Problem Set 1

Integrantes:
Juan Pablo Bermudez Cespedes
Lina Maria Bautista Salcedo
Esteban Mesa Gomez
Pharad Sebastian Escobar Carreño

Universidad de los Andes
BDYML
Bogotá, Colombia
18 de Septiembre 2023

1. Link Repositorio

Para el presente taller se utilizó el siguiente repositorio Predicting_Income_BDML, el cual contiene codigos completamente reproducibles en la carpeta scripts. Las bases de datos se encuentran en la carpeta stores y el documento final en la carpeta document.

2. Introducción

Los ingresos individuales y sus diferentes componentes son importantes para determinar el nivel de riesgo de una persona en caer en pobreza, los precios de la mano de obra en el mercado laboral, la determinación de precios en el agregado de la economía y entre otras mediciones económicas. En Colombia, la estimación de los ingresos de los individuos y hogares es utilizado para identificar la proporción de la población que está en pobreza monetaria, la estimación de los grupos de ingresos para identificar las clases sociales y entre otras estadísticas relacionadas con el mercado laboral. Sin embargo, a pesar de la sofisticación de las mediciones en una muestra específica, diferentes instrumentos de política pública para ser efectivos, necesitan predecir el comportamiento de los ingresos al conjunto de la población, especialmente en políticas sociales focalizadas o políticas de recaudo tributario.

En ese sentido, en este ejercicio de aplicación esperamos desarrollar modelos de regresión lineal con el objetivo de estimar : (i) la relación lineal y cuadrática de la edad con los ingresos salariales medidos por hora, (ii) estimar la edad pico y sus respectivos intervalos de confianza utilizando el método bootstrap, (ii) estimar la brecha salarial de género condicional y no condicional utilizando la regresión por MCO y teorema FWL, (iv) estimar la edad pico para hombres y mujeres con sus respectivos intervalos de confianza utilizando el método bootstrap y (iv) estimar el poder predictivo de los modelos de de ingresos individuales salariales por hora. Para esto utilizaremos un subconjunto de 32177 observaciones de la GEIH para la ciudad de Bogotá que fueron utilizados en la elaboración del “Informe de medición de la Pobreza Monetaria y Desigualdad, 2018”.

Los resultados de las estimaciones nos permiten evidenciar que (i) El incremento de un año de vida, en promedio, está asociado a un incremento del 5,83 % del salario por hora y a su vez, la edad tiene una relación decreciente en un punto denominado edad pico, (ii) la brecha condicional de género es de 10, 21 % a favor de los hombres y la brecha no condicional es de 4,71 % a favor de los hombres, (iv) la edad pico de las mujeres se encuentra entre los 42,49 años y los 47,11 años; y la edad pico de los hombres está entre los 44,88 años y los 50,56 años y (v) el modelo con 20 predictores tiene el menor MSE y puede predecir con suficiencia los ingresos laborales por hora.

3. Datos

El presente caso práctico utilizará 32177 observaciones de la Gran Encuesta Integrada de Hogares (GEIH) 2018 filtrados para la ciudad de Bogotá. Los datos proporcionan información relevante sobre las características del individuo, la vivienda, la inserción en el mercado laboral, los componentes de los ingresos y el total de ingresos laborales. Los datos provienen de una muestra probabilística, estratificada y multietapa por unidades primarias y secundarias de muestreo. La encuesta es aplicada a personas de 10 años y más que trabajen o estén buscando trabajo. Los tamaños de la muestra se calculan con una precisión deseada de la variable de tasa de desempleo del 10 % y un error estándar relativo del 5

3.1. Proceso de adquisición y procesamiento de los datos

Los datos necesarios para el presente estudio se encuentran disponibles en el siguiente enlace (https://ignaciomsarmiento.github.io/GEIH2018_sample/). Allí, existen 10 chunks (trozos) de datos con 3218 observaciones cada uno, salvo los trozos 1, 9 y 10 en donde había una observación menos; es decir, 3217 observaciones. De manera que, al finalizar el ejercicio de recolección se tendrían 32177

observaciones.

Para poder llevar a cabo dicha tarea, se usó el método del scraping. Esta metodología se amoldaba perfectamente a nuestras necesidades pues, mediante código en R, era posible automatizar la recolección de datos de cada uno de los trozos de la página web (mencionada previamente). Así, al inspeccionar la página web, se encontró que estos elementos estaban codificados como 'li a', por lo que para acceder a ellos utilizamos el comando html_elements('li a'), generando una lista con 11 elementos. Sin embargo, como previamente fue mencionado, solo había 10 URLs para los trozos de los datos, entonces, estamos capturando un elemento ajeno a los deseados. Para poder filtrar este objeto indeseado, fue necesario utilizar la función grepl('Data chunk'), de la que esperamos que tome los elementos de la lista que coincidan con las cadenas de carácteres "Data chunk", ya que nuestros links de interés están nombrados de esta manera. Ahora bien, al observar detalladamente nuestra lista, encontramos que los objetos que filtramos y que son de nuestro interés, tienen atributo href, por lo que al usar el comando html_attr ('href') accedemos a 10 objetos de tipo texto que al juntarlos con nuestro link principal a través del comendo paste0(), nos permiten acceder a cada uno de los links que contienen los chunks de los datos.

Seguidamente, explorando cada uno de los links, en busca de elementos de tipo tabla con el comando <html_table()>, no fue posible encontrar una forma de extraer la tabla de datos de cada link de forma automática. Así, fue necesario realizar la búsqueda del Xpath de la tabla encontrada al inspeccionar cada página web a través del siguiente comando: html_nodes(xpath='/html/body/div/div/div[2]/div'). Este procedimiento en un primer momento derivó en diferentes problemáticas. Sin embargo, posteriormente, se encontró otro atributo en el objeto buscado por lo cual se usó html_attr("w3-include-html"))). De esta manera, se logró acceder a una cadena de caracteres que contenía parte del URL que debíamos usar para acceder a las tablas. En consecuencia, realizamos este proceso de forma iterativa, de tal manera que tuviéramos una lista con 10 cadenas de caracteres (una para cada uno de los chunks)

Aquí encontramos, que de la misma forma que el paso anterior, al unir la url inicial con cada uno de los objetos de nuestra lista, accedemos a 10 páginas nuevas, en las que al utilizar el comando <html_table()>era posible extraer cada tabla de forma completa (proceso que no fue posible realizar usando en los links que encontramos previamente). Posteriormente, guardamos cada una de las tablas en un objeto de una lista y procedemos a unirlas en un único data frame a través de las filas. A este último dataframe lo llamamos base.datos.original, como se esperaba, se construyó una base de datos con 32177 observaciones y 178 variables. Finalmente, guardamos esta base de datos tipo .Rdata en la carpeta Stores de nuestro repositorio de GitHub.

El proceso de limpieza de datos puede ser consultado a detalle en el script del repositorio llamado "02_Manejo_datos". Allí, se cargó la base de datos y se inició filtrando por individuos mayores de edad. Dicho filtro redujó la muestra de observaciones de 32177 a 24054.

Posteriormente, se procede a limpiar la base de datos de variables innecesarias para nuestro estudio. La selección de variables que verá a continuación responde a una consulta de bibliografía y análisis descriptivos de los datos obtenidos.

Variables de interés:

- **Age:** (discreta): provee una proxy de la experiencia laboral y da amplia información sobre las características del individuo
- **Urban=clase**: índica si el individuo se encuentra en zona rural (0) o urbana (1), esto se debe tener en cuenta al evaluar los efectos del salario. Sin embargo, dado que estamos estudiando a la población Bogotana (urbana) puede que no tenga variabilidad significativa
- **College**: dummy que toma el valor de 1 para aquellos individuos con educación terciaría. Esta variable es fundamental para estudiar los retornos crecientes (esperados) de la educación en el

salario.

