Universidad de los Andes
Big Data and Machine Learning for Applied Economics
Taller 1

Integrantes:
David Santiago Caraballo Candela [1]
Sergio David Pinilla Padilla [2]
Juan Diego Valencia Romero [3]

5 de septiembre de 2022

GitHub: Repositorio Taller 1

1 Fraude Fiscal

En el sector público, la declaración exacta de los ingresos individuales es fundamental para calcular los impuestos. Sin embargo, el fraude fiscal de todo tipo siempre ha sido un problema importante. Según datos del Servicio de Impuestos Internos (IRS, por sus siglas en inglés), cerca del 83,6% de los impuestos se pagan voluntariamente y a tiempo en los Estados Unidos.[4]. Una de las causas de este desfase es el subreporte de ingresos por parte de los particulares. Luego, un modelo de predicción de los ingresos podría ayudar a detectar los casos de fraude que podrían conducir a la reducción de la brecha entre ingresos y egresos del Estado. Además, puede ayudar a identificar a personas y familias vulnerables que requieran más asistencia social.

El caso latinoamericano es diametralmente opuesto. Particularmente, Colombia es uno de los países que más pierde recursos por la evasión de impuestos. En cifras, se dejan de recolectar 3,4 puntos porcentuales del PIB (Producto Interno Bruto) por concepto de renta de las empresas, seguido por el IVA (Impuesto del Valor Agregado) equivalente al 1,4% del PIB y la renta de las personas naturales con 0,7%. Así las cosas, planificar políticas y establecer programas sociales es un proceso que se ve entorpecido por la escasez y, naturalmente, el detrimento del presupuesto público. La entidad encargada de planear, levantar, analizar y difundir las estadísticas oficiales del país, DANE (Departamento Administrativo Nacional de Estadística), provee la GEIH (Gran Encuesta Integrada de Hogares), una encuesta mediante la cual se solicita información sobre las condiciones de empleo de las personas además de las características generales de la población a nivel nacional, cabecera-resto, regional, departamental, y para cada una de las capitales de los departamentos. Este trabajo constituye una aplicación práctica de los tópicos aprendidos en el curso de *Big Data and Machine Learning for Applied Economics* sobre predicción y estimación.

El documento se organiza de la siguiente manera: la sección 2 contiene una descripción detallada de los datos a utilizar así como su procesamiento; en la sección 3 se construye un perfil de edad-ingresos para los individuos de la muestra; la sección 4 provee evidencia empírica sobre una potencial brecha salarial entre géneros y la sección 5 evalúa el poder predictivo de las especificaciones utilizadas. Las consideraciones finales son presentadas en la sección 6.

2 Datos

Este trabajo utiliza la información de cada persona ocupada y mayor a 18 años encuestada en la GEIH del DANE, dentro de Bogotá D.C. y para todos los meses del año 2018, respectivamente. Tales datos son relevantes para el estudio debido que en esta encuesta se indaga acerca de las condiciones de vida y los índices del mercado laboral en Colombia.

[1] Código: 201813007.
[2] Código: 201814755.
[3] Código: 201815561.
[4] Ver https://www.irs.gov/newsroom/the-tax-gap.

Esta encuesta, que utilizaba el marco geoestadístico del Censo Nacional de Población y Vivienda (CNPV) de 2005, a la cual se refiere como GEIH-M05, se actualizó para estar acorde con el marco y los resultados del CNPV de 2018. Tal actualización permitió tener en cuenta los cambios en la evolución demográfica, organización del territorio y distribución de la población que ha experimentado el país entre las dos (2) operaciones censales. Adicionalmente, la nueva encuesta, GEIH-M18, introdujo recomendaciones de organismos como la Organización Internacional del Trabajo (OIT) y la Organización para la Cooperación y el Desarrollo Económico (OCDE), en materia de caracterización de la Población en Edad de Trabajar (PET) y la visibilidad estadística de otros grupos, lo que mantuvo las estadísticas del mercado laboral colombiano acorde a los estándares internacionales. Por lo tanto, esta información es un insumo robusto para analizar el comportamiento de los ingresos y el mercado laboral en la capital.

Los datos de la muestra fueron recolectados de un enlace de GitHub[5] mediante un procesamiento de *data-scraping* en `RStudio`, un entorno de desarrollo integrado para el lenguaje de programación `R`. Debido a que la web donde se encontraba la información consistía en una pagina dinámica, para extraer los *data chunks* requeridos fue necesario entrar a inspeccionar la pagina dentro de la sección de `Red`, en donde se encontraron los enlaces estables que contenían cada una de las tablas de la base de datos. La base principal (*master*) constaba de 32.177 observaciones a nivel individual para un total de 178 variables.

Posteriormente, se realizó la limpieza y organización utilizando el entorno de `Visual Studio Code (VSC)` con el lenguaje de programación `Python`. La base se filtró para incluir únicamente a los individuos ocupados y mayores de edad (es decir, a partir de los 18 años en adelante), lo cual redujo el número de observaciones a 16.452. Luego, de las 178 variables iniciales se realizó la selección de 19 variables consideradas como relevantes para cumplir a cabalidad con los objetivos, las cuales pueden ser organizadas dentro de tres (3) macro-categorías diferentes:

1. **Socio-económicas (tradicionales):** Aquellas que permiten identificar características básicas de los individuos, i.e.; i) edad (`age`), como variable contínua; ii) género (`sex`), como variable dicótoma que diferencia entre hombres (1) y mujeres (0); iii) nivel de educación (`maxEducLevel`), como variable categórica que identifica el máximo nivel educativo (del rango entre sin educación y educación terciaria), y; iv) estrato socio-económico (`estrato1`), como variable categórica. Además, también se incluyó a otra variable metodológica "clásica": el mes (`mes`), gracias a que esta sirve para desglosar y controlar por los efectos de la estacionalidad sobre las demás variables.

