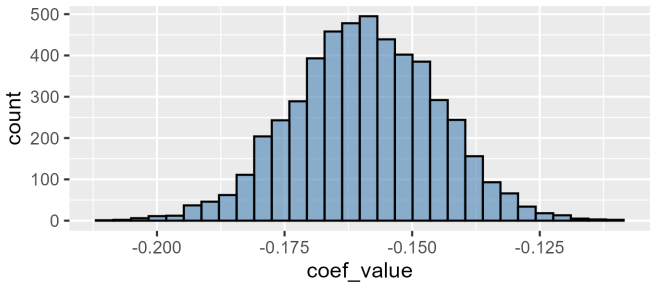
El Cuadro 6 muestra que los coeficientes de $female$ y $resid\_fem$ son idénticos. Sin embargo, los errores estándar no lo son. Esto ocurre debido a la forma en que se realiza la estimación en dos etapas: al reducir el número de variables en la segunda ecuación, los grados de libertad utilizados también disminuyen.

Otra manera de estimar el error estándar es a través del método bootstrap. Este procedimiento consiste en realizar un remuestreo de la muestra n veces y reestimar el coeficiente en cada repetición. El error estándar se obtiene a partir de la dispersión de esta distribución de estimaciones.

Cuadro 6. Comparación de las estimaciones por MCO y FWL

	Variable dependiente:	
	log(y_ingLab_m)	resid_ing
	(1)	(2)
female	−0.163***	
	(0.015)	
age	0.091***	
	(0.004)	
age_sqr	−0.001***	
	(0.00005)	
resid_fem		−0.163***
		(0.017)
Constant	12.338***	−0.000
	(0.071)	(0.007)
Observations	9,892	9,892
R^2	0.069	0.012
Adjusted R^2	0.069	0.012
Residual Std. Error	0.739 (df = 9888)	0.739 (df = 9890)
F Statistic	245.548*** (df = 3; 9888)	119.495*** (df = 1; 9890)
Note:		*p<0.1; **p<0.05; ***p<0.01

Figura 4. Coeficiente de $female$



Nota: La figura muestra el histograma de los coeficientes de $female$ obtenidos a través del bootstrap con 5000 remuestreos.

La Figura 4 muestra la distribución del coeficiente de $female$. Al calcular el error estándar, obtenemos un valor de 0.01472, prácticamente idéntico al obtenido mediante la estimación por MCO.

Al estimarmos el modelo 2, no hemos considerado ningún control. Es posible que existan variables omitidas influenciando nuestros resultados. De esta manera, a fin de verificar la robustez agregamos los siguientes controles a nuestra estimación:

- **Capital humano**: Incluimos edad y su cuadrado como aproximación a la experiencia laboral, así como el nivel educativo más alto alcanzado. Estas variables capturan diferencias en productividad potencial.
- **Intensidad laboral**: En las especificaciones con salarios mensuales, se consideran las

horas usuales de trabajo y la existencia de un segundo empleo. En las especificaciones con salarios por hora, estas variables se omiten deliberadamente para evitar un ajuste redundante.

- **Características del empleo**: Incorporamos ocupación, tamaño de la firma, tipo de relación laboral, formalidad del empleo y la antigüedad en el puesto. Estos factores permiten aproximar la comparabilidad entre trabajos, tal como lo exige la noción de "igual trabajo".

La estrategia empírica consiste en estimar primero la brecha salarial sin controles y, posteriormente, añadir secuencialmente los bloques de covariables. De este modo, se puede observar cómo evoluciona el coeficiente asociado a la variable de género, lo que permite interpretar en qué medida el diferencial incondicional se explica por diferencias observables en características de los trabajadores y de sus empleos. En esta parte estimamos apenas como variable dependente la variable de ingreso de la ocupación principal por hora.
Nota: En la ecuación (5) se omitieron las variables relacionadas con el oficio y el número de personas que trabajan en la empresa.

