Prediciendo el ingreso laboral por hora de los empleados de la ciudad de Bogotá para 2018

Carrillo, Andrés

Hernández, Lizbeth

Higuera , Diana

Vallejo , Juan

12 de febrero de 2023

1. Introducción

La estimación de salarios esperados o conocer los retornos salariales de los invidious dadas ciertas caracteriticas se ha abordado durante año, no solo como por conocer los salarios de los individuos, si no tambien por enteder que caracteriscas hacen que estos salarios cambien, permitiendo tomar a los gobiernos y hacedores de politicas publicas mejores dediciones. Por ejemplo, la teoría de capital humano desde la teorín a de Micer (1974) postula que la inversión en capital humano a través de la educación representa ganancias futuras que se reflejan a través de incrementos en los ingreso laborals del orden del 10 %, dicho de otro modo, cada año de estudio adicional que tenga un individuo va a representar un incremento del 10 % en su ingreso laboral. De manera similar Becker (1964) postula que el capital humano incluye características innatas al individuo que generan ganancias en ingreso laborals; de hecho, las inversiones que hace un individuo en salud, bienestar, educación, desde esta perspectiva, se plantea buscando un retorno a la inversión realizada bien sea a través de ingreso laboral o incrementos adicionales a sus ingresos.

Igualmente, en el sector publico es importante conocer las declaraciones exactas de los ingresos de las personas, sin embargo, muchos de estos datos son autoreportados, donde los individuos suelen negarse o subestimar sus ingresos lo que con lleva a declaraciones de renta inexactas o en su defecto fraude fiscal. Seún el Internacional Revenue Service (IRS), alrededor del 83,6 % de los impuestos en Estados Unidos se paga voluntariamente y a tiempo. Es aqui donde un modelo de prediccion de ingresos puede ayudar a detectar casos de fraude y disminuir la brecha de evación fiscal. Ademas, un modelo de predicicon de ingresos puede servir para identiciar a potenciasles beneficiariso y hogares vulnerables que puedan necesitar ayuda gubernamental.

2. Datos

La información utilizada en el análisis corresponde a la base de datos de la Gran Encuesta Integrad de Hogares -GEIH- de la ciudad de Bogotá para el año 2018, realizada por el DANE de manera mensual. Esta encuesta tiene como objetivo proporcionar información sobre la estructura de la fuerza de trabajo (empleo, desempleo e inactividad) y sus características, de esta encuesta se obtienen los principales indicadores de mecarcado laboral como: la tasa global de participación (TGP), la tasa de ocupación (TO), la tasa de desempleo (TD) entre otras, lo que permite a las entidades gubernamentales tomar decisiones y llevar a cabo investigaciones correspondientes al mercado laboral del país. Además de las variables de mercado laboral, la base contiene información acerca de las características sociodemográficas de las personas como sexo, edad, educación, estado civil y características generales relacionadas con vivienda, servicios públicos, y seguridad social.

2.1. Proceso de adquisición de datos

El conjunto de datos fue obtenido del sitio web https://ignaciomsarmiento.github.io/GEIH2018_sample/ , al inspeccioanr el conjunto de datos se aprecia que no se encuentran en un formato descargable, sino que estan incrutrados en una tabla en la página en 10 subconjuntos de datos, para lo cual se procede hacer uso de la herramienta de **web-scraping** que permite extraer datos de sitios web de forma automatizada, y asi poder hacer la recopilació de los datos. Para el proceso de recopilación y posteriormente manipualción de los datos se hace uso del entorno de ***RStudio*** pra el lengguaje de programacion en **R**

2.2. Limpieza , manipulacion y proceso de imputación de datos

El análisis de los datos se restringe a la sub muestra de los empleados mayores de 18 años de la ciudad de Bogotá en el período 2018. De lo anterior, se obtiene una muestra de 16.542 observaciones que representa el 51 % de los individuos de Bogotá.

El proceso de manipulación de los datos inicio con una inspección de las variables de ingreso que contiene el conjunto de datos, se observan las varaibles de ingreso laboral mensual e ingreso total mensual, en ambos casos estas varaibles presentan valores perdidos en sus observaciones, en el caso del ingreso laboral mensual tiene un 40 % de observaciones no reportadas, mientras la varaible inngreso total mensual tiene un 11 % de observaciones no reportadas.

Lo anterior plantea un problema importante sobre nuestra variable principal a predecir más adelante que es el ingreso laboral por hora, para lo cual se procede hacer dos cosas: 1) Construir la varaible de ingreso laboral según la metodología de pobreza del DANE y para lo cual contamos con las varaibles inputadas por la entidad en nuestor conjunto de datos y 2) Imputar la variable de ingreso total mensual utilizando la mediana.

En el primer caso de la construcción de la variable de ingreso mesual por metodología DANE, corresponde a la sumatoria de las variables ingreso por primera actividad,ingreso por segunda actividad y ingreso en especie, con las respectivas variables de ingreso tanto de primera y segunta actividad como de especie imputadas por el DANE.Además, se encuentra que esta variable contruida no posee valores perdidos y que el 1 % de las observaciones reportan un ingreso laboral de cero.

En el segundo caso, correspondiente a la inputación de la variable ingreso total se realizó se realizo utilizando la mediana de la variable ingreso total y esto debido a que la media esta afectada por valore muy altos en nuestro conjunto de datos.En conclusión tanto la variable construida por la metodologá del DANE como la imputada son casi identicas en

sus distribuciones. A partir de esta analisis sobre la variable ingreso se contruye tanto la variable ingreso laboral semanal y la variable ingreso laboral por hora.

2.3. Estadísticas descriptivas

Iniaciamos con un analisis gráfico donde se observa que en promedio el ingreso laboral por hora de los hombres y las mujeres no parece ser diferente. No obstante, cuando observamos este mismo grafico pero para el ingreso laboral mensual se evidencia una brecha positiva a favor de los hombres, es decir, los hombres tiene un ingreso laboral mensual en promedio más alto que las mujeres. Estas díferencias puenden deber a que en general en Colombia se pagan salarios mensuales en ves de salarios por hora.

