Taller 1.

Predicción de Ingresos.

Integrantes: Isabella Mendez Pedraza. Cód.: 201814239
Manuela Ojeda Ojeda. Cód.: 201814476
Juan Sebastian Tellez Melo. Cód.: 201513710
Andres Mauricio Palacio Lugo. Cód.: 201618843

1. Introducción.

Para determinar el valor de los impuestos que debe pagar cada persona es fundamental la exactitud en la declaración de los ingresos. Sin embargo, el fraude fiscal de todos los tipos siempre ha sido un problema que está muy presente. Menos del 90% de los impuestos son pagados voluntariamente y en el tiempo estimado en USA. La razón de esta brecha es que muchas personas no reportan correctamente sus ingresos.

Un modelo predictivo del ingreso podría ayudar a señalar casos de fraude para reducir la brecha. Además, podría ayudar a identificar personas en condiciones de vulnerabilidad que puedan necesitar un apoyo o ayuda adicional.

2. Datos.

La Gran Encuesta Integrada de Hogares (GEIH) del DANE es una encuesta mediante la que, además de las características generales de la población (como la edad, sexo, nivel educativo, entre otros), se obtiene información acerca de las condiciones de empleo de las personas (si están empleados, cantidad de horas que trabajan, si tienen seguridad social) y su nivel y fuentes de ingresos. Con esta información se generan indicadores del mercado laboral en Colombia como la tasa de ocupación, la rama laboral en la que se desempeñan los colombianos y la remuneración, y el comportamiento del mercado laboral para grupos poblacionales específicos como los jóvenes.

Scrapping

Para el análisis del presente documento, se usaron datos extraídos de la GEIH con una serie de modificaciones hechas por profesores de la Facultad de Economía de la Universidad de Los Andes. Estos datos fueron publicados en 10 páginas de internet diferentes por lo que fue necesario acceder a ellos mediante métodos de *web scraping*, utilizando el software estadístico de R y el código utilizado para ello se encuentra en el repositorio descrito al inicio del

documento en el script denominado "*2_Data_cleaning*". En este procesamiento de los datos desde las páginas de internet se identificó que los datos, aunque no tenían restricciones de seguridad, si se encontraban almacenados en objetos que requirieron revisar en detalle los recursos de red de cada página de internet. Una vez se logró obtener los datos, se creó un *loop* en R que permitió descargarlos y almacenarlos unidos en tables locales.

Limpieza

En un primer acercamiento con los datos optamos por tomar las variables que consideramos relevantes para el análisis, aquellas que pudieran explicar o relacionarse con el nivel de ingresos de las personas y también aquellas variables características de las personas o del contexto de la muestra. Iniciamos la limpieza eliminando las personas que son desempleadas o menores de 18 años. Para tener mayor claridad sobre los datos a manejar y depurar la base, calculamos el porcentaje de missing values por variable y eliminamos aquellas que tuvieran un porcentaje mayor al 50%. Luego, eliminamos los NA's restantes. Finalmente, para dar inicio al análisis, generamos las variables que necesitábamos, como "Ingresos_laborales" y definimos las varaibles catégoricas.

