Presentado por Stiven Peralta, Jazmine Galdos, Andrea Clavijo, Sergio Jiménez, Nicolás Barragán

Nota: la base de datos usada, al igual que el script de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace: https://github.com/stivenperalta/Problem_set_1

Introducción

La subdeclaración de ingresos es un fenómeno que se presenta como una forma de reducir la carga fiscal de los contribuyentes: empresas y personas. Se estima que Colombia pierde anualmente cerca del 0,7% del PIB por este concepto en personas naturales. Comparado con America Latina, se encuentra por arriba de sus pares (Concha et al., 2017). Las disparidades del sistema tributario colombiano y vacíos en la norma han posibilitado que este fenómeno se mantenga. Se estima que la subdeclaración es superior al 50% en lo relacionado con el impuesto a la renta de personas naturales. A lo largo de este *set Problem* se desarrolla un modelo de predicción basado en características individuales usando datos de la *Gran encuesta integrada de hogares* (GEIH) del Departamento Administrativo Nacional de Estadística (DANE) para realizar una estimación de los ingresos. De este modo, se espera conocer, por esta vía, que variables determinan el ingreso. Esto podría, a futuro, servir como insumo para una mejor categorización de las personas que deben declarar como también poder establecer mejores mecanismos de recaudo.

Entre los principales resultados, se encontró que, en general, variables como la experiencia, la educación, la edad y el estrato socioeconómico sirven para explicar el salario con un alto nivel de confianza. Respecto a la edad, se encontró un comportamiento de orden cuadrático, con limite en una edad promedio de 57 años. Así mismo, confirmamos la existencia de una brecha salarial por género, en donde en promedio, las mujeres devengan un salario 0.9% menor que los hombres, controlando por características generales de los individuos y del empleo. Finalmente, se abordaron modelos predictivos para el salario, en donde se identificó que la capacidad predictiva aumenta en la medida en que se incluyen más variables en el modelo de entrenamiento, llegando hasta tener una predicción con un error estándar distribuido normalmente y con un valor promedio de 0.34. al analizar la distribución de esta predicción no se identificaron características atípicas que dieran cuenta de valores extremos a revisar en el marco de una política de control fiscal. Finalmente, se realizó la revisión de la bondad de ajuste de los modelos, en donde se identificó que el modelo predictivo que incluye condiciones como la experiencia, la educación, la edad y el estrato

socioeconómico resulta consistente tanto a través del método de validación cruzada, como a través del método LOOCV.

1) Contexto

La subdeclaración de impuestos es un problema que aqueja a la mayoría de los países del mundo y Colombia no es la excepción: de acuerdo con Fedesarrollo anualmente el país pierde por concepto de la brecha tributaria (elusión y evasión) cerca del 5,4% del PIB, lo que se traduce en dejar de recibir alrededor de 68 billones de pesos para financiamiento de gasto público e inversión social. Al mismo tiempo, esto crea un incentivo perverso en aquellos que tienen una motivación a pagar. Sobre esto, la mayor pérdida de recaudo se da por la evasión del impuesto de renta de las empresas, con un 3,4% del PIB; seguida de la evasión del IVA, que representa el 1,3% del PIB, y finalmente la evasión por concepto de impuesto de renta a personas, con cerca de un 0,7% del PIB, de este porcentaje el 0,06% es de rentas por ingresos laborales (La República, 2022).

En un sistema tributario complejo como el colombiano, donde existen disparidades en la carga tributaria que asumen empresas y personas, la literatura identifica dos tipos de desigualdades que podrían explicar, en parte, las causas de este fenómeno: las verticales y horizontales. Las primeras suceden cuando las personas o empresas de mayor ingreso contribuyen menos que el resto como proporción de su base gravable; las segundas, cuando teniendo la misma cantidad de ingresos, unas pagan más que otras (Concha *et al*., 2017).

Al respecto, la DIAN calcula que, en ambas formas de elusión, la subdeclaración en personas naturales es superior al 50%. Una posible explicación de estas desigualdades es la variedad de exenciones especiales existentes como ingresos no gravables, rentas especiales, sobretasas, compensaciones, deducciones extraordinarias, rentas exentas y créditos tributarios. Aunque el cálculo exacto de lo que pierde el país por este concepto no se tiene, de acuerdo con las aproximaciones realizadas por la Cepal, este porcentaje comparado con otros países de América Latina es superior en cerca de un 20% si se compara con Chile, México y Perú (Concha *et al*, 2017). El recaudo de los impuestos en personas naturales mejoraría si las tarifas efectivas de tributación incrementan y si se da un trato más uniforme con muy pocas tarifas y tratamientos diferenciales por cada fuente de ingreso.

2) Datos

Programa estadístico

Todos los análisis estadísticos presentados se realizaron en R versión 4.3.0 (R Core Team, 2023).

Recolección de datos

Para el análisis del presente trabajo, se tomó como referencia la *Gran encuesta integrada de hogare*s del 2018 (GEIH). La encuesta recolecta información acerca de las condiciones de empleo vigentes en el país: trabajo, tipo de trabajo, salarios, ingresos, entre otros; de igual forma, recolecta variables demográficas de los encuestados como el sexo, la edad, el nivel educativo, entre otros. La GEIH, es recolectada por el Departamento Administrativo Nacional de Estadística (DANE), entidad encargada de "producir y difundir información estadística oficial, como bien público, con altos estándares de calidad y rigor técnico para la toma de decisiones a nivel nacional y territorial" (DANE, s.f., p 1). Dicho esto, se evaluó la GEIH del 2018 como un insumo relevante para la evaluación de un perfil económico de los bogotanos, lo cual permita analizar la capacidad de contribución tributaria, dadas sus características.

Acceso a los datos

La obtención de los datos se importó de la página web https://ignaciomsarmiento.github.io/GEIH2018_sample/. Allí, se emplearon técnicas de Web scraping, las cuales consisten en la extracción de datos de páginas web tanto dinámicas como estáticas. La página tiene los datos almacenados en 10 data *chunks* que comparten una estructura similar en sus datos. No se encontró restricciones para acceder a las 10 bases de datos. Todas las submuestras están almacenadas en una página dinámica en formato HTML. Con el paquete *rvest* versión 1.0.3 Wickham H (2022) se realizó el scraping de los datos. Específicamente, se empleó un bucle en el que se extrajo la información de las 10 bases y se almacenaron en un vector. Posteriormente, se concatenó la información de las submuestras en una base maestra que tuviera toda la información disponible. En total, la base maestra constó de 32177 observaciones y 179 variables.

