Problem Set 1: Predicting Income

Vivian Cabanzo Fernández
Cristian Felipe Muñoz Guerrero
Laura Daniela Diaz Torres
Zenneth Olivero Tapia

Septiembre 2025

Resumen

Este estudio analiza los determinantes del ingreso laboral en Bogotá utilizando la GEIH 2018, combinando un enfoque econométrico.Se examinan las diferencias salariales según edad y genero controlando por otras variables como educación, tipo de ocupación, formalidad laboral, tamaño de empresa. Los resultados muestran que los ingresos aumentan con la edad hasta alrededor de los 42 años, en línea con la teoría del capital humano, y que persiste una brecha salarial de género de aproximadamente 15 % a favor de los hombres.Además, se estimaron ocho modelos de regresión y se validaron con técnicas de cross-validation, encontrando que el modelo más completo, que incorpora un conjunto amplio de variables, ofrece el mejor desempeño predictivo. Estos hallazgos resaltan la utilidad de aplicar herramientas de Big Data y Machine Learning para comprender la dinámica salarial y aportar evidencia para políticas públicas orientadas a reducir desigualdades en el mercado laboral.

Palabras clave: Ingresos laborales, Brechas salariales, GEIH, Predicción, Validación cruzada

Clasificación JEL: C81, C55, J31, O15, H26

El repositorio con el código reproducible en el Software R, se encuentra disponible en:

`https://github.com/LauraDaniela17/Taller1_BigData_MachineLearning`

1. Introducción

La distribución de los ingresos derivados del trabajo y las desigualdades que la atraviesan constituyen uno de los temas más relevantes en la economía contemporánea. Comprender cómo influyen factores como la edad, la educación, la ocupación, la formalidad laboral y el género en los salarios no solo permite explicar las diferencias existentes en el mercado laboral, sino también aporta elementos clave para el diseño de políticas públicas orientadas a una asignación más equitativa de los recursos.

Este estudio analiza la dinámica salarial de la población ocupada en Bogotá a partir de la Gran Encuesta Integrada de Hogares (GEIH) 2018, elaborada por el DANE. La muestra utilizada incluye información socioeconómica y laboral de los individuos, lo que permite explorar tanto los determinantes observables del ingreso como los patrones generales de desigualdad. A partir de esta base, se abordan tres dimensiones centrales: el perfil edad–salario, la brecha salarial de género y la capacidad predictiva de distintos modelos econométricos.

En primer lugar, se encuentra que los salarios aumentan con la edad y la experiencia hasta alcanzar un máximo alrededor de los 42 años, para luego estabilizarse y decrecer, en línea con la teoría del capital humano. En segundo lugar, se evidencia la persistencia de una brecha salarial de género: aun después de controlar por educación, tipo de empleo y tamaño de empresa, los hombres ganan en promedio un 15 % más que las mujeres, lo que sugiere la presencia de factores estructurales y posibles mecanismos de discriminación. Finalmente, se evalúa la predicción del salario horario mediante técnicas de validación cruzada —Validation Set Approach, K-Fold y Leave-One-Out—, comparando ocho modelos que incorporan diferentes combinaciones de variables. Los resultados muestran que los modelos más completos, aquellos que incluyen características laborales y de caracterización poblacional, alcanzan menores valores de error cuadrático medio (RMSE), reflejando una mayor capacidad predictiva frente a las especificaciones más simples.

En conjunto, los hallazgos no solo confirman la existencia de desigualdades salariales en Bogotá, sino que también ponen de relieve la utilidad de aplicar herramientas de Big Data y Machine Learning al estudio de fenómenos económicos. Este enfoque, además de ser metodológicamente robusto, permite capturar con mayor precisión las dinámicas del mercado laboral y aporta evidencia valiosa para la formulación de políticas orientadas a reducir brechas y mejorar la equidad en el acceso a ingresos laborales.

2. Datos

Los datos de este estudio provienen de la *Gran Encuesta Integrada de Hogares* (GEIH), desarrollada por el Departamento Administrativo Nacional de Estadística (DANE). Esta encuesta es la principal fuente oficial de información sobre el mercado laboral colombiano y las mediciones de pobreza y desigualdad. Recopila información socioeconómica de los hogares en diferentes regiones del país y, extrapola los resultados con el fin de reflejar la situación real de la población residente. [1]

En este trabajo se emplea la base correspondiente al año 2018. Aunque coincide con la realización del *Censo Nacional de Población y Vivienda 2018*, el diseño muestral de la GEIH aún no incorporaba los ajustes derivados de dicho censo, continuando con el marco basado en 2005. Para el análisis se restringió la muestra a los individuos encuestados en Bogotá (32.177 observaciones), una de las áreas urbanas con mayor número de observaciones, lo que permite obtener una aproximación representativa de la dinámica del mercado laboral y de la desigualdad en ingresos del país.

2.1. Obtención de los datos

Cabe aclarar que la información empleada no proviene de un archivo de descarga directa de la web oficial del DANE. Se utilizó la información publicada en el portal de GitHub del profesor Ignacio Sarmiento de la Universidad de los Andes[2],la cual se encuentra organizada en diez páginas `HTML`, cada una de las cuales contiene una tabla de datos.

Para construir la base de datos completa fue necesario implementar un procedimiento de *web scraping* en `R Studio`. Este proceso consistió en: (1.) generar las direcciones web de las diez paginaciones, (2.) extraer la primera tabla de cada página mediante funciones de la librería `rvest`, (3.) convertir dichas tablas en data tables, y por último (4.) limpiar los nombres y los tipos de variables con el paquete `janitor`. Finalmente, se integraron todas las tablas en un único conjunto de datos para su análisis respectivo.

Es importante resaltar que este ejercicio de recolección no presenta restricciones de acceso, pues la información está publicada de manera abierta y además, no representa riesgos penales según la normativa colombiana sobre el *web scraping*.

[1] Véase DANE, *Gran Encuesta Integrada de Hogares 2018: Descripción de la operación estadística*, disponible en:`https://microdatos.dane.gov.co/index.php/catalog/547/study-description`.

[2] Disponible en: `https://ignaciomsarmiento.github.io/GEIH2018_sample/`.

2.2. Limpieza y preparación de los datos

Una vez consolidada la base de datos, se procedió a realizar un proceso de filtrado con el fin de garantizar el propósito del análisis. En primer lugar, se restringió la muestra a individuos de 18 años o más que declararon estar ocupados. Con esta depuración, el tamaño muestral se redujo a 16.542 observaciones.

Posteriormente, se realizó una depuración de la base conservando únicamente las variables relevantes para el análisis, tanto de interés principal como de control, las cuales fueron empleadas en los modelos econométricos presentados más adelante. En particular, se incluyeron variables como: sexo, edad, tipo de ocupación, formalidad laboral, tamaño de la firma, nivel educativo, horas usuales trabajadas por semana y el ingreso mensual de la primera actividad.

La selección de las variables de control responden a la necesidad de capturar los principales determinantes de los ingresos laborales, de acuerdo con la teoría del capital humano y la segmentación del mercado de trabajo. El nivel educativo es uno de los factores más relevantes del capital humano, dado que mayores años de estudio suelen traducirse en mayores ingresos. El tipo de ocupación y la formalidad laboral capturan diferencias institucionales y contractuales que influyen en la remuneración. El tamaño de la firma es importante porque empresas más grandes tienden a pagar salarios más altos debido a economías de escala o estructuras salariales más formales.

Dado que la variable interés se define como el ingreso por hora y la GEIH no la reporta directamente, fue necesario construir una aproximación. Para ello, se empleó la variable `impa`, que según la documentación oficial del DANE, corresponde al ingreso mensual de la actividad principal realizada. A partir de este valor se estimó el ingreso/hora dividiéndolo entre las horas usuales trabajadas en dicha actividad, multiplicadas por 4,33 (promedio de semanas en un mes).

$$w_i = \frac{Ingresoprimeraactividad_i}{Horastrabajadassemana_i \times 4{,}33} \tag{1}$$

Es necesario resaltar que, si bien asumir que todos los individuos trabajaron exactamente 4,33 semanas al mes constituye un supuesto fuerte, esta aproximación se justifica debido a que permite estandarizar la medición del ingreso por hora a partir de la información disponible, facilitando la comparabilidad entre observaciones y garantizando la consistencia del análisis.