Estadísticas descriptivas

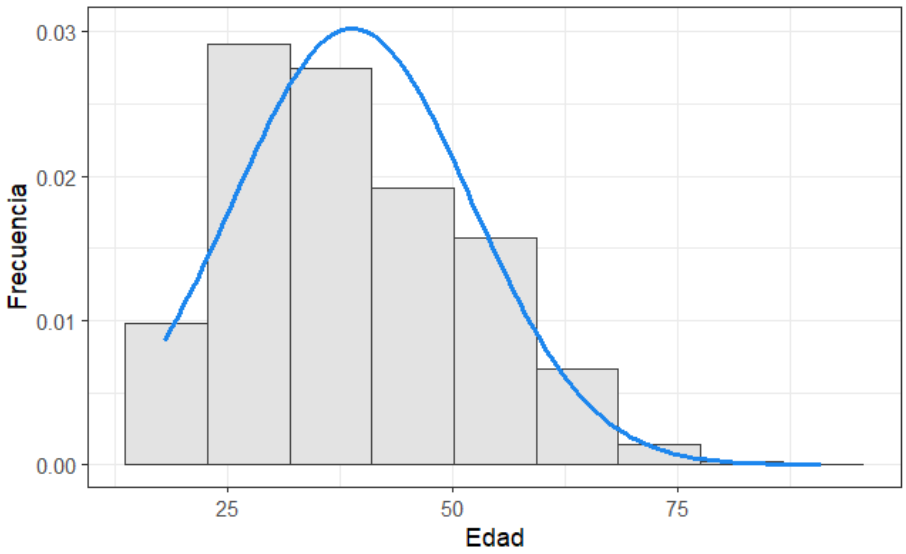
Tabla 1. Estadísticas descriptivas

Statistic	N	Mean	St. Dev.	Min	Max
Edad	14,763	38.888	13.198	18	91
Educación terciaria	14,763	0.326	0.469	0	1
Cuenta Propia	14,763	0.297	0.457	0	1
Desempleado	14,763	0.000	0.000	0	0
Estrato	14,763	2.521	0.987	1	6
Formal	14,763	0.602	0.489	0	1
Horas trabajadas	14,763	47.198	15.056	1	130
Informal	14,763	0.398	0.489	0	1
Ingtot	14,763	1,775,332.000	2,654,431.000	20,000.000	85,833,333.000
MicroEmpresa	14,763	0.426	0.495	0	1
Experiencia	14,763	61.970	88.028	0	720
Ocupado	14,763	1.000	0.000	1	1
Sexo	14,763	0.527	0.499	0	1
y_total_m_ha	14,763	8,542.185	13,866.550	0.472	350,583.300
y_total_m	14,763	1,617,605.000	2,431,392.000	84.000	70,000,000.000
Ingresos_laborales	14,763	8.624	0.833	−0.752	12.767
Edad cuadrado	14,763	1,686.488	1,126.269	324	8,281
Mujer	14,763	0.473	0.499	0	1
MaxEducLevel3	14,763	0.045	0.207	0	1
MaxEducLevel4	14,763	0.091	0.288	0	1
MaxEducLevel5	14,763	0.117	0.322	0	1
MaxEducLevel6	14,763	0.326	0.469	0	1
MaxEducLevel7	14,763	0.414	0.493	0	1

Como parte de la limpieza de datos, se restringió la información a sujetos con edades a partir de los 18 años y que estuviesen ocupados (utilizando la variable "ocu =1"). Así, llegamos a una muestra de 14,763 observaciones. De esta muestra, el 47.3% corresponde a mujeres mientras que el 52.7% son hombres.

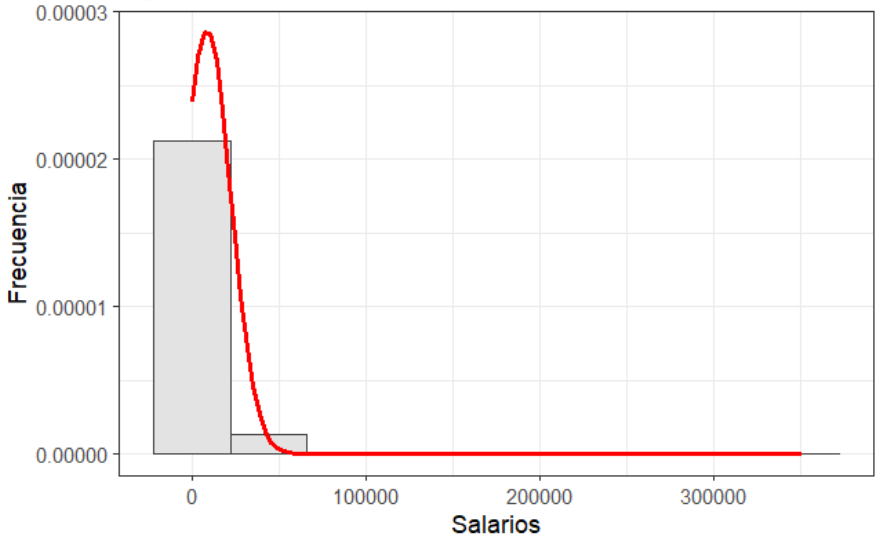
La edad promedio de las personas es 39 años con una desviación estándar de 13 años y la persona ocupada con mayor edad tiene 91 años.