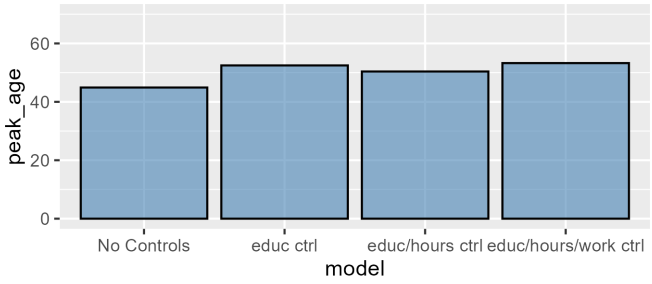
Los resultados en el Cuadro 7 muestran que el coeficiente de $female$ sigue siendo negativo y significativo. Además, la inclusión de controles incrementa la magnitud del coeficiente. El modelo con todos los controles indica que las mujeres llegan a ganar aproximadamente un 10 % menos que los hombres, dado que ambos tengan la misma experiencia, edad y ejerzan profesiones similares.

Un aspecto interesante sería estimar la edad pico en la que las mujeres alcanzarían su máximo salarial. Esto puede calcularse mediante un problema sencillo de maximización:

$$pico = -\frac{\beta_2}{2\beta_3},$$

donde β_2 es el coeficiente de age y β_3 es el coeficiente de age^2.

Figura 5. Edad pico según nuestros modelos



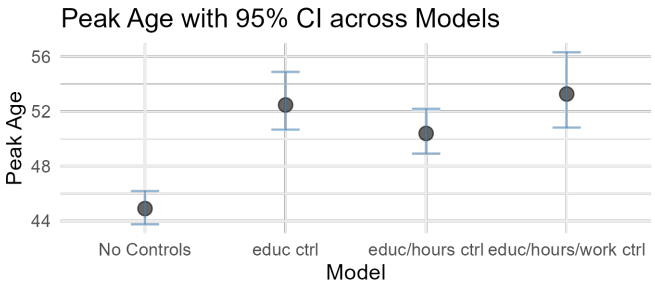
La Figura 5 muestra los resultados de la edad pico. En general, los modelos indican que la edad pico estaría entre los 45 y 54 años. La inclusión de controles parece incrementar

Cuadro 7. Resultados de las regresiones con controles

	Variable dependiente:				
	log(y_ingLab_m_ha)				
	(1)	(2)	(3)	(4)	(5)
female	−0.045***	−0.058***	−0.142***	−0.179***	−0.106***
	(0.015)	(0.014)	(0.012)	(0.012)	(0.012)
age		0.068***	0.062***	0.068***	0.038***
		(0.004)	(0.003)	(0.003)	(0.003)
age_sqr		−0.001***	−0.001***	−0.001***	−0.0003***
		(0.00004)	(0.00004)	(0.00004)	(0.00003)
primarios incompleto			0.199**	0.223**	0.167**
			(0.094)	(0.093)	(0.076)
as.factor(primario completo)4			0.289***	0.303***	0.203***
			(0.091)	(0.090)	(0.073)
as.factor(secundario incompleto)5			0.347***	0.355***	0.229***
			(0.091)	(0.089)	(0.073)
as.factor(secundario completo)6			0.565***	0.575***	0.292***
			(0.090)	(0.088)	(0.072)
as.factor(terciario)7			1.253***	1.238***	0.557***
			(0.089)	(0.088)	(0.073)
totalHoursWorked				−0.009***	−0.010***
				(0.0005)	(0.0004)
cuentaPropia					
microEmpresa					−0.392***
					(0.044)
formal					0.261***
					(0.015)
Constant	8.747***	7.392***	6.583***	6.940***	8.793***
	(0.010)	(0.068)	(0.104)	(0.104)	(0.164)
Observations	9,892	9,892	9,891	9,891	9,891
R^2	0.001	0.046	0.335	0.358	0.576
Adjusted R^2	0.001	0.045	0.335	0.358	0.572
Residual Std. Error	0.727 (df = 9890)	0.711 (df = 9888)	0.593 (df = 9882)	0.583 (df = 9881)	0.476 (df = 9798)
F Statistic	9.317*** (df = 1; 9890)	158.028*** (df = 3; 9888)	623.580*** (df = 8; 9882)	612.986*** (df = 9; 9881)	144.719*** (df = 92; 9798)
Note:					*p<0.1; **p<0.05; ***p<0.01