- **cotPension**: dummy que toma valor de 1 en el caso de estar cotizando pensión. Sirve para evaluar los efectos que tiene la formalidad en el trabajo (funciona como una proxy).
- **cuentaPropia**: captura a aquellos individuos que se emplean de forma propia.
- depto: departamento, ya que todas las observaciones son de Bogotá, esta variable se puede eliminar.
- **directorio**: forma en la que el DANE nombra a cada uno de los hogares encuestados con el fin de controlar diferentes variables sin tener que comprometer información privada de estos.
- **dsi**: controlamos a partir de una dummy de desempleo. Dado que queremos utilizar observaciones del salario, todos aquellos individuos desempleados deben ser eliminados de la muestra.
- **estrato=estrato1**: variable numérica categórica del estrato (1-6). Nos permite capturar la diferencia o similitud salarial entre individuos que viven en el mismo nivel de estrato socioeconómico.
- **formal**: dummy para determinar si el individuo tiene un trabajo formal o no, esta variable puede capturar las diferencias en términos salariales de pagar todas las prestaciones de ley como salud y pensión a comparación de aquellos que no lo hacen.
- **inac**: controlamos a partir de una dummy de inactividad. Dado que queremos utilizar observaciones del salario, todos aquellos individuos inactivos deben ser eliminados de la muestra.
- **ingtot**: ingreso total de los individuos. Es importante observar la correlación de estavariabe con el salario, dado que individuos que reciban ingresos por otras actividades (e.g. rentas), tienen menos incentivos para trabajar. Sin embargo, es importante denotar la alta correlación de esta variable con nuestra variable de interés.
- **hoursworkUsual**: para observar la variable mensual de salario es necesario contar con el número de horas trabajadas y controlar.
- **maxEducLevel**: variable categórica del mayor nivel de educación alcanzado. Toma valores de 1 a 7 incluyendo categorías de ningún tipo de educación, preescolar, primaria incompleta, primaria completa, bachillerato incompleto, bachillerato completo y educación terciaria. Esta variable permite controlar los efectos en el salario de un mayor nivel educativo.
- **oficio**: variable cualitativa con 99 categorías en las que se identifica la actividad en la cual se desempeña el individuo. Este control puede proveer bastante información, ya que se esperaría que trabajadores que realizan actividades similares obtengan salarios similares
- **orden**: también es una variable indicativa que el DANE genera (igual que el directorio), para identificar los hogares encuestados y diferenciarlos.
- **salud=p6090**: dummy, responde a la pregunta "¿ Está afiliado, es cotizante o es beneficiario de alguna entidad de seguridad social en salud?. 1 sí, 2 no y 9 no sabe. Esta variable también funciona como control del desempeño de los empleados, dado que contar con un sistema de salud puede mejorar las condiciones físicas de los empleados haciéndolos más productivos. De tal manera, ciudadanos que controlen sus problemas de salud pueden trabajar más horas y ser más productivos, recibiendo mejores salarios.
- **seguridadsocial=p6100**: Esta variable captura el tipo de régimen de seguridad social al que el individuo se encuentra afiliado. Denotado por contributivo (eps), Especial y subsidiado (eps-s).
- **sex**: una dummy de sexo permite capturar las diferencias salariales entre hombres y mujeres.

- **microEmpresa**: Dummy que determina si el empleado trabaja en una micro empresa (con menos de 5 empleados). Esperamos que empresas más grandes puedan proveer salarios más altos a sus empleados, por lo que debería encontrarse una relación negativa de los salarios y las microempresas.
- **sizeFirm**: Variable categórica que determina el tamaño de la empresa en la que el individuo se emplea, de la misma manera que la variable anterior, entre más alta sea la categoría (mpas grande la empresa), esperamos que los individuos cuenten con mejores salarios.
- **y_salary_m**: Salario nominal mensual incluyendo comisiones y propinas, es una de nuestras variables de interés para analizar el comportamiento de los salarios.
- **y_salary_m_hu**: Salario real por hora incluyendo comisiones y propinas, es nuestras variable de interés principal, ya que nos permite estudiar el margen intensivo del empleo (por horas).
- **sub.alimentación, sub.transporte, sub.familiar, sub.educativo** : Estas 4 variables son dummys que capturan si los individuos recibieron subsidios de cada uno de estos tipos en el último mes. Estas variables son seleccionadas para poder comprobar la hpótesis de que personas que reciben transferencias son menos propensos a trabajar para suplir estas necesidades.

Al analizar depto y urban, encontramos que no tienen una cantidad de observaciones significativas por lo que preferimos eliminarlas. Por otro lado, la variable salud tiene 3 labels: 1, 2 y 9. Donde 9 indica: no sabe, no informa, sin embargo menos del 0.001 % de los datos entra en esta categoría por lo que se decide imputarlos por la media. En este sentido, todas las observaciones tomarían valores de 1 y 2, sin embargo, para hacer un análisis más adecuado como dummies, hacemos que todas las observaciones con valores 2 se igualen a 0, tal que 1 significa estar afiliado a salud y 0 no estarlo.

De la misma forma que se realizó con la variable categórica anterior (salud), para las variables de subsidio se interpolan por la moda aquellas observaciones que toman la clasificación de "no sabe, no responde", sabiendo que no representan un número tan significativo en la muestra. Igual que se realizó previamente, las observaciones que tomaban el valor de 2 (no subsidio) las hacemos iguales a 0. Construyendo dummies con valores de 0 y 1 en donde 1 representa el haber recibido un subsidio, sea familiar, educativo, de transporte o de alimentación.

Posteriormente, se analizan los datos faltantes o NA de la base de datos, por lo que se evalúa qué porcentaje de estas variables presenta faltantes:

Variable	Porcentaje de NA
cotPension	0.32
formal	0.32
hoursworkUsual	0.32
oficio	0.32
seguridadsocial=p6100	0.1
microEmpresa	0.32
sizeFirm	0.32
y_salary_m	0.59
y_salary_m_hu	0.59
sub.alimentación	0.57
sub.transporte	0.57
sub.familiar	0.57
sub.educativo	0.57

Nota: Las variables que no se muestran no tienen faltantes (NA)

Cuadro 1: Porcentaje de observaciones faltantes por cada variable

Teniendo en cuenta los porcentajes previamente mostrados, decidimos eliminar las observaciones que no cuentan datos para <cotPension><formal><hoursWorkUsual><oficio><microEmpresa>y <sizeFirm>. Con la eliminación de estos NA no se perdió ningún dato del salario pero sí se aseguró que hubiera suficiente información en las variables explicativas.

Podemos observar que al realizar esta transformación, nuestra nueva base de datos cuenta con 16397 observaciones y que ahora las únicas variables que tienen valores faltantes son: (Cuadro 2)

Variable	Porcentaje de NA
y_salary_m	0.4
y_salary_m_hu	0.4
sub.alimentación	0.36
sub.transporte	0.36
sub.familiar	0.36
sub.educativo	0.36

Nota: Las variables que no se muestran no tienen faltantes (NA)

Cuadro 2: Porcentaje de observaciones faltantes por cada variable

Sin embargo, en esta última eliminación de NA, las variables dsi (desempleo) e inac (inactivos), quedaron obsoletas tomando únicamente valores de 0. Lo anterior nos indica que nuestra muestra ahora solo incluye a nuestros individuos de interés: personas ocupadas y asalariadas.

Ahora bien, para el siguiente paso, decidimos crear dos bases de datos diferentes:

1. Una base de datos eliminando todos los NA: Al realizar este procedimiento filtrando por los valores del salario, finalmente llegamos a una base de datos con 9785 observaciones.
2. Una base de datos imputando los NA con la media para las variables numéricas y con la moda para las categóricas: Aquí logramos mantener 16397, sin embargo, es importante tener en cuenta que para la variable de salarios casi que el 40 % de los datos ha sido imputado.

La imputación de esta gran cantidad de datos nos puede llevar a tener conclusiones erróneas, por lo que en el desarrollo de este trabajo se utilizará preferentemente la base de datos que eliminó los NA, pero se llevarán a cabo comparaciones con la otra base de datos imputada para observar si existen cambios significativos entre ambas estimaciones. Después de imputar por las medias, devolvemos las variables categóricas a tipo factor, para poder tratarlas como tal en las próximas regresiones. Finalmente, dado que nuestros modelos en adelante utilizan el salario como logaritmo, agregamos una nueva columna a nuestras dos bases de datos que calcula el logaritmo de esta variable, para no tener que acudir a su transformación cada vez que se desee correr algún modelo.

3.2. Estadística descriptiva

Para comenzar, proponemos una visualización de las correlaciones entre todas las variables de nuestra base de datos. De esta manera, es posible identificar relaciones, esperadas y no esperadas según la literatura, y determinar variables de interés. Esto puede apreciarse en la siguiente gráfica (Figura 1):