Adicionalmente, se llevó a cabo una revisión de la calidad de los datos con el fin de identificar inconsistencias en la base. Se identificó que tienen resultados durante 12 meses, y que, en este período, ningún encuestado

presenta más de una observación. Esto significa que la información disponible corresponde a una única medición por individuo, por lo que el estudio se centrará en analizar cómo se comportaron los salarios en promedio durante el año, y no en la continuidad mes a mes de cada encuestado

Sobre los valores faltantes, se encontró una sola observación para la variable correspondiente al máximo nivel educativo alcanzado, por lo cual se decidió utilizar la moda de la variable para la imputación del faltante. Por otro lado, en la variable de interés, se encontró que aproximadamente el 1,5 % de las observaciones presentan valores faltantes y cerca del 9 % reportan un ingreso laboral por hora igual a cero. Estos casos pueden reflejar tanto problemas de reporte como situaciones de trabajo no remunerado; sin embargo, no corresponden al objeto de análisis de este estudio, centrado en los determinantes de los ingresos laborales positivos y la brecha de género. Por esta razón, dichas observaciones fueron excluidas de la muestra final.

Asimismo, se transformó el ingreso por hora mediante el logaritmo natural, estrategia ampliamente utilizada en la literatura. En general, las distribuciones de ingresos presentan colas pronunciadas debido a la presencia de valores atípicos, ya sea muy altos o muy bajos, que pueden distorsionar los resultados de los modelos econométricos. El uso del logaritmo permite suavizar parcialmente el peso de estos valores extremos, reduciendo la asimetría de la distribución y acercándola a una forma más próxima a la normalidad. Esta decisión metodológica, además, exige la exclusión de ingresos con valor a cero por su misma naturalidad matemática.

2.3. Estadísticas descriptivas

En el cuadro 1 se refleja un resumen descriptivo de las variables que se tuvieron en cuenta en el análisis. La edad promedio de la población encuestada de mujeres es de 38.8 años y de los hombres fue de 38.9, los cuales reportaron trabajar 44.29 y 49.86 horas respectivamente, con una diferencia media de 5.56 horas trabajadas. El ingreso por hora promedio de las mujeres fue de $ 8.402 pesos y de los hombres de $ 8.849, representado una diferencia media de $ 807.3 pesos. Adicionalmente, en términos de formación educativa, se pueden obsevar diferencias entre los niveles máximos de educación entre hombres y mujeres desde preescolar hasta secundaria completa, donde los hombres superan a las mujeres en dichos niveles alcanzados. Sin embargo, resulta interesante que existe una mayor proporción de mujeres que lograron un nivel terciario de educación que los hombres. Por último, relacionado a la formalidad laboral, el 60 % de la poblacion se encuentra en trabajo formal, proporcion que se mantiene muy similar entre hombres y mujeres. Dentro de la población muestreada, 3 de cada 10 personas trabajan en empreas con más de 15 empleados, mientras que 2 de cada 10 son autoempleados.

Cuadro 1: Tabla de descriptivas por sexo

	Mujer (N=6973)		Hombre (N=7772)		Diff. in Means	Std. Error
	Mean	Std. Dev.	Mean	Std. Dev.		
Ingreso por hora	8042.18	11332.16	8849.57	14702.55	807.39	215.01
Edad	38.80	12.81	38.97	13.53	0.17	0.22
Horas trabajadas	44.29	14.71	49.86	14.82	5.56	0.24
Nivel educativo máximo	N	Pct.	N	Pct.		
Ninguno	54	0.8	43	0.6		
Preescolar	0	0.0	0	0.0		
Primaria incompleta	286	4.1	376	4.8		
Primaria completa	585	8.4	758	9.8		
Secundaria incompleta	728	10.4	1004	12.9		
Secundaria completa	2142	30.7	2665	34.3		
Terciaria	3178	45.6	2926	37.6		
N/A	0	0.0	0	0.0		
Formalidad	N	Pct.	N	Pct.		
Informal	2761	39.6	3095	39.8		
Formal	4212	60.4	4677	60.2		
Tamaño de la firma	N	Pct.	N	Pct.		
Auto-empleado	1789	25.7	1741	22.4		
2–5 empleados	1175	16.9	1575	20.3		
6–510 empleados	440	6.3	592	7.6		
11–15 empleados	868	12.4	1019	13.1		
Más de 15	2701	38.7	2845	36.6		
Tipo de trabajo	N	Pct.	N	Pct.		
Empleado empresa particular	4102	58.8	4651	59.8		
Empleado del gobierno	272	3.9	299	3.8		
Empleado doméstico	541	7.8	22	0.3		
Cuenta propia	1901	27.3	2476	31.9		
Empleador	150	2.2	322	4.1		
Sin remuneración (familiar)	0	0.0	0	0.0		
Sin remuneración (No fam.)	0	0.0	0	0.0		
Jornalero	0	0.0	1	0.0		
Otro	7	0.1	1	0.0		

Sobre la variable de interés de este estudio, es decir, el logaritmo del ingreso por hora, se realizó una gráfica la cualmuestra cómo se distribuye la variable para hombres y mujeres. En general, las dos curvas tienen un gran parecido, donde la mayoría de las personas, sin importar el sexo, se concentran en un rango bastante similar de ingresos.

Sin embargo, hay algunos matices importantes. En los niveles más bajos de ingreso, se nota una ligera mayor presencia de mujeres. En cambio, en la parte alta de la distribución, donde se ubican los salarios más elevados, la curva de los hombres se extiende un poco más, lo que indica que ellos tienen más probabilidades de alcanzar ingresos altos. Lo anterior quiere decir que, aunque hombres y mujeres comparten un patrón de ingresos muy parecido, las mujeres tienden a estar más representadas en los tramos bajos y los hombres en los más altos.

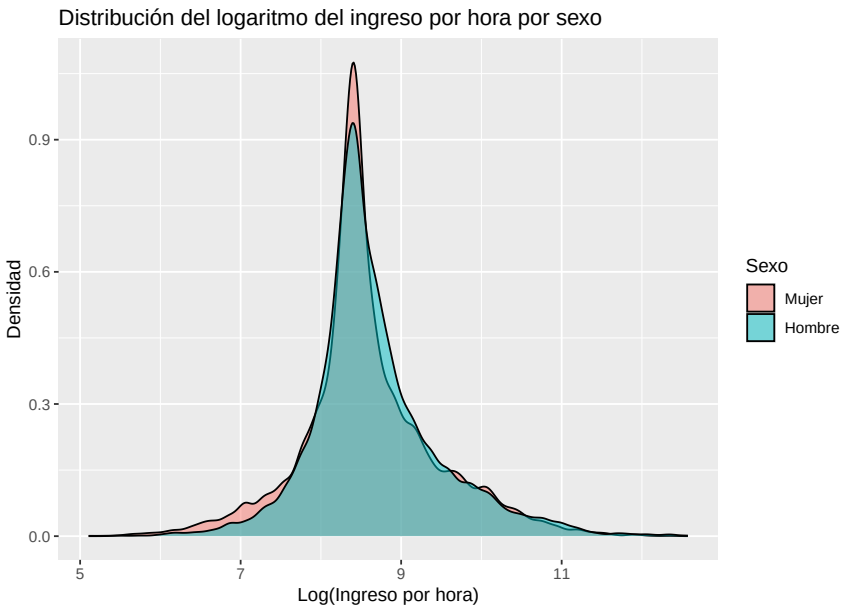

Figura 1: Gráfica de densidad del logaritmo del ingreso por hora para mujeres y hombres

La figura 2 muestra cómo cambian los ingresos por hora en logaritmos a medida que las personas alcanzan distintos niveles educativos. Como era de esperarse, quienes tienen más años de estudio suelen tener un ingreso mayor, y esta tendencia se aprecia de manera consistente tanto en hombres como en mujeres. Sin embargo, lo llamativo es que la diferencia de género se mantiene en todos los niveles: incluso cuando mujeres y hombres tienen la misma educación, los hombres siguen concentrándose en los tramos más altos de ingreso.