Figura 1. Salario por hora entre hombres y mujeres

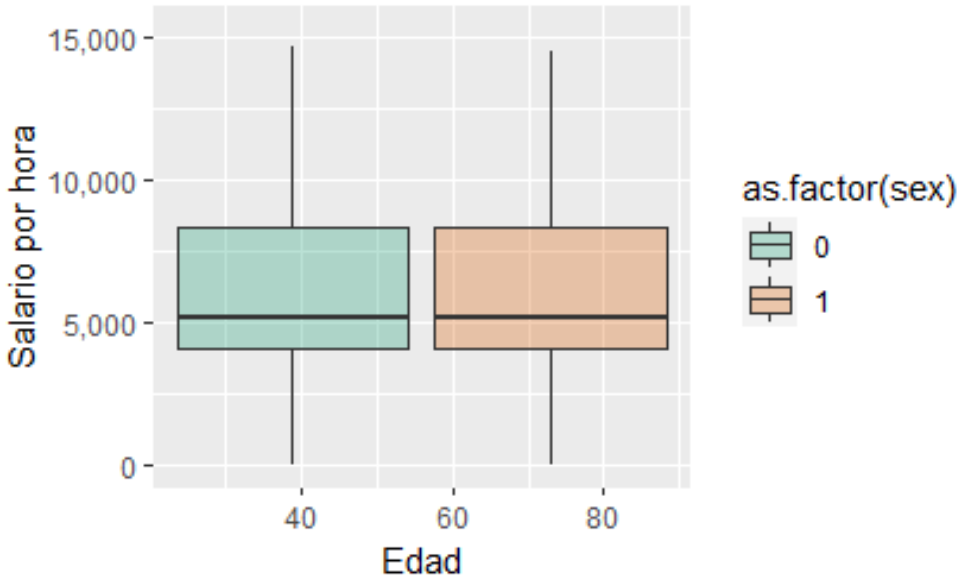
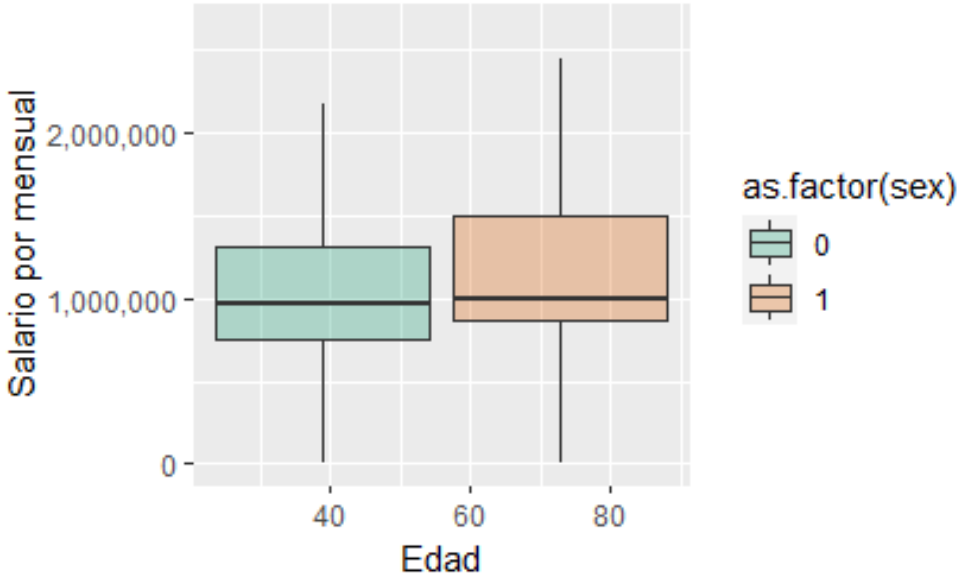


Figura 2. Salario mensual entre hombres y mujeres



En el cuadro 1 se presentan los estadisticos descriptivos , la edad promedio de los individuos de la muestra es de 36 años , la proporcion de mujeres es de 51 % y el salario promedio por hora es de 9.106 pesos con una desviación estandar de 3.872, el 56 % de los individuos son trabajadores formales, en promedio trabajan 40 horas en total, en cuanto a nivel educativo el 23 % tiene nivel educativo primaria, el 20 % secundaria, mientras que el 38 % tiene educación media y el 46 % eduación superior.

Cuadro 1. Estadisticas descriptivas generales

Statistic	Media /Proporción	Dev. Est
Edad	36.674	18.608
Sexo (Mujer)	0.509	0.014
Salario por hora	9,106	3,872
Formal	0.561	0.115
Total horas trabajadas	40.16	22.62
Educación primaria	0.239	0.146
Educación secundaria	0.209	0.143
Educación media	0.388	0.107
Educación superior	0.466	0.043

3. Identificación

Para realizar una aproximacion a la prediccion del ingreso salarial este trabajo tiene como objetivo pricipal construir un modelo de ingreso laborals individuales por hora de los empleados mayores de 18 años de la ciudad de Bogotá. Teniendo en cuenta esto, la estrategia empirica se basa en el siguiente modelo de regresión lineal

$$w = f(X) + u \tag{1}$$

Donde w es el ingreso laboral por hora para el individuo i, y X es una matriz que incluye las posibles variables explicativas o predictoras del ingreso laboral por hora y se asume x como una vector de los atributos individuales y laborales (edad, el sexo del trabjador, el nivel educativo, la ocupación) entre otras. Nos centraremos en

$$f(x) = X\beta$$

Dado que nuestro objetivo es poder encontrar la funcion adecuada que ajuste y permita predecir de la mejor forma w a partir de la equación (1). En este caso β es un vector de parámetros estimado que resume la relacion de cada determinante x sobre el ingreso laboral.

4. Análisis y Resultados

4.1. Perfil ingreso laboral-Ead

Iniciamos primero estimando los parametros para la predicción del ingreso laboral por hora de los empleados en Bogotá . Para este primer apartado utilizamos como variable predictiva del ingreso laboral por hora, la edad y edad al cuadrado, para poder entender como cambia el salaria por hora de los individuos , en la medida en que aumentan su edad, la razón de incluir la variable de edad al cuadrado se debe a que estudios relacionados con la economía laboral sugieren que el perfil de edad-ingreso laboral de los trabajadores sigue una función concanva, los ingresos laborales tienden hacer bajos cuando el trabajador es joven y luego van aumentando en la medida en que el trabajado vaaumentando su edad hasta alcanzar un maximo en torno a los 50 años, y posteriormente tienden a disminuir.

Para estimar esta relación , se realiza un modelo de regresión simple, como se muestra en la siguiente ecuación

$$log(Ingreso - laboral_{hora}) = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 \tag{2}$$

El cuadro 1 muestra las estimaciones del ingreso laboral por hora en función de la edad

y edad al cuadrado,se observa que el estadístico F es muy pequeño y estadísticamente significativo, lo que indica que al menos una de las varaibles predictoras del modelo esta relacionada con la variable predicha. El ajuste del modelo se puede hacer a partir de los valores del RSE o por medio del R2, el RSE nos dice que en promedio cualquier predicción del modelo se aleja 0.80 unidades del valor verdadero, mientras que el R2 nos dice que las variables del modelo son capaz de explicar solo el 1 % de la variaciones observadas en el ingreso laboral por hora. En general se observa que todos los coeficientes incluido el de la constante son estadísticamente significativo a un nivel de confianza del 95 %. En el caso de la constante se interpreta como el valor promedio salarial por hora cuando una persona tiene 18 años que es la edad cero en nuestra muestra, lo que indica que una persona recien ingresa al mercado laboral tiene un ingreso laboral de 8.03 pesos, en el caso del coefiente de la edad se interpreta como la variacion promedio del ingreso laboral por hora, un aumento de un año en la edad el ingreso laboral por hora aumenta en 4 %, lo que muestra que a mayor edad mayor sera el ingreso laboral por hora recibido por una persona. En el caso de la edad al cuadrado esta captura el efecto decreciente en el ingreso laboral por hora en cierto punto de mayoria de edad.