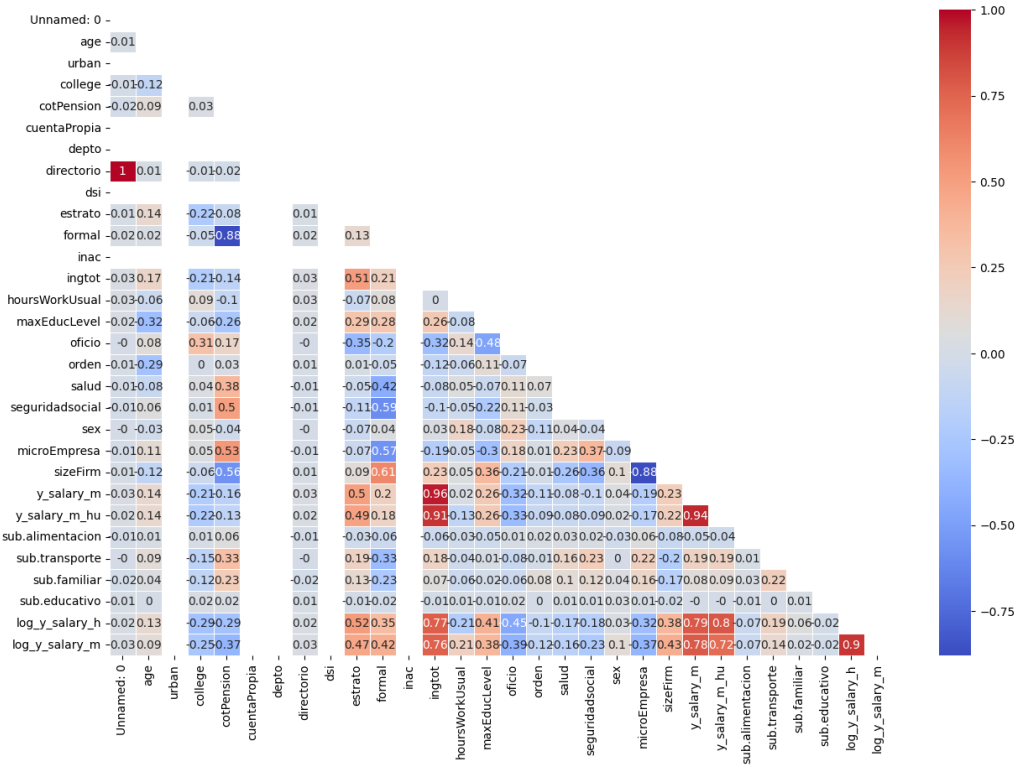


Figura 1: Matriz de correlación de variables de interés

El presente estudio tomará el salario mensual por hora (y_salary_m_hu) y el salario mensual (y_salary_m) como variables objetivo. Para dichas variables se encuentran correlaciones esperadas salvo para college. Esta es una variable categórica que resume si los individuos completan su educación terciaria o no. De acuerdo con la literatura, el salario se correlaciona positivamente con la educación, sin embargo, en la base estudiada este valor es negativo. En concreto, la correlación entre college y y_salary_m_hu es -0,22; similarmente, la correlación entre college y y_salary_m es -0,23. Es decir, las personas que no completan su educación terciaria tienen, en promedio, salarios mensuales más altos. Este dato atípico podría reflejar problemas en la representatividad y/o calidad de los datos. Por ello, los siguientes análisis serán realizados utilizando la variable maxEducLevel, variable categórica ordenada que resume el nivel educativo máximo alcanzado por cada individuo.

Ahora bien, con el objetivo de profundizar en el conocimiento de los datos estudiados, el presente estudio adopta el enfoque diferencial de ciclo de vida propuesto por DANE (2020). Dicho enfoque concuerda con la literatura que ha estudiado el comportamiento de los salarios pues reconoce las necesidades y las capacidades diferenciales a lo largo del ciclo de vida. Dicho ciclo de vida es definido de la siguiente manera (Cuadro 3):

Rango de Edad	Categoría
0-5	Primera Infancia
6-11	Infancia
12-18	Adolescencia
14-28	Juventud
29-44	Adultos jóvenes
45-59	Adultos maduros
>=60	Adulto mayor

Cuadro 3: Categorización del ciclo de vida. Fuente: DANE

Así, la población a estudiar tiene una distribución aproximadamente igual por sexo. Las mujeres representan un 50,2 % de las observaciones mientras que los hombres representan un 49,8 %. Esto

refleja las estadísticas nacionales actuales de distribución por sexo. Dicha distribución es constante a lo largo de los grupos de edades como se aprecia en la siguiente tabla (Cuadro 4):

Participación	Sexo	Jóvenes	Adultos jóvenes	Adultos maduros	Adultos mayores
Dentro de la categoría	Mujeres	53.48 %	49.49 %	47.42 %	52.05 %
	Hombres	47.51 %	50.50 %	52.57 %	47.94 %
Dentro del total muestral	Mujeres	14.87 %	19.96 %	9.88 %	5.44 %
	Hombres	13.46 %	20.37 %	10.95 %	5.01 %

Cuadro 4: Participación de hombres y mujeres según el ciclo de vida por categoría (sexo) y dentro del total de la muestra

En contraste, la formalidad, dentro de la base de datos estudiada, tiene una brecha inusualmente alta. En concreto, la población formal representa el 77,3 % de las observaciones; por su parte, el 22,7 % restante corresponde a la población informal. Dicha diferencia podría responder a la definición de formalidad la cual, en este caso, se define como personas que tienen seguridad social. Aun así, los datos sugieren una mayor formalidad, tanto para hombres como para mujeres, en las edades de adultos jóvenes a adultos maduros (ver Cuadro 5). Esta situación refleja el comportamiento típico del salario a lo largo de la edad pues, según la literatura, una persona, en promedio, llega al máximo de sus ingresos en este rango de edades.

Sexo	Tipo de empleado	Jóvenes	Adultos jóvenes	Adultos maduros	Adultos mayores
Hombres	Formal	72.53 %	80.74 %	73.04 %	67.61 %
	Informal	27.46 %	19.25 %	26.95 %	32.38 %
Mujeres	Formal	70.46 %	83.26 %	83.97 %	78.04 %
	Informal	29.53 %	16.73 %	10.95 %	21.95 %

Cuadro 5: Formalidad de hombres y mujeres según el ciclo de vida.

Seguidamente, estudiamos el comportamiento del nivel máximo de educación a lo largo del ciclo de vida. Para ello, la siguiente tabla resume la media y la desviación estándar del nivel máximo de educación alcanzado por cada uno de los ciclos de vida (Cuadro 6)

Sexo	Medida	Jóvenes	Adultos jóvenes	Adultos maduros	Adultos mayores
Hombres	Media	6.27	6.08	5.63	5.68
	Desviación Estándar	0.79	1.08	1.29	1.40
Mujeres	Media	6.53	6.31	5.72	5.73
	Desviación Estándar	0.63	0.96	1.32	1.44

Cuadro 6: Educación de hombres y mujeres según el ciclo de vida.

Aunque los promedios son similares, el nivel promedio máximo de educación tiende a decrecer ligeramente a medida que la edad aumenta. En concreto, jóvenes y adultos jóvenes, en promedio, completan la secundaria; por su parte, adultos maduros y adultos mayores tienden a tener la secundaria incompleta como nivel máximo de educación alcanzado, en promedio. Sin embargo, los niveles altos de desviación estándar por grupos nos sugieren que, en la base de datos, la educación máxima no es una característica propia de un grupo, sino, más bien es una característica heterogénea. Dicha varianza en los datos de educación puede ser explicada debido a la participación mayoritaria de jóvenes y adultos jóvenes al igual que la mayor frecuencia de niveles de educación terciarios y secundarios.

Finalmente, se ahondó en el comportamiento del salario teniendo en cuenta los anteriores hallazgos. Para comenzar, tanto hombres como mujeres tienen un salario promedio aproximadamente constante para cada uno de sus ciclos de vida. Adicionalmente, la varianza del salario nos permite inferir que el salario no presenta brechas entre sexo ni ciclo de vida, en promedio (ver cuadro 7). Este comportamiento puede explicarse debido a que la media del logaritmo del salario mensual es 13,91 y su desviación estándar es 0,75.

Sexo	Medida	Jóvenes	Adultos jóvenes	Adultos maduros	Adultos mayores
Hombres	Media	13.67	13.96	13.84	13.68
	Desviación Estándar	0.52	0.66	0.77	0.87
Mujeres	Media	13.73	14.07	14.12	14.09
	Desviación Estándar	0.60	0.79	0.90	0.90

Cuadro 7: Promedio y desviación estándar del logaritmo natural del salario para hombres y mujeres según el ciclo de vida.

Por otro lado, el salario parece tener un comportamiento diferencial en la categoría de formalidad. Con el objetivo de comprobar matemáticamente la diferencia de salarios entre formales e informales, se propone una validación de diferencia de medias. Debido a que no existe normalidad entre los datos, se propone una prueba no paramétrica. El estadístico de la prueba de Kruskal-Wallis es 1682,1078 y su p-valor es 0,0. Así con una significancia del 1 %, es posible afirmar que hay diferencia en el promedio de los grupos formales e informales.
Sin embargo, el comportamiento del salario no es claro para otro tipo de variables. De manera ilustrativa, la Figura 2 resume el comportamiento de las variables edad, nivel máximo de educación, sexo y salario mensual. Teóricamente, las variables categóricas como el estrato deberían agrupar con mayor eficiencia los datos de ingresos, ya que, por definición, reflejan las características físicas de los inmuebles y las condiciones productivas de las familias (DANE, 2023). Empero, la Figura 3 ejemplifica la variabilidad de los datos de la muestra estudiada. Por lo tanto, es posible que la base de datos estudiada presente errores en términos de recolección de datos y/o representatividad, lo que podría estar contribuyendo a la falta de claridad en la relación entre las variables y los salarios. Es importante destacar que los modelos de machine learning posteriores tendrán en cuenta estos errores con el objetivo de clarificar la relación entre las variables y los salarios y, así, predecir los salarios de manera efectiva.