El nivel terciario (técnicos, tecnólogos,profesionales y postgrado) es el que refleja con mayor claridad esta brecha, ya que allí los hombres no solo tienen ingresos promedio más altos, sino también una mayor dispersión hacia valores elevados. En otras palabras, la educación ayuda a mejorar el ingreso, pero no alcanza por sí sola para cerrar la brecha salarial entre mujeres y hombres.

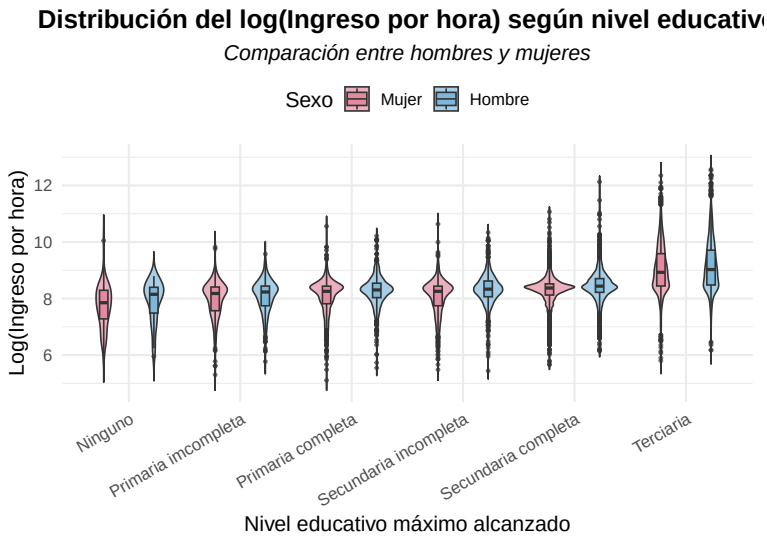

Figura 2: Gráfica de violín distribucón del logarítmo del ingreso por nivel educativo

Por último, la figura 3 nos permite observar con mayor detalle cómo varían los ingresos por hora, dependiendo del tipo de trabajo que desempeñan las personas. A simple vista, se observa que las diferencias no solo dependen de la ocupación, sino también del sexo. Por ejemplo, quienes trabajan en el gobierno presentan, en promedio, los ingresos más altos dentro de las categorías analizadas; sin embargo, incluso allí se mantiene la ventaja para los hombres. En el extremo opuesto, el trabajo doméstico aparece como el sector con los menores ingresos y con muy poca dispersión, lo que sugiere una homogeneidad en la precariedad que afecta de manera especial a las mujeres.

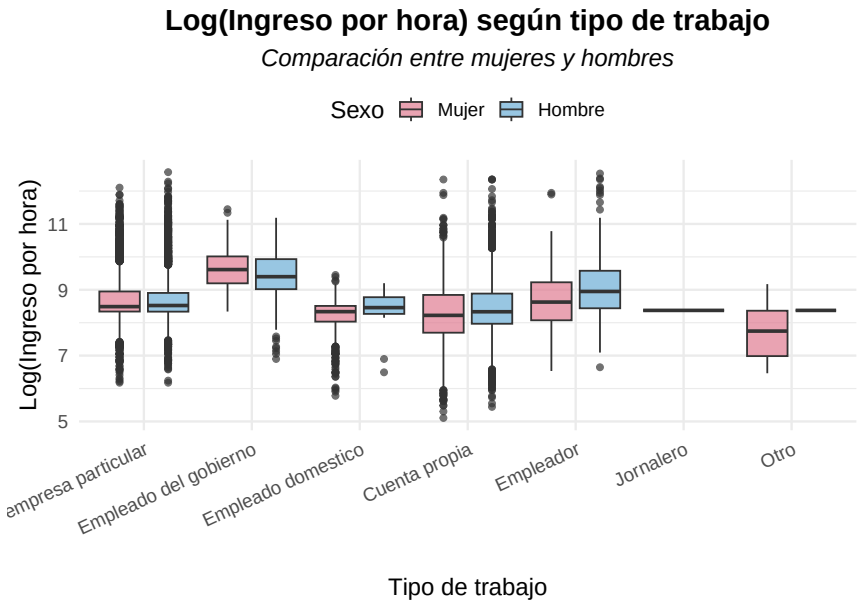

Figura 3: Gráfica boxplot distribución del logarítmo del ingreso por tipo de trabajo

Las ocupaciones por cuenta propia y los empleadores muestran una historia diferente: en estos grupos los ingresos son mucho más heterogéneos, con una dispersión amplia que refleja tanto casos de ingresos muy bajos como de ingresos particularmente altos. Aquí también son los hombres quienes logran acceder a la parte superior de la distribución. Finalmente, en categorías con menor presencia, como jornaleros u "otros", los ingresos se concentran en niveles bajos y la brecha de género parece menos pronunciada, aunque ello responde más a las limitaciones de las oportunidades que a una mayor equidad.

3. Perfil de salarios según edad

En el cuadro 2 se presentan los resultados de la regresión del logaritmo del ingreso por hora sobre la edad y la edad al cuadrado. Los coeficientes son significativos al 1 % y muestran el patrón esperado: el de la edad es positivo y el de la edad al cuadrado es negativo. Esto implica que los salarios crecen con la edad en las primeras etapas de la vida laboral, pero después de cierto punto los incrementos se reducen hasta volverse negativos.

Cuadro 2: Regresión OLS del perfil edad–salario

	Coef.	**EE**	**p-valor**
Constante	7.582***	(0.060)	<0.001
Edad	0.055***	(0.003)	<0.001
Edad^2	-0.00065***	(0.00003)	<0.001
Observaciones	14,745		
R^2	0.023		
F-statistic	173.9		<0.001

Notas: Errores estándar robustos entre paréntesis. $^{***}p < 0{,}01$, $^{**}p < 0{,}05$, $^{*}p < 0{,}1$.

Al tratarse de un modelo Log-Lin, el coeficiente de edad (0.005) se puede interpretar que en promedio, un año adicional aumenta en 5,55 % el ingreso, sin embargo, dichos efedtos se van reduciendo teniendo en cuenta el coeficiente al cuadrado de edad. Ambos coeficientes son altamente significativos (p <0.001), lo cual confirma que la relación entre edad e ingresos no es lineal sino de tipo "curva". En otras palabras, la edad tiene un efecto positivo al inicio, pero a medida que se acumulan más años, existen rendimientos decrecientes en el aumento de los ingresos.

El modelo tiene un R^2 de 0.023, lo que significa que la edad por sí sola explica alrededor del 2.3 % de las diferencias en el logaritmo del ingreso horario. Aunque este valor es bajo, es esperable porque los salarios también dependen de muchas otras variables como educación, ocupación, género, sector o experiencia laboral. A pesar de esto, el test F es significativo (p <0.001), lo que indica que el modelo en conjunto sí aporta información útil para entender la relación entre edad y salarios.

Cuadro 3: Edad pico del perfil edad–salario (OLS, bootstrap percentil)

Estadístico	**Valor**
Edad pico (puntual)	42.28
IC 95 % percentil (LI)	41.48
IC 95 % percentil (LS)	43.15
Réplicas válidas	1000.00

Nota: Intervalos de confianza construidos con bootstrap percentil (1,000 réplicas). Modelo OLS: $\log(w)$ sobre Edad y Edad^2.

De acuerdo a lo evidenciado en el cuadro 3, se calculó la edad pico usando los coeficientes de la regresión. La edad pico corresponde al punto máximo de la curva edad–salario. El valor estimado fue de 42.3 años, con un intervalo de confianza al 95 % de entre 41.5 y 43.1 años, obtenido mediante bootstrap.