Limpieza de los datos

Para la limpieza de los datos se realizó una primera selección de variables potencialmente relevantes, de acuerdo con la literatura. En el apartado de variables seleccionadas se especificará cada una de las variables a considerar en los análisis y la justificación teórica por la cual se

consideran relevantes. En total, se seleccionaron 33 variables[1]. Todas las variables fueron renombradas, excepto los identificadores de vivienda, hogar, persona y factores de expansión. De igual forma, se reorganizó la estructura de la variable parentesco con el jefe del hogar (parentesco_jhogar) como una dicótoma que toma el valor de 1 si es jefe del hogar y 0 de lo contrario; residencia en zona urbana (urbano) que toma el valor de 1 si reside en cabecera urbana y 0 de lo contrario y; sexo del encuestado (mujer) que toma el valor de 1 si es mujer y 0 de lo contrario. Además, se creó una variable de edad al cuadrado (edad2). Finalmente, se filtró la base para eliminar a los menores de 18 años y a la población desempleada. Hasta este punto, la base contaba con 22640 observaciones y 34 variables.

Imputación de valores en la variable de salario por hora. Como variable de interés se seleccionó el salario real por hora (salario_real_hora). Esta variable tenía 12748 valores faltantes (NA's). Por esta razón, se optó por evaluar la posibilidad de imputar valores de salario. Una primer alternativa, consistía en dar un valor promedio del salario por hora a todos los NA's; no obstante, este método presenta un riesgo bastante alto de sesgar los resultados ya que empujaría los datos hacia el promedio dada la magnitud de NA's. Por consiguiente, se optó por imputar salarios por hora con base en la media del ingreso del hogar. Es decir, dado que las personas restantes en la muestra reportaron estar empleadas, pero no tenían un valor en la variable de salario por hora, se cree que esto se pudo deber a un error en el diligenciamiento de la encuesta. Esto, porque son personas que trabajan y, por consiguiente, reciben un salario en su mayoría.

Con base en lo anterior, se optó por imputar un salario por hora de la siguiente manera: primero, se creó un salario promedio por hogar, con la omisión de los valores NA's, de acuerdo con los ingresos de cada integrante del hogar; segundo, se creó una nueva base en la que estaban las variables de identificadores por hogar (directorio y secuencia_p) y el salario promedio del hogar; tercero, se unió la base con la base maestra por identificadores de hogar y; finalmente, a todas las personas que reportaron estar trabajando y tenían más integrantes que percibían salario en el hogar, se les imputó el salario promedio del hogar como valor del salario por hora que percibían. Así, el valor de NA's quedó en 6784, el cuál corresponde a personas con NA's sin más integrantes en el hogar. Al comparar la variable de salario por hora sin y con valores imputados, se encontró que la

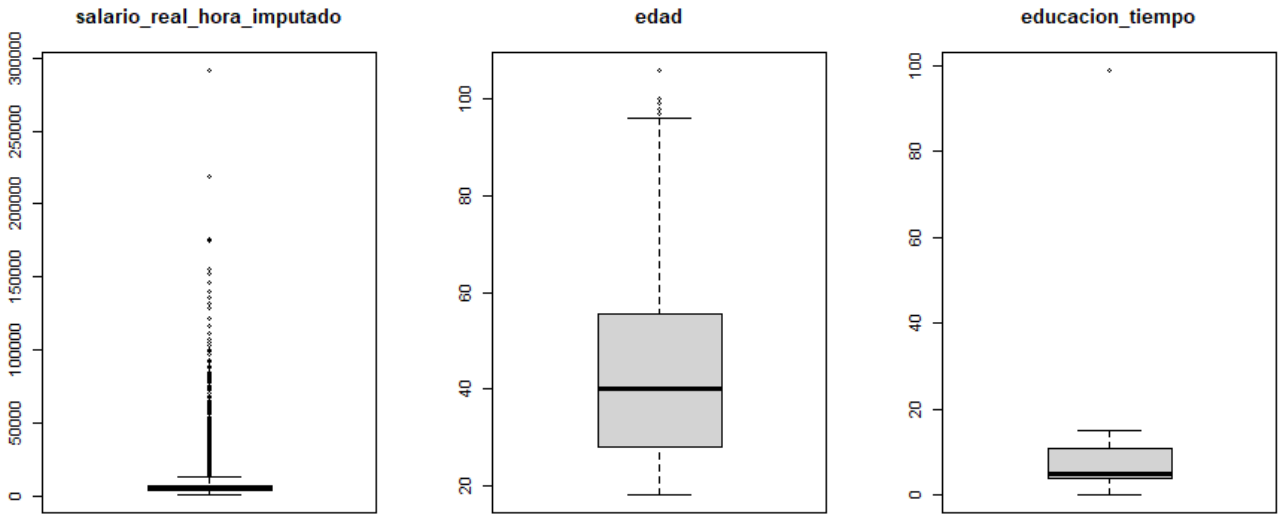
[1] "age", "sex", "maxEducLevel", "p6210s1","cuentaPropia", "pea", "wap", "p6240", "relab", "sizeFirm", "dsi", "estrato1", "formal", "p6050", "p6426", "totalHoursWorked", "y_bonificaciones_m", "y_gananciaIndep_m", "y_salary_m_hu", "y_vivienda_m", "ingtot", "p7040", "cotPension", "regSalud", "clase", "directorio", "dominio","fex_c", "fex_dpto", "fweight", "depto", "secuencia_p", "orden"

variable de interés pasó de tener un 43% al 70% de las observaciones. Con respecto a la evaluación de la media, no se encontraron diferencias significativas en el salario por hora sin (M =7946, DS = 11607) y con (M =7970, DS = 11158) imputación.

Outliers de las variables continuas de interés. Con el objetivo de evaluar valores atípicos (outliers), se optó por realizar una gráfica de bigotes y una prueba de outliers estandarizada. En el Gráfico 1.1 se presenta la caja de bigotes para las variables continuas de interés:

Gráfico 1.1.

Outliers de las variables de interés



Fuente: elaboración propia.

Como se observa, visualmente se encontraron valores extremos, especialmente en el margen superior de las variables de salario por hora, edad y tiempo de educación. Por consiguiente, se realizó el outliers_test del paquete *car* versión 3.1.2 (Fox y Weisberg, 2019). Esta prueba identifica valores atípicos mediante pruebas estadísticas para cada observación, donde señala los valores que difieren significativamente de los demás datos. Con respecto a la variable de salario por hora, se encontraron 10 valores atípicos, no obstante, al evaluar la coherencia de estos, se encontró que los datos son acordes a las características de las personas, razón por la que no se consideran errores en la recolección de la información. Por ejemplo, todas las personas tenían una educación terciaria, de los cuales 7 son trabajadores formales y 3 tienen NA's. Con respecto a las variables de edad (una observación) y tiempo de educación (2 observaciones), no se encontró coherencia en la información dado que tenían varios valores en NA. por consiguiente, los 3 valores fueron omitidos de la muestra.