2. **Condiciones laborales**: Aquellas que permiten identificar características laborales de los individuos, i.e.; i) formalidad (`formal`), como variable dicótoma que indica si se esta inscrito (1) o no (0) en la seguridad social; ii) tipo de oficio (`oficio`), como variable categórica que distingue entre 99 tipos diferentes de actividades económicas; iii) tamaño de la firma (`sizeFirm`), como variable que agrupa a las firmas dentro de cinco (5) categorías según el tamaño de su nomina; iv) régimen de Salud (`regSalud`), como variable categórica que distingue entre personas afiliadas al régimen contributivo, el especial y el subsidiado; v) tiempo en la empresa (`p6426`), que corresponde a una variable continua que cuenta el numero de meses que lleva trabajando un individuo dentro de su principal empresa; vi) relación laboral (`relab`), como variable con nueve (9) categorías que describen el tipo de relación laboral de cada individuo con su empresa/actividad económica (e.g. empleado del gobierno, empleador, trabajador por cuenta propia, entre otros); vii) número de horas -usualmente- trabajadas a la semana (`p6426`), como una variable continua que permite medir el ingreso por hora trabajada, y; viii) una pregunta acerca de si se quiere trabajar más horas a la semana (`p7090`), como una variable dicótoma que serviría como *proxy* de la oferta laboral no asignada.

[5] Ver https://ignaciomsarmiento.github.io/GEIH2018_sample/.

3. **Formas de ingreso:** Aquellas que permiten identificar las distintas mediciones del ingreso que perciben los individuos, i.e.; i) ingreso total (`ingtot`), es una variable continua que corresponde al total bruto de los ingresos (sea observado o imputado) de una persona -esto incluye ingresos laborales, en especie, de renta e intereses, entre otros-; ii) el valor total de otras fuentes de ingresos (`p7510s7a1`), como las ganancias de juegos de azar, las liquidaciones, la venta de propiedades como acciones o vehículos, entre otros, y; iii) el ingreso monetario -compilado- de tanto de la primera (`impa` $\cup$ `impaes` $=$ `impacomp`) como de la segunda (`isa` $\cup$ `isaes` $=$ `isacomp`) actividad laboral de cada encuestado, el cual puede ser un dato observado o imputado.

Tabla 1. Valores faltantes.

Missing values	Nivel educativo	Régimen de salud	IMPA compilado	ISA compilado
Cuenta	1	1420	248	31
Participación (%)	0,006	8,58	1,49	0,18

Tercero, es posible evidenciar a partir de la Tabla 1 que algunas de las variables seleccionadas tienen valores faltantes (*missing values*). En unos casos la presencia de estos *missings* es menor al 1.5% de las observaciones, como en el caso de las variables `maxEducLevel`, `impacomp` e `isacomp`, y tan solo en el caso de `regSalud` la presencia de *missings* llega al 8.6% de las observaciones. Así las cosas, para resolver este problema de falta de información en algunas variables relevantes, se realizó un ejercicio de imputación de valores utilizando un modelo de Vecinos Más Cercanos (KNN, por sus siglas en inglés).

Debido que el modelo KNN permite aproximar el valor y/o la clase de cada *missing value* según el comportamiento de sus k vecinos más cercanos, fue necesaria la selección de un valor especifico para el hiper-parámetro k. Entonces, se realizó una revisión general de literatura relacionada con el proceso de selección de tal k, y se encontró que en investigaciones de diferentes áreas del conocimiento que trabajaban con bases de datos independientes (Ramírez et al., 2020; Javaheri, 2021; Mwangi, 2010; Wang et al., 2017), recurrentemente se llegaba a un k^* cercano a 19. Por lo tanto, partiendo de la bibliografía estudiada y tomando en cuenta que este proceso únicamente modificaría una pequeña sección de la muestra, se concluyó que era razonable imputar los datos faltantes utilizando un modelo KNN ($k^* = 19$).

Para finalizar, la Tabla 2 muestra las estadísticas descriptivas de algunas variables relevantes luego de haber realizado la imputación de los *missing values*. Con respecto al ingreso total (`ingtot`), se evidencia que en esta muestra los individuos pueden obtener ingresos mensuales dentro de un rango de cero (0) hasta los COP $85,8 millones. Pero, además, se concluye que la distribución de esta variable esta sesgada hacia la cola izquierda, con valores atípicos presentes en la cola derecha. Debido a que la media ($\approx$ COP $1,8 millones) es mayor a la mediana ($\approx$ COP $1,1 millones), y el mínimo tan solo está a 0,39 desviaciones estándar de la mediana mientras que el máximo está a 31,69 desviaciones estándar de la mediana. También, dentro de esta muestra el rango de edades (`age`) va desde 18 años hasta 94 años, con una desviación estándar de 13,5 años. La distribución de esta variable esta más centrada respecto al ingreso total, pero sigue ligeramente sesgada hacia la izquierda. La media (39,4 años) es un poco mayor a la mediana (38 años), y con la diferencia absoluta entre la mediana y el mínimo (18 años) es menor a la diferencia absoluta entre la mediana y el máximo (56 años). Para las horas trabajadas (`p6426`), como todos los individuos seleccionados fueron clasificados como ocupados, dentro de la muestra mínimo se trabajó una (1) hora a la semana y máximo 130 horas, con una desviación estándar de 15,5 horas. Sorprende que la mediana de horas trabajadas a la semana sea de 48 horas, debido a que, para el momento en que se tomo la muestra, en Colombia el máximo legal de horas semanales de trabajo a tiempo completo correspondía a estas mismas 48 horas.