Esto confirma que los ingresos aumentan en edades tempranas y medias, alcanzan un máximo en los 40s y luego empiezan a disminuir. Aunque la literatura suele ubicar el pico alrededor de los 50 años, en esta muestra se encuentra antes, lo que puede deberse a características específicas del mercado laboral analizado.

4. Brecha salarial de género

4.1. Brecha salarial sin controles

El modelo presentado estima la brecha salarial entre hombres y mujeres sin incluir variables de control adicionales que expliquen diferencias individuales o laborales. La variable dependiente es el logaritmo del salario, y la única variable explicativa es el sexo (hombre y mujer) del individuo, codificada como 1 para hombres.

El coeficiente estimado en el cuadro 4 para la variable `sexo` es de 0,09, lo que indica que, en promedio, los hombres ganan aproximadamente un 9.4 % más que las mujeres. Esta diferencia se calcula como $\exp(0{,}09) - 1 \approx 0{,}094$, lo que representa la brecha salarial *incondicional*, es decir, sin ajustar por características como edad, educación o tipo de empleo.

El coeficiente es estadísticamente significativo al nivel del 1 % ($p < 0{,}01$), lo que muestra que la diferencia observada no es producto del azar. El intervalo de confianza del 95 % para este coeficiente va de 0,06 a 0,11, lo que refuerza la robustez de la estimación.

Cuadro 4: Estimación de Brecha Salarial

Modelo log-lineal sin controles

Término	**Esti.**	**EE**	**Estad.**	**p-valor**	**Conf. Inf.**	**Conf. Sup.**
Intercept	8.58	0.01	874.8	0.00	8.56	8.60
Sex (1 = Hombre)	0.09	0.01	6.37	0.00	0.06	0.11

Notas: Modelo sin controles. Intervalos de confianza al 95 %. Valores en log-salario.

Este resultado refleja una penalización salarial promedio para las mujeres en el mercado laboral, pero no muestra si la brecha se debe a diferencias en características observables (como experiencia o nivel educativo) o a factores como discriminación o clasificación ocupacional. Para ello, es necesario adicionar controles en el modelo, como se hace en la siguiente sección.

4.2. Brecha salarial con controles

El modelo log-lineal con controles permite estimar la brecha salarial entre hombres y mujeres ajustando por características laborales relevantes. En contraste con el modelo incondicional, este enfoque considera factores que pueden explicar parte de la disparidad salarial observada.

$$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \theta X + u$$

El coeficiente estimado en el Cuadro 5 para la variable `sexo` es de 0,14, lo que indica que, manteniendo constantes las demás variables incluidas en el modelo, los hombres ganan en promedio un 15 % más que las mujeres. Esta diferencia se interpreta como $\exp(0{,}14) - 1 \approx 0{,}150$, y es estadísticamente significativa al nivel del 1 % ($p < 0{,}01$), con un intervalo de confianza del 95 % entre 0,12 y 0,16.

Este muestra que, después de controlar por edad, nivel educativo, tipo de empleo, formalidad laboral y tamaño de empresa, continua una penalización salarial para las mujeres. En otras palabras, la brecha salarial no se explica completamente por diferencias en características observables, siendo la presencia de factores caractización o institucional que afectan el salario en mujeres.

Modelo log-lineal con controles

Cuadro 5: . Estimación de Brecha Condicional con Controles

TERM	**Esti.**	**Std. error**	**Stad.**	**p.value**	**conf. low**	**conf. high**
Intercept	6,48	0,08	79,10	0,00	6,32	6,54
Sex (1 = Hombre)	0,14	0,01	13,07	0,00	0,12	0,16
Age	0,04	0,00	17,90	0,00	0,04	0,05
Primaria Incompleta	0,14	0,07	2,01	0,04	0,00	0,28
Primaria Completa	0,24	0,07	3,64	0,00	0,11	0,38
Secundaria Incompleta	0,32	0,07	4,73	0,00	0,18	0,45
Secundaria Completa	0,42	0,07	6,39	0,00	0,29	0,55
Terciaria	1,00	0,07	15,09	0,00	0,87	1,13
Empleado Público	0,32	0,03	11,20	0,00	0,26	0,38
Empleado Doméstico	0,14	0,03	4,20	0,00	0,08	0,21
Cuenta Propia	0,09	0,02	5,04	0,00	0,06	0,13
Empleador	0,51	0,03	15,79	0,00	0,45	0,57
Jornalero	0,59	0,64	0,92	0,36	−0,66	1,84
Otro tipo de relación laboral	−0,37	0,23	−1,65	0,10	−0,82	0,07
Formalidad Laboral	0,33	0,02	21,68	0,00	0,30	0,36
Empresas: 2-5 empleados	0,07	0,02	3,48	0,00	0,03	0,12
Empresas: 6-10 empleados	0,19	0,03	6,89	0,00	0,14	0,25
Empresas: 11-15 empleados	0,27	0,03	6,89	0,00	0,14	0,25
Empresas: más de 15 empleados	0,38	0,02	15,53	0,00	0,33	0,42

Notas: Errores estándar robustos entre paréntesis. $^{***}p < 0{,}01$, $^{**}p < 0{,}05$, $^{*}p < 0{,}1$.

Se estima el siguiente modelo log-lineal del ingreso horario en logaritmo:

$$\log(w_i) = \beta_0 + \beta_1 \cdot \text{sexo}_i + \beta_2 \cdot \text{edad}_i + \beta_3 \cdot \text{edad}_i^2 + \beta_4 \cdot \text{educación}_i + \beta_5 \cdot \text{tipo\_empleo}_i + \beta_6 \cdot \text{formalidad}_i + \beta_7 \cdot \text{tamaño\_empresa}_i + \varepsilon_i$$

Además, los coeficientes de las variables de control muestran patrones esperados: la educación tiene un efecto positivo y creciente sobre el salario, la formalidad laboral está asociada con mayores ingresos, y el tamaño de la empresa también influye positivamente. La edad presenta un efecto positivo, aunque no lineal, como lo indica el coeficiente de la variable cuadrática.

Este modelo ofrece una visión más precisa de la brecha salarial, al aislar el efecto del sexo sobre el ingreso de otros factores relevantes. Sin embargo, la persistencia de una brecha significativa sugiere que podrían existir mecanismos de discriminación o caracterización ocupacional que no se capturan completamente en el modelo.

4.2.1. Estimación con FWL

El modelo estimado mediante el método de Frisch-Waugh-Lovell (FWL) permite identificar el efecto parcial de la variable `res_sex` sobre el ingreso, una vez se resta el impacto de otras covariables. El coeficiente estimado para `res_sex` es de 0.1414, lo que quiere deicir, que manteniendo constantes las demás variables, los hombres presentan un ingreso aproximadamente 15.2 % mayor que las mujeres, según la transformación exponencial del coeficiente ($\exp(0{,}1414) - 1 \approx 0{,}152$).

Este resultado es altamente significativo (valor t = 13.07, p $< 2 \times 10^{-16}$), lo que muestra que la diferencia observada no es producto del azar. El intercepto, cercano a cero y no significativo, tiene sentido con la estructura de una regresión entre residuos, donde el promedio de los errores es nulo por construcción. Los residuos del modelo muestran una distribución centrada en cero, sin evidencia de sesgos extremos, aunque no se evalúan aquí supuestos como normalidad o homocedasticidad.