Finalmente, se borraron los datos que no tuvieran un valor de salario por hora. Así, la muestra total quedó conformada por 15853 observaciones.

Variables de interés para el análisis. Existen diferentes rasgos y características que pueden explicar los ingresos de una persona. A continuación se señalan las variables seleccionadas y la justificación teórica por la cual se considera que son relevantes en el análisis:

salario_real_hora_imputado: el salario muestra el pago que recibe una persona por su trabajo. Se agrega el salario porque es la principal variable de interés en el modelo a estimar del mercado laboral, ya que interesa saber el efecto de las variables explicativas sobre el salario. En este caso, se dispone del valor nominal del salario por una hora de trabajo.

log_salario_hora_imputado: es la función logaritmo aplicada a salario_real_hora_imputado, que se utiliza con el fin de evaluar el efecto de las variables explicativas en términos porcentuales sobre el salario por hora. Como el logaritmo natural es una transformación monótona creciente, se puede aplicar para analizar el comportamiento del salario por hora.

edad: esta variable refleja los intereses, preferencias, oficios, necesidades del individuo. Conocer la edad de los individuos es fundamental para hacer un filtro y saber cuál es la población objetivo con el fin de hacer predicciones.

edad2: la transformación de elevar la edad al cuadrado tiene el objetivo de observar un posible efecto de la edad sobre el salario que no sea lineal, especialmente, para observar si el salario presenta rendimientos decrecientes respecto a la edad.

mujer: esta variable representa el género (o sexo), al ser una variable dicótoma, por lo cual es fundamental para observar diferencias salariales entre hombres y mujeres, ya que en el contexto colombiano existe una brecha salarial, relacionada generalmente con un contexto de discriminación hacia la mujer.

educacion_alcanzada y educación_tiempo: esta variable es importante dado que la educación o capital humano de un individuo mide sus conocimientos frente a uno o varios temas específicos, que pueden ser útiles en el empleo. La función de Mincer (1958) explica el salario mediante la educación y la experiencia adquirida. La variable educación_tiempo cumple el mismo objetivo, pero como variable continua.

emprendedor: esta variable refleja si el trabajador se desempeña por cuenta propia o no.

formal_informal: esta variable puede ser de importancia puesto que en el 2018 la informalidad en Colombia fue del 48,2% (El tiempo, 2019), lo cual puede estar directamente relacionado con el ingreso o salario que recibe cada individuo. Es decir, dependiendo de si su vinculación es informal, sus ingresos serán, con mayor probabilidad más inestables y menores a quienes se ecnuentran vinculados al sector formal.

parentesco_jhogar: dado el contexto de discriminación en el trabajo, esta variable permite evaluar si las diferencias en brechas salariales, en caso de existir, se ven acentuadas cuando la jefa del hogar es mujer.

relacion_laboral: las características del trabajo también son importantes para tener en cuenta a la hora de mirar el salario por hora. Ser empleado o empleador representa diferencias en el margen salarial.

tamaño_empresa*:* el tamaño puede influir en el flujo de ingresos que tienen sus empleados, especialmente en la mano de obra especializada.

Estadísticas descriptivas

Se realizó un análisis de las estadísticas descriptivas de las principales variables de interés. La Tabla 1.1 presenta los resultados descriptivos:

Tabla 1.1

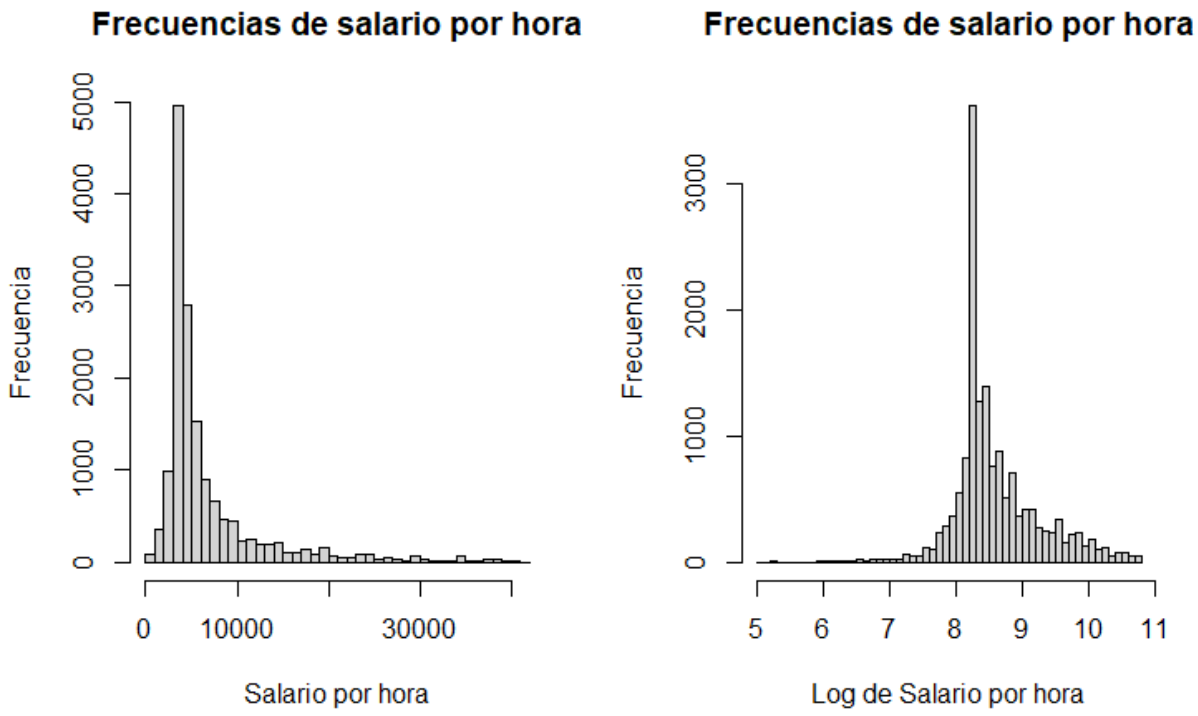
Estadísticas descriptivas de las variables de interés

variable	n_observaciones	promedio	sd	mín	max
salario_real_hora_imputado	15,853	7,970.382	11,158.580	152	291,667
edad	15,853	40.003	15.750	18	99
educacion_tiempo	15,853	6.582	3.568	0	15
mujer	15,853	0.528	0.499	0	1
emprendedor	15,853	0.141	0.348	0	1
formal_informal	12,798	0.661	0.473	0	1
parentesco_jhogar	15,853	0.409	0.492	0	1

Fuente: elaboración propia.