dicha edad. Cabe señalar que, al estimar la edad pico a partir de los coeficientes de MCO, no es posible calcular sus errores estándar ni los intervalos de confianza. Para generar dichos intervalos, empleamos el método bootstrap. En particular, remuestreamos los datos 2000 veces y registramos la distribución de la edad pico. Con esta distribución construimos el intervalo de confianza al 5 % para nuestros modelos.

Figura 6. Intervalos de confianza de la edad pico



La Figura 6 muestra los resultados de los intervalos de confianza. Como era de esperar, la inclusión de más controles amplía dichos intervalos; sin embargo, el incremento no supera los 2 años con respecto al valor central de la estimación.

En nuestras estimaciones encontramos que el coeficiente de mujer es negativo y significativo en todas las especificaciones, lo que confirma la existencia de una brecha salarial en contra de las mujeres. En el modelo más simple, esta brecha es incondicional y refleja la diferencia promedio de ingresos entre géneros. Al incorporar controles de edad, educación y características laborales, la magnitud de la brecha disminuye, lo que indica que parte de la diferencia se explica por la distinta composición de hombres y mujeres en términos de capital humano y tipo de empleo. Sin embargo, el hecho de que la brecha permanezca significativa aun en las especificaciones más completas señala la existencia de un componente no explicado. Esto puede interpretarse como evidencia de discriminación de género en el mercado laboral, aunque también podría reflejar variables no observadas o problemas de selección. En cualquier caso, los resultados sugieren que la brecha salarial no se reduce únicamente a diferencias observables y requiere una interpretación cuidadosa.

5. Prediciendo Salarios

Para evaluar y comparar el desempeño predictivo, reportamos el Error Cuadrático Medio de Predicción (RMSE) de las especificaciones de los dos puntos anteriores y lo contrastamos con cinco modelos adicionales que incorporan no linealidades y mayor complejidad. Los modelos seleccionados fueron los siguientes:

- **Modelo 1 (Edad cuadrática):**

$$\ln w_i = \beta_0 + \beta_1 \,\text{Edad}_i + \beta_2 \,\text{Edad}_i^2 + u_i$$

- **Modelo 2 (Brecha salarial por género incondicional):**

$$\ln w_i = \beta_0 + \beta_1 \,\text{Mujer}_i + u_i$$

- **Modelo 3 (Brecha salarial por género condicional):**

$$\begin{aligned}\ln w_i = {} & \beta_0 + \beta_1 \,\text{Mujer}_i + \beta_2 \,\text{Edad}_i + \beta_3 \,\text{Edad}_i^2 + \beta_4 \,\text{Educ}_i \\ & + \beta_5 \,\text{Horas}_i + \beta_6 \,\text{Microemp}_i + \beta_7 \,\text{Formal}_i + \beta_8 \,\text{TamañoFirma}_i + \beta_9 \,\text{Oficio}_i + u_i\end{aligned}$$