Modelo de regresión lineal: `res_w` sobre `res_sex`

Resumen de residuos

Mínimo	1er Cuartil	Mediana	3er Cuartil	Máximo
-2.9902	-0.3611	-0.0359	0.3254	3.8200

Coeficientes del modelo

Variable	Estimación	Error estándar	Valor t	Pr(>|t|)
Intercepto	$1{,}064 \times 10^{-17}$	0.005244	0.00	1.000
`res_sex`	0.1414	0.01082	13.07	$< 2 \times 10^{-16}$

Códigos de significancia:
*** $< 0{,}001$, ** $< 0{,}01$, * $< 0{,}05$, . $< 0{,}1$, $> 0{,}1$

En cuanto a los estadísticos globales, el R^2 ajustado es bajo (0.0114), lo cual es esperable en este tipo de regresión, ya que el objetivo no es explicar la varianza total del ingreso, sino aislar el efecto neto del sexo. Sin embargo, el estadístico F (170.9) y su valor-p extremadamente bajo confirman que la variable `res_sex` tiene capacidad explicativa dentro del modelo. En conjunto, estos resultados sugieren la existencia de una brecha salarial condicional

significativa entre hombres y mujeres, incluso después de controlar por factores como edad, educación, tipo de empleo y tamaño de empresa. El uso del método FWL refuerza la validez de esta estimación al eliminar posibles sesgos por colinealidad o efectos indirectos de otras covariables.

Estadísticos del modelo

- Error estándar residual: 0.6367 (gl = 14,742)
- R^2 múltiple: 0.01146
- R^2 ajustado: 0.0114
- Estadístico F: 170.9 (gl = 1 y 14,742)
- Valor-p del modelo: $< 2{,}2 \times 10^{-16}$

4.2.2. Estimación con FWL y bootstrap

El intervalo de confianza del 95 % para el coeficiente estimado fue calculado mediante el método percentil, utilizando 1000 réplicas bootstrap. Este enfoque no paramétrico permite evaluar la incertidumbre de la estimación sin dar por sentada la normalidad en la distribución de los errores, lo cual es especialmente útil en contextos donde los supuestos clásicos pueden no cumplirse.

El intervalo obtenido en la escala original del modelo fue (0,1197, 0,1627), lo que indica que, con un 95 % de confianza, el efecto condicional del sexo sobre el ingreso se encuentra dentro de ese rango. Dado que el intervalo no incluye el valor cero, se refuerza la evidencia de una diferencia salarial significativa entre hombres y mujeres, incluso bajo el remuestreo que captab la variabilidad empírica de los datos. Esta robustez metodológica complementa la inferencia clásica y aporta mayor credibilidad a la estimación obtenida mediante el método FWL.

Llamado de función:

```
boot.ci(boot.out = boot_res, type = "perc")
```

Intervalo de confianza (95 %):

Nivel de confianza	Intervalo percentil
95 %	(0.1197, 0.1627)

4.3. Picos de ingreso por género

El análisis de la relación entre edad y salario permite identificar el momento en el que los ingresos alcanzan su punto más alto antes de comenzar a descender. A este momento se le conoce como edad pico salarial y refleja una etapa importante dentro del ciclo de vida laboral, que no necesariamente ocurre a la misma edad en hombres y mujeres.

Para reflejar lo anterior, se estimó la relación ingreso-edad por género en la siguiente gráfica, cuyos resultados muestran que, en promedio, tanto hombres como mujeres alcanzan su pico de ingresos entre los 40 y 50 años, con la pequeña diferencia de que el pico en los hombres está mas cercano a los 50 y el de las mujeres a los 40. Esto sugiere que las trayectorias salariales masculinas tienden a extenderse durante más tiempo, mientras que las de las mujeres suelen estabilizarse y empezar a descender antes.

Ahora bien, desde el punto de vista estadístico, las diferencias no son tan claras como parecen. Los intervalos de confianza de las estimaciones se superponen, lo que significa que no podemos asegurar con total certeza que hombres y mujeres alcancen su máximo en momentos realmente distintos.

Donde sí hay una diferencia evidente es en el nivel de salario alcanzado en ese punto máximo: los hombres, en promedio, llegan a un ingreso más alto que las mujeres. Este hallazgo confirma que la brecha salarial de género no solo se refleja en los promedios generales, sino también en el tope de las trayectorias laborales.

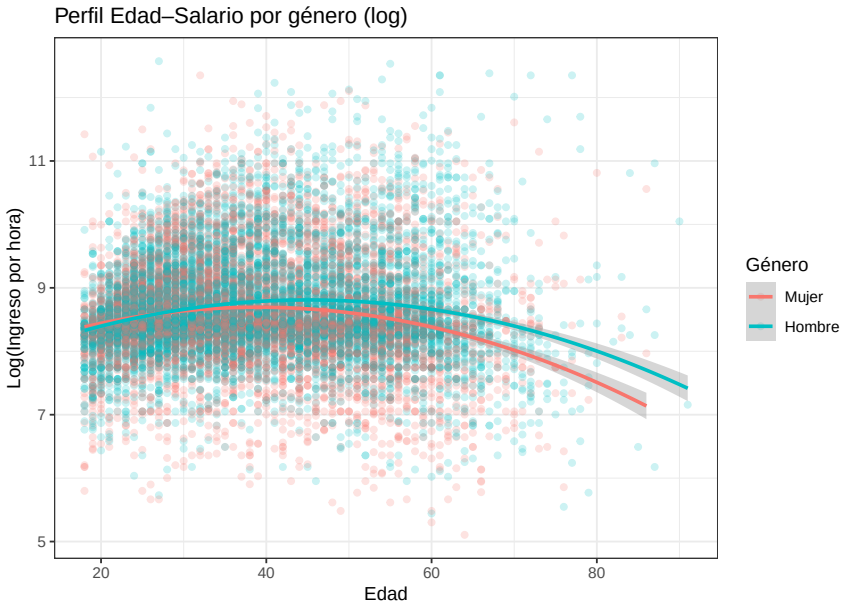

Figura 4: Perfil edad–salario por género (log)

En conclusión, aunque la edad en la que hombres y mujeres alcanzan su máximo ingreso puede no diferir de manera significativa, la desigualdad se mantiene en los niveles de salario. Esto muestra que la brecha de género sigue presente y se manifiesta incluso en el mejor momento de la carrera laboral.

5. Predicción de salarios

En el análisis de modelos, la evaluación de su desempeño predictivo es fundamental para analizar la información de manera adecuada. Para ello, se emplean diversas técnicas de validación que permiten estimar el error de predicción de diferentes modelos para seleccionar de manera adecuada el que mejor poder de predicción tiene, en este caso, los salarios por hora. En el presente análisis se comparan tres enfoques ampliamente utilizados, a saber:

Validation Set Approach (VSA)

Este método consiste en dividir el conjunto de datos en dos partes: un conjunto de entrenamiento y un conjunto de validación. El modelo se entrena sobre la primera parte y se evalúa sobre la segunda. Aunque es simple de implementar, su principal desventaja es que los resultados pueden variar considerablemente dependiendo de cómo se realice la división. Para el ejercicio, los datos se segregaron en una muestra de entrenamiento que contendrá el 70 % de la información y una muestra de prueba que contendrá el 30 % restante.

K-Fold Cross-Validation (K-Fold CV)

Este enfoque divide el conjunto de datos en K subconjuntos o Folds de tamaño similar. El modelo se entrena K veces, cada vez utilizando K - 1 folds como conjunto de entrenamiento y el restante como muestra de prueba. Los errores de cada iteración se promedian para obtener una estimación más robusta del desempeño del modelo.

Leave-One-Out Cross-Validation (LOOCV)

Es un caso particular de K-Fold CV donde K es igual al número total de observaciones. En cada iteración, el modelo se entrena con todos los datos menos uno, que se utiliza como validación. Aunque proporciona una estimación casi sin sesgo, este método es computacionalmente costoso y más sensible al sobreajuste si existe colinealidad entre variables.

A través de las técnicas previamente descritas, se estima la Raiz del Error Cuadrático Medio (RMSE) como métrica principal para evaluar el ajuste y la capacidad de generalización de los modelos de regresión lineal utilizados.

El RMSE mide el promedio del error al cuadrado entre los valores reales

y los predichos. Al medir su raíz, el resultado se presenta en las mismas unidades que la variable dependiente.

De acuerdo con lo anterior, al realizar comparación de RMSE de diferentes modelos, se prefieren aquellas mediciones más pequeñas, las cuales implican un mejor desempeaño predictivo del modelo, o lo que es lo mismo, menor error.