Lo anterior implica que cerca del 50% de los encuestados: o semanalmente trabajan más de lo que deberían en su trabajo principal, o poseen un segundo trabajo que les consume horas adicionales durante la semana o, en su defecto, ambos. Incluso, se evidencia que la muestra esta balanceada en términos de género (`sex`), ya que el 53% son hombres y el 47% son mujeres, pero no esta tan balanceada en términos de formalidad (`formal`), porque la mayoría de los encuestados (el 59%) son trabajadores formales. En último, vale la pena indicar que la mayor parte de la muestra proviene de individuos pertenecientes a los estratos uno (1), 11%, dos (2), 42%, y tres (3), 36%, lo cual agrega a poco menos del 90% de la base de datos en uso (*using*) para las estimaciones ($\approx$ 14.600 individuos).

Tabla 2. Estadísticas descriptivas (sin agrupar).

VARIABLE	Obs.	Media	Desviación	Mínimo	Mediana	Máximo
Ingreso total	16542	1769379	2675628	0	1051160	85833330
Edad	16542	39,44	13,48	18	38	94
Horas	16542	47,01	15,54	1	48	130
Antigüedad	16542	63,76	89,49	0	24	720
Género	16542	0,53	0,50	0	1	1
Formal	16542	0,59	0,49	0	1	1
Horas extra	16542	0,10	0,30	0	0	1
Estrato 1	16542	0,11	0,31	0	0	1
Estrato 2	16542	0,42	0,49	0	0	1
Estrato 3	16542	0,36	0,48	0	0	1
Estrato 4	16542	0,07	0,25	0	0	1
Estrato 5	16542	0,02	0,14	0	0	1
Estrato 6	16542	0,03	0,16	0	0	1
Régimen contributivo	16542	0,75	0,43	0	1	1
Régimen especial	16542	0,03	0,16	0	0	1
Régimen subsidiado	16542	0,14	0,35	0	0	1
Sin estudios	16542	0,01	0,09	0	0	1
Primaria incompleta	16542	0,05	0,21	0	0	1
Primaria completa	16542	0,09	0,29	0	0	1
Secundaria incompleta	16542	0,11	0,32	0	0	1
Secundaria completa	16542	0,32	0,47	0	0	1
Terciaria/Superior	16542	0,42	0,49	0	0	1

3 Perfil Edad-Ingreso

La presente sección aborda uno (1) de los temas más trabajados dentro de la "sub-disciplina" de *labour economics*. Distintos exponentes de la literatura relacionada como David Card, Daron Acemoglu y Marco Tabellini han considerado que, potencialmente, existe una relación cóncava entre los ingresos laborales y la edad de los trabajadores, la cual no solo es sostenida en el tiempo sino que también se acrecienta al pasar el mismo. En efecto, para analizar la veracidad de esta hipótesis dentro del mercado laboral de Bogotá D.C., se estimó una regresión lineal en la cual el nivel de ingresos laborales (o *earnings*) es función de la edad lineal y cuadrática del trabajador.

Como medida del nivel de ingresos laborales de cada trabajador, se utilizó una sumatoria horizontal de las variables de la base de datos que poseen información acerca de los ingresos laborales tanto de la primera actividad económica (`impacomp`) como de la segunda (`isacomp`). Note que esto fue explicado en la sección 2 cuando se trató acerca de los valores faltantes. Luego, ingresos laborales se define como:

$$Ingreso\ laboral_i = IMPA\ compilado_i + ISA\ compilado_i$$

De manera que, el modelo a estimar se especifica como:

$$Ingreso\ laboral_i = \beta_0 + \beta_1 Edad_i + \beta_2 Edad_i^2 + \eta_{1,i} \quad (1)$$

Tal definición garantiza que, condicional en los datos; por una parte, se están teniendo en cuenta todas las fuentes de ingreso laboral de un individuo, lo cual es imperante si se tiene en cuenta que casi el 50% de la muestra trabaja semanalmente más de las 48 horas de un trabajo tiempo completo; y, por otro lado, intencionalmente se están excluyendo otras fuentes de ingreso que no están directamente relacionadas con el trabajo -motivado por un salario-, como la venta de activos, las transferencias o los intereses del ahorro. Ciertamente, esto permite estudiar con precisión la influencia de la edad de un trabajador sobre sus ingresos por actividades laborales.

Tabla 3. Estadísticas descriptivas del ingreso laboral.

VARIABLE	Obs.	Media	Desviación	Mínimo	Mediana	Máximo
Ingreso laboral	16542	1608618	2346376	0	991333	5000000

Antes de realizar cualquier estimación, se debe hacer hincapié comportamiento de la variable de interés que, en este caso, presenta valores atípicos hacia la cola derecha de la distribución. Así como se evidencia en la Tabla 3, la media (≈ COP $1,6 millones) es superior a la mediana (≈ COP $1,0 millones) y el mínimo esta a 0,42 desviaciones estándar de la mediana, mientras que el máximo está a 22,37 desviaciones estándar de la mediana. Por tanto, para refinar el proceso de MCO de la regresión se decidió truncar a la variable dependiente hasta valores correspondientes a su percentil 95 (COP $5,0 millones).

Tabla 4. Incidencia de la edad sobre el ingreso laboral.