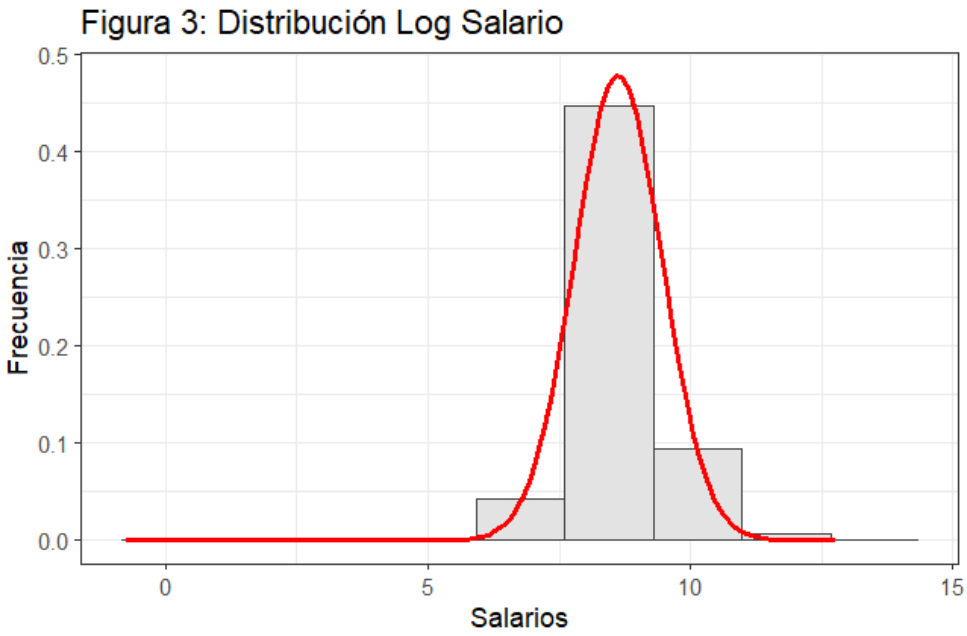
Figura 1: Distribución de edad



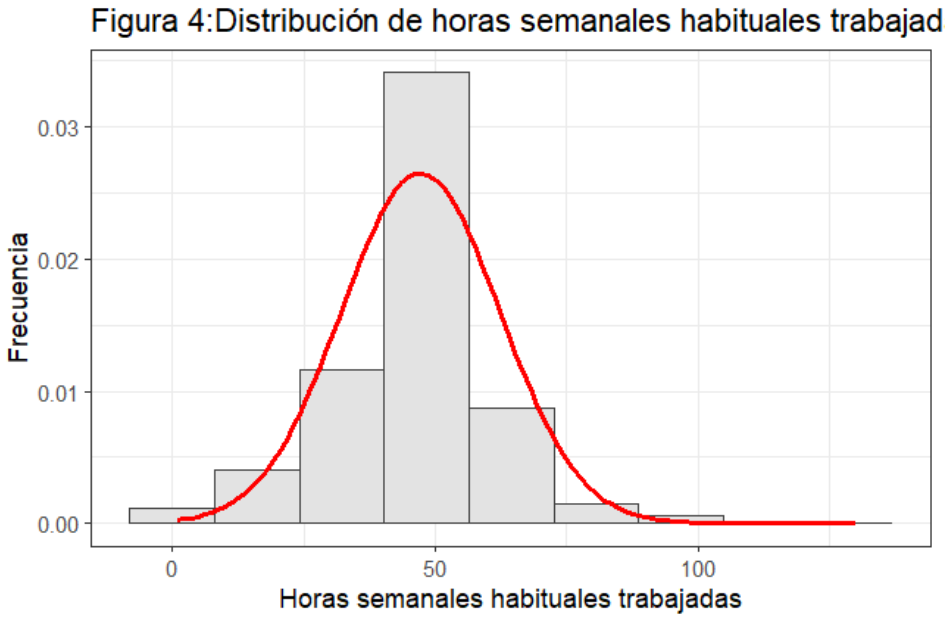
En la figura 1 observamos la distribución por edad, allí se evidencia que solo tenemos personas mayores de edad y hay muy pocas personas que tienen más de 75 años. También se observa que las edades con mayor frecuencia están entre los 20 años y 30 años.