- **Modelo 4 (Modelo 3 con polinomio de tercer grado en edad e interacciones de este polinomio con el resto de controles):**

$$\begin{aligned}\ln w_i = {} & \beta_0 + \sum_{m=1}^{3} \beta_1^m \,\text{Edad}_i^m + \beta_2 \,\text{Mujer}_i + \beta_3 \,\text{Educ}_i + \beta_4 \,\text{Horas}_i \\ & + \beta_5 \,\text{Microemp}_i + \beta_6 \,\text{Formal}_i + \beta_7 \,\text{TamañoFirma}_i + \beta_8 \,\text{Oficio}_i \\ & + \sum_{m=0}^{3} \Big[\gamma_1^m (\text{Mujer}_i \,\text{Edad}_i^m) + \gamma_2^m (\text{Educ}_i \,\text{Edad}_i^m) + \gamma_3^m (\text{Horas}_i \,\text{Edad}_i^m) \\ & \quad + \gamma_4^m (\text{Microemp}_i \,\text{Edad}_i^m) + \gamma_5^m (\text{Formal}_i \,\text{Edad}_i^m) + \gamma_6^m (\text{TamañoFirma}_i \,\text{Edad}_i^m) \\ & \quad + \gamma_7^m (\text{Oficio}_i \,\text{Edad}_i^m)\Big] + u_i\end{aligned}$$

- **Modelo 5 (Modelo 4 pero con el polinomio de edad de grado 5):**

$$\begin{aligned}\ln w_i = {} & \beta_0 + \sum_{m=1}^{5} \beta_1^m \,\text{Edad}_i^m + \sum_j \beta_j \,\text{Control}_{ji} \\ & + \sum_{m=0}^{5} \sum_j \gamma_j^m \,(\text{Control}_{ji} \,\text{Edad}_i^m) + u_i\end{aligned}$$

Controles: Mujer, Educ, Horas, Microemp, Formal, TamañoFirma, Oficio.

- **Modelo 6 (Modelo 3 añadiendo Eduación Terciaria como control):**

$$\begin{aligned}\ln w_i = {} & \beta_0 + \beta_1 \,\text{Mujer}_i + \beta_2 \,\text{Edad}_i + \beta_3 \,\text{Edad}_i^2 + \beta_4 \,\text{Educ}_i \\ & + \beta_5 \,\text{Horas}_i + \beta_6 \,\text{Microemp}_i + \beta_7 \,\text{Formal}_i + \beta_8 \,\text{Tamaño}_i + \beta_9 \,\text{Oficio}_i \\ & + \beta_{10} \,\text{College}_i + u_i\end{aligned}$$

- **Modelo 7 (Modelo 6 con polinomios de grado 5 para edad y total de horas trabajadas):**

$$\begin{aligned}\ln w_i = {} & \beta_0 + \sum_{m=1}^{5} \beta_1^m \,\text{Edad}_i^m + \sum_{m=1}^{5} \beta_2^m \,\text{Horas}_i^m \\ & + \beta_3 \,\text{Mujer}_i + \beta_4 \,\text{Educ}_i + \beta_5 \,\text{Microemp}_i + \beta_6 \,\text{Formal}_i + \beta_7 \,\text{Tamaño}_i + \beta_8 \,\text{Oficio}_i + \beta_9 \,\text{College}_i + u\end{aligned}$$

- **Modelo 8 (Modelo 6 con polinomios de grado 3 para edad y educación, así como interacciones de estos polinomios con el resto de controles):**

$$\ln w_i = \beta_0 + \sum_{m=1}^{3} \beta_1^m \,\text{Edad}_i^m + \sum_{m=1}^{3} \beta_2^m \,\text{Horas}_i^m + \sum_{m=0}^{3} \sum_{j} \gamma_{1j}^m \,(\text{Edad}_i^m \,\text{Control}_{ji}) + \sum_{m=0}^{3} \sum_{j} \gamma_{2j}^m \,(\text{Horas}_i^m \,\text{Control}_{ji}) + u_i$$

Controles: Mujer, Educ, Microemp, Formal, Tamaño, Oficio, College.