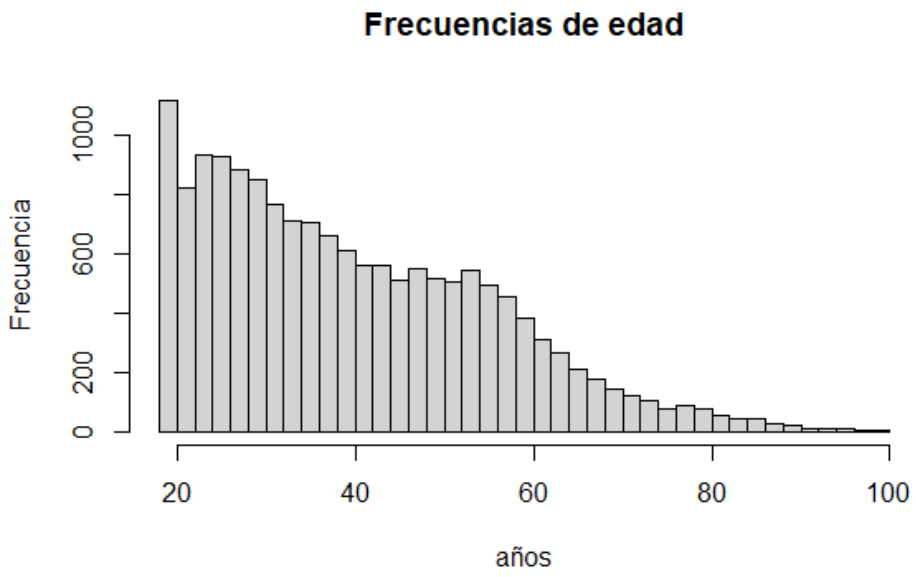
En primer lugar, se observa que el salario real por hora promedio de los encuestados es de $7970 (SD = 11158) pesos colombianos, con un valor mínimo de $152 y un máximo de 291,667. El Gráfico 1.2 presenta un histograma de frecuencias del salario por hora. Como se puede observar, la mayoría de las personas reportó un salario por hora aproximado al salario mínimo por hora del 2018 ($3255), lo cual indica que la mayoría de la población percibía ingresos mensuales cercanos al salario mínimo.

Gráfico 1.2



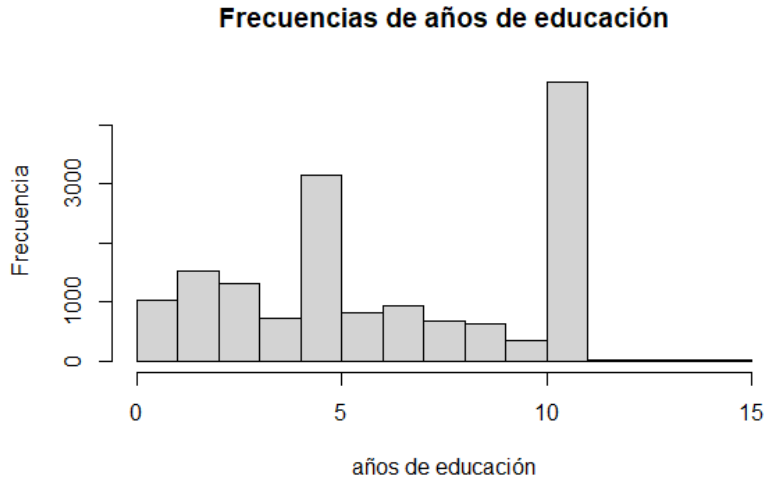
Fuente: elaboración propia.

Con respecto a la edad, se observa que los empleados en Colombia tienden a ser en su mayoría jóvenes, con una cola asimétrica hacia la derecha. La media de edad es de 40 años (DS = 15,75) con una edad máxima de 99 años. Especialmente, se observa que los bogotanos mayores de 60 que trabajan solo representan el 11,2% de la población económicamente activa empleada:



Fuente: elaboración propia.

Ahora bien, con respecto a los años de educación, la media fue de 6.58 años (DS = 3.56) con un mínimo de 0 años y máximo de 15 años. Especialmente, la mayoría de los encuestados respondieron haber estudiado 11 años, lo cual estaría relacionado a la educación básica secundaria en Colombia, que consiste en esa misma cantidad. Probablemente, esta podría ser una de las razones que explique el alto porcentaje de trabajadores que reportan un salario cercano al mínimo.



Fuente: elaboración propia.

Finalmente, con respecto a las variables dicótomas y categóricas de interés, el 52% son mujeres. Por otra parte, se observa que el número de trabajadores independientes (emprendedores) es bastante bajo, con solo un 14%; no obstante, no se observa una relación clara con respecto al número de trabajadores informales (34%). Es decir, probablemente, un alto porcentaje de los trabajadores independientes son formales, lo cual podría explicar la diferencia porcentual. Ahora bien, es importante tener en consideración que existen 3055 datos faltantes con respecto a la variable de trabajo formal. Probablemente, esta pueda ser otra de las diferencias que explique los cambios significativos.

Punto 3

Se tiene el modelo semilogarítmico Log-Lin de mercado laboral para estimar el salario:

$$log\ (w) = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 + u$$

Donde $log(w)$ se refiere al salario por hora devengado, $Edad$ a la edad de cada individuo en la muestra y $Edad^2$ es la variable $Edad$ elevada al cuadrado.

Después de realizar la estimación de los parámetros $\widehat{\beta_1}, \widehat{\beta_2},$ y $\widehat{\beta_3}$ por mínimos cuadrados ordinarios (MCO) se obtiene el siguiente resultado:

Tabla 3.1: Regresión Salario-Edad

	Dependent variable:
	Ln(salario)
Edad	0.020***
	(0.002)
Edad2	-0.0002***
	(0.00002)
Observations	15,853
R^2	0.014
Note:	*p<0.1; **p<0.05; ***p<0.01

En la tabla anterior se puede evidenciar que $\widehat{\beta_2}$= 0,020 y $\widehat{\beta_3}$= -0,0002 son estadísticamente significativos; es decir, con respecto a $\widehat{\beta_2}$, un (1) año más de edad aumenta en promedio el salario

por hora en 2%. Por su parte, el coeficiente β_3 es negativo, y muestra que si la edad al cuadrado aumenta 1 año entonces el salario disminuye 0,02%. Es importante resaltar que, aunque este valor es significativo, el cambio es marginal.

Un modelo como este, que muestra el salario en términos de la edad y la edad al cuadrado es una aproximación empírica para representar el efecto de la edad sobre el salario, lo cual explica que este crece mientras la edad aumenta. Así, se observa que se llega a un punto máximo y, posteriormente, a medida que la edad crece, se presenta una disminución en el salario (es decir, presenta rendimientos decrecientes con respecto a la edad).