Figura 2: Distribución Salario





En la figura 2 se observa la distribución de los salarios sin ningún tratamiento, sin embargo, en la figura 3 se evidencia cómo es la distribución del logaritmo del salario de los individuos, que es la variable que se utilizará en las estimaciones de este documento. Se observa que la media del logaritmo del salario está entre $8 y $9 mil pesos por hora.



A partir de la tabla 1 y la figura 4 se observa que en promedio los individuos trabajan 47 horas habitualmente en la semana, la persona que menos trabaja es 1 hora y la que más hora trabaja

llega a 130 horas. Los individuos trabajan con mayor frecuencia entre 40 y 50 horas habitualmente en la semana.

3. Perfil Edad-Salario

Para estimar el perfil edad-salario de los individuos vamos a estimar:

$$\lg(w) = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u$$

Tomamos nuestra variable Ingresos_totales que es el logaritmo de y_total_m_ha (ingresos asalariados + independientes total - nominal por hora). Al estimar el modelo encontramos que el incremento de 1 año de edad es asociado a un cambio en los ingresos de 5.5%. Observamos que nuestro modelo tiene un R^2 de 0.02, es decir, que solamente el 20% de la varianza total de nuestro resultado objetivo está siendo explicado por nuestro modelo, por lo que podemos pensar que no es el mejor modelo.

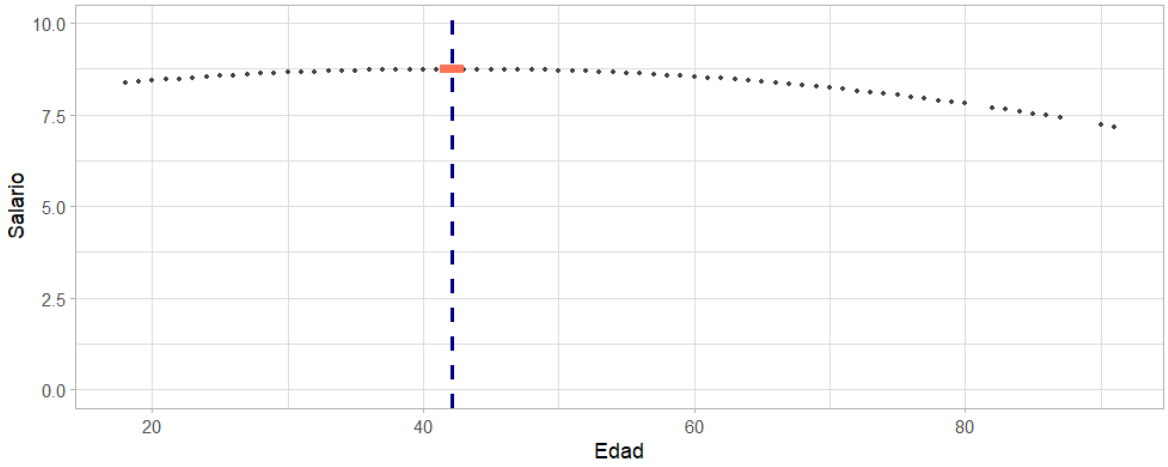
Tabla 2. Perfil edad-salario

	Dependent variable:
	Ingresos_laborales
Edad	0.055***
	(0.003)
Edad cuadrado	−0.001***
	(0.00004)
Constant	7.578***
	(0.060)
Observations	14,763
R^2	0.023
Adjusted R^2	0.023
Residual Std. Error	0.823 (df = 14760)
F Statistic	173.004*** (df = 2; 14760)
Note:	*p<0.1; **p<0.05; ***p<0.01

En la figura 5 observamos que como era de esperarse y según la evidencia de la economía laboral los salarios tienden a ser bajos cuando el trabajador es joven y aumentan a medida que el trabajador crece llegando a un máximo, en donde el salario empieza a disminuir.

Continuando con el modelo anterior y para la construcción del bootstrap utilizamos intervalos de confianza en los cuales encontramos que la edad mínima es 41.24 años y la máxima es 42.95 años. Adicionalmente, se observa que la edad máxima es 42.09 años. Es decir, que en promedio aproximadamente a los 42 años los individuos obtienen su logaritmo de salario máximo de aproximadamente $8.750 por hora con una confianza del 95%.