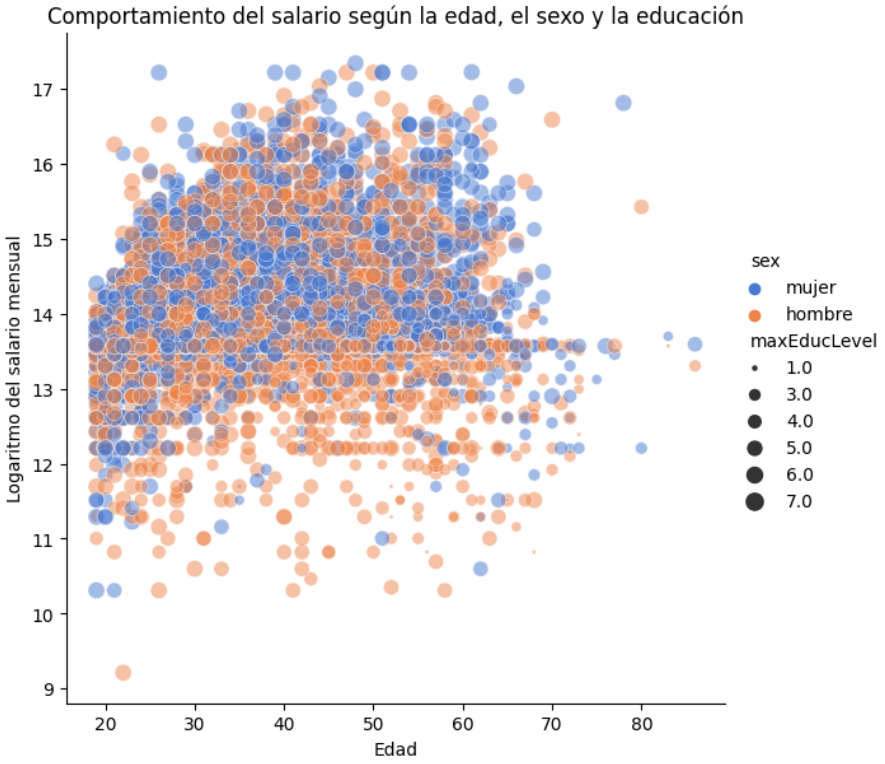


Figura 2: Dispersión del logaritmo natural del salario según la edad, el sexo y el nivel de educación.

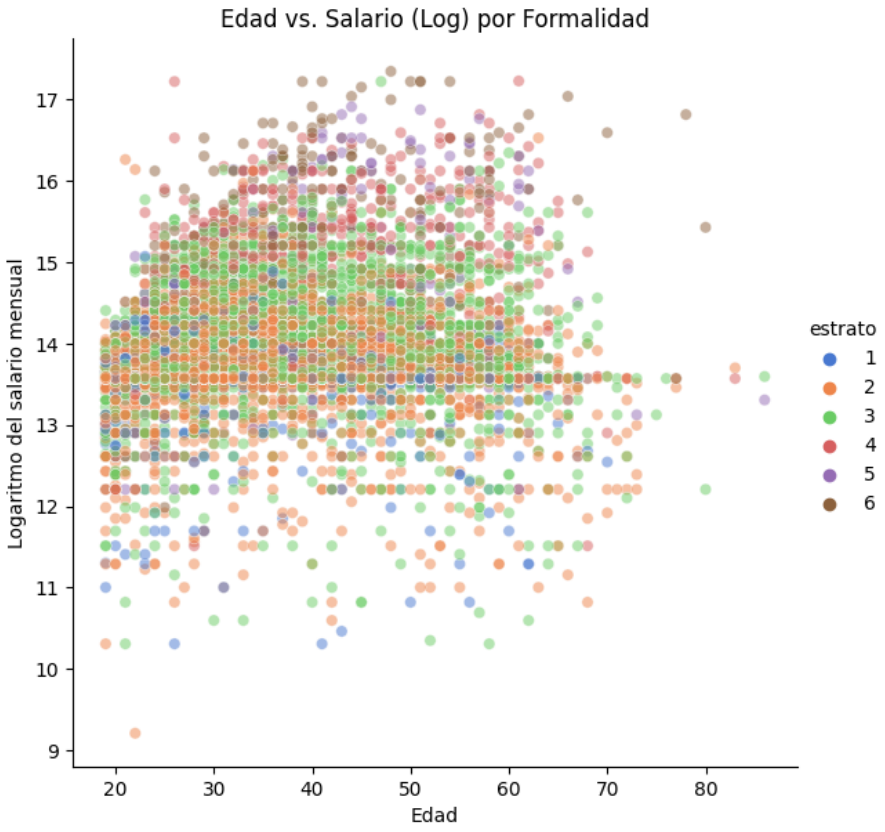


Figura 3: Dispersión del logaritmo natural del salario según edad y estrato

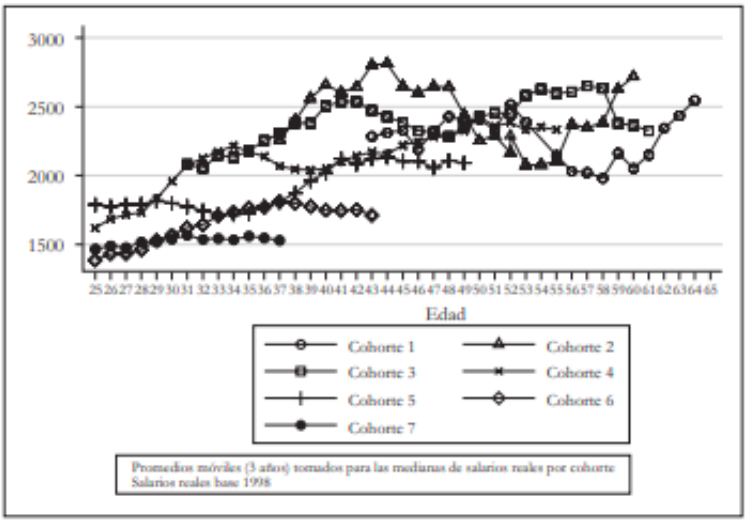
4. Perfil edad- salario

Para comenzar, el modelo propuesto es el siguiente:

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

Dicho modelo responde a la teoría económica y evidencia empírica del mercado laboral. Para ilustrar lo anterior Borjas, G. (2016) en su libro elabora un modelo económico para la oferta de trabajo basado, entre otras, en un trade-off de los agentes quienes deciden entre el placer por descansar hoy y los beneficios económicos derivados del trabajo en un futuro. Borjas, G. (2016) argumenta que entre los veinte y treinta años los salarios suelen crecer rápidamente; por lo cual, el costo de no trabajar es mayor durante estos años. Posteriormente, los salarios continúan creciendo marginalmente hasta los cincuenta años, aproximadamente, punto en el cual comienzan a reducirse hasta la edad de retiro.

Por su parte, Robbins, Salinas, D., & Manco, A. (2009) desarrollan una investigación con este enfoque para el caso colombiano. En su estudio, los autores, revisan los datos de salarios por hora para diferentes cohortes de nacimiento con el objetivo de calcular las elasticidades de oferta del trabajo en Colombia. Durante este ejercicio, los investigadores destacan que en las medianas de salarios por hora, para cada uno de los cohortes de nacimiento analizados, "un patrón de salarios relativamente cóncavos a través del ciclo de la vida, lo cual es consistente con resultados para otros países y es la principal explicación de la concavidad de participación y oferta de horas a través del ciclo de la vida." (pp. 149-150) Según los autores, dicha concavidad, se presenta entre los cuarenta y los cincuenta años de vida para las mujeres en Colombia como se resume en el siguiente gráfico.



Fuente: cálculos propios. Encuestas de Hogares (DANE)

Gráfico 2. *Mediana de salario por hora para mujeres por cohorte de nacimiento*

Figura 4: Fuente: Robbins, Salinas, D., & Manco, A. (2009)

Con base en la teoría expuesta previamente, se procede a calcular el modelo para nuestra base de datos. Los resultados de dicha estimaciòn se muestran a continuación.

Para comenzar, el modelo (1) fue calculado usando la base de datos sin missing values. Así, el modelo propone un coeficiente positivo y significativo al 1 % de significancia. En concreto, el incremento de un año de vida, en promedio, está asociado a un incremento del 5,8 % del salario por hora. Sin embargo, acorde con la teoría, el modelo señala que el componente cuadrático de la variable edad es significativo al 99,9 %. Es decir, el salario aumenta conforme aumenta la edad (β_2), pero este incremento es marginalmente decreciente (β_3).

	Dependent variable:	
	log_y_salary_h	
	(1)	(2)
age	0.058***	0.032***
	(0.004)	(0.002)
age2	−0.001***	−0.0003***
	(0.00005)	(0.00002)
Constant	7.429***	7.997***
	(0.070)	(0.040)
Observations	9,785	16,397
R^2	0.035	0.039
Adjusted R^2	0.035	0.039
Residual Std. Error	0.708 (df = 9782)	0.573 (df = 16394)
F Statistic	176.313*** (df = 2; 9782)	329.354*** (df = 2; 16394)
Nota 1:		*p<0.1; **p<0.05; ***p<0.01
Nota 2:	(1)con datos que eliminan NA, (2) con datos que imputan NA	

Cuadro 8: Estimación del Logaritmo del salario segun la edad

En adición a lo anterior, el modelo permite explicar el 3,5 % de la varianza de la variable logaritmo del salario por hora y, a su vez, el estadístico F es congruente pues, con una significancia del 1 %, es posible afirmar que en su conjunto los coeficientes estimados son diferentes de cero.