	(1)
VARIABLES	Ingreso laboral
Edad	$120554,48^{***}$
	(4903,83)
$Edad^2$	$-1414,98^{***}$
	(59,243)
Constante	$-900640,75^{***}$
	(94855,15)
Observaciones	16542
R^2	0,036
Estadístico de Fisher	310,2***
rECM	1176051
Controles	No
Otros	No

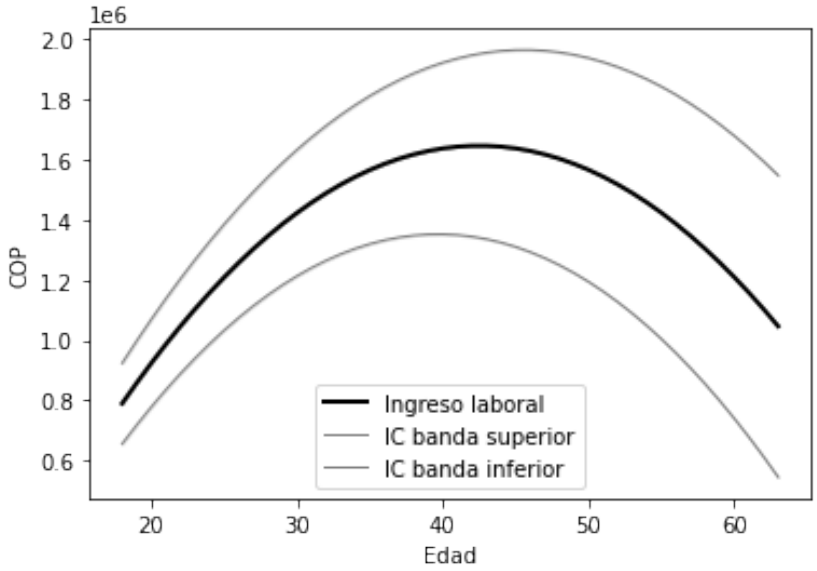
Errores estándar entre paréntesis.
***p<0,01, **p<0,05, *p<0,1

La Tabla 4 exhibe los resultados de estimar el modelo (1) por Mínimos Cuadrados Ordinarios (MCO). De acuerdo con el estadístico R^2, el modelo explica tan solo en un 3,6% la variación de los ingresos laborales y la raíz cuadrada del Error Cuadrático Medio (rECM) muestra que en promedio la estimación dentro de muestra del ingreso laboral va a estar COP $1,18 millones desfasada de su valor observado. Sin embargo, la significancia global de este modelo es robusta, debido a que el Estadístico de Fisher presenta un elevado valor de 310,2 y su *p-value* es del 0,0%. Lo cual implica que, aunque las variables explicativas utilizadas no le permiten al modelo tener un ajuste dentro de muestra deseable, en conjunto estas si son muy relevantes a la hora de intentar explicar el comportamiento del ingreso laboral.

Más en detalle, el coeficiente asociado a la edad lineal es estadísticamente significativo al 1% y sugiere que, a partir de los 18 años, un (1) año adicional en edad incrementa en COP $120.554 el ingreso laboral; resultado que es consiste con la intuición económica ya que mientras más años de ocupación laboral tenga un individuo, se tendrá una mayor acumulación de capital humano (mediante una mayor experiencia en el oficio o un mayor tiempo de educación) y esto se reflejará en una mayor productividad marginal del trabajo, lo cual -en principio- permitirá que mejore el ingreso laboral de individuos con mayor edad.

Note que el coeficiente asociado a la edad cuadrática es estadísticamente significativo al 1%. Su signo, negativo, implica que la edad tiene rendimientos marginales decrecientes sobre el ingreso laboral por lo que es una función cóncava (presenta un máximo global); es decir, que existe un nivel de edad óptimo (edad*) donde el ingreso laboral se maximiza. Intuitivamente este resultado tiene sentido económico, ya que mientras un individuo va a entrando en una edad avanzada hay presiones biológicas que van mermando sus capacidades físicas y cognitivas, lo que tiene un efecto negativo en sobre la productividad marginal del trabajo y sobre la remuneración laboral. En consecuencia, el nivel de edad óptimo para que un individuo maximice su ingreso laboral, corresponderá a aquella edad en la cual mientras se presenta una elevada acumulación de experiencia y capital humano, las presiones físicas del envejecimiento no son muy significativas.

Figura 1. Relación entre la edad y el ingreso laboral.



La Figura 1 complementa el análisis anterior de manera gráfica pues es posible dilucidar que a los 42,6 años se alcanza el máximo nivel de ingreso laboral en cerca de COP $1,67 millones. En los momentos previos a este punto se observa un acelerado ascenso del ingreso durante los primeros años de trabajo, para luego comenzar a caer rápido cuando se superan los 50 años de vida. Adicionalmente, al incluir intervalos de confianza al 95% estimados a través del uso de errores estándar calibrados con 1.000 repeticiones de *bootstrap*: hay evidencia estadística suficiente para afirmar que el nivel de ingreso en su punto máximo ($IC_{95}^{43} = [1.18, 2.11]$) va a ser superior al nivel de ingreso laboral a los 18 años ($IC_{95}^{18} = [0.50, 1.08]$).

Así pues, se concluye que el modelo (1) ayudó a esbozar que la edad sí posee una relación estadística significativa de carácter cóncavo respecto al ingreso laboral que percibe el individuo. Pero, surgen cuestionamientos acerca de de la capacidad de esta forma funcional para ajustarse a las variaciones dentro de muestra de la variable dependiente y de predecir los ingresos laborales de una persona. La siguiente sección evalúa nuevas especificaciones con otras baterías de variables explicativas, con el objetivo de encontrar otros factores determinantes que permitan mejorar la capacidad tanto explicativa como predictiva del modelo sobre la variable de interés.