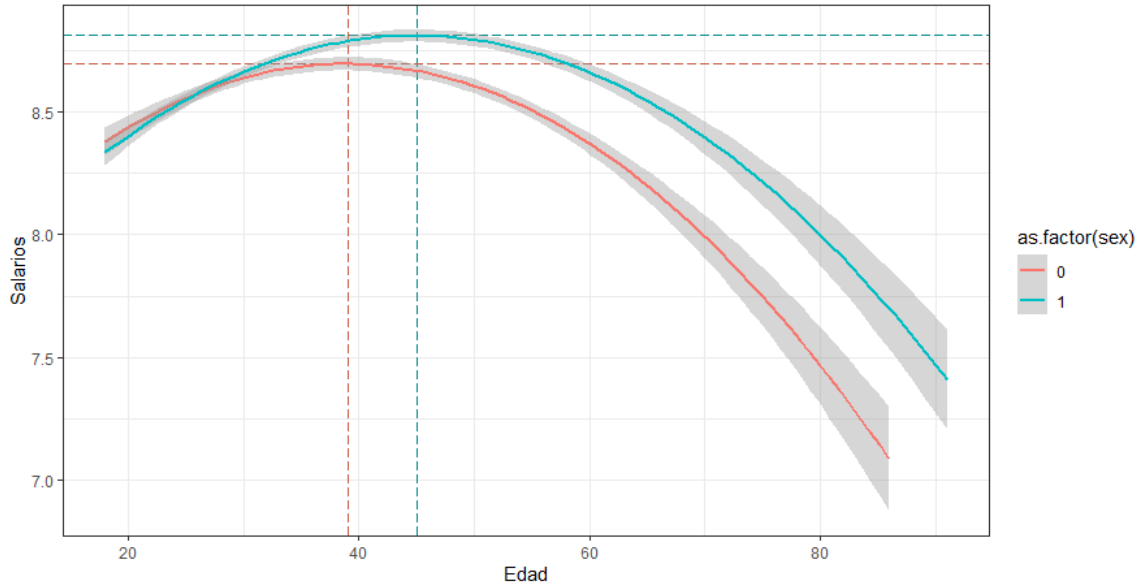
Figura 5:Perfil edad-salario



4. Brecha salarial por género

Al analizar la distribución de los salarios por género, encontramos que las mujeres obtienen, en promedio, el mayor salario hacia los 45 años, mientras que los hombres obtienen su mayor salario, promedio, hacia los 39 años. La figura 8, a continuación, muestra que las edades pico para cada caso (mujer-hombre) son estadísticamente diferentes, notando que con un intervalo de confianza del 95% estas dos estimaciones no se cruzan en la edad más alta en cada género.

Figura 8. Perfil Edad vs. Salario por Género



Ahora bien, para el análisis se utilizará el siguiente modelo y estimaremos el impacto del género sobre los ingresos laborales, utilizando el Teorema de FWL (Frisch-Waugh-Lovell):

El análisis consistirá en introducir variables de control en este modelo y corroborar el efecto que tiene el género sobre los ingresos evaluados, dentro de estos controles se tienen las siguientes variables: Edad, persona independiente o no, es informal o no, máximo nivel de educación, si es una microempresa o no y el tiempo que lleva trabajando.

$$\lg(w) = \beta_1 + \beta_2 Female + u$$

La regresión inicial consideró solamente la relación entre los ingresos laborales y el hecho de ser mujer o no (Modelo 1) y con ello encontramos un coeficiente de -0.09032, esto muestra que el hecho de ser mujer tiene un impacto negativo en los salarios, estimado en dicha magnitud.

Tabla 3. Brecha Salarial Incondicional

	Variable dependiente (y):
	Ingresos laborales
Mujer	-0.0903160^{***}
	(0.0137120)
Constante	8.6666820^{***}
	(0.0094345)
Observations	14,763
R^2	0.0029305
Adjusted R^2	0.0028629
Residual Std. Error	0.8318438 (df = 14761)
F Statistic	43.3841200^{***} (df = 1; 14761)
Note:	*p<0.1; **p<0.05; ***p<0.01

Para los análisis siguientes se estudiaron los siguientes modelos:

- Modelo 2. Relación de los ingresos laborales frente al hecho de ser mujer y las variables de control mencionadas previamente.
- Modelo 3. Para este modelo se realizó una regresión entre la variable que indica si se es mujer o no y las variables de control (R1), una regresión entre los ingresos laborales y las variables de control (R2) y, finalmente, estimamos la regresión de R2 en R1.