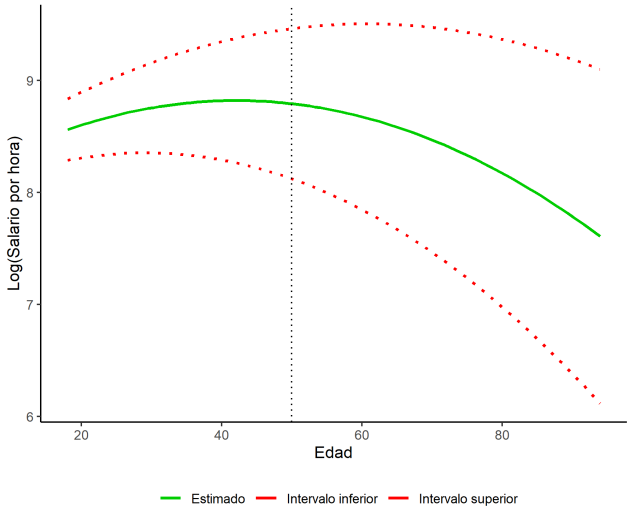
Cuadro 2. Estimación modelo de regresión ingreso laboral - Edad

	Variable dependiente:
	log (ingreso laboral hora)
Edad	0.04^{***}
	(0.003)
Edad al cuadrado	-0.0004^{***}
	(0.0000)
Constante	8.03^{***}
	(0.06)
Observaciones	15,121
R^2	0.01
R^2 Ajustado	0.01
RSE	0.80
Estadístico F	98.84^{***}
***Nota**: Errores estándar robusto en paréntesis*	*p<0.1; **p<0.05; ***p<0.01

Para poder obtener la edad en donde el ingreso laboral por hora de un individuo empieza a caer, derivamos la equación (2) remplazamos los parametros β de edad por los coeficientes obtenidos en el Cuadro 1 e igualamos a cero, despejando se obtiene que a partir de los 39 años empieza a descender el ingreso laboral por hora. La Figura 3 muestra el perfil estimado de la edada y el ingreso laboral por hora, utilizando la metodología de Bootstrap para construir los intervalos de confianza que se muestran con las líneas de color rojor Se puede observar como la línea estimada verde se encuentra dentro de los intervalos de confianza, ademas que el valor promedio del salio es creciente entre los 18 y 50 años años, y posterior a esta edad los individuos empiezan a tener una caida en sus ingreso laboral, ya sea bien por que estan llegando a la etapa de la vejez, o empiezan a gozar de su pensión,

valor que es menor al ingreso laboral recibido por trabajar.

Figura 3. Bootstrap para estimación de errores robustos



4.2. Brecha salarial de género

Las estimaciones de la brecha salarial entre hombres y mujeres es importante ya que evidencia la disparidad en ingreso dependiendo del género, y esto se traduce, por ejemplo, menos participación en el mercados laboral de las mujeres y mayor tiempo en la busqueda de empleo. De acuerdo con datos del DANE entre 2013 y 2018 la brecha salarial por género en Colonbia habia disminuido 5.3 puntos porcentuales, no obstante, entre 2018 y 2019 esta brecha aumento en 0.8 puntos porcentuales ubicandose en 12.9, lo que quiere decir, que en 2019 las mujeres ganaban 12.9 % menos que sus colegas hombres.

Siguiendo la identificación del modelo expuesto en la parte 3 de este documento, se estiam la brecha salaria incondicional expresada en la siguiente ecuación:

$$log(ingresolaboral_{hora}) = \beta_1 + \beta_2 Mujer + u \tag{3}$$

Donde $log(ingresolaboral_{hora})$ mide el ingreso laboral por hora del individuo i, mientras

que la variable explicativa *mujer* es una varaible binaria que toma el valor de uno si la persona en la base de datos es una mujer y cero si es un hombre, en este caso si el coeficiente de la variable predictora mujer es negativo , las mujeres ganan menos que en % en relacion a los hombres.

El cuadro 2 muestra las estimación de la brecha salaria por hora, se observa que el ingreso laboral por hora cae en un 7 % por ser mujer, siendo esto estadisticamente significativo, lo que evidencia la existencia de una brecha en el mercado laboral en contra de las mujeres, es decir, en promedio una mujer recibe 7 % menos de ingreso laboral por hora que un hombre,los estadísticos de ajuste del modelo muestran un estadsitico F significativo que indica que las varaibles incluidas en el modelo explican un 30.47 % los cambios en el ingreso por hora, mientras que el sexo del trabajador explica solamente un 2 % las varaiciones del ingreso laboral por hora.

Cuadro 3. Estimación brecha salarial de género

	Variable dependiente:
	log (ingreso laboral hora)
Mujer = 1	-0.07^{***}
	(0.01)
Constante	8.77^{***}
	(0.01)
Observaciones	15,121
R^2	0.002
R^2 Ajustado	0.002
RSE	0.80
Estadístico F	30.47^{***}

Nota: Errores estándar robusto en paréntesis *p<0.1; **p<0.05; ***p<0.01

Por otra parte, es muy común escuchar el eslogan " ***a igual trabajo, igual ingreso laboral***", para poder contrastar esta hipotesis, se deberia encontrar que para empleados con características laborales y profesionales similares, no deberia existir diferencias salariales entre hombres y mujer. Para corroborar esto estimamos el siguiente modelo de regresión:

$$log(Ingreso - laboral_{hora}) = \beta_1 + \beta_2 Mujer + \beta_3 Edad + \beta_4 Edad2 + u \tag{4}$$

El cuadro 3 muestra las estimación de la brecha salaria por hora sin control y con controles donde se incluyeron las variables de edad, si es trabajador formal, el nivel de estracto,, nivel educativo y total de horas trabajadas,se observa que el ingreso laboral por hora , controlando por estas variables, el salario por hora cae un 18 % por ser mujer, siendo esto estadisticamente significativo y un coeficiente mayor que el modelo sin controles, igualmentem, se observa que ser trabajador formal aumenta el salario por hora en un 41 %, ademas, a un nivel socioeconomico más alto y niveles educativos más altos, es mayor es el retonro en el salario por hora,los estadísticos de ajuste del modelo muestran un estadsitico F significativo que indica que las varaibles incluidas en el modelo explican los cambios en el ingreso por hora, ademas, el ajuste del modelo es del 42 %.

Cuadro 4. Estimación brecha salarial de género condicionada

	Variable dependiente:	
	log (ingreso laboral hora)	
	(1)	(2)
Mujer = 1	-0.07^{***}	-0.18^{***}
	(0.01)	(0.01)
Edad		0.04^{***}
		(0.002)
Edad al cuadrado		-0.0004^{***}
		(0.0000)
Formal		0.44^{***}
		(0.01)
Educación secundaria		0.09^{***}
		(0.02)
Educación media		0.21^{***}
		(0.02)
Educación Superior		0.74^{***}
		(0.02)
Total horas trabajadas		-0.01^{***}
		(0.0004)
Constante	8.77^{***}	7.81^{***}
	(0.01)	(0.05)
Observaciones	15,121	15,121
R^2	0.002	0.35
R^2Ajustado	0.002	0.35
RSE	0.80	0.65
Estadístico F	30.47^{***}	$1{,}028.65^{***}$

Nota: Errores estándar robusto en paréntesis *p<0.1; **p<0.05; ***p<0.01

4.2.1. Estimación de la brecha de género bajo el teorema FWL

Una de las tecnicas utilizadas para limpiar las distorciones o relaciones entre variables explicativas se utilizo el teorema de Frisch-Waugh-Lovel, consiste en estimar regresiones a partir de los residuos tanto de la variable explicativa como de la variable predicha. El

cuadro 4 muestra que tanto el coeficiente de mujer en la regresión con controles es igual que el coeficiente de la regresión de FWL. Se puede concluir que limpiado los efectos de las relaciones entre variables el Teorema de FWL muestra que sigue existiendo una brecha negativa para las mujeres y este coeficiente se mantiene igual.

Cuadro 5. Estimación brecha salarial de género condicionada

	Variable dependiente:	
	log (ingreso laboral hora)	log (ingreso laboral hora) - Residual
	(1)	(2)
Mujer = 1	-0.18^{***}	
	(0.01)	
Mujer - Residuos		-0.18^{***}
		(0.01)
Constant	7.81^{***}	0.00
	(0.05)	(0.01)
Observaciones	15,121	15,121
R^2	0.35	0.02
R^2 Ajustado	0.35	0.02
RSE	0.65	0.65
Estadístico F	$1{,}028.65^{***}$	283.04^{***}
Note:		*p<0.1; **p<0.05; ***p<0.01

El cuadro 5 muestra los parametros estimados a partir del Teorema de FWL por el método de Bootstrap que tiene como objetivo estimar la incertidumbre y ajsute del modelo, se puede apreciar en la columna 3 que el error estandar no cambia mucho por lo cual los resultados encontrados no se estan sobre estimando y las pruebas de hipotesisi sobre los parametros va ser valida.