Tanto $\widehat{\beta_2}$ y $\hat{\beta}_3$ son estadísticamente significativos, teniendo en cuenta distintos niveles de significancia (1%, 5% y 10%) ya que el p-valor es menor a 0,01 para ambos coeficientes. Esto implica que existe suficiente evidencia estadística para afirmar que los parámetros estimados de $Edad\ y\ Edad^2$ influyen en el salario. Al analizar el R^2, este tiene un valor de 0,014, lo cual se puede considerar como un bajo porcentaje de variabilidad en el salario explicados por $Edad$ y $Edad^2$. En otras palabras, el modelo tiene un bajo ajuste. Esto se puede deber a que hay variables importantes que explican el salario devengado que no han sido incluidas en el modelo. Por ejemplo, este modelo excluye variables como educación y experiencia, claves para el modelo de Mincer (1958), además de otras variables como el género.

Para hallar la edad a la cual se maximiza el salario, se deriva y se iguala a 0 la función del logaritmo del salario respecto a la edad:

$$\frac{\partial log(w)}{\partial\ Edad} = \beta_2 + \beta_3 Edad^2 = 0$$

$$Edad* = \frac{-\widehat{\beta_2}}{2\hat{\beta}_3}$$

$$Edad* = \frac{-0.0203207}{2(0.0001771)}$$

$$\boldsymbol{Edad} * = 57{,}37$$

Esto quiere decir que la edad a la cual se obtiene el máximo salario por hora es a los 57 años.

Por el método de remuestreo Bootstrap, se pueden obtener los errores estándar del parámetro $\widehat{\frac{-\beta_2}{2\beta_3}}$. En este caso, se usará R=1000, donde R es el número de submuestras utilizadas en el método.

Aplicando el comando boot en R:

Bootstrap Statistics:

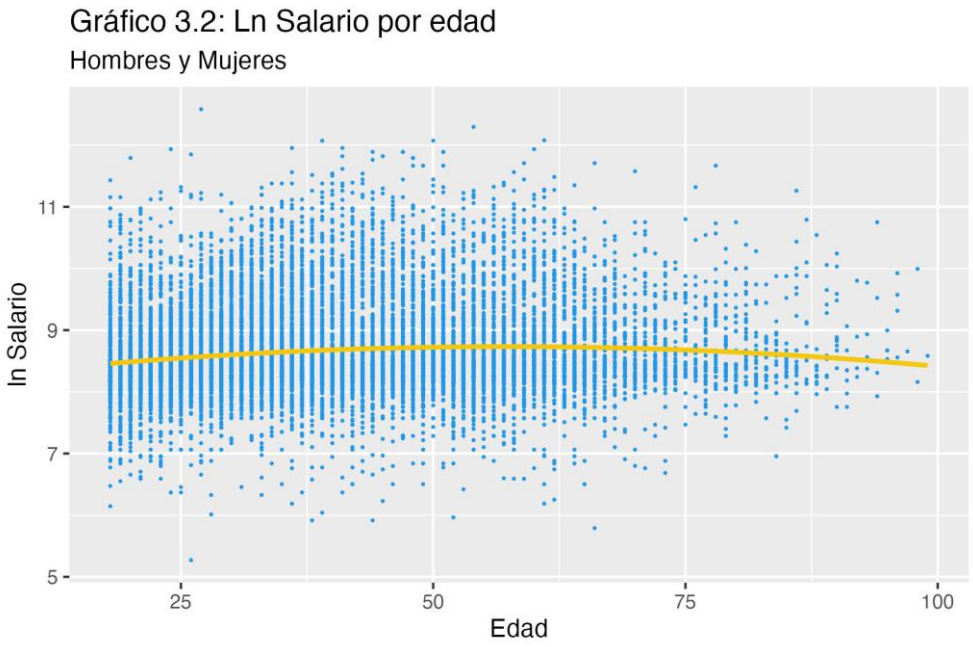
	original	bias	std.error
t1*	57.4	0.223	1.72

Lo que quiere decir que el error estándar del parámetro $\widehat{\frac{-\beta_2}{2\beta_3}}$ es igual a 1,72. Ahora con este valor se calculan los intervalos de confianza de la Edad*. En este caso se calcularán con una confianza del 95%:

IC_1= 57,37- 1,96*1,72= 53,8.
IC_2= 57,37+ 1,96*1,72= 60,5.

Esto quiere decir que, con 95% de probabilidad, la edad que maximiza el salario está entre 53,8 años y 60,5 años en la población. Intuitivamente esto tiene mucho sentido, puesto que, el mercado laboral requiere individuos en edad productiva y con rápido nivel de aprendizaje, especialmente personas que puedan adaptarse a cambios tecnológicos estén actualizados en el conocimiento y tengan las habilidades que requieren las empresas; por lo cual, a mayor edad, hay un punto en el cual el salario llega al máximo y luego el salario no crece con la edad.

El Gráfico 3.2 muestra toda la distribución muestral (puntos azules) y la línea modelo predicho (línea amarilla).



Punto 4

Se tiene el modelo no condicional:

Modelo 1: $$log(w) = \beta_1 + \beta_2 Mujer + u$$

Donde $log(w)$ se refiere al salario por hora devengado, $Mujer$ es una variable dummy que toma el valor de 1 si el individuo es mujer y 0 si es hombre.

Ahora, con el fin de estimar los parámetros, se estima el siguiente modelo condicional:

Modelo 2: $$log\ (w) = \alpha_1 + \alpha_2 Mujer + \alpha_3 Edad + \alpha_4 Edad^2 + \alpha_5\, Educación + \alpha_6 Relacion\_Laboral + \alpha_7 Tamaño\_empresa + u$$

Donde las variables $Mujer, Educación$, $Edad$ y $Edad^2$ son variables de control que reflejan las características del trabajador, mientras que las variables $Relacion\_Laboral$ y $Tamaño\_empresa$ son variables que se controlan por características del empleo.

La variable $Educación$ se refiere a los años de educación de individuo; $Relacion\_Laboral$ se refiere al tipo de trabajo que tiene el individuo, la variable $Tamaño\_empresa$ da información sobre la cantidad de personas que trabajan en la empresa del individuo.

Las variables $Relacion\_Laboral$ y $Tamaño\_empresa$ son variables categóricas mientras que $Edad$ y $Edad^2$son variables continuas.

Tabla 4.1: Regresión Salario-Mujer

	Dependent variable:		
	Ln(salario) (1)	Ln(salario) (2)	Ln(salario) (3)
Mujer	-0.030**		
	(0.013)		
Mujer FWL		-0.009	
		(0.012)	
Mujer FWL (Bootstrap)			-0.009
			(0.013)
AIC	27854.3	25350.1	
Variables de Control	No	Si	Si
Observations	12,798	12,798	12,798
R^2	0.0004	0.00005	

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

El modelo FWL (2) ha sido calculado con las variables de control, edad, educación, ocupación y tamaño de la empresa.