Con el objetivo de iniciar con el análisis de predicción, se realiza un reconocimiento de las variables contenidas en la base de datos seleccionada para formular de manera precisa los modelos a evaluar.

Cuadro 6: Nombres de las variables en el dataset `datos_lim`

N°	**Nombre de la variable**
1	directorio
2	secuencia_p
3	orden
4	sex
5	age
6	relab
7	mes
8	formal
9	size_firm
10	max_educ_level
11	impa
12	hours_work_usual
13	ingreso_hora_1
14	id_individuo
15	sex_label
16	log_ingreso_hora
17	age2

Notas: Esta tabla resume las variables disponibles en el conjunto de datos seleccionado `datos_lim`.

Es importante resaltar que la base de datos completa, presenta un total de 14.745 observaciones y 17 variables como se observa en el Cuadro 6.

Por otra parte, a lo largo del presente estudio, se ha tomado como variable dependiente el salario por hora, por lo que se realizará el análisis de predicción sobre los modelos formulados con dicha variable. Sin embargo, debido a que los datos de salarios usualmente presentan una distribución asimétrica a la derecha, se tomará en cuenta la variable con su respectiva transformación con

logaritmo natural. A continuación, se presenta graficamente las diferencias entre la variable antes y despues de la transformación.

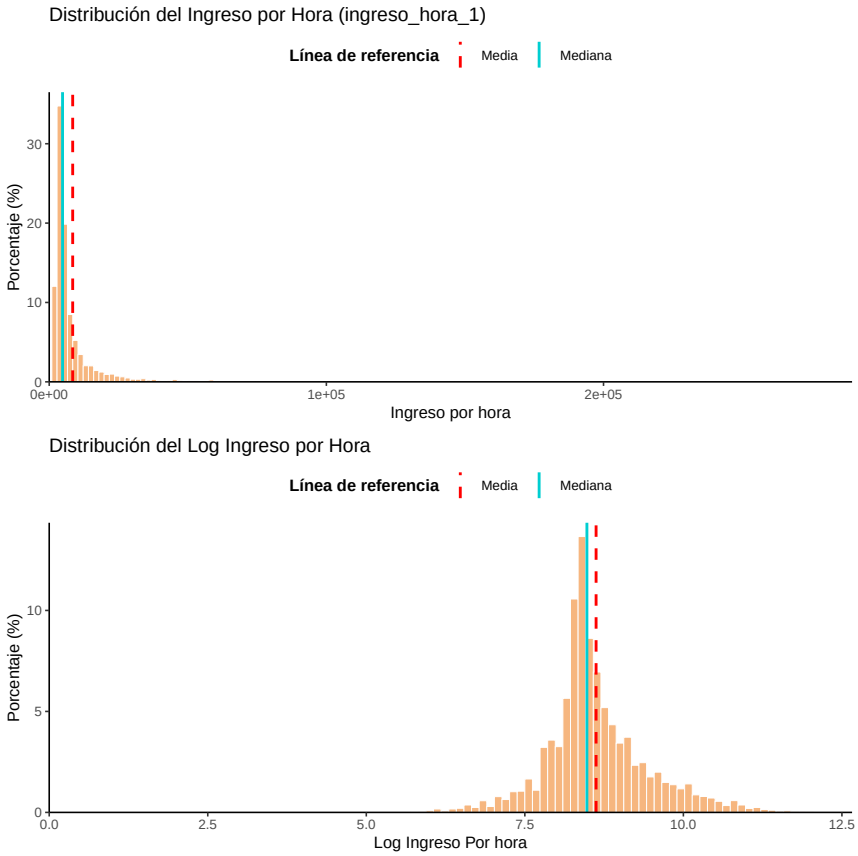

Figura 5: Distribución del ingreso horario antes y después del logaritmo

La Figura 5 muestra la distribución del ingreso por hora antes y después de aplicar la transformación logarítmica. En el panel superior, correspondiente a la variable *ingreso_hora_1*, se observa una clara asimetría positiva, con una alta concentración de observaciones en valores bajos y una larga cola derecha que refleja la presencia de valores atípicos elevados. Esta distribución sesgada genera una diferencia sustancial entre la media y la mediana, dificultando el cumplimiento de supuestos clásicos como la normalidad y la homocedasticidad en modelos estadísticos.

En contraste, el panel inferior presenta la variable *log_ingreso_hora*, cuyo histograma evidencia una distribución mucho más simétrica y cercana a la normal. La media y la mediana se encuentran alineadas, lo que indica una reducción significativa del sesgo. Esta transformación permite estabilizar la varianza, mitigar el efecto de los valores extremos y facilitar la interpretación

de los coeficientes, por lo que resulta especialmente útil para el análisis de ingresos laborales.

Con el objetivo de analizar el poder de prediccion de los modelos que se ajustan con la base de datos seleccionada, se procederá a realizar un estudio de los resultados de RMSE por cada uno de los tres metodos descritos con anterioridad para su posterior comparación. A continuación, se relacionan los modelos seleccionados:

Cuadro 7: Especificaciones de los Modelos de Regresión

Modelo	**Ecuación**
Modelo 1	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{age}^2 + u$
Modelo 2	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + u$
Modelo 3	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{age} + \beta_4 \cdot \text{age}^2 + \beta_5 \cdot \text{educ} + \beta_6 \cdot \text{relab} + \beta_7 \cdot \text{formal} + \beta_8 \cdot \text{size\_firm} + u$
Modelo 4	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex} + \beta_4 \cdot (\text{age} \cdot \text{sex}) + \beta_5 \cdot \text{age}^2 + \beta_6 \cdot \text{educ} + u$
Modelo 5	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{educ} + \beta_4 \cdot (\text{sex} \cdot \text{educ}) + \beta_5 \cdot \text{age} + u$
Modelo 6	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{relab} + u$
Modelo 7	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{formal} + \beta_4 \cdot \text{educ} + \beta_5 \cdot (\text{formal} \cdot \text{educ}) + u$
Modelo 8	$\ln(w) = \beta_1 + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{relab} + \beta_4 \cdot (\text{educ} \cdot \text{relab}) + \beta_5 \cdot \text{age} + u$

5.1. Muestra de Entrenamiento y de Prueba

De acuerdo con la metodología de Validation Set Approach (VSA), se debe dividir la variable dependiente *log_ingreso_hora* en dos grupos, a saber, una muestra de entrenamiento correspondiente al 70 % y una muestra de prueba correspondiente al 30 % de la información, como se evidencia en la Figura 6.

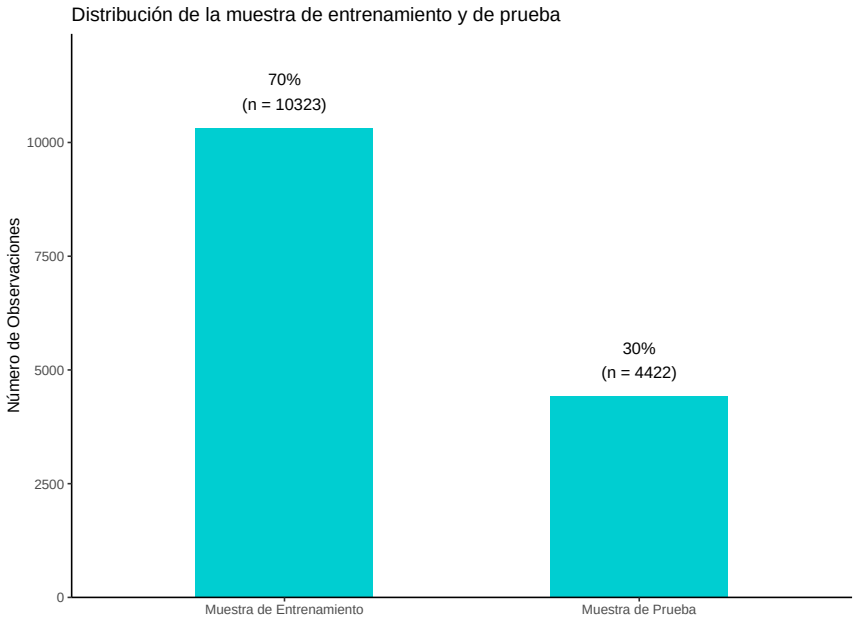

Figura 6: Distribución de la muestra de entrenamiento y de prueba

En la Figura 6, se evidencia que del total de observaciones (14.745), corresponden 10.323 observaciones a la muestra de entrenamiento, mientras que 4.422 pertenecen a los datos seleccionados para testear los resultados.