4 Brecha Salarial

Otro elemento que desde la literatura relacionada y el *policy-making* se considera que se debe tener en cuenta cuando se busca tanto explicar como predecir los ingresos de un individuo corresponde a su género. En Colombia, la población femenina cuenta con menor remuneración laboral en comparación con los hombres, a pesar de tener los mayores niveles de educación y un aumento en la participación el los años recientes.

Si bien el país ha tenido ciertos avances en la transición de la educación, formación e intermediación del sector femenino i.e., la Ley 1496 de 2011 –la cual tiene como objetivo garantizar la igualdad salarial entre hombres y mujeres– que fija los mecanismos necesarios para que dicha igualdad sea real, los resultados aún no son significativos. A la luz de los estudios de la CEPAL (2017) esta incongruencia puede verse reflejada por circunstancias culturales o factores como el cuidado infantil (teniendo presente el enfoque académico de la mujer); sin embargo, otros estudios por parte de DANE (2006) apuntan hacia un problema estructural desde el ecosistema del empleo inclusivo y los actores institucionales implicados.

Dicho lo anterior, la presente sección buscará evaluar si en el mercado laboral de Bogotá D.C. el género es un factor determinante que, tanto condicional como incondicionalmente, genera brechas en la remuneración laboral de los individuos. En línea con lo comentado en la sección 3, la distribución de la variable de ingreso laboral sigue estando concentrada hacia la cola izquierda, con valores atípicos (*outliers*) en la cola derecha. Así, se le realizó una transformación logarítmica a la variable dependiente, $\log(\textit{Ingreso laboral})_i$, con el objetivo de mejorar su comportamiento para cumplir con supuestos clásicos de la regresión lineal como la normalidad de los errores y facilitar la interpretación por medio de una especificación semi-elástica en la variable dependiente. De manera que, el modelo a estimar se especifica como:

$$\log(\textit{Ingreso laboral})_i = \beta_0 + \beta_1 Mujer_i + \eta_{2,i} \tag{2}$$

Tabla 5. Brecha de género sobre el ingreso laboral.

	(2)
VARIABLES	$\log(\textit{Ingreso laboral})$
Mujer	$-0,2345^{***}$
	(0,013)
Constante	$13,958^{***}$
	(0,009)
Observaciones	16542
R^2	0,019
Estadístico de Fisher	$316,0^{***}$
rECM	0,839
Controles	No
Otros	No

Errores estándar entre paréntesis.
***p<0,01, **p<0,05, *p<0,1

La Tabla 5 muestra los resultados de estimar el modelo (2) por MCO de la brecha de género sobre el nivel de ingreso laboral. Este nuevo modelo no equivale a una mejora significativa en el ajuste con respecto al (1), dado que solo explica en un 0,19% la variación del logaritmo de los ingresos laborales y su rECM es de 0,839 (representando un desfase promedio entre observación y estimación de $\approx$ COP \$1,25 millones). Sin embargo, posee significancia global dado por el Estadístico de Fisher (316,0) el cual es significativo al 1%, por lo tanto esta nueva especificación sí posee información relevante para entender el comportamiento del ingreso laboral.

Fíjese que el coeficiente asociado a la variable indicadora de género es estadísticamente significativo al 1% y sugiere que, en promedio, las mujeres en Bogotá ganan un 23,45% menos que los hombres. Por su parte, el coeficiente asociado a la constante del modelo también es estadísticamente significativo al 1%. Su magnitud captura el ingreso laboral promedio que posee un hombre al momento de la encuesta (≈ COP $1,15 millones). Luego, los resultados no condicionales de este modelo muestran que, el efecto diferencial de ser una mujer trabajadora en Bogotá implica una brecha salarial de género de COP $241.045,4. Las razones subyacentes que pueden estar detrás de ello se dividen en tres (3) hipótesis: i) productividades marginales del trabajo heterogéneas, esto es que por diferencias preexistentes en educación, habilidad, fuerza bruta o inteligencia, los hombres pueden estar accediendo a trabajos mejor remunerados (e.g. ejecutivos) o que requieren de capacidades innatas de su género; ii) concentración del ingreso, dado el facto de ser la capital del país, naturalmente los conglomerados financieros, las fábricas y empresas se radican allí por facilidad de acceso al crédito, los trámites y papeleos e, incluso, las oportunidades de crecimiento por la masa de población; iii) economía del cuidado, habría motivos para sospechar que, basado en las estadísticas descriptivas de los estratos, hay un fuerte efecto de las actividades no remuneradas del hogar que en alta medida llevan a cabo las mujeres sopesando que pueden ser madres cabeza de hogar.

Entonces, considerando que la sección 3 encontró que para la población bogotana la edad es una variable relevante para entender el comportamiento de los ingresos laborales, y que durante esta ejercicio se concluyó que también existe una brecha de género incondicional en la remuneración laboral. Se plantea una nueva especificación funcional que busca estudiar los efectos de ambas variables en conjunto (incluyendo sus interacciones) sobre la variable de interés. De manera que, el modelo a estimar se especifica como:

$$\log(\textit{Ingreso laboral})_i = \beta_0 + \beta_1 Edad_i + \beta_2 Mujer_i + \beta_3 Edad_i^2 + \beta_4 EM_i + \beta_5 E^2 M_i + \eta_{3,i} \quad (3)$$

Donde EM_i corresponde a la interacción entre la edad lineal y el género ($Edad_i \times Mujer_i$), y $E^2 M_i$ corresponde a la interacción entre la edad cuadrática y el género ($Edad_i^2 \times Mujer_i$).

Tabla 6. Efecto diferencial de la edad y género sobre el ingreso laboral.