De acuerdo a los resultados obtenidos en el modelo no condicional (Modelo 1), el parámetro $\widehat{\beta_2}$ indica que, en promedio, el salario por hora de una mujer es 3 % menos que el de un hombre, ceteris paribus. Por otro lado, al utilizar los controles (Modelo 2) de $Edad$, $Edad^2$, $Educación$, $Relacion\_Laboral$ y $Tamaño\_empresa$, se usa el Teorema de Frisch-Waugh-Lovell para hallar el parámetro $\widehat{\alpha_2}$, el cual da un valor de 0,009. Es decir, que el salario por hora de las mujeres es 0,9% más bajo en comparación con los hombres (ceteris paribus) cuando se estima el modelo 2.

Se puede observar que este efecto es menor con los controles incorporados en comparación con el modelo 1.

Aunque los parámetros $\beta y\, \widehat{\alpha_2}\, Mujer$ de ambos modelos sea negativo, el coeficiente $\widehat{\alpha_2}$ (del modelo con controles) no es significativo (a ningún nivel), por lo cual estadísticamente no es posible afirmar que ser mujer influya en el salario por hora obtenido respecto a los hombres.

Al comparar ambos modelos con el estadístico R^2, podemos observar que el segundo modelo tiene un ligero mejor ajuste en comparación al primer modelo. Además, se ha incluido el estadístico Akaike Information Criterion (AIC), el cual es utilizado para comparar los ajustes de diferentes modelos, y poder elegir el "mejor". Este estadístico busca que haya un balance en el ajuste del modelo y la complejidad, penalizando los modelos con muchos predictores (complejidad del modelo). Al hacer la comparación por AIC, el mejor modelo sería el (2), que contiene controles.

Para obtener los errores estándar, se usa la técnica Bootstrap de remuestreo. En este caso, se definirá R= 1000, donde R es el número de submuestras utilizadas para calcular el error estándar. De acuerdo con los resultados en la Tabla 4.1, el error estándar de Bootstrap (0,013) es ligeramente mayor al que es simplemente obtenido por FWL (0,012). Esto puede suceder debido a que la metodología Bootstrap robustece los errores, asumiendo heteroscedasticidad.

Tabla 4.2: Regresión Salario-Edad por Género

	Dependent variable:		
	Ln(salario)		
	(1)	(2)	(3)
Edad	0.037^{***}	0.038^{***}	0.037^{***}
	(0.003)	(0.004)	(0.004)
Edad2	-0.0004^{***}	-0.0004^{***}	-0.0003^{***}
	(0.00003)	(0.00005)	(0.00004)
AIC	25378.7	12237.4	13127.1
Variables de Control	Si	Si	Si
Observations	12,798	6,187	6,611
R^2	0.178	0.209	0.153
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		
	Las variable de control empleadas son edad, educación, ocupación y tamaño de la empresa.		

La tabla 4.2 presenta los resultados de las diferencias en el salario por edad en las mujeres y hombres. El modelo (1) presenta el modelo global, mientras el modelo (2) muestra los resultados para los mujeres y el modelo (3) para hombres. Todos estos modelos contienen los mismos controles: edad, educación, ocupación y tamaño de la empresa.

Tal y como se hizo en el punto 3 (derivando la función de salario respecto a la edad e igualando a 0 y luego reemplazando por los parámetros estimados), se obtuvo que la edad que maximiza el salario devengado es 52,4 años. En el caso de las mujeres, estas llegan al salario máximo a la edad de 50,7 (aprox 51) años de edad, mientras que los hombres a los 53,7 (aprox 54) años de edad. Esto también se puede visualizar en el Gráfico 4.2.

Edades max en salario con controles

	General	**Mujeres**	**Hombres**
Edad	52.4	50.7	53.7

Calculando por el método de Bootstrap (con R=1000) los errores estándar de la edad máxima para hombres y mujeres, lo cual da un error estándar de 1,9022 para la mujer y 2,2093 para hombre.

Los intervalos de confianza (del 95%) de la edad máxima tanto para hombres como para mujeres son los siguientes:

IC_1- Mujer: 50,7 – 1,96*(1,936301652)= 46,90.
IC_2- Mujer: 50,7 + 1,96*(1,936301652)= 50,49.

IC_1- Hombre: 53,7 – 1,96*(2,438478959)= 48,92.
IC_2- Hombre: 53,7 + 1,96*(2,438478959)= 58,47

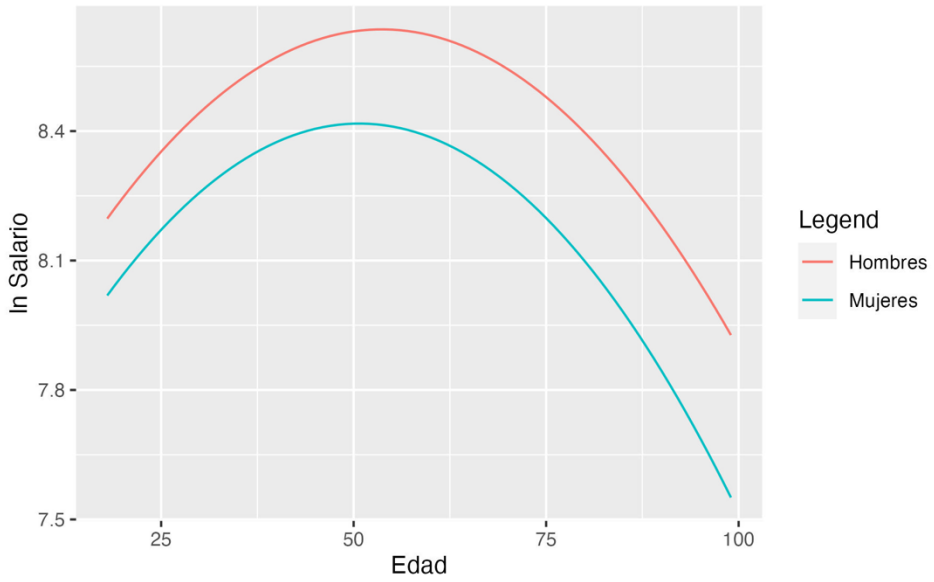
Con el fin de estimar la validez estadística de cada uno de los resultados de las edades "pico" se usa el estadístico t, con el fin de mirar la significancia de la edad máxima para mujeres y hombres.