Cuadro 6. Estimación brecha salarial de género condicionada

Estadísticos FWL Bootstrap			
	Original	Sesgo	Error Estándar
β_0	0	-0.0001	0.005
Mujer	-0.18	-0.0001	0.014

Por otra parte, tambien se calcularon las edades máximas donde el ingeso por hora empieza a caer entre mujeres y hombres como su intervalo de confianza, se observa que los hombres tienen una edad promedio más altas en 3 años, antes de que su salario por hora empiece a caer en comparación con las mujeres, si observamos el limite superior de las mujesres que es de 54 años , es la edad máxima de los hombres, sin embargo, hay hombres con edad 10 años adicional donde su salario aun no cae sino hasta alcanzar los 65 años. Contextualmente, podriamos asumir que como son empleados, hay un marco legal diferencial en edades entre hombres y mujeres para pensionarse y esto puede estar envidenciandose en los resultados del cuadro 6.

Cuadro 7. Estimación de las edades máximas de Mujeres y Hombres e Intervalos

Edades máximas por Bootstrap entre (Mujeres y Hombres)			
	Lim. Inferior	Edad máx	Lm. Superior
Mujeres	49	51	54
Hombres	49	54	65

Los resultados del Cuadro 7 se pueden mostrar tambien de forma gráfica. La Figura 4 muestra el comportamiento del logaritmo del salario dado diferentes niveles de edad de las mujeres, mientras la Figura 5 muesta este mismo comportamiento pero para los hombres. Si superpusieramos ambos graficos podrimos ver como los salarios por hora a diferentes niveles de edad de los hombres se encuentran por encimas de los salarios por hora de las muejres a diferentes niveles de edad.

Figura 4. Bootstrap para cambios en el salario los salarios por hora dada la edad de los hombres

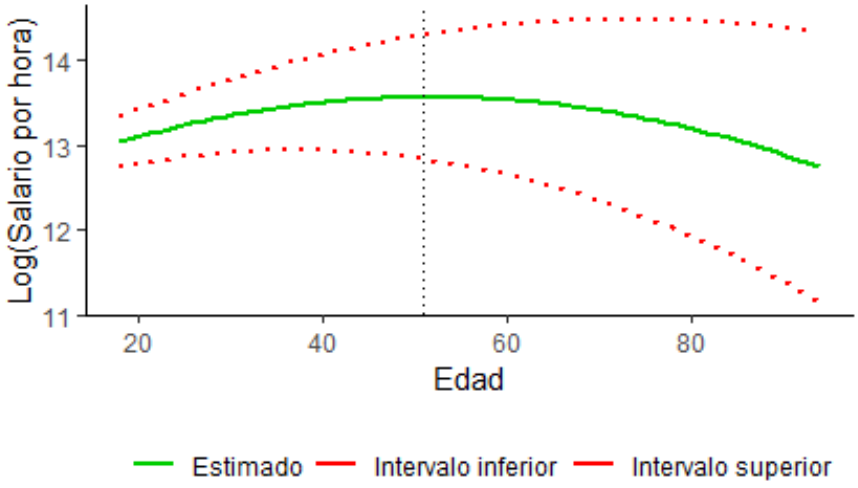
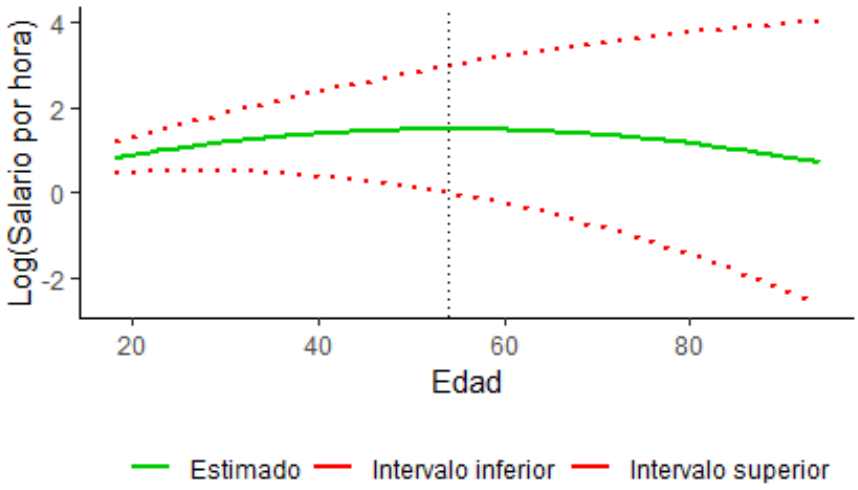


Figura 5. Bootstrap para estimar la predicción de los salarios por hora dada la edad de los hombres



4.3. Prediccción ganancias

Para esta sección se corrieron 9 modelos que varían en los predictores elegidos, forma funcional incluidas y modelo de ajuste. Para la comparación de los diferentes modelos, se compararon los RMSE, buscando llegar al mínimo valor posible. Se optó por tomar

el RMSE ya que es simple de entender e interpretar, siendo el promedio del error de las predicciones. Adicionalmente, penaliza de forma más fuerte los errores más grandes relativo a los más pequeños. Por último, es una de las medidas más populares para medir bondad de ajuste, por lo que hay bastante material de apoyo disponible para hacer mejores análisis de los modelos evaluados.

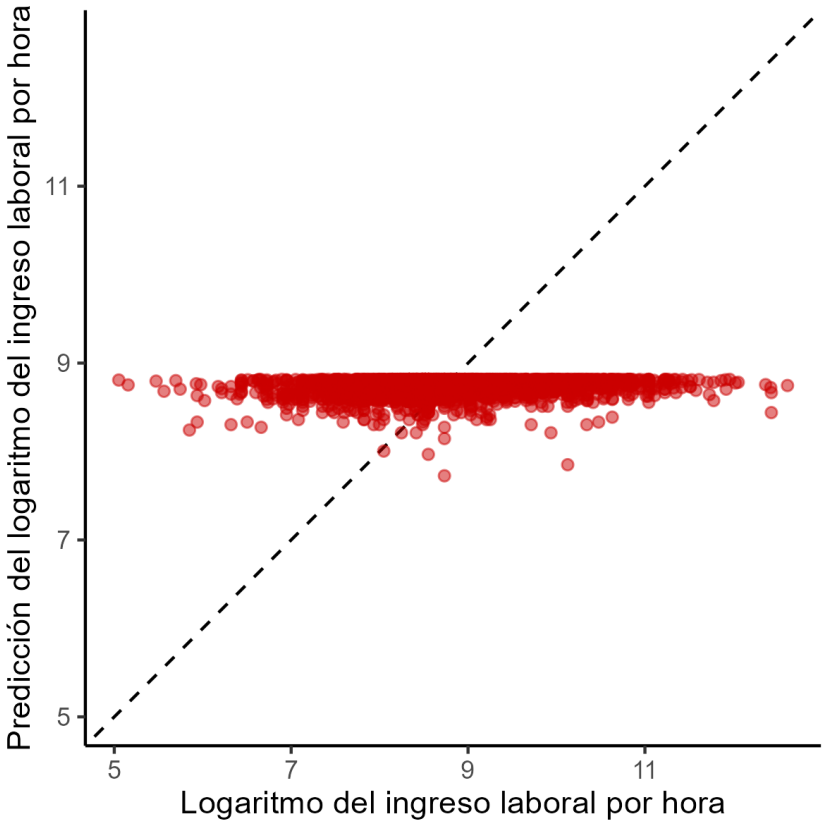
Teniendo en cuenta lo anterior, a continuación, se hará una descripción general de los modelos evaluados y un análisis del que obtuvo el RMSE más bajo.