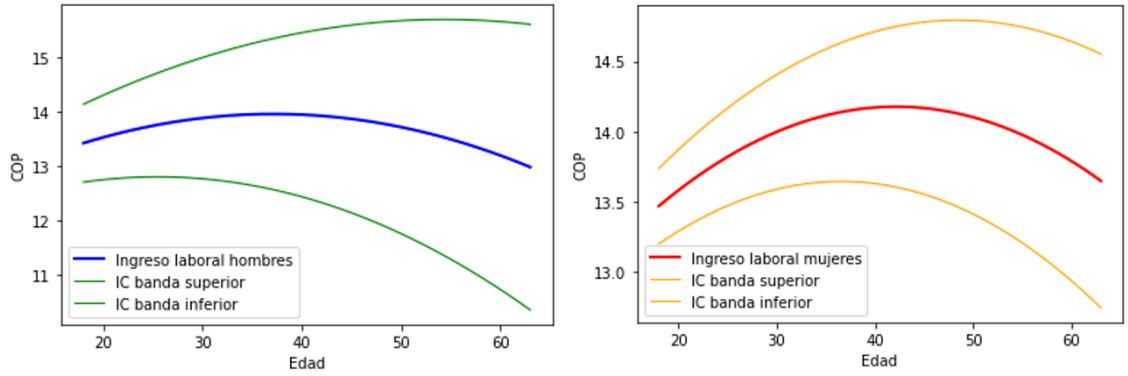
	(3)
VARIABLES	log(*Ingreso laboral*)
Edad	$0,1022^{***}$
	(0,005)
Mujer	$-0,0885$
	(0,132)
Edad^2	$-0,0012^{***}$
	(0,000)
Edad×Mujer	$0,0066$
	(0,007)
Edad^2×Mujer	$-0,0003^{***}$
	(0,000)
Constante	$12,025^{***}$
	(0,089)
Observaciones	16.542
R^2	0,089
Estadístico de Fisher	317,0***
rECM	0,809
Controles	No
Otros	No

Errores estándar entre paréntesis.
***p<0,01, **p<0,05, *p<0,1

La Tabla 6 evidencia el resultado de estimar el modelo (3) por MCO. Su ajuste dentro de muestra presenta una mejora pequeña relativa los modelos (1) y (2); ya que, el R^2 sube a 0,089 (se esta explicando en un 8,9% la variación del ingreso laboral) y el rECM se reduce a 0,809 (el desfase promedio entre el dato observado y el predicho es de ≈ COP $1,22 millones). Además, la significancia global del modelo sigue siendo siendo robusta con un Estadístico de Fisher igual a 317,0 y significativo al 1%. Lo cual corrobora la hipótesis de que al menos una de las variables explicativas propuestas (ya sea edad, género o alguna de sus interacciones) es relevante para entender el comportamiento del ingreso laboral.

Posteriormente, al revisar en minucia la significancia individual de cada variable, se llega a tres (3) conclusiones: i) la edad es una variable explicativa que continúa siendo robusta tanto en su componente lineal como en su componente cuadrático, significativa al 1% en ambos casos; ii) así como en el modelo (1), los signos de los coeficientes de edad evidencian una relación cóncava (y maximizable globalmente) entre los años de un individuo y su nivel de ingreso laboral, y; iii) la variable dicótoma de género pierde significancia individual cuando se incluye en una misma especificación con la edad, y además la única interacción que es significativa (con un signo negativo) corresponde a la interacción entre la dicótoma de género y edad cuadrática. Este resultado estaría indicando que, en Bogotá D.C., la brecha salarial de genero no es independiente del nivel de edad de una persona, es despreciable para los primeros años de trabajo, pero mientras los trabajadores se van volviendo mayores, el mercado laboral termina penalizando más rápido el envejecimiento de las mujeres que el envejecimiento de los hombres.

Figura 2. Relación diferencial de la edad y género sobre el ingreso laboral.



La Figura 2 incluye elementos que corroboran las conclusiones de presentadas en la Tabla 6, pero también hay características que contribuyen a poner en duda la significancia individual de la variable mujer sobre la determinación del ingreso laboral. En una visión, se puede observar que la senda del ingreso laboral para los hombres esta por encima de la de las mujeres, y que el punto máximo del ingreso laboral llega más temprano para las mujeres (≈ COP $1,10 millones a los 36,3 años) que para los hombres (≈ COP $1,47 millones a los 42,6 años). Sin embargo, luego de realizar la misma calibración de la sección anterior sobre los intervalos de confianza mediante 1.000 repeticiones de *bootstrap*, se encuentra que en ningún punto de la senda de *earnings* los hombres y las mujeres presentan ingresos laborales estadísticamente distintos. Entonces, surge la pregunta si es estrictamente necesario mejorar la especificación del modelo (3) para obtener resultados que permitan discernir si la variable dicótoma de mujer es estadísticamente relevante para entender el comportamiento del ingreso laboral en Bogotá D.C.