Cuando estimamos el Modelo 1 encontramos una relación entre los salarios y el hecho de ser mujer de -0.09032 y cuando incluimos variables de control (Modelo 2), ésta relación inicial pasa a -0.1413, esto ocurre porque dichas variables se encuentran correlacionadas. Por lo tanto, a partir del Teorema de FWL, podemos deducir que existe un efecto de las variables de control que se debe tener en cuenta al momento de analizar las predicciones de los salarios, pues la disminución sobre estos salarios no ocurre solamente por el hecho de ser mujer, aquí se puede evidenciar que las demás variables también generan efectos que deben ser limpiados de los modelos en el momento de realizar las predicciones, es decir, existe un efecto que no es explicado por la variable mujer sobre los salarios y que puede corresponder a variables no observadas. Nótese que al estimar el Modelo 2 encontramos un coeficiente exactamente igual al coeficiente que tendría la variable mujer cuando se regresiona únicamente contra todos los controles.

Ahora bien, teniendo en cuenta que al incorporar variables de control en los modelos el coeficiente que relaciona el hecho de ser mujer con los ingresos laborales aumenta, es posible inferir que esta muestra presenta un problema económico de selección y no de discrimación, pues cuando se utiliza solo la variable mujer sobre los salarios, la diferencia entre mujeres y hombres es más pequeña. Así, las diferencias salarias obedecen a otras variables que condicionan la selección y asignación de salarios.

Tabla 4. Comparación de Modelos

	Variable dependiente (y):		
	Ingresos laborales		Modelo Residuos
	(1)	(2)	(3)
Mujer (1)	-0.090^{***}	-0.141^{***}	
	(0.014)	(0.011)	
Controles	(Datos en código)	(Datos en código)	
Modelo Mujer vs. Controloes			-0.141^{***}
			(0.011)
Constant	8.667^{***}	7.230^{***}	0.000
	(0.009)	(0.083)	(0.005)
Observations	14,763	14,763	14,763
R^2	0.003	0.381	0.011
Adjusted R^2	0.003	0.381	0.011
Residual Std. Error	0.832 (df = 14761)	0.655 (df = 14750)	0.655 (df = 14761)
F Statistic	43.384^{***} (df = 1; 14761)	757.827^{***} (df = 12; 14750)	167.850^{***} (df = 1; 14761)
Note:			*p<0.1; **p<0.05; ***p<0.01

Para corroborar estos datos se utilizó un *boostrapping* sobre los datos de la distribución de la muestra, con el fin de validar el coeficiente encontrado de relación de la variable cuando se es mujer y los controles frente a los ingresos estimados, con esto se estima la incertidumbre en dicho coeficiente.

El resultado del Bootstrap mostró un error de estimación de 0.01559 y un sesgo de 0.00060 y se identificó exactamente el mismo coeficiente estimado de los Modelos 2 y 3.

Los cálculos de este punto se encuentras en el script guardado en el repositorio como *4_Gender_Earnings.*

5- Predicción de ingresos

En esta sección se realizó un análisis de validación cruzada para determinar la capacidad predictiva de los métodos estadísticos cuando se enfrentan a un conjunto de datos fuera de la muestra de entrenamiento de estos modelos. El objetivo de esta sección es determinar el mejor modelo para predecir los salarios mediante características observables de los individuos y sus trabajos. Con este fin se tomaron dos aproximaciones, el "enfoque de "*conjunto de variación*", y la validación cruzada "*Leave-One-Out*" (LOOCV). A continuación, discutiremos la especificación de los modelos y los resultados de rendimiento al probarse con la muestra de testeo.

En la primera metodología, "*conjunto de validación*", se divide la muestra total con una relación 70-30 porciento donde se apartó el 70% para el entrenamiento de los modelos y el restante 30% para el testeo de estos. Tomando la muestra de entrenamiento se estimó en total 9 los modelos que se observan en la Tabla 5. Los modelos estimados se organizaron de menor complejidad a mayor complejidad. Una vez entrenados los modelos, se procedió a utilizarlos para predecir el salario con la muestra de testeo y comparar sus rendimientos.