Este modelo muestra que existen diferencias entre los salarios de las mujeres y hombres, incluso cuando se controla por las características de las personas como edad, educación, ocupación, además de los tipos de trabajos. Sin embargo, cabe resaltar que hay variables explicativas que podrían estar sesgando este modelo. En primera instancia, se está obviando el tiempo de experiencia laboral, el cual se sabe por el modelo de Mincer, que influye de manera sustancial el salario esperado. Esta variable viene acompañada de Experiencia^2 puesto que se espera que, de la misma forma que la variable edad, tenga una relación positiva inicialmente con el salario, hasta

llegar a un punto máximo donde empiece a decrecer. Además, se esperaría que otras variables omitidas pudieran también influir en el salario, como la raza.

Debido a esto, pese a que el modelo de la tabla 4.2 mejoró en comparación al presentado en la tabla 4.1, debido a que el ajuste, medido con el R^2 mejoró de manera importante, no podemos

Gráfico 4.2: Salario por edad con controles



afirmar que existe efectivamente un problema de discriminación.

Por otro lado, debido a la limpieza de datos realizada, se eliminaron las observaciones que contaban con valores faltantes en las variables de interés. Esto podría haber causado un sesgo de selección, si es que un grupo importante de la población haya sido eliminado al realizar la limpieza.

Edades Max Salario IC

	Mujeres	Hombres
	50.7	53.7
IC inferior	46.7	48.6
IC superior	54.3	58.2

5. Predicción del ingreso laboral

Anteriormente se desarrollaron diferentes modelos para inferir los efectos de diferentes variables en el ingreso laboral de los individuos. Sin embargo, uno de los intereses principales de este trabajo es, a partir del análisis de las diferentes variables, estudiar modelos predictivos del salario, para asistir en la identificación de posibles inconsistencias en los reportes del ingreso.

Para asegurar la reproducibilidad del ejercicio se establece una semilla, en este caso set.seed(777). Para empezar, se divide la base de datos en: una muestra de entrenamiento (que representa el 70% de los datos) y otra de prueba (30%). Como resultado, la muestra de entrenamiento (*train*) se compone de 8.984 observaciones y la de prueba (*test*) de 3.812 observaciones.

Dado que el principal interés es predecir bien fuera de la muestra, es necesario evaluar los modelos en los datos de prueba. Así, posteriormente se utiliza el coeficiente estimado con el set de entrenamiento y luego se usa como predictor en los datos de prueba. Se procede a estimar nuevamente los modelos expuestos en los puntos anteriores más cinco (5) modelos predictivos, los cuales exploran la complejidad a partir de no linealidades y la interacción de variables:

Tabla 5.1

Especificación de los modelos

Modelo	Forma funcional
1	$log\ (w) = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 + u$
2	$log(w) = \beta_1 + \beta_2 Mujer + u$
3	$log(w) = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 + \beta_4 Educación + \beta_5 Tamaño\_empresa + u$
4	$\log(w) = \beta_1 + \beta_2 Mujer + \beta_3 Edad + \beta_4 Edad^2 + \beta_5 Educación + \beta_6 edad * Mujer + u$
5	$\log(w) = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 + \beta_4 Educación + \beta_5 Estrato + \beta_6 Edad * Mujer + u$
6	$\log(w) = \beta_1 + \beta_2 Mujer + \beta_3 Edad + \beta_4 Edad^2 + \beta_5 Educación + \beta_6 Educación^2 + \beta_7 estrato + \beta_8 Mujer * edad + u$
7	$\log(w) = \beta_1 + \beta_2 Mujer + \beta_3 Edad + \beta_4 Edad^2 + \beta_5 Edad^3 + \beta_6 Eduación + \beta_7 Educación^2 + \beta_8 Estrato + \beta_9 Edad * Mujer + u$

Nota: la variable $Mujer$ identifica el género.

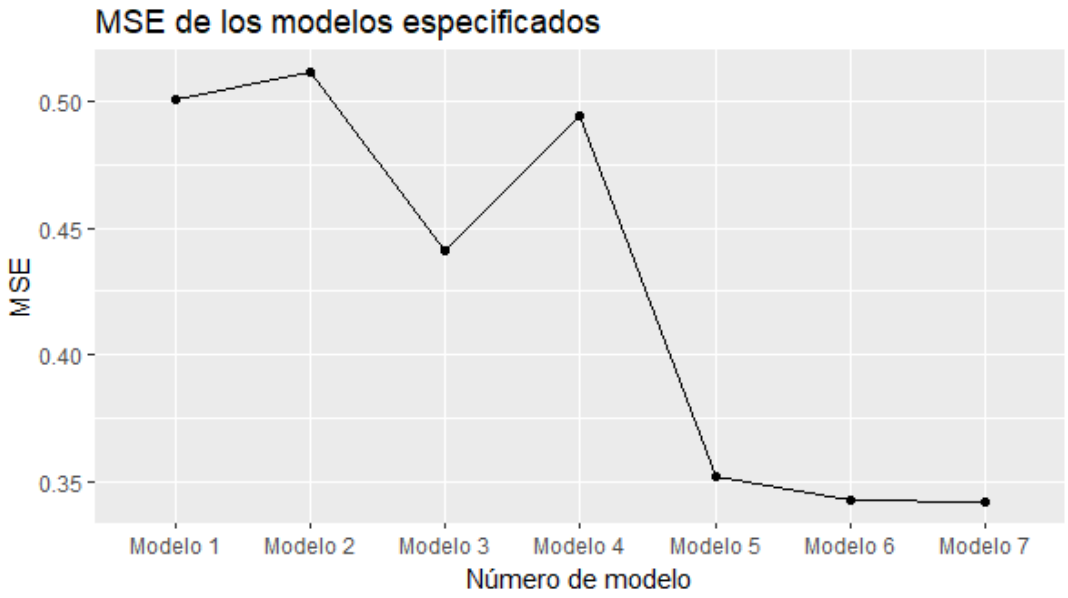
Para el desarrollo del análisis predictivo se usa el método de validación cruzada. Se entrenaron siete modelos con las especificaciones antes desarrolladas, y posteriormente se evaluó su capacidad de predicción respecto a la muestra de prueba. Como métrica de performance se usará el error cuadrático medio (MSE). A continuación, se presentan los resultados para los modelos evaluados:

Tabla 5.2 Errores de predicción

	Modelo	MSE
1	Modelo 1	0.501
2	Modelo 2	0.511
3	Modelo 3	0.441
4	Modelo 4	0.494
5	Modelo 5	0.352
6	Modelo 6	0.343
7	Modelo 7	0.342

A medida que se aumenta la complejidad del modelo, incluyendo variables adicionales, interacciones y no linealidades, se aumenta la capacidad predictiva del modelo, como lo evidencia la tabla 5.2. A medida que el modelo se complejiza, el error de predicción se va reduciendo, sin embargo, puede que se dé un punto en que demasiada complejidad implique un sobreajuste del modelo en la muestra de entrenamiento y una capacidad predictiva deficiente. No obstante, de acuerdo con los resultados, no parece ser el caso de ninguno de los modelos puesto que su MSE en el set de test no incrementa (Gráfico 5.1).