En este orden de ideas, es pertinente aclarar que el proceso de estimación del impacto de la edad y del genero sobre el ingreso laboral en los modelos (1), (2) y (3) corresponde a un proceso de estimación ingenuo. Dado que, los estimadores presentados representan al efecto promedio de estas variables explicativas dentro de toda la muestra, pero no están condicionados a la presencia de características relevantes de los individuos o del mercado laboral que pueden estar creando heterogeneidad. Esto implica que, las *gender wage gaps* encontradas en los modelos (2) y (3) representan brechas incondicionales entre los ingresos de todos los hombres y de todas las géneros en la muestra, pero no necesariamente son evidencia que permita generar conclusiones razonables acerca de la existencia de brechas salariales que rompan con la máxima social de *"equal pay for equal work"*. En consecuencia, se revisó si la brecha de género salarial entre hombres y mujeres persiste bajo conocimientos, desempeños laborales, sectores de trabajo y mercado similares. De manera que, el modelo a estimar se especifica como:

$$\log(\textit{Ingreso laboral})_i = \beta_0 + \beta_1 E_i + \beta_2 M_i + \beta_3 E_i^2 + \beta_4 EM_i + \beta_5 E^2 M_i + X\Gamma + \eta_{4,i} \qquad (4)$$

Donde X es una batería de controles agregada a la especificación del modelo (3), con el objetivo de evaluar si en el mercado laboral de Bogotá D.C. existe (o no) *"equal pay for equal work"*. En efecto, para probar la consistencia de los resultados encontrados mediante la estimación del modelo (4), se realizó la estimación de dos (2) modelos adicionales. En principio, el modelo (4.1) utilizará el Teorema de Frisch-Waugh-Lovell (FWL) para recalcular los coeficientes de las variables de interés presentadas en el modelo (3). Y, adicionalmente, el modelo (4.2) se usará para refinar los errores calculados por el modelo (4.1) mediante MCO.

Tabla 7. Efecto diferencial de la edad y género sobre el ingreso laboral bajo diferentes métodos de estimación.

	(4)	(4.1)	(4.2)
VARIABLES	log(*Ingreso laboral*)	log(*Ingreso laboral*)	log(*Ingreso laboral*)
Edad	$0,0539^{***}$	$0,0539^{***}$	$0,0539^{***}$
	(0,003)	(0,003)	(0,003)
Mujer	$-0,1583^{*}$	$-0,1583^{*}$	$-0,1583$
	(0,093)	(0,093)	(0,097)
Edad^2	$-0,0006^{***}$	$-0,0006^{***}$	$-0,0006^{***}$
	(0,00004)	(0,00004)	(0,00004)
Edad×Mujer	$0,0040$	$-0,0040$	$-0,0040$
	(0,005)	(0,005)	(0,005)
Edad^2×Mujer	$-0,0001^{*}$	$-0,0001^{*}$	$-0,0001$
	(0,000059)	(0,000059)	(0,000062)
Constante	$10,9644^{***}$	-	-
	(0,100)	-	-
Observaciones	16.542	16.542	16.542
R^2	0,554	0,043	0,043
Estadístico de Fisher	179,4***	145,5***	145,5***
rECM	0,568	0,568	0,568
Controles	Sí	Sí	Sí
Método	MCO	FWL	FWL
Bootstrap	No	No	Sí

Controles: (4),(4.1) y (4.2); tamaño de la firma, horas semanales de trabajo, formalidad, antigüedad, nivel educativo, relación laboral, oficio y mes.

Otros: (4.2) Errores estándar de *bootstrap* (1.000 repeticiones).

Notas: Bajo el método de estimación FWL, los modelos (4.1) y (4.2) no poseen intercepto. Errores estándar entre paréntesis.

***p<0,01, **p<0,05, *p<0,1

A partir de la Tabla 7, es posible dilucidar que ante la presciencia de controles, si bien la relación cóncava entre edad e ingreso laboral sigue siendo significativa al %1, la variable dicótoma de género termina presentando una débil significancia (del %10) y la magnitud del coeficiente de la interacción entre edad cuadrática y mujer pierde robustez. Inclusive, se puede evidenciar que en el modelo (4.2) la inclusión de errores estándar estimados mediante *bootstrap* llevó a descartar a la variable de género (junto con cualquiera de sus interacciones) de la lista de características significativas que permiten explicar el comportamiento del ingreso laboral de los individuos. Por lo tanto, es razonable concluir que: aunque si hay evidencia de la existencia de un *gender wage gap* no condicional dentro del mercado laboral de la capital colombiana, no hay suficiente evidencia estadística para afirmar que esta brecha existe de forma condicional al tipo de trabajo desempeñado, luego Bogotá no se estaría incumpliendo el principio de *"equal pay for equal work"*.

El mecanismo por el cual podría estar ocurriendo la pérdida de significancia estadística de la variable de género está ligado con el sesgo de selección. Esto es, porque la forma en cómo se logró recolectar la información del proceso de *data-scraping* presentó restricciones al acceso a más locaciones del país pues en un inicio se comentó en la sección 2 que gran parte de la muestra se acumulaba en los estratos más bajos. En este sentido, los grupos (hombres y mujeres) no son comparables cuando no se condiciona en las variables explicativas. Esto se conecta, de igual manera, con la discusión del origen de la brecha salarial pues Bogotá D.C. es una ciudad que no es representativa de Colombia al ser poco homogénea en condiciones laborales por lo que subsisten motivos para apuntar hacia nuestras hipótesis iniciales de la sección 4 pues en el modelo (3) el estimador de interés tampoco resultó estadísticamente significativo.

Para finalizar, vale la pena resaltar que la inclusión de la batería de controles mejoró el ajuste dentro de muestra de los modelos. En el modelo (4) el R^2 llegó hasta el 55,4%, lo que indica que la inclusión de las nuevas características realzó sustancialmente la capacidad del modelo de explicar las variaciones del ingreso laboral. También, aunque el Estadístico de Fisher en los modelos (4) y (4.1) bajó a 179,4 y 145,5 (respectivamente), su significancia se mantuvo por debajo del 1%. De hecho, y aun más importante, los rECM de los nuevos modelos (4) y (4.1) presentaron una reducción sustancial relativo los rECM de los modelos (2) y (3). En promedio, las predicciones del ingreso del modelo (4) están desfasadas en COP $863.124 y las del modelo (4.1) están desfasadas en COP $863.098, lo cual representa una mejoría potencial de al menos COP $360.000 en el ajuste de cada predicción dentro de muestra, con respecto al modelo (3).