En contraste, el modelo (2) fue estimado usando la base de datos cuyos valores faltantes fueron imputados. Así, la primera observación es que se pasó de una estimación de 9785 observaciones a un modelo con 16397. Si bien ambos modelos son consistentes en los signos y la significancia de los estimadores, la segunda estimación presenta ligeros cambios en cuanto a magnitud de los efectos. En particular, para esta estimación, el aumento de un año de vida, en promedio, está asociado a un incremento del 3,2 % del salario por hora. Al igual que en el modelo anterior, el modelo (2) sugiere un componente cuadrático de la variable edad significativo al 99,9 %; sin embargo, su magnitud es más grande. De manera que el cambio marginal en el salario a lo largo de la vida es mayor.

Por si fuera poco, el modelo (2) explica el 3,9 % de la varianza de la variable objetivo. Esto sugiere una ligera mejoría en la predicción que se puede hacer con el modelo. Así mismo, el estadístico F sugiere, con una significancia del 1 %, que los coeficientes del modelo en su conjunto son diferentes de cero.

Por otro lado, la edad pico promedio estimada es de 46,08 años de edad. El intervalo de confianza estimado con una significancia del 95 % permite identificar que la edad pico de la muestra está entre los 44,45 años y los 47,71 años. En ese sentido, resulta interesante evidenciar que para la muestra, la edad pico está por debajo de los cincuenta años, lo que puede implicar que existen otros factores relacionados con la estructura del mercado laboral que hace que el intervalo se encuentre por debajo de lo estimado en otras aplicaciones empíricas.

De esta manera, se concluye que el modelo estimado para la relación edad y salario es coherente con la teoría económica de oferta de trabajo (Borjas, G., 2016) al igual que con la evidencia empírica para el caso colombiano (Robbins, Salinas, D., & Manco, A., 2009). Como se mencionó en la sección de estadísticas descriptivas, los datos sugieren la existencia de un nivel de ingresos máximo a lo

largo del ciclo de vida. Dicha edad pico fue corroborada por el análisis econométrico que encontró un efecto positivo de la edad en el salario, pero con rendimientos marginales decrecientes. Lo cual refleja la acumulación de experiencia laboral, construcción de redes, educación, entre otras variables determinantes para el salario que se adquiere a lo largo de la vida laboral.

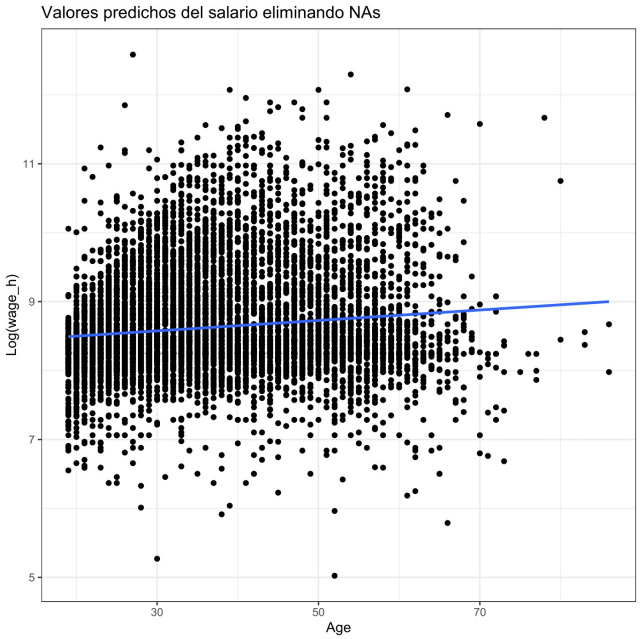


Figura 5: Valores predichos y ajuste del modelo de regresión con la edad (proxy de experiencia)

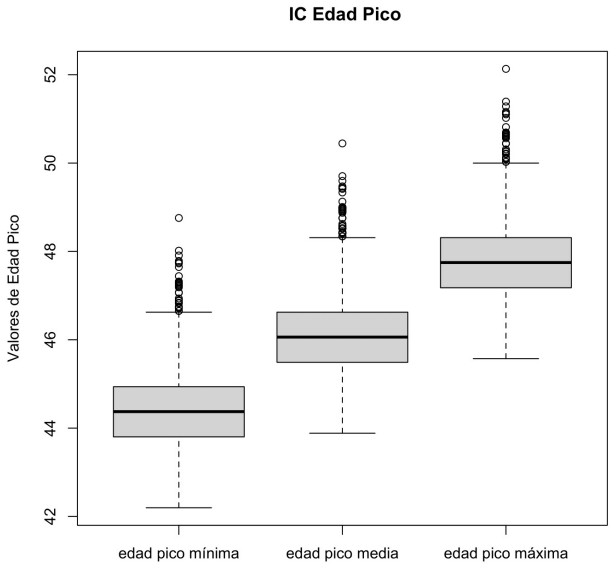


Figura 6: Intervalos de confianza de la edad pico por bootstrap

5. Brecha salarial de género

La evidencia empírica ha demostrado que los hombres ganan más que las mujeres en diferentes mercados laborales. Según Ñopo et al (2009) encuentran que para 18 países de América Latina los hombres ganan aproximadamente 20 % más que las mujeres después de controlar por diferentes características. Para Colombia, según estimaciones del DANE para el año 2021, del grupo de mujeres

ocupadas, ellas ganaron 6,3 % menos que los hombres. La mayores brechas de ingresos salariales se dieron en las mujeres con bajos niveles educativos, que viven en la ruralidad, que tienen 45 y 54 años y que vivían en hogares con presencia de menores de 18 años (DANE, 2022).

Por su parte, la estimaciones del DANE concluyen que las mujeres que trabajan a tiempo parcial tuvieron una brecha del 34,2 % respecto a los hombres y para aquellas que trabajaron de tiempo completo se enfrentan a una brecha a favor de las mujeres de -7,9 %. Por el nivel educativo, a pesar de que la brecha laboral disminuye a medida que aumenta el nivel educativo, las mujeres con nivel de educación profesional tienen un brecha de 21,6 % respecto al mismo grupo de hombres asalariados. Así mismo, el tamaño de la empresa influye en la tendencia de la brecha salarial. En las empresas con más de 200 trabajadores la brecha está a favor de los hombres con 2,5 % y las empresas con 10 y 50 trabajadores tienen una brecha a favor de las mujeres con 19,8 % (DANE, 2022).

En ese sentido, los ingresos salariales están afectados, en el nivel del empleado, por el nivel educativo, la edad y la clase social (Blattman et al., 2014, 2016; Blattman & Dercon, 2018; Mckenzie & Puerto, 2017; McKenzie & Woodruff, 2014), y en el nivel de la empresa, por su tamaño y el tipo de relación laboral(Henrique De Andrade et al., 2013; McKenzie, 2017; McKenzie & Woodruff, 2014).

En consecuencia, para estimar la brecha salarial condicional se controló por la edad, la edad al cuadrado, por educación terciaría (0 = no tenía y 1= tenia), el estrato socioeconómico (0 = no pertenece al estrato 6 y 1 = pertenece al estrato 6), la formalidad del empleo (0 = es informal y 1 = es formal) , cantidad de horas trabajadas y el tamaño de la empresa (0 = tiene más de 5 empleados y 1 = tiene 5 o menos empleados).

En el modelo de brecha no condicional se encontró que la variable sexo es significativa al 99,9 % y de signo positivo, lo que estaría explicando una brecha salarial por hora a favor de los hombres de 4,71 %. La estimación es cercana a la brecha promedio estimada por el DANE (2022) del 6,3 % a favor de los hombres entre el grupo de ocupados.

Por su parte, el modelo de estimación de la brecha condicional permitió encontrar que la variable de interés es significativa al 99,9 % y con signo positivo, lo que explica que existe una brecha salarial del 10,21 % a favor de los hombres. La brecha aumentó cuando se controla por otros factores relacionados con la variación de los ingresos laborales, esta tendencia es coherente con estudios realizados en otras muestras en países en desarrollo, donde se evidencia que la brecha salarial está a favor de los hombres (Bager & Schøtt, 2004;Arroyo, Fuentes & Jiménez, 2016; DANE, 2022)

Respecto al ajuste de los modelos, el modelo de brecha no condicional explica el 0,1 % de la variación del modelo. Respecto al Error Residual Estándar, el modelo tiene un error porcentual de 0,083 %, lo que puede indicar que los datos tienen un buen nivel de ajuste al modelo predictivo. Por otro, al analizar la gráfica de dispersión entre los valores predichos y residuos para casa categoría de sexo se encuentran que el modelo para algunas observaciones sobrestimo y subestimo la variable dependiente, pero en un rango entre 2 y -2 unidades.