Gráfico 5.1

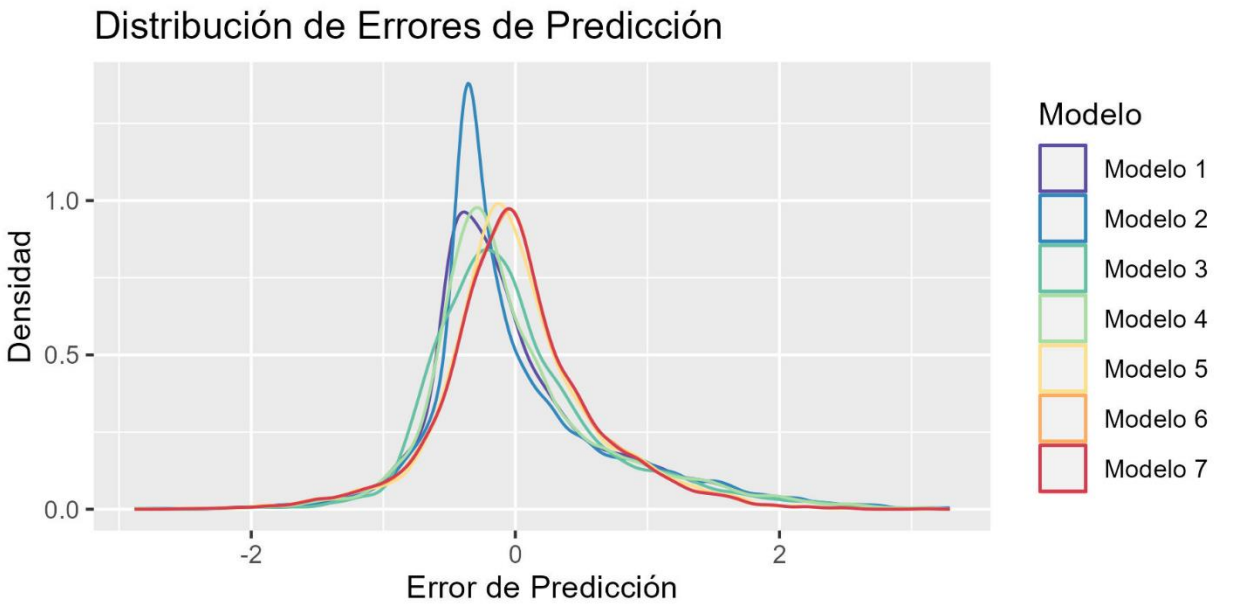


De acuerdo con los resultados de la tabla, los modelos 6 y 7 son aquellos que presentan un menor error de predicción. Al considerar variaciones de orden cuadrático respecto a la educación, esto se puede deber a que haya un comportamiento similar de esta variable con el comportamiento de la edad: a medida que un individuo aumenta su nivel educativo, su salario estimado aumenta, sin embargo, existe un punto límite en donde mayor educación no representa un aumento en los ingresos. Así mismo, la inclusión de los demás controles como el estrato y el género robustece la estimación. La diferencia entre el modelo 6 y el modelo 7 está en la inclusión de una variación cubica respecto a la edad, sin embargo, esto no genera mayor variación en la predicción y no se cuenta con evidencia que pueda justificar un comportamiento de este tipo.

Adicionalmente, para hacer un análisis de presencia de outliers, se procede a graficar la distribución de los errores de predicción de los modelos planteados, se identifica que los modelos 6 y 7 tienen una distribución estándar con media cercana a 0, lo cual da cuenta de su bondad predictiva, mientras que los demás modelos, si bien pueden tener una distribución similar tienden a alejarse de la media 0. Con el fin de identificar si los valores en las colas de distribución corresponden a valores extremos de los datos o al error propio del modelo, se realizó el análisis de las observaciones con errores de predicción en los percentiles 5% y 95% de la distribución para el modelo 7, si bien corresponden a valores extremos de los datos, se encuentran dentro de los valores normales de las variables, por lo que no se identifican condiciones que puedan sugerir

inconsistencias en el reporte de ingreso laboral, estos errores corresponden al error de predicción propio del modelo (Gráfico 5.2).

Gráfico 5.2



Fuente: elaboración propia.

Finalmente, para validar los modelos 6 y 7 con la mejor capacidad predictiva, se corrió la regresión bajo el modelo LOOCV. Los resultados obtenidos para el modelo 7 fueron de un MSE mayor al del enfoque de validación cruzada, el cual era de 0.342. De igual forma, esto ocurrió para el modelo 6, para el que se obtuvo un MSE de 0.582, mientras que con validación cruzada fue de 0.343. Esto confirma que, al entrenar el modelo con todos los datos, se aumenta el MSE, es decir, el modelo pierde poder de predicción. Tiene sentido según la teoría, ya que, al entrenar el modelo con todos los datos, este pude aproximarse a la forma funcional del proceso generador de datos, no obstante, para este ejercicio, esto no ocurrió de esta manera en el caso de la validación cruzada convencional. Esto explica la diferencia entre los dos métodos, mostrando que, posiblemente el modelo escogido no sea el mejor y se deba explorar otro tipo de variables y de formas funcionales.

Referencias

Concha, T.; Ramírez, J.; Acosta, O. (2017). *Tributación en Colombia: reformas, evasión y equidad*. Estudios y perspectivas. Cepal.

DANE., (2018). Mercado laboral (Empleo y desempleo) Históricos. Recuperado de: https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo https://www.dane.gov.co/index.php/estadisticas-por-tema/mercado-laboral/empleo-y-desempleo/geih-historicos

El tiempo, 2019. "Informalidad de trabajadores no cede y es de 48,2 %", consultado en: https://www.eltiempo.com/economia/cifras-de-la-informalidad-laboral-en-colombia-a-enero-de-2019-326116

Fox J, Weisberg S (2019). _An R Companion to Applied Regression_, Third edition. *Sage, Thousand Oaks CA*. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.

La República, (2022). *La evasión de impuestos le estaría quitando a Colombia cerca de $80 billones al año*. Recuperado de: https://www.larepublica.co/economia/la-evasion-deimpuestos-le-estaria-quitando-a-colombia-cerca-de-80-billones-al-ano-3418446

Mincer, J. (1958). "Investment in Human Capital and Personal Income Distribution". *Journal of Political Economy.*

R Core Team (2023). _R: A Language and Environment for Statistical Computing_. *R Foundation for Statistical Computing*, Vienna, Austria. <https://www.R-project.org/>.

Wickham H (2022). _rvest: Easily Harvest (Scrape) Web Pages_. *R package version 1.0.3,* <https://CRAN.R-project.org/package=rvest>.