El primer modelo evaluado, fue una regresión lineal de la forma:

$$log(ingresolaboral_{hora}) = \beta_{i1} + \beta_{i2} + Edad + \beta_3 Edad^{i2} + \epsilon i \tag{5}$$

El modelo obtuvo un r cuadrado ajustado del 1 % el tercero más bajo de los 9 modelos evaluados. En materia de métricas de ajuste fuera de la muestra, el modelo obtuvo un RMSE de 0,787 el tercero más alto de los modelos evaluados. Esto se puede apreciar gráficamente en la Figura 1.

Figura 6. Modelo 1:Precisión de las predicciones



Fuente: Cálculos propios con datos del DANE 2018.

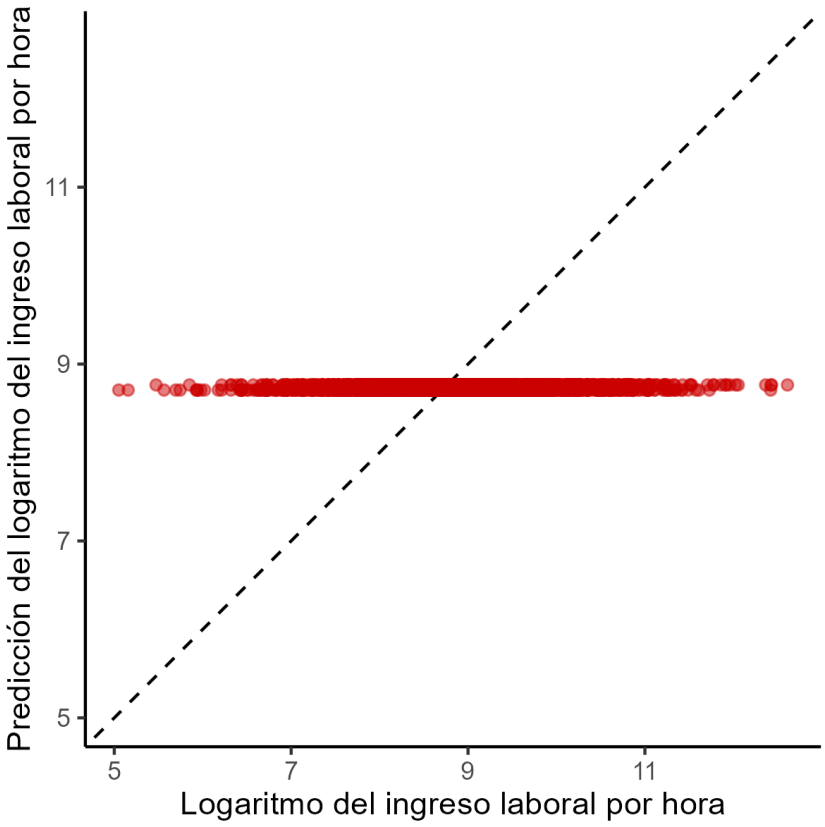
Prediciones más acertadas mostrarían una relación positiva entre los valores observados y los valores predichos, es decir, más alineada a la diagonal demarcada. Sin embargo, como se puede observar,los valores predichos no parecen mostrar relación alguna con los valores observados. Lo anterior, sumado a la poca variación explicada de los predictores medida por el r cuadrado ajustado, nos permite afirmar que este modelo no resulta deseable para hacer buenas predicciones del ingreso laboral por hora.

Con respecto al segundo modelo, se corrio una regresión lineal con la siquiente especificación:

$$log(ingresolaboral_{hora}) = \beta_{i1} + \beta_{i2} Sexo + \epsilon i \tag{6}$$

Como era de esperarse, al ser un modelo que solo cuenta con un predictor, fue el que menor desempeño tuvo tanto en las métricas dentro de muestra como en las fuera de muestra. Según el R cuadrado ajustado, el modelo explica el 0,1 % de la variación de los datos. Alineado con lo anterior, esto se ve reflejado en el RMSE más alto de todos los modelos evaluados con 0,8 y con una nula relación entre los valores observados y los valores predichos por nuestro modelo (ver Figura 2)

Figura 7. Modelo 2: Precisión de las predicciones



Fuente: Cálculos propios con datos del DANE 2018.

Para el tercer modelo, se corrió una especificación más robusta que la anterior, incluyendo controles de máximo nivel educativo, tipo de trabajador, un indicador de si es formal, el tamaño de la firma y la ocupación u oficio de la persona. Adicionalmente, se incluyeron las variables de edad y edad al cuadrado de la primera especificación.

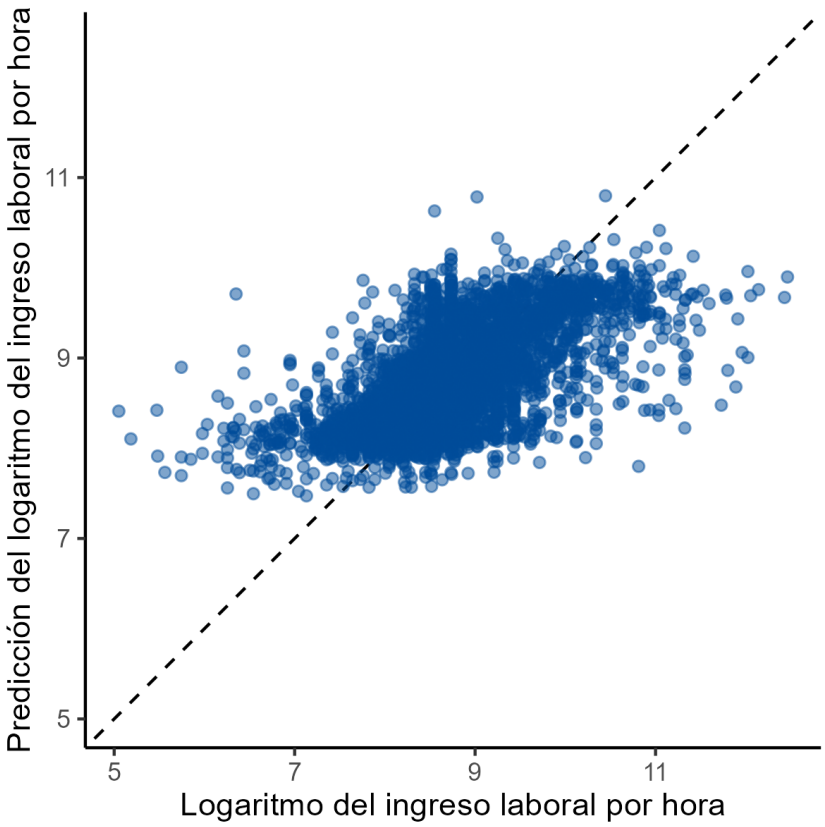
La elección de los controles se sustenta en una intuición económica detrás de las dinámicas socioeconómicas que afectan el ingreso laboral y en literatura que explora esta relación. De esta manera, se espera que personas con mayores niveles educativos tengan acceso a mejores ingreso laborals, por lo que se incluye para controlar por esta diferencia. En materia de formalidad, aunque hay evidencia de que las distribuciones de ingreso no distan mucho entre mercados, puede actuar como una proxy de cuenta propismo y pobreza. Con respecto al tipo de empleo, es claro que la profesión y por ende el sector en el que se desenvuelve el empleado tiene una relación directa con el ingreso recibido, por lo que la variable de oficio se incluye para controlar por esa variación.