La métrica de rendimiento escogida para este análisis fue el error cuadrático medio (MSE), el cual nos permite medir que tan lejos están las predicciones de los valores observados. Esta métrica tiende a disminuir a medida que se aumentan la complejidad del modelo, con un punto de infección en el que se invierte la trayectoria y aumento, por lo cual podría inducirnos a sobre ajustar el modelo con el fin de obtener errores más bajos y afectando su capacidad predictiva. Por tal motivo, al combinarlo con validación cruzada, y calculando sobre la muestra de testeo, provee un estimador de la tasa de error en prueba.

El error cuadrático medio resultante para cada una de las pruebas a los modelos se muestra en la Figura 10. Se observa que a medida que aumenta la complejidad del modelo y el número de regresores disminuye considerablemente la tasa de error de prueba. Sin embargo, a partir del modelo 8 se alcanza un nivel de complejidad que induce al incremento de la tasa de error. Adicionalmente, entre el modelo 5 y el modelo 7 no se observan cambios significativos en el error. En consecuencia, se eligen los modelos 5 y 6 como los de mejor rendimiento dados sus menores errores cuadráticos medios sobre la muestra de prueba.

En la Figura 11 y la Figura 12, podemos observar la distribución de los errores de acuerdo con el nivel de ingresos. Se evidencia que a medida que se mueva a ingresos bajos y a ingresos altos la tasa de error se mueve significativamente. Sin embargo, podría deberse a causas diferentes. Por un lado, las personas de ingresos bajos están siendo sobre estimadas dadas

sus características y las de sus trabajos. Mientras que las personas de ingresos altos están siendo subestimadas.

Esto podría indicar que los modelos tienen dificultad para predecir los ingresos de la parte inferior y superior de la distribución de ingresos. Por lo cual, desde el punto de vista de la política y la evasión fiscal, los modelos estimados no permiten identificar que personas dadas sus características están reportando menos ingresos.

Siguiendo con el análisis de validación cruzada, En segunda metodología implementada para evaluar el rendimiento de los modelos, se aplicó la distribución LOOCV y se estimó la tasa de error promedio para los dos modelos de mejor rendimiento escogidos anteriormente (modelo 5 y 6). En la Figura 13. Se encuentra la comparación de los errores cuadráticos medios tanto para la aproximación de conjunto de validación como la de LOOCV. Se reporta un incremento de la tasa de error de prueba con la aproximación LOOCV, lo cual implicaría que el sesgo es más alto de lo que se había estimado con la primera aproximación.

El incremento de los errores cuadráticos medios de la prueba puede estar relacionado con la división de las observaciones entre muestra de entrenamiento y muestra de testeo. Esto debido a la importancia relativa que tiene cada variable sobre el ajuste del modelo. Mientras que en la aproximación de conjunto de validación se escoge aleatoriamente la muestra de entrenamiento y la de prueba, lo cual podría llevar a que el ajuste del modelo cambie dependiendo de que conjunto de observaciones entro en entrenamiento o en prueba; en la aproximación de LOOCV se corrige esto eligiendo a cada observación como prueba y entrenamiento, llevando a un ajuste del modelo más preciso dada la importancia de cada observación.

Tabla 5. Especificación de los modelos estimados

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7	Modelo 8	Modelo 9
Salarios	SI	SI	SI	SI	SI	SI	SI	SI	SI
female	SI		SI	SI	SI	SI	SI	SI	SI
Edad		SI	SI	SI	SI	SI	SI	SI	SI
Edad^2		SI	SI	SI	SI	SI	SI	SI	SI
cuentaPropia				SI	SI	SI	SI	SI	SI
informal				SI	SI	SI	SI	SI	SI
Nivel Educativo				SI	SI	SI	SI	SI	SI
microEmpresa				SI	SI	SI	SI	SI	SI
experiencia				SI	SI	SI	SI	SI	SI
Oficio					SI	SI	SI	SI	SI
female* cuentaPropia* informal						SI	SI	SI	SI
Polinomio Experiencia (1 hasta 4 grado)							SI	SI	SI
female* cuentaPropia* informal* Experiencia								SI	SI
female* cuentaPropia* informal* Experiencia* microEmpresa									SI