Cuadro 9: Brecha salarial de género condicional y no condicional

	Dependent variable:	
	log_y_salary_h	
	(1)	(2)
sex1	0.1021***	0.0471***
	(0.0125)	(0.0146)
age	0.0462***	
	(0.0032)	
age2	−0.0005***	
	(0.00004)	
college1	−0.3619***	
	(0.0130)	
estrato_socioeco	−0.2570***	
	(0.0201)	
formal1	0.3773***	
	(0.0180)	
hoursWorkUsual	−0.0134***	
	(0.0005)	
microEmpresa1	−0.3197***	
	(0.0181)	
Constant	8.2061***	8.6007***
	(0.0641)	(0.0103)
Observations	9,785	9,785
R^2	0.3004	0.0011
Adjusted R^2	0.2998	0.0010
Residual Std. Error	0.6033 (df = 9776)	0.7207 (df = 9783)
F Statistic	524.6084*** (df = 8; 9776)	10.4684*** (df = 1; 9783)
Nota1:		*p<0.1; **p<0.05; ***p<0.01
Nota 2:		(1) Brecha condicional, (2) brecha no condicional

Cuadro 10: Estimación de la brecha salarial por género condicional y no condicional

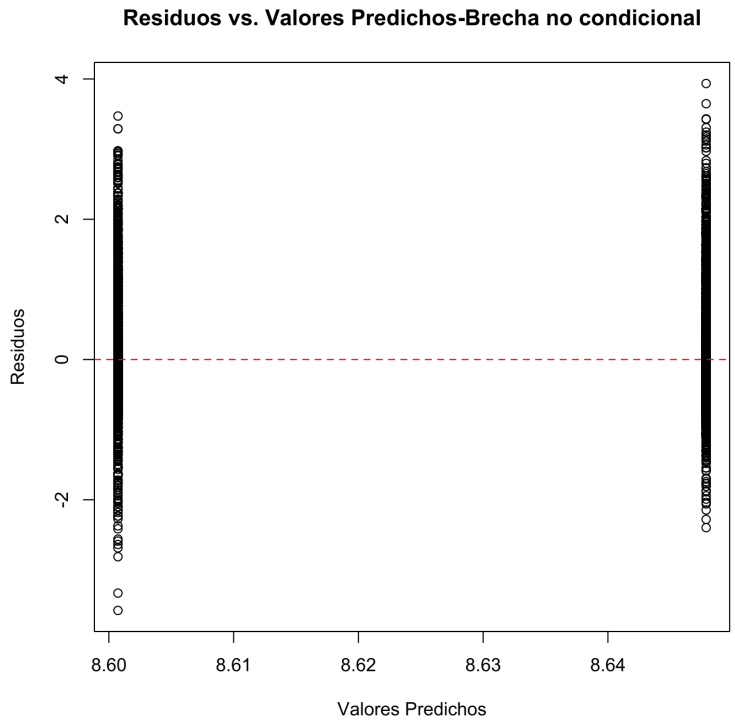


Figura 7: Residuos vs valores predichos del modelo de brecha no condicional

El modelo de brecha condicional explica aproximadamente el 29,98 % de la variación del modelo, 29,8 % más explicación que el modelo anterior debido al aumento de variables explicativas. Respecto al Error Residual Estándar, el modelo tiene un error porcentual de 0,069 %, lo que puede indicar que los datos tienen un buen nivel de ajuste al modelo predictivo. En ese sentido, la gráfica de dispersión entre los valores predichos y los residuos del modelo muestra que una parte de la muestra puede ajustarse al modelo, sin embargo, en otra parte de la muestra (círculo rojo) el modelo subestimó los valores de la variable de interés.

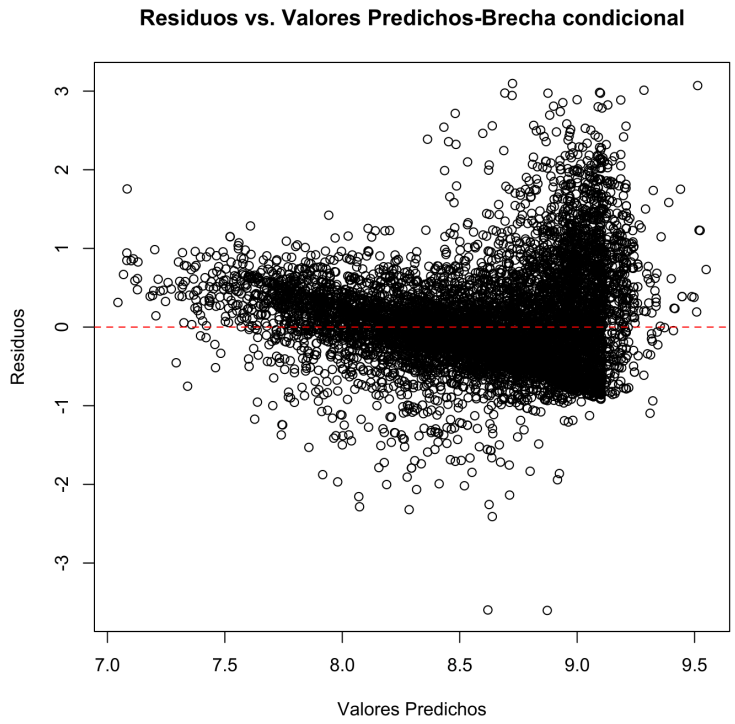
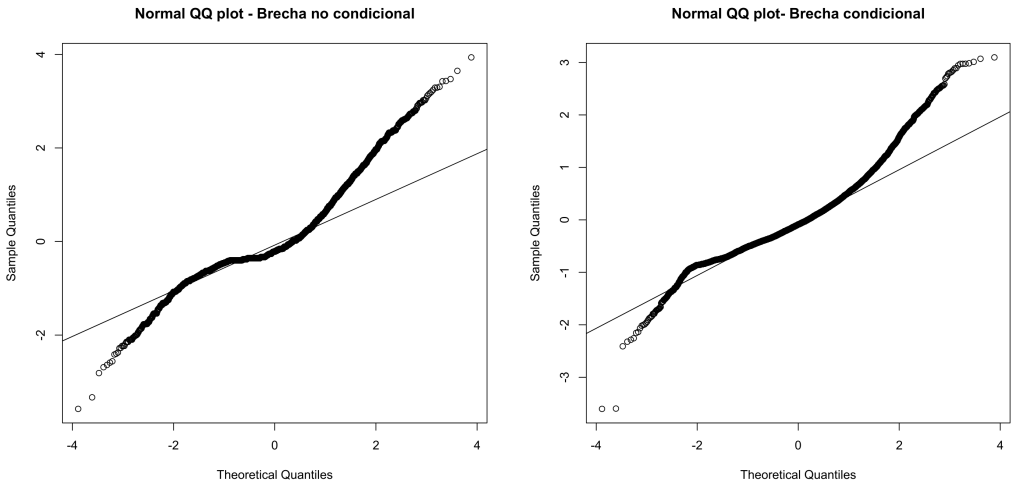


Figura 8: Residuos vs valores predichos del modelo de brecha condicional

En comparación, ambos modelos permiten ajustarse adecuadamente a los datos, sin embargo, la predicción del ingreso salarial es mejor en el modelo de brecha salarial condicional debido a la incorporación de más variables predictoras (7 variables adicionales) que aumentan la variación del modelo (29,8 % adicional) y disminuyen el sesgo 12,3 p.p. respecto al modelo de brecha no condicional. En ese sentido, las diferencias en la estimación de la brecha salarial en ambos modelos plantea un problema de selección de variables.

En el modelo de brecha condicional al controlar por otras características se puede evidenciar que la proporción de la brecha aumento a favor de los hombres, es decir, otros factores independientes al sexo influyen en el incremento de la variable salario por hora. La selección de las variables estuvo dada por la disponibilidad de las variables y su distribución para la muestra analizada y su relación con la evidencia empírica externa que nos mostró cuáles podrían ser los mejores predictores sin entrar a un modelo sobre especificación del modelo, y a su vez, aumentando la variabilidad del mismo.