Por su último, hay evidencia que demuestra la estrecha relación que hay entre las características de la firma y las condiciones de pobreza de sus empleados. Se espera que firmas más pequeñas o unipersonales cuenten con ingreso laborals más bajos que empresas más grandes y consolidadas en el mercado.

$$\begin{aligned} log(ingresolaboral_{hora}) &= \beta_{i1} + \beta_{i2}Sexo + \beta_{i3}MaxEduc + \beta_{i4}TipoTrabajador + \\ &= \beta_{i5}Formal + \beta_{i6}TipoOficio \\ &= \beta_{i7}TamañoFirma + \epsilon i \end{aligned} \tag{7}$$

Esta especificación mejoró significativamente el R cuadrado ajustado del modelo, llegando al 40 %. Esto se ve reflejado adicionalmente en un mejor desempeño fuera de muera, reduciendo el RMSE a 0,62. En la Figura 3, se puede observar que, relativo a los otros dos modelos anteriores, ya se empieza a ver una relación positiva mucho más clara, alineándose más con la diagonal de la gráfica.

Figura 8. Modelo 3: Precisión de las predicciones



Fuente: Cálculos propios con datos del DANE 2018.

Para el tercer modelo se realiza una especificación alterna, haciendo uso del teorema Frisch–Waugh–Lovell (FWL). Para ello, se utiliza la versión de Mostly Harmless Econometrics en la que se residualizan solo los predictores y se regresan con la variable dependiente original.

$$log(ingresolaboral_{hora}) = \beta_{i1} + \beta_{i2} Resid.Sexo + \epsilon i \tag{8}$$

Para este caso, se residualiza la variable sexo, regresándola con el resto de controles incluidos en la especificación inicial. Como se puede ver en la tabla 2, se obtienen los mismos coeficientes que en la estimación original, sin embargo, las métricas de ajuste dentro de la

Cuadro 8. Estimación por MCO y su variación por el teorema FWL.

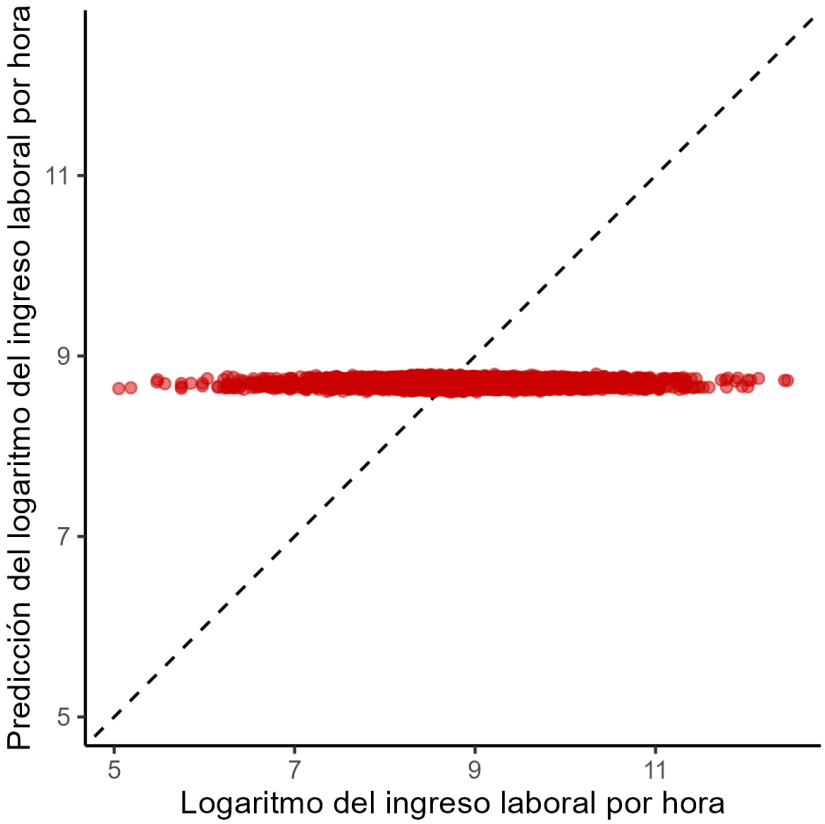
	Variable Dependiente	
	Log.Hourly.Wage	Log.Hourly.Wage
	(1)	(2)
Sexo	1.1***	
	(0.0142)	
Edad	1.03***	
	(0.0026)	
Edad2	1.00***	
	(0.0000)	
Residuo.Sexo		1.10***
		(0.0184)
Constante	7,667	6,000
Observaciones	11,577	11,577
R2	0.0407	0.0023
R2 Ajustado	0.0402	0.0022

Nota: i) Errores estándar en paréntesis *p <0.1; **p <0.05; ***p <0.01.
ii) En la tabla no se incluyen los valores de los controles especificados arriba.
iii) Los coeficientes se encuentran exponenciados.

muestra caen significativamente.

El R cuadrado ajustado pasa de 40 % a 0.2 % dado que ahora volvemos a una especificación de regresión simple. La poca varianza que explica nuestro segundo modelo se ve reflejado en la calidad de las predicciones fuera de muestra. Como se puede ver en la Figura 4, relativo a lo encontrado en el modelo 3 por regresión lineal, no se muestra relación alguna entre los valores estimados y los observados. Este resultado se complementa con un RMSE fuera de muestra de 0,791, el más alto de todas las especificaciones analizadas. Estos resultados nos sugieren que este tipo de variación no resulta deseable en un contexto de predicción.

Figura 9. Modelo 3 - FWL: Precisión de las predicciones



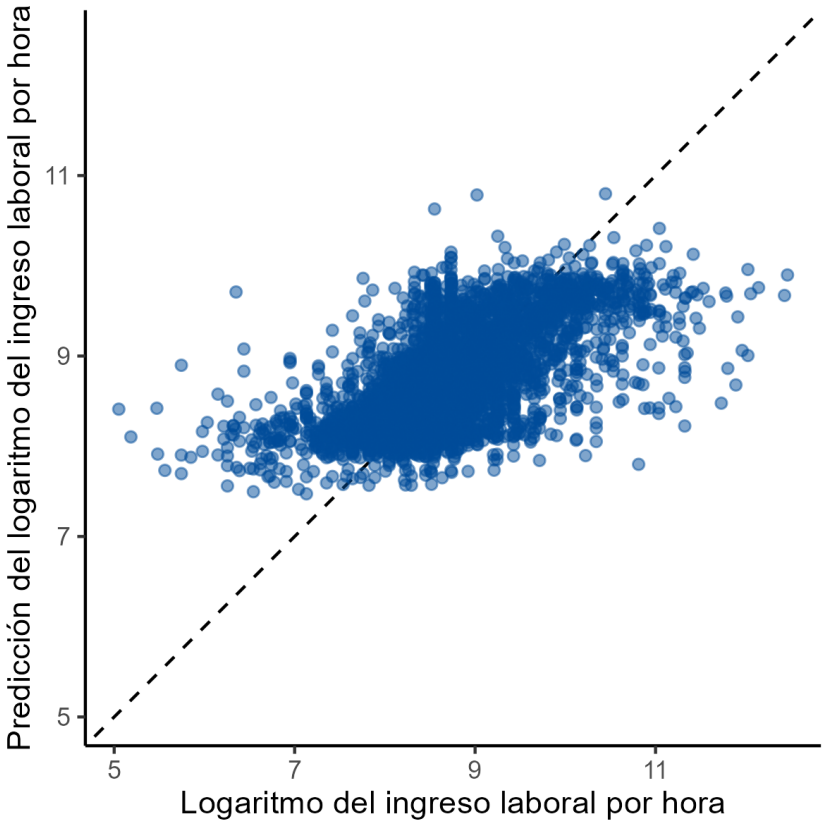
Fuente: Cálculos propios con datos del DANE 2018.