Una vez se dividen los datos en los dos grupos mecionados con antelación, se procede a formular los respectivos modelos para calcular la medición de RMSE.

5.2. Análisis de RMSE

Con el fin de seleccionar el modelo que mejor poder de predicción posee, se debe calcular su desempeño en términos de los resultados de RMSE de 8 los modelos seleccionados de acuerdo con las variables contenidas en la base de datos.

5.2.1. RMSE por el método Validation Set Approach - VSA

En primera instancia, se calcularán las mediciones de RMSE por el método de Validation Set Apprach - VSA, el cuál usará la muestra de entrenamiento y de prueba presentada en el punto 5.1.

Cuadro 8: Resultados del RMSE por (VSA)

Modelo	**Ecuación**	**RMSE (VSA)**
Modelo 1	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{age}^2 + u$	0.8116
Modelo 2	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + u$	0.8180
Modelo 3	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{age} + \beta_4 \cdot \text{age}^2 + \beta_5 \cdot \text{educ} + \beta_6 \cdot \text{relab} + \beta_7 \cdot \text{formal} + \beta_8 \cdot \text{size\_firm} + u$	0.6311
Modelo 4	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex} + \beta_4 \cdot (\text{age} \cdot \text{sex}) + \beta_5 \cdot \text{age}^2 + \beta_6 \cdot \text{educ} + u$	0.6857
Modelo 5	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{educ} + \beta_4 \cdot (\text{sex} \cdot \text{educ}) + \beta_5 \cdot \text{age} + u$	0.6940
Modelo 6	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{relab} + u$	0.7780
Modelo 7	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{formal} + \beta_4 \cdot \text{educ} + \beta_5 \cdot (\text{formal} \cdot \text{educ}) + u$	0.6506
Modelo 8	$\ln(w) = \beta_1 + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{relab} + \beta_4 \cdot (\text{educ} \cdot \text{relab}) + \beta_5 \cdot \text{age} + u$	0.6704

Como se muestra en el Cuadro 8, se calcularon los valores del error cuadrático medio (RMSE) utilizando el método de validación por partición simple (Validation Set Approach - VSA). Los resultados indican que el **Modelo 3** presenta el mejor desempeño predictivo, al incorporar una amplia gama de variables explicativas del salario por hora, como edad, educación, tipo de ocupación, formalidad y tamaño de la empresa. En segundo lugar, se encuentra el **Modelo 7**, el cual introduce una interacción relevante entre la formalidad laboral y el nivel educativo de los individuos encuestados.

En contraste, el **Modelo 2**, que considera únicamente el sexo como variable explicativa, presenta el valor de RMSE más elevado entre todas las especificaciones analizadas. Esto sugiere que el sexo, por sí solo, no constituye un determinante significativo en la predicción del ingreso horario dentro de esta muestra.

5.2.2. RMSE por el método K-FOLD

Cuadro 9: Resultados del RMSE por K-FOLD

Modelo	**Ecuación**	**RMSE K-FOLD**
Modelo 1	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{age}^2 + u$	0.8104
Modelo 2	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + u$	0.8187
Modelo 3	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{age} + \beta_4 \cdot \text{age}^2 + \beta_5 \cdot \text{educ} + \beta_6 \cdot \text{relab} + \beta_7 \cdot \text{formal} + \beta_8 \cdot \text{size\_firm} + u$	0.6373
Modelo 4	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex} + \beta_4 \cdot (\text{age} \cdot \text{sex}) + \beta_5 \cdot \text{age}^2 + \beta_6 \cdot \text{educ} + u$	0.6876
Modelo 5	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{educ} + \beta_4 \cdot (\text{sex} \cdot \text{educ}) + \beta_5 \cdot \text{age} + u$	0.6956
Modelo 6	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{relab} + u$	0.7793
Modelo 7	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{formal} + \beta_4 \cdot \text{educ} + \beta_5 \cdot (\text{formal} \cdot \text{educ}) + u$	0.6563
Modelo 8	$\ln(w) = \beta_1 + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{relab} + \beta_4 \cdot (\text{educ} \cdot \text{relab}) + \beta_5 \cdot \text{age} + u$	0.6762

De acuerdo con los resultados presentados en el Cuadro 9, se observa que el **Modelo 3** presenta el menor error cuadrático medio (RMSE = 0.6373) bajo la validación cruzada tipo K-FOLD, lo cual sugiere que esta especificación ofrece el mejor desempeño predictivo entre los modelos considerados. Este modelo incluye variables demográficas (sexo, edad y edad al cuadrado), educación, tipo de ocupación, formalidad y tamaño de la empresa, lo cual indica que la combinación de características individuales y del entorno laboral contribuye significativamente a explicar el salario por hora.

Por el contrario, los modelos más simples, como el **Modelo 1** (sólo edad y edad al cuadrado) y el **Modelo 2** (sólo sexo), presentan valores de RMSE más altos (0.8104 y 0.8187 respectivamente), lo que evidencia una menor capacidad explicativa. Asimismo, otros modelos con interacciones específicas, como el **Modelo 5** (sexo*educación) y el **Modelo 8** (educación*relación laboral), también muestran un buen desempeño predictivo, pero sin superar al Modelo 3.

Estos resultados destacan la importancia de incorporar múltiples dimensiones del capital humano y del contexto laboral para mejorar la precisión en la predicción del ingreso.

5.3. Comparación entre RMSE por VSA y K-FOLD

Cuadro 10: Resultados del RMSE por VSA y K-FOLD

Modelo	**Ecuación**	**RMSE (VSA)**	**RMSE (K-FOLD)**
Modelo 1	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{age}^2 + u$	0.8116	0.8104
Modelo 2	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + u$	0.8180	0.8187
Modelo 3	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{age} + \beta_4 \cdot \text{age}^2 + \beta_5 \cdot \text{educ} + \beta_6 \cdot \text{relab} + \beta_7 \cdot \text{formal} + \beta_8 \cdot \text{size\_firm} + u$	0.6311	0.6373
Modelo 4	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex} + \beta_4 \cdot (\text{age} \cdot \text{sex}) + \beta_5 \cdot \text{age}^2 + \beta_6 \cdot \text{educ} + u$	0.6857	0.6876
Modelo 5	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{educ} + \beta_4 \cdot (\text{sex} \cdot \text{educ}) + \beta_5 \cdot \text{age} + u$	0.6940	0.6956
Modelo 6	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{relab} + u$	0.7780	0.7793
Modelo 7	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{formal} + \beta_4 \cdot \text{educ} + \beta_5 \cdot (\text{formal} \cdot \text{educ}) + u$	0.6506	0.6563
Modelo 8	$\ln(w) = \beta_1 + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{relab} + \beta_4 \cdot (\text{educ} \cdot \text{relab}) + \beta_5 \cdot \text{age} + u$	0.6704	0.6762

De acuerdo con los resultados presentados en el Cuadro 10, se observa que el **Modelo 3** presenta el menor error cuadrático medio (RMSE) tanto bajo el enfoque de Validación Simple (VSA) como bajo el método de validación

cruzada (K-FOLD). Este modelo incluye un conjunto más completo de variables explicativas del salario por hora, entre ellas edad, sexo, educación, tipo de ocupación, formalidad laboral y tamaño de la empresa, lo que parece contribuir a una mayor capacidad predictiva.