Como se puede observar, para explorar nuevos modelos predictivos propusimos añadir polinomios a las variables continuas disponibles (edad y horas trabajadas), así como interactuar dichos polinomios con el resto de predictores del modelo. De esta manera, capturamos posibles relaciones no lineales y heterogeneidades en los retornos a la edad y a las horas según las características individuales y del mercado laboral. Esta estrategia nos permite evaluar si una mayor flexibilidad funcional se traduce en mejoras de desempeño predictivo. Los resultados de RMSE para los 8 modelos se presentan a continuación.

Cuadro 8. Comparación del RMSE entre los modelos seleccionados

Modelo	RMSE
1	0.67766
2	0.69275
3	0.44956
4	0.44769
5	0.54028
6	0.44956
7	0.44763
8	0.62034

Al comparar los ocho modelos estimados, observamos que las especificaciones más simples (Modelos 1 y 2) presentan los RMSE más altos, con valores de 0.678 y 0.693 respectivamente. Esto indica que los modelos que consideran únicamente edad o género, sin incorporar más información, son poco precisos en la predicción de los salarios. A medida que incluimos más controles y formas funcionales flexibles, el RMSE disminuye de manera importante, lo que evidencia la relevancia de capturar la heterogeneidad de las variables individuales y laborales para disminuir el sesgo.

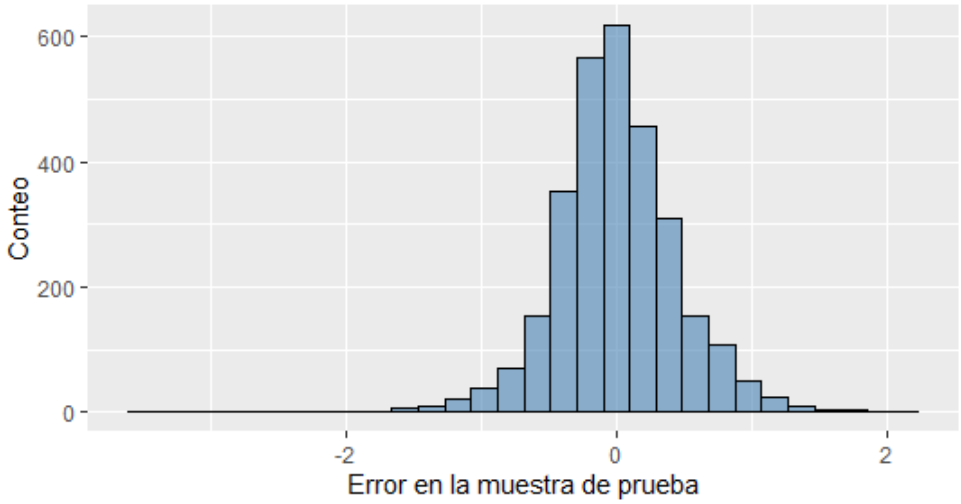
Por otro lado, los Modelos 3, 4, 6 y 7 alcanzan los RMSE más bajos, todos en torno a 0.448–0.450. Dentro de este grupo, destacan especialmente el Modelo 4 y el Modelo 7. El Modelo 4 introduce polinomios cúbicos en la edad junto con interacciones con variables clave como género, educación, formalidad u oficio, lo que permite capturar diferencias en cómo la edad impacta los salarios dependiendo de estas características. El Modelo 7, por su parte, amplía la flexibilidad al incorporar polinomios de quinto grado tanto en edad como

en horas trabajadas, además de los controles seleccionados para el Modelo 3, lo que le da capacidad para representar no linealidades más complejas en dos dimensiones fundamentales del mercado laboral. El hecho de que ambos modelos obtengan RMSE muy similares a los de especificaciones más sencillas (3 y 6) indica que gran parte de la capacidad predictiva ya está capturada por los controles, pero que la incorporación de estas formas funcionales adicionales ayuda a refinar la predicción en los márgenes donde las relaciones son más complejas.