Para el cuarto modelo, se mantienen los predictores del tercer modelo y se agrega el estrato, el régimen de salud al que pertenece y el parentezco con el jefe/fa del hogar. La inclusión de estas variables se sustenta en la intuición de la relación entre la capacidad de gasto de una persona y su lugar de residencia, así como el régimen de salud al que pertenece, por definición, se espera que mayores ingresos laborales impliquen una mayor probabilidad de estar en el régimen contributivo y de residir en estratos más altos. Por otro lado, el parentezco se incluye bajo el supuesto de que, una persona cabeza de hogar, puede tener mayores responsabilidades en los gastos del hogar y por ende mayores necesidades de generar un ingreso para cubrirlas.

$$
\begin{aligned}
log(ingresolaboral_{hora}) &= \beta_{i1} + \beta_{i2}Sexo + \beta_{i3}MaxEduc + \beta_{i4}TipoTrabajador + \beta_{i5}Formal + \\
&= \beta_{i6}TipoOficio + \beta_{i7}TamañoFirma\beta_{i8}Estrato + \\
&= \beta_{i9}RegimenSalud + \beta_{i10}Parentezco + \epsilon i
\end{aligned}
\tag{9}
$$

Bajo esta especificación el R cuadrado ajustado sube a 44 %, lo cual implica un aumento en la variación explicada con respecto al mejor modelo hasta ahora (especificación 3). En materia de desempeño de predicción fuera demuestra, el RMSE baja de 0.622 a 0.593, lo cual se ve reflejado en una mayor relación posivita entre las prediciones y los valores observados (ver Figura 5)

Figura 10. Modelo 4 - MCO: Precisión de las predicciones



Fuente: Cálculos propios con datos del DANE 2018.

El modelo 5 explora alguna interacciones entre variables que pueden verse relacionadas. De esta manera, el modelo 5 incluye todas las variables del 4 más unas interacciones entre la edad y el sexo, la edad y el estrato y el sexo y el estrato. Estas interacciones se incluyen con el objetivo de capturar ciertas interrelaciones entre predictores que pueden estar potenciando el efecto final sobre el ingreso laboral. Inicialmente, se estima que no solo el sexo, sino su relación con la edad pueden afectar de manera distinta la brecha de género en términos de ingreso laboral. Igualmente, pueden haber variaciones en el ingreso laboral al interior de cada estrato en función del sexo y la edad que no están siendo capturadas actualmente.

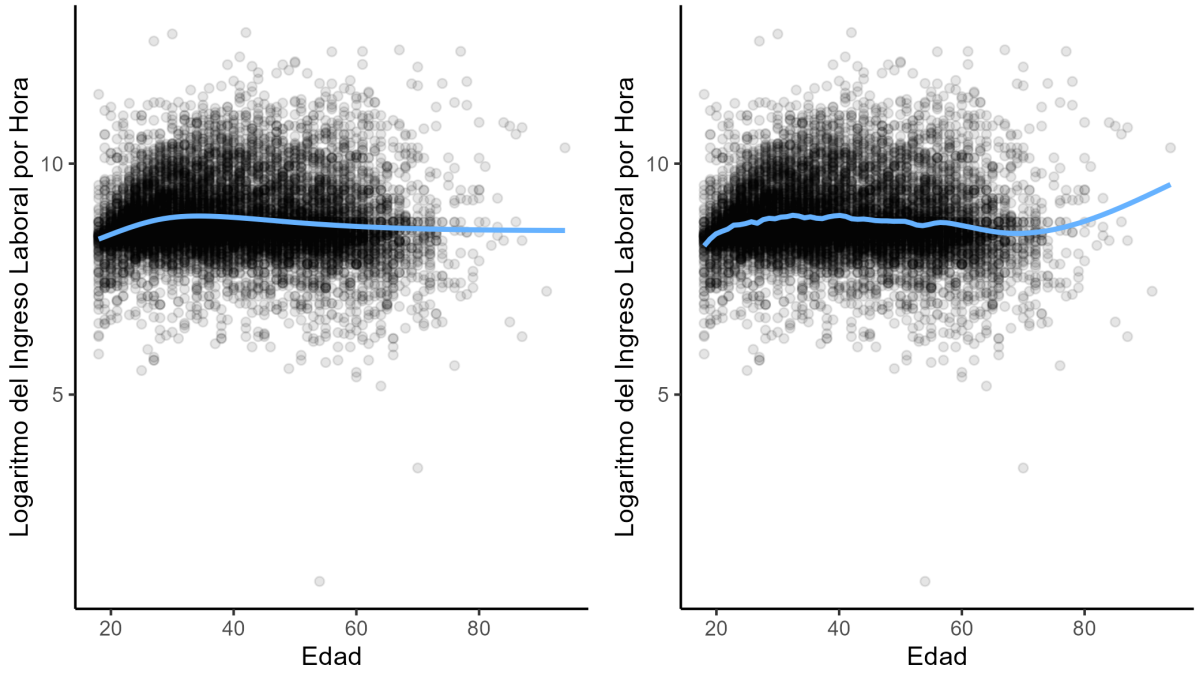
$$
\begin{aligned}
log(ingresolaboral_{hora}) &= \beta_{i1} + \beta_{i2}Sexo + \beta_{i3}MaxEduc + \beta_{i4}TipoTrabajador + \beta_{i5}Formal + \\
&= \beta_{i6}TipoOficio + \beta_{i7}TamañoFirma\beta_{i8}Estrato + \\
&= \beta_{i9}RegimenSalud + \beta_{i10}Parentezco + \beta_{i11}Sexo * Age + \\
&= \beta_{i12}Sexo * Estrato + \beta_{i13}Age * Estrato + \epsilon i
\end{aligned}
\tag{10}
$$

En materia de métricas al interior de la muestra, no se observa una mejora en el R cuadrado ajustado. Sin embargo, cuando se revisa el RMSE, se observa una leve caída, pasando de 0.593 a 0.592, probablemente no significativa. Por dicha razón, no se muestra la gráfica para este modelo.

El modelo 6 no se incluyen las interacciones pasadas y se exploran interacciones adicionales. En este modelo se exploran con mayor detalle las complementariedades potenciales que pueden haber entre el sexo, la posición laboral, la profesión y la educación, entendiendo las potenciales variaciones que pueden haber al interior de cada una de dichas categorías sobre el ingreso laboral. Sin embargo, los efectos encontrados no varían significativamente con respecto a modelos pasados, e incluso el RMSE aumenta a 0.596.

Para el modelo 7 no se incluyen ninguna de las interacciones exploradas anteriormente debido a su nulo efecto o efecto negativo sobre el RMSE. En su lugar se explora la introducción de polinomios de la edad. Lo anterior con el objetivo de capturar relaciones no lineales con el logaritmo del ingreso laboral por hora. Sin embargo, se descarta la idea debido a la poca variación de los datos. Como se puede ver en la Figura 6, la diferencia del ajuste entre un polinomio de grado 4 y otro de grado 30 no es muy alta, probablemente por la transformación logarítmica del ingreso laboral por hora.