Figura 10. Comparación de MSE y Complejidad de los Modelos

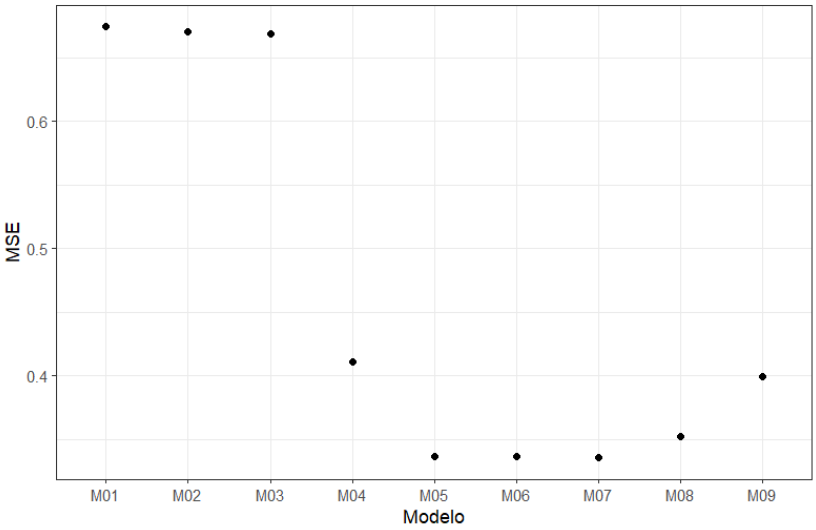


Figura 11. Distribución de los errores modelo 5

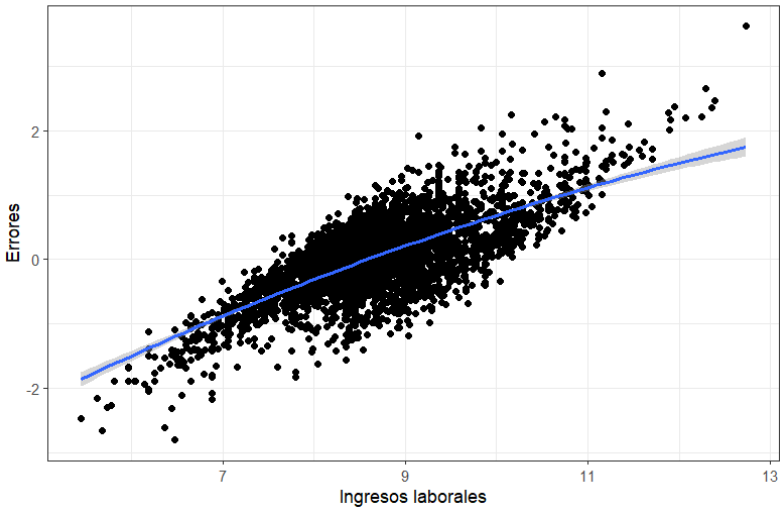


Figura 12. Distribución de los errores modelo 6

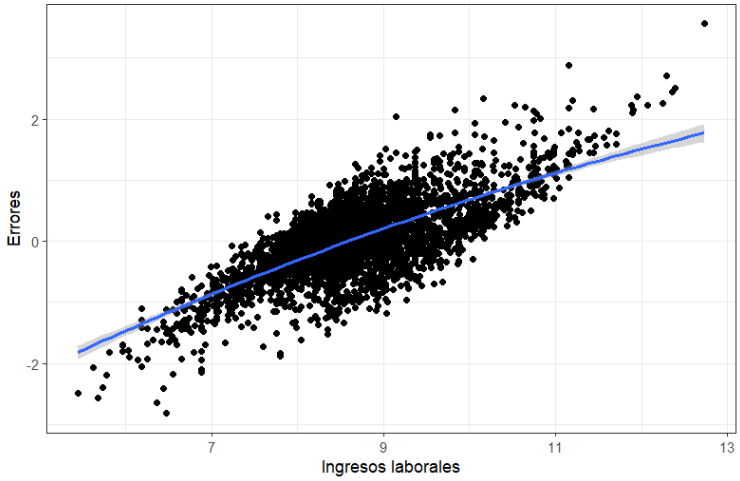


Figura 13. Comparación MSE de LOOCV y conjunto de validación