Por último, los Modelos 5 y 8 muestran un RMSE más alto (0.540 y 0.620, respectivamente) en comparación con los modelos de complejidad intermedia. Esto refleja que, aunque las interacciones y polinomios de alto grado pueden capturar patrones no lineales, también pueden inducir sobreajuste y empeorar la capacidad predictiva fuera de muestra. En conjunto, nuestros resultados resaltan la importancia de un balance: los modelos con controles adecuados y cierta flexibilidad no lineal son los que logran el mejor desempeño, mientras que un aumento excesivo en complejidad resulta contraproducente.

Ahora, para la especificación con el menor error de predicción, conviene examinar las observaciones con los errores más altos (en valor absoluto) para determinar si esos desajustes se deben a limitaciones del modelo o, por el contrario, a posibles subdeclaraciones de ingreso por parte de algunos individuos. Con ese fin, analizamos la distribución de los errores.

Figura 7. Histograma del error de predicción en el "testing set"



Al examinar la distribución de los errores de predicción en la muestra de prueba, observamos que estos se concentran fuertemente alrededor de cero y siguen una forma aproximadamente simétrica. Esto indica que, en promedio, el modelo no presenta un sesgo sistemático: predice tanto por encima como por debajo del valor observado con una frecuencia similar. La mayoría de los errores se ubica en un rango relativamente estrecho, lo

que refleja un buen ajuste global, aunque se identifican algunas observaciones en las colas de la distribución. Estas colas corresponden a los casos en los que el modelo "falla" con mayor magnitud y, por lo tanto, resultan los más relevantes de analizar en detalle.

Al concentrarnos en las colas de la distribución —específicamente los percentiles 1 y 99— encontramos que estas observaciones extremas comparten ciertas características diferenciales frente al resto de la muestra. En particular, tienden a registrar menos horas trabajadas (43.97 en promedio, lo cual está por debajo del percentil 25 en la muestra total), menor grado de formalidad y una mayor presencia de trabajadores en microempresas. Esto sugiere que los errores más grandes no provienen de perfiles totalmente atípicos, sino de segmentos del mercado laboral donde la variabilidad en los ingresos es mayor y más difícil de capturar con el modelo. Además, se trata de grupos menos presentes en la muestra, lo que limita el ajuste y explica en parte por qué el modelo no logra predecirlos con precisión.

Cuadro 9. Descriptivas — Observaciones en las colas de la distirbución del error

Variable	**N = 60**
Age	36.98 (13.41)
Horas trabajadas	43.97 (19.89)
Educación	
1	1/60 (1.7 %)
3	0/60 (0 %)
4	4/60 (6.7 %)
5	5/60 (8.3 %)
6	13/60 (22 %)
7	37/60 (62 %)
Mujer	35/60 (58 %)
Micro empresa	23/60 (38 %)
Formal	32/60 (53 %)
Tamaño firma	
1	5/60 (8.3 %)
2	18/60 (30 %)
3	4/60 (6.7 %)
4	11/60 (18 %)
5	22/60 (37 %)
Educación Terciaria	13/60 (22 %)

Nota: Para variables continuas, media (desv. est.). Para categóricas, frecuencia (porcentaje).

Dentro de estas 60 observaciones extremas, identificamos cuatro casos con un leverage aproximadamente tres veces menor que la media, pero que se ubican por encima del percentil 99 de los errores de predicción. Este patrón resulta llamativo porque se trata de individuos con perfiles relativamente comunes —es decir, no atípicos en sus características observables— cuyos ingresos declarados superan de forma importante lo que el modelo predice. En este sentido, podrían constituir alertas para la DIAN, ya que los errores no parecen deberse a rarezas en el perfil, sino a posibles discrepancias entre lo esperado y lo reportado.