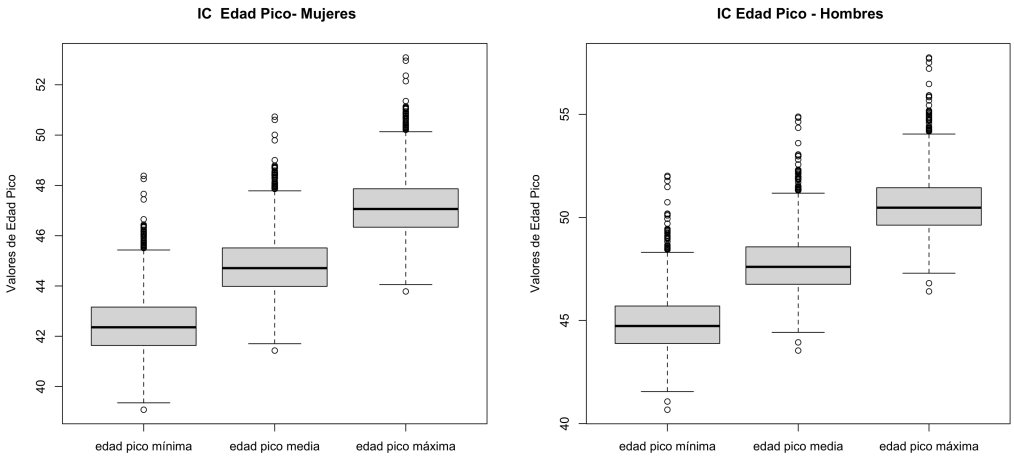
(a) Brecha no condicional (b) Brecha Condicional

Figura 9: Gráfico Normal Q-Q

En el anterior gráfico (Figura 7) se puede evidenciar que existe mayor sesgo de predicción el modelo de brecha no condicional. Las dos colas de la distribución de los errores están alejadas de la línea de distribución normal. Por otra parte, para el modelo de brecha condicional podemos observar que una mayor parte de la muestra se ajusta a la distribución muestral normal, sin embargo, en los dos extremos se pueden observar que valores predichos de la muestra que no se ajustan a la normalidad, en menor medida que en el modelo no condicional.

Edad pico por género

La edad pico estimada para las mujeres de la muestra utilizando el modelo de predicción lineal fue de 44.68 años. La estimación del intervalo de confianza con el método bootstrap fue de 42.49 años la edad pico mínima y 47. 11 años es la edad pico máxima. En el caso de los hombres, la edad media pico estimada fue de 47,60 años y el intervalo de confianza estimado está entre los 44,80 años y los 50,58 años, en promedio. Las estimaciones del intervalo de confianza se realizaron con una significancia del 95 % y un error estándar estimado de 1.19 años para las mujeres y 1.46 años para los hombres.



(a) Para mujeres (b) Para hombres

Figura 10: Intervalo de edad pico por género

6. Prediciendo ingresos

Elegimos predecir los siguientes modelos a través del validation set approach en el que dividimos la muestra en 70 % para entrenamiento y 30 % para prueba :

- **Modelo sex**
  $log(y\_salary\_h) = \beta_0 + \beta_1 sex$
- **Modelo age**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2$
- **Modelo sexage**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sex$
- **Modelo 1**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sex + \beta_4 maxEducLevel + \beta_5 estrato + \beta_6 hoursWorkUsual + \beta_7 microEmpresa$
- **Modelo 2**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sex + \beta_4 maxEducLevel + \beta_5 estrato + \beta_6 hoursWorkUsual + \beta_7 microEmpresa + \beta_8 salud + \beta_9 seguridadsocial$
- **Modelo 3**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sex + \beta_4 maxEducLevel + \beta_5 estrato + \beta_6 hoursWorkUsual + \beta_7 microEmpresa + \beta_8 salud + \beta_9 seguridadsocial + \beta_{10} sub.transporte + \beta_{11} sub.familiar + \beta_{12} sub.educativo + \beta_{13} sub.alimentacion$
- **Modelo 4**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sex + \beta_4 maxEducLevel + \beta_5 estrato + \beta_6 hoursWorkUsual + \beta_7 microEmpresa + \beta_8 salud + \beta_9 seguridadsocial + \beta_{10} sub.transporte + \beta_{11} sub.familiar + \beta_{12} sub.educativo + \beta_{13} sub.alimentacion + \beta_{14} sex * maxEducLevel + \beta_{15} sex * salud + \beta_{16} sex * microEmpresa + \beta_{17} sex * formal$
- **Modelo 5**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sex + \beta_4 maxEducLevel + \beta_5 estrato + \beta_6 hoursWorkUsual + \beta_7 microEmpresa + \beta_8 salud + \beta_9 seguridadsocial + \beta_{10} sub.transporte + \beta_{11} sub.familiar + \beta_{12} sub.educativo + \beta_{13} sub.alimentacion + \beta_{14} sex * maxEducLevel + \beta_{15} sex * salud + \beta_{16} sex * microEmpresa + \beta_{17} sex * forma + \beta_{18} relab$
- **Modelo 6**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 sex + \beta_4 maxEducLevel + \beta_5 estrato + \beta_6 hoursWorkUsual + \beta_7 microEmpresa + \beta_8 salud + \beta_9 seguridadsocial + \beta_{10} sub.transporte + \beta_{11} sub.familiar + \beta_{12} sub.educativo + \beta_{13} sub.alimentacion + \beta_{14} sex * maxEducLevel + \beta_{15} sex * salud + \beta_{16} sex * microEmpresa + \beta_{17} sex * forma + \beta_{18} relab + \beta_{19} sizeFirm + \beta_{20} formal$
- **Modelo 7**
  $log(ysalaryh) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 age^4 + \beta_5 age^5 + \beta_6 sex + \beta_7 maxEducLevel + \beta_8 estrato + \beta_9 hoursWorkUsual + \beta_{10} microEmpresa + \beta_{11} salud + \beta_{12} seguridadsocial + \beta_{13} sub.transporte + \beta_{14} sub.familiar + \beta_{15} sub.educativo + \beta_{16} sub.alimentacion + \beta_{17} sex * maxEducLevel + \beta_{18} sex * salud + \beta_{19} sex * microEmpresa + \beta_{20} sex * formal + \beta_{21} relab + \beta_{22} sizeFirm + \beta_{23} formal$
- **Modelo 8**
  $log(y\_salary\_h) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 age^3 + \beta_4 age^4 + \beta_5 age^5 + \beta_6 age^6 + \beta_7 age^7 + \beta_8 age^8 + \beta_9 age^9 + \beta_10 age^{10} + \beta_{11} sex + \beta_{12} maxEducLevel + \beta_{13} estrato + \beta_{14} hoursWorkUsual + \beta_{15} hoursWorkUsual^2 + \beta_{16} hoursWorkUsual^3 + \beta_{17} hoursWorkUsual^4 + \beta_{18} hoursWorkUsual^5 + \beta_{19} hoursWorkUsual^6 + \beta_{20} hoursWorkUsual^7 + \beta_{21} hoursWorkUsual^8 + \beta_{22} hoursWorkUsual^9 + \beta_{23} hoursWorkUsual^{10} + \beta_{24} microEmpresa + \beta_{25} salud + \beta_{26} seguridadsocial + \beta_{27} sub.transporte + \beta_{28} sub.familiar + \beta_{29} sub.educativo + \beta_{30} sub.alimentacion + \beta_{31} sex * maxEducLevel + \beta_{32} sex * salud + \beta_{33} sex * microEmpresa + \beta_{34} sex * formal + \beta_{35} relab + \beta_{36} sizeFirm + \beta_{37} formal$

En primer lugar, vemos que los modelos están ordenados del menos complejo al más complejo, teniendo en el primero apenas un predictor, y en el último 37. Los primeros tres modelos son los que se estimaron en los puntos anteriores, y a partir del cuarto modelo agregamos variables que pensamos ayudarían para la predicción de los salarios. De este modo, llegamos a un modelo tan complejo que incluía un polinomio de grado 10 tanto para la edad como para las horas trabajadas usualmente. Además, para agregar complejidad a los modelos, realizamos interacciones entre sexo con algunas variables como nivel máximo de educación, salud y formal.

En la siguiente tabla podemos ver el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE), medidas que se utilizan para verificar el desempeño de un modelo por fuera de muestra

model	MSE	RMSE	R2
modelage	0.52	0.72	0.00
modelsex	0.50	0.71	0.03
modelsexage	0.50	0.71	0.03
model 1	0.24	0.49	0.53
model 2	0.24	0.49	0.54
model 3	0.22	0.47	0.58
model 4	0.22	0.47	0.59
model 5	0.21	0.46	0.60
model 6	0.20	0.45	0.62
model 7	0.20	0.45	0.62
model 8	0.28	0.53	0.63

Las tres especificaciones estudiadas previamente se encuentran en esta tabla nombradas como “Modelo age”,”Modelo sex” y “Modelo sex age”. Podemos ver que los tres modelos cuentan con un RMSE aproximado de 0.7 y un MSE cercano a 0.5. Lo anterior nos lleva a determinar que las métricas de prueba son muy altas en relación con las adquiridas en los modelos que se estimaron posteriormente, demostrando entonces que el modelo no logra predecir de forma exitosa en los datos de prueba como lo hacen los demás modelos que tienen un MSE promedio de 0.2.

Podemos observar que en general, a partir del Modelo 1, los errores cuadráticos medios (de prueba) tienen una baja magnitud, por lo que los modelos predicen de buena manera el salario. Aparte, podemos ver que a medida que el modelo es más complejo, se va a reduciendo el MSE, por lo que vamos comprobando que los modelos más complejos al reducir el sesgo, reducen también el MSE. Pero además, observamos que desde el modelo 7, los modelos muy complejos están agregando mucha varianza, por lo que el MSE aumenta.