El segundo mejor desempeño lo alcanza el **Modelo 7**, que incorpora una interacción entre la formalidad y el nivel educativo de los individuos, lo que sugiere que el efecto de la educación sobre los ingresos puede variar según el grado de formalidad del empleo.

En contraste, el **Modelo 2**, que únicamente considera el sexo como variable explicativa, arroja el RMSE más alto en ambas metodologías, lo que indica que esta variable por sí sola no ofrece una capacidad predictiva adecuada del salario por hora.

Estos hallazgos refuerzan la importancia de considerar múltiples factores y sus interacciones al modelar el ingreso por hora.

5.3.1. Análisis del Modelo 3 (Menor RMSE)

De acuerdo con los resultados del Cuadro 10, el **Modelo 3** presenta el menor error de predicción, lo que sugiere que es la especificación más robusta entre las consideradas.

Para evaluar aquellas observaciones que el modelo no predice adecuadamente, se calcularon los errores de predicción en la muestra de prueba, definidos como la diferencia entre el valor observado y el valor predicho del logaritmo del salario por hora. Al analizar la distribución de estos errores mediante histogramas y medidas de dispersión, se identificaron valores ubicados en las colas, es decir, observaciones con errores particularmente altos (positivos o negativos).

Estas observaciones extremas podrían interpretarse como potenciales outliers, pero no necesariamente representan casos que deban ser investigados por la Dirección de Impuestos y Aduanas Nacionales - DIAN. Su presencia puede atribuirse a limitaciones del modelo, como variables omitidas, especificaciones funcionales inadecuadas o ruido inherente a los datos.

En consecuencia, si bien es importante monitorear los errores extremos, no es posible concluir que estas observaciones sean producto de evasión o comportamiento atípico deliberado. Más bien, constituyen oportunidades para mejorar la especificación del modelo y considerar posibles factores no observados que inciden en el ingreso laboral.

5.4. RMSE por el método Leave One Out Cross Validation - LOOCV

El analisis de RMSE por los metodos de VSA y K-FOLD, evidenciaron que, los modelos con mejor capacidad de predicción son el **Modelo 3** y el **Modelo 7**. Por lo anterior, a partir del metodo de LOOCV se realizará el calculo de RMSE para su respectiva comparación y evaluación.

Cuadro 11: Comparación de RMSE para los Modelos 3 y 7

Modelo	**Ecuación**	**RMSE (VSA)**	**RMSE (K-FOLD)**	**RMSE (LOOCV)**
Modelo 3	$\ln(w) = \beta_1 + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{age} + \beta_4 \cdot \text{age}^2 + \beta_5 \cdot \text{educ} + \beta_6 \cdot \text{relab} + \beta_7 \cdot \text{formal} + \beta_8 \cdot \text{size\_firm} + u$	0.6311	0.6373	0.6376
Modelo 7	$\ln(w) = \beta_1 + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{formal} + \beta_4 \cdot \text{educ} + \beta_5 \cdot (\text{formal} \cdot \text{educ}) + u$	0.6506	0.6564	0.6566

El método *Leave-One-Out Cross-Validation* (LOOCV) permite evaluar la capacidad predictiva de un modelo al ajustar $n-1$ observaciones y predecir la que queda por fuera, repitiendo este proceso para todas las observaciones del conjunto de datos. Este enfoque es útil para mitigar el sobreajuste y obtener una métrica robusta de error de predicción.

En el Cuadro 11, se observa que el **Modelo 3** presenta el menor valor de RMSE bajo LOOCV (0,6376), lo que confirma su superior capacidad predictiva respecto al **Modelo 7**, cuyo RMSE es ligeramente mayor (0,6566). Esta diferencia, aunque no extrema, sugiere que el Modelo 3, al incorporar una mayor cantidad de variables explicativas (incluyendo edad, sexo, educación, formalidad, tipo de ocupación y tamaño de empresa), es más eficaz al capturar las complejidades de la determinación del salario por hora.

Cabe destacar que LOOCV es especialmente sensible a observaciones atípicas. Por lo tanto, las predicciones con mayor error en este esquema podrían corresponder a individuos con características inusuales dentro del conjunto de datos (por ejemplo, trabajadores con combinaciones poco comunes de edad, educación y tipo de empleo). Estas observaciones, en lugar de ser descartadas, pueden constituir casos de interés para estudios posteriores o incluso para acciones de supervisión en políticas públicas.

5.5. Comparación de RMSE entre los Modelos 3 y 7

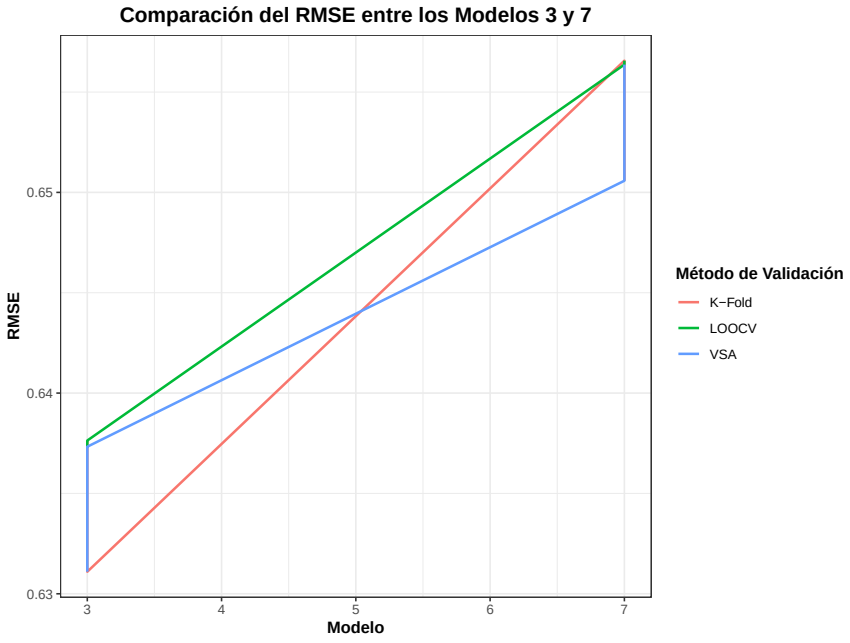

Figura 7: Distribución del ingreso horario antes y después del logaritmo

La Figura 7 muestra una comparación visual del *Root Mean Squared Error* (RMSE) entre los Modelos 3 y 7, utilizando tres enfoques distintos de validación: **Validation Set Approach (VSA)**, **K-Fold Cross Validation** y **Leave-One-Out Cross Validation (LOOCV)**.

- **Consistencia del Modelo 3:** El Modelo 3 presenta consistentemente los valores de RMSE más bajos en los tres métodos de validación, lo cual indica un mejor desempeño predictivo frente al Modelo 7. Esto sugiere que el Modelo 3 es la especificación más robusta para predecir el logaritmo del salario por hora.
- **Tendencia ascendente del RMSE:** En ambos modelos, el RMSE tiende a aumentar ligeramente desde VSA hacia LOOCV, lo cual es esperable ya que LOOCV representa una validación más estricta frente al sobreajuste.
- **Modelo 7 como alternativa competitiva:** Aunque el Modelo 7 tiene un mayor RMSE, sus resultados siguen siendo razonables, especialmente bajo los métodos VSA y LOOCV. La inclusión de una interacción entre *formalidad laboral* y *educación* permite capturar relaciones relevantes para ciertos subgrupos poblacionales.
- **Estabilidad entre K-Fold y LOOCV:** Los valores de RMSE obtenidos mediante K-Fold y LOOCV son muy similares para ambos modelos, lo cual indica una buena estabilidad del desempeño predictivo bajo diferentes esquemas de validación cruzada.

En conjunto, estos resultados permiten concluir que el **Modelo 3 ofrece la mayor precisión predictiva** entre las especificaciones evaluadas. Su riqueza en variables explicativas —como sexo, edad, educación, tipo de ocupación, formalidad y tamaño de empresa— le permite capturar con mayor fidelidad la heterogeneidad en los salarios por hora de los individuos.