Figura 11. Ajuste de polinomios grado 4 y grado 30 al logaritmo del ingreso laboral por hora

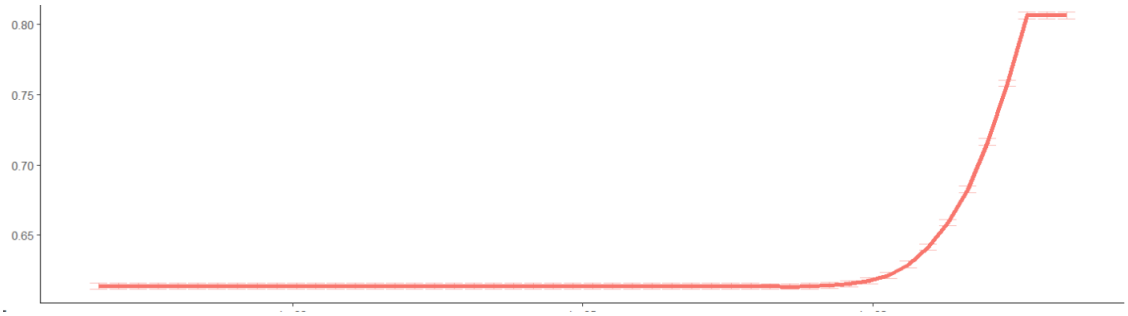


Fuente: Cálculos propios con datos del DANE 2018.

Por último se corre un modelo diferente con la especificación que arrojó el RMSE más bajo (Modelo 5). Específicamente se corre un Lasso para ver si nuestro modelo se ve beneficiado por un proceso de reducción de varianza y selección de variables.

Para ello, se normalizan las variables lineales dado que el Lasso es sensible a la escala de las mismas y se hace un proceso de resampleo por bootstrap de la base de entrenamiento para hallar el lambda óptimo que minimiza el RMSE del modelo. Sin embargo, al examinar los resultados, no se encuentra una mejora significativa en el RMSE al aumentar el lambda, por el contrario, este aumenta de manera acelerada después de 0.01 (ver Figura 7).

Figura 12. Lambdas y RMSE de los modelos ajustados por bootstrap



Fuente: Cálculos propios con datos del DANE 2018.

Lo anterior sugiere una restricción casi que no vinculante en el proceso de minimización de los residuales al cuadrado y por ende, un modelo muy similar al de MCO. De todas maneras, se corre el modelo con el lambda óptimo para revisar sus métricas dentro y fuera de muestra. Desafortunadamente, se encuentra que, a pesar de que la restricción no resulta muy limitante, sí afectó el RMSE, pasando de 0.592 en el modelo 5 a 0.674.

Revisando la matríz de lambdas evaluados en el proceso de optimización, se identifica que los valores tomados no evaluaron el 0 como posibilidad, por lo que puede ser la razón para que el proceso no arrojara una regresión tipo MCO como modelo óptimo.

Teniendo en cuenta el análisis anterior, el modelo 5 fue el modelo que obtuvo el menor RMSE de los 9 modelos analizados. Lo anterior se puede justificar principalmente por la inclusión de variables relevantes que ayudaron a explicar una mayor parte de la variación del modelo y por ende a ajustarlo mejor. Adicionalmente, a pesar de que se probaron modelos más complejos con polinomios, variables adicionales e interacciones, estas variables no aportaron a explicar una mayor variación y, por el contrario, generaron más ruido, lo que se vió reflejado en aumento paulatinos en el RMSE. Por último, se intentó probar un modelo Lasso que nos permitiera hacer selección de variables frente a la especificación consolidada hasta el momento con el mejor desempeño fuera de muestra. Sin embargo, la optimización del parámetro lambda para dicho modelo resultó ser muy cercana a cero,

sugiriendo al modelo 5 como mejor modelo.

En este orden de ideas, para el punto siguiente solo se volverá a correr el modelo 5 utilizando el método de validación LOOCV.

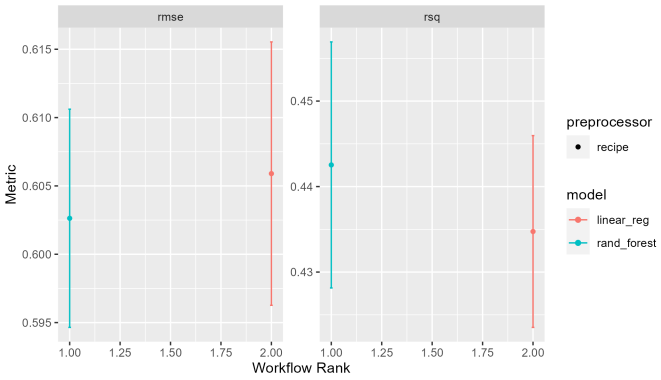
4.4. Validación Cruzada por LOOCV y K- Fold

Para esta sección se decide hacer la validación cruzada por medio de la metodología K-Fold, no solo porque la literatura la aconseja por dar resultados muy parecidos al LOOCV con una demanda computacional mucho más baja, sino por el tiempo que implica completar la especificación de forma manual debido a la cantidad de niveles que tienen las categorías de control. Sin embargo, en el código del taller se encuentra adjunta la forma como se llevaría a cabo dicha metodología.

En este orden de ideas, en aras de la costo eficiencia se correrá el modelo 5 aplicando la metodología K-Fold de validación cruzada a la base de entrenamiento, la cual, en lugar de representar el 70 % de la base original, se dividió por medio de una estratificación de la variable dependiente. Esta estratificación permite mantener una distribución similar de la variable dependiente entre la base entrenamiento y evaluación.

Una vez especificadas las bases de datos con las que se va a ajustar y evaluar el modelo, se define la especificación usada en el modelo 5 y se ajusta en las 10 submuestras generadas por el proceso K-Fold. El ajuste se hace por medio de un modelo de MCO y otro de Bosques Aleatorios para fines comparativos. En la Figura 8 se puede el promedio del RMSE con los intervalos de confianza, derivado de la metodología de validación cruzada por K-Fold = 10 para un modelo de MCO y uno de Bosques Aleatorios.

Figura 13. Promedio RMSE e intervalos de confianza



Fuente: Cálculos propios con datos del DANE 2018.

A pesar de que el promedio del modelo de bosques aleatorios dio menor, se toma la especificación del MCO, el cual arrojó un RMSE promedio de 0.606. Dicho valor resulta, en promedio más alto que el observado en el modelo 5 (0.592). Sin embargo, estas estimaciones, resultan ser más cercana al valor real del RMSE, puesto que reduce la varianza que se da por un cambio en los datos con los que se entrena el modelo. Por último, aunque esta estimación no reduce el sesgo tanto como en teoría se esperaría con una metodología LOOCV, se ha demostrado que las estimaciones del error de evaluación por medio de K-Fold no sufren ni por alto sesgo (partición inicial 70/30) ni por alta varianza (potencialmente con LOOCV por las n muestras casi identicas que se promedian).

5. Nota: Github

En el siguiente link https://github.com/ra-carrillo/Big_Data_Problem_Set1 se encuentran el acceso al repositorio del Problem set 1, donde se almacena tanto el documento en PDF, como el código y los gráficos.

Referencias

[1] DANE(2020)., *Brecha salarial de géenro en Colombial*

[2] Deming, D(2022)., *Four facts about human capital*, Nberworking paper series.

[3] Internal Revenue Service(2016)., *Tax Gap Estimates for Tax Years 2008-2010.*

[4] Mincer, J.,(1974)., *Schooling, experience, and earning.*Human Behavior Social Institutions.

[5] Nova, D. , Bermúdez, R. , Gordillo, C. y Sanchez, D, *Algoritmo para la conformación del ingreso per cápita para la medicion de pobreza.*