En este sentido, podemos ver en la siguiente gráfica el comportamiento del RMSE de prueba y el RMSE de entrenamiento para cada uno de los modelos que estimamos según su complejidad. Tanto el MSE de prueba como el de entrenamiento tienen el comportamiento que se espera. Como ya se mencionó anteriormente, el MSE de prueba se va reduciendo a medida que aumenta la complejidad de un modelo, pero llega un punto en el que el modelo es tan complejo que se sobre ajusta a los datos, por lo que va perdiendo capacidad de predicción fuera de muestra, aumentando el MSE. Por otro lado, el MSE de entrenamiento siempre decrece, debido a que si se sobre ajusta a los datos, en realidad está ganando capacidad de predicción dentro de la muestra.

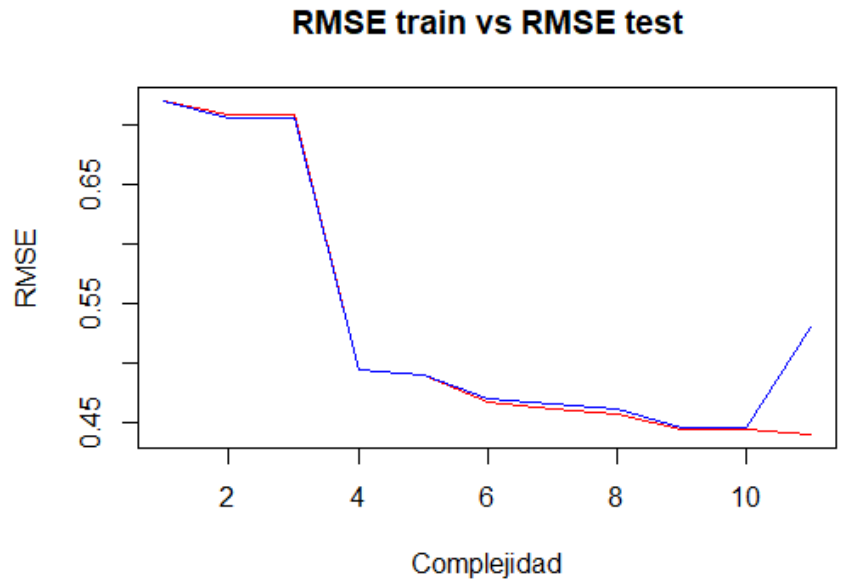


Figura 11: RMSE entrenamiento y de prueba según la complejidad del modelo
Fuente: elaboración propia

Tanto en la tabla como en la gráfica, podemos observar que el modelo con el menor MSE es el modelo 6, el cual incluye todas las variables que nos interesan, por lo que podemos decir que son suficientes para una buena predicción y no es necesario agregar más complejidades y no linealidades (como se realiza en los modelos 7 y 8). Sin embargo, sería interesante observar la distribución de los errores de predicción de este modelo, para ver si hay posibles valores atípicos. En este sentido, en el siguiente gráfico se presenta la distribución de los errores de predicción, donde se puede ver que tiene media cercana a 0, pero tiene muchos valores atípicos en la cola izquierda. Lo anterior implica que el valor pronosticado para esas observaciones es mayor que el valor del salario real. Es importante entonces cuestionarnos de la causa de los valores atípicos: ¿es posible que indiquen un grave problema que debería revisar la DIAN?

Para responder la anterior pregunta, tenemos dos hipótesis. La primera es que algunas personas están mintiendo acerca de sus niveles de salario, lo cual es un potencial problema para la DIAN, quien debería revisar de cerca que está sucediendo, dado que los individuos reportan ingresos salariales menores a los estimados. En segundo lugar, creemos que los individuos pueden estar en estado de subocupación o subempleo, en el que las personas estén trabajando menos de las 48 horas legales (para 2018), o en condición de empleo inadecuado en términos de competencias o ingresos (DANE, s.f.). Teniendo en cuenta esto, nuestra hipótesis se encamina de forma particular a que los individuos con situaciones de salarios atípicos (concentradas en la parte izquierda de la distribución), pueden estar siendo remunerados de forma inadecuada en contraste con sus capacidades, recibiendo salarios menores a los estimados. Lo anterior supone una investigación más profunda por parte del departamento de estadística para poder determinar las razones por las que estos individuos toman empleos inadecuados.

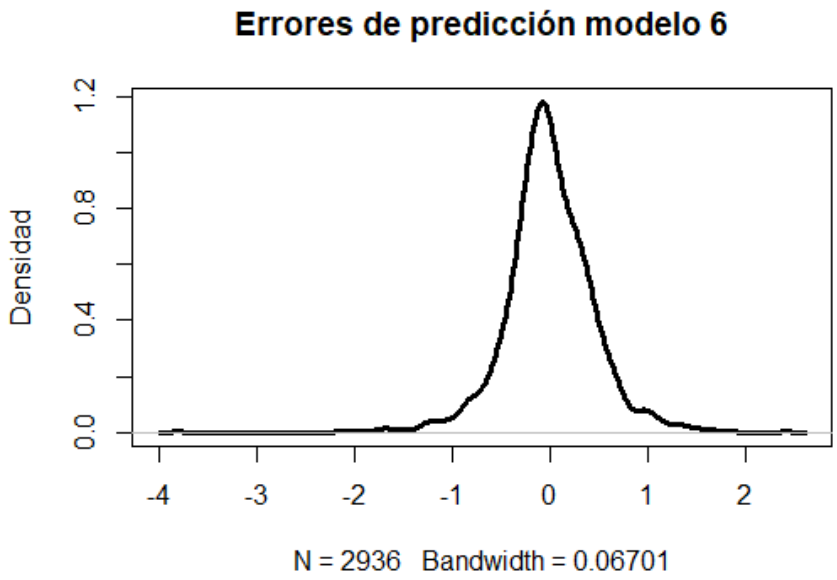


Figura 12: Gráficos de densidad de los errores de predicción

Por último, con el fin de comprobar los resultados anteriores, calculamos el RMSE de los dos modelos con mejor predictibilidad (modelos 6 y 7) usando el método Leave One Out Cross Validation (LOOCV). Obtuvimos que para el modelo 6, el RMSE es 0.446695, mientras que para el modelo 7 es 0.446693. Comparando con los RMSE que dan con el método de validación cruzada simple vemos que no cambian significativamente, y que aparte ninguno de los dos modelos es mejor que el otro, ya que la diferencia es de apenas 0.000002.

7. Referencias:

Blattman, C., Fiala, N., Martinez, S. (2014). Generating skilled self-employment in developing countries: Experimental evidence from Uganda. Quarterly Journal of Economics, 129(2), 697–752. https://doi.org/10.1093/qje/qjt057

Blattman, C., Green, E. P., Jamison, J., Christian Lehmann, M., Annan, J. (2016). The returns to microenterprise support among the ultrapoor: A field experiment in postwar Uganda. American Economic Journal: Applied Economics, 8(2), 35–64. https://doi.org/10.1257/app.20150023

Blattman, C., Dercon, S. (2018). The impacts of industrial and entrepreneurial work on income and health: Experimental evidence from Ethiopia. American Economic Journal: Applied Economics, 10(3), 1–38. https://doi.org/10.1257/app.20170173

Borjas, G. (2016). Labor Supply. En Borjas, G. (Ed.) Labor Economics.(pp. 21-83). Mc Graw-Hill.

DANE. (2020). Guía para la inclusión del enfoque diferencial e interseccional. Recuperado de https://www.dane.gov.co /files/investigaciones/genero/guia-inclusion- enfoque-difencias-intersecciones-produccion-estadistica-SEN.pdf

DANE. (2022). Nota estadística sobre la brecha salarial de género. Recuperado de: https://www.dane.gov.co/file estadisticas/dic-brecha-salarail-genero-2022-v3.pdf

DANE. (2023). Estratificación socioeconómica para servicios públicos domiciliarios. Recuperado de https://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-informacion/estratificacion-socioeconomica#

Henrique De Andrade, G., Bruhn, M., The, D. M., Bank, W. (2013). A Helping Hand or the Long Arm of the Law? Experimental Evidence on What Governments Can Do to Formalize Firms. http://econ.worldbank.org.

McKenzie, D., Woodruff, C. (2014). What are we learning from business training and entrepreneurship evaluations around the developing world? World Bank Research Observer, 29(1), 48–82. https://doi.org/10.1093/wbro/lkt007

McKenzie, D. (2017). Identifying and spurring high-growth entrepreneurship: Experimental evidence from a business plan competition.American Economic Review, 107(8), 2278–2307. https://doi.org/10.1257/aer.20

Mckenzie, D., Puerto, S. (2017). Growing Markets through Business Training for Female Entrepreneurs A Market-Level Randomized Experiment in Kenya. http://econ.worldbank.org.

Robbins, Salinas, D., Manco, A. (2009). La oferta laboral femenina y sus determinantes: evidencia para Colombia con estimativas de cohortes sintéticas. Lecturas de Economía, 70(70), 137–163. https://doi.org/10.17533/udea.le.n70a2258