En conjunto, nuestro análisis muestra que la mayor parte de los errores extremos se asocian a segmentos con baja formalidad, microempresas y menos horas trabajadas, lo cual refleja limitaciones del modelo en grupos poco representados y con alta variabilidad de ingresos. Sin embargo, los pocos casos con leverage bajo y errores muy grandes merecen especial atención, ya que no se explican por perfiles raros y podrían corresponder a situaciones en las que los ingresos declarados se apartan de manera significativa de lo que cabría esperar. Así, los errores de predicción resultan útiles tanto para identificar limitaciones del modelo como para señalar posibles candidatos de investigación por parte de la DIAN.

Finalmente, con el fin de contrastar la estabilidad de nuestros resultados, aplicamos la técnica de Leave-One-Out Cross-Validation (LOOCV) a los dos modelos que previamente habían mostrado el menor error predictivo bajo el enfoque de validation set. Esta comparación nos permite evaluar si el desempeño de los modelos se mantiene cuando cada observación es excluida sucesivamente, lo cual otorga una medida más robusta de error fuera de muestra y se relaciona estrechamente con la sensibilidad del modelo frente a observaciones influyentes.

Para los dos modelos con mejor desempeño en validación simple, encontramos que el RMSE mediante el enfoque de validation set fue de 0.4477 en el Modelo 4 y de 0.4476 en el Modelo 7. Cuando aplicamos Leave-One-Out Cross-Validation (LOOCV), el error cambió ligeramente: el Modelo 4 obtuvo un RMSE de 0.4440, mientras que el Modelo 7 alcanzó 0.4504. Esto significa que, con LOOCV, el Modelo 4 mejora marginalmente su desempeño, mientras que el Modelo 7 empeora.

Las diferencias entre ambos métodos son pequeñas, pero informativas. El hecho de que el Modelo 4 mantenga un error estable e incluso algo menor bajo LOOCV sugiere que es más robusto y generaliza mejor a nuevas observaciones. En contraste, el Modelo 7 parece más sensible al esquema de validación: aunque su desempeño era casi idéntico al del Modelo 4 con validation set, en LOOCV su error aumenta, lo que indica mayor variabilidad y posible sobreajuste a la partición inicial de entrenamiento/validación.

En el caso del Modelo 4, la complejidad proviene de la combinación de polinomios cúbicos en la edad con interacciones con todas las demás variables del modelo. Esta estrategia le permite ajustar la relación entre edad y salario de manera diferenciada según género, educación, formalidad u oficio. Al incorporar la heterogeneidad de forma más estructurada,

el modelo logra mantener un desempeño estable incluso bajo un esquema exigente como el LOOCV, donde cada observación se deja por fuera en la estimación.

Por su parte, el Modelo 7 introduce polinomios de quinto grado en la edad y en las horas trabajadas, lo que genera curvas más sensibles y con oscilaciones más marcadas, en especial en los extremos del soporte. Al no estar acompañada de tantas interacciones, esta flexibilidad resulta menos organizada y más vulnerable a la exclusión de observaciones. En consecuencia, el error de predicción aumenta con LOOCV, lo que sugiere que el Modelo 7 es más inestable frente a datos atípicos o poco frecuentes.

En conclusión, el vínculo con la estadística de influencia es directo: en LOOCV, cada observación se deja por fuera una vez, y aquellas con alto leverage o con residuos grandes (PRESS residuals) afectan más el error total. En este sentido, es probable que el Modelo 7 sea más vulnerable a observaciones influyentes, lo que explica el aumento en su RMSE bajo LOOCV. El Modelo 4, en cambio, parece menos sensible a estas observaciones, lo que lo convierte en una especificación más estable frente a posibles outliers o casos de alto apalancamiento.

Referencias

Berkeley D-Lab (2024). Using big data for development economics. https://dlab.berkeley.edu/news/using-big-data-development-economics. Accedido en septiembre 2025.

Hossin, M. (2023). Big data-driven public policy decisions: Transformation, benefits, and challenges. *SAGE Open*, 13(4):1–12.