5 Predicción de Ingresos

Las conclusiones extraídas de las secciones 3 y 4 permiten entender mejor como funciona el mercado laboral en Bogotá D.C., y esto provee insumos conceptuales para poder plantear el mejor modelo posible para predecir el ingreso de los individuos. En tales casos, según los resultados de los modelos (1) al (4.2), si se desea predecir de forma efectiva el ingreso laboral de un individuo, para detectar posible fraude en la recaudación de impuestos, un requisito es tener cuenta la naturaleza cóncava de la relación entre la edad y el ingreso laboral, omitiendo la variable dicótoma de sexo. Es en este orden de ideas que, se desarrollaron seis (6) especificaciones adicionales que se ocupan de potenciales relaciones no lineales e interacciones existentes entre las variables de la base de datos (*using*), a fin de elaborar un modelo de predicción laboral con el menor rECM posible (aquel que minimiza la función de pérdida).

De manera que, los modelos a estimar se especifican como:

$$\log(\textit{Ingreso laboral})_i = \beta_0 + \beta_1 E_i + \beta_2 E_i^2 + \beta_3 H_i + \beta_4 H_i^2 + \beta_5 EH_i + \beta_6 EH_i^2 + Z\Lambda + \eta_{5,i} \quad (5)$$

Donde H_i corresponde a las horas de trabajo semanales, H_i^2 a las horas de trabajo semanales cuadrática, EH_i a la interacción entre la edad lineal y las horas de trabajo semanales ($Edad_i \times Horas_i$), E^2H_i corresponde a la interacción entre la edad y horas de trabajo semanales ($Edad_i \times Horas_i^2$) y Z una batería de controles ($\neq X$) que excluye aquellas variables dentro del modelo.

$$\log(Ingreso\ laboral)_i = \beta_0 + \beta_1 E_i + \beta_2 E_i^2 + \beta_3 M_i + \beta_4 D_i + \gamma R + \alpha S + Z\Lambda + \eta_{6,i} \quad (6)$$

Donde R corresponde al vector de variables dicótomas del régimen de salud, S al vector de variables dicótomas del estrato socio-económico, D_i es la variable dicótoma sobre la disposición a trabajar más horas durante la semana.

$$\log(Ingreso\ laboral)_i = \beta_0 + \beta_1 E_i + \beta_2 E_i^2 + \beta_3 M_i + \beta_4 D_i + \gamma MEduc + \alpha MRela + Z\Lambda + \eta_{7,i} \quad (7)$$

Donde $MEduc$ corresponde a la interacción entre el género y el vector de variables dicótomas del nivel educativo ($Mujer_i \times Educ$) y $MRela$ corresponde a la interacción entre el género y el vector de variables dicótomas del tipo de relación laboral ($Mujer_i \times Rela$).

$$\log(Ingreso\ laboral)_i = \beta_0 + \beta_1 E_i + \beta_2 E_i^2 + \beta_3 D_i + \beta_4 G + \beta_5 H_i + \beta_6 GH_i + Z\Lambda + \eta_{8,i} \quad (8)$$

Donde G_i corresponde a la variable dicótoma sobre otros ingresos por juegos de azar, venta de activos y vehículos, GH_i a la interacción entre otros ingresos por juegos de azar, venta de activos y vehículos, y las horas de trabajo semanales ($Gambling_i \times Horas_i$).

$$\log(Ingreso\ laboral)_i = \beta_0 + \beta_1 E_i + \beta_2 E_i^2 + \beta_3 D_i + \beta_4 A_i^2 + \beta_5 A_i^3 + \sum_{j=1}^{3} (\gamma_j EA^j) + Z\Lambda + \eta_{9,i} \quad (9)$$

Donde A_i corresponde a los meses de antigüedad en el trabajo y EA_i^j a la interacción entre la edad lineal y antigüedad elevada al j-ésimo elemento de la sumatoria ($E_i \times A_i^j$).

$$\log(Ingreso\ laboral)_i = \beta_0 + \beta_1 E_i + \beta_2 E_i^2 + \gamma EEduc + Z\Lambda + \eta_{10,i} \quad (10)$$

Donde $EEduc$ corresponde a la interacción entre la edad lineal y el vector de variables dicótomas del nivel educativo ($Edad_i \times Educ$).

En las secciones 3 y 4, únicamente se utilizó al rECM dentro de muestra como criterio de información para evaluar el ajuste del modelo. No obstante, esto puede ser problemático debido a que, por si solo el rECM no es capaz de evaluar la existencia de potenciales problemas de *overfitting* del modelo dentro de la muestra. Lo cual implica que, si se llegara a escoger un modelo a partir de este criterio, se corre el riesgo de seleccionar una especificación que se comporta bien dentro de la muestra seleccionada, pero que no funciona bien a la hora de predecir cuando se le presentan nuevos datos. Por ende, para evaluar el ajuste de estas diez (10) especificaciones, se dividió aleatoriamente la muestra entre un 70% para el *training* de los modelos y un 30% para el *testing* de estos. El criterio rECM sobre la muestra de *testing* va a permitir evidenciar cuan efectivo es cada especificación al momento de tener que enfrentarse a un problema de predicción por fuera de muestra (validez externa/replicabilidad poblacional del modelo).

Se decidió utilizar el criterio de información rECM por dos (2) razones concretas: i) esta medida de riesgo es consistente con una función de perdida cuadrática y, por lo tanto, sus evaluaciones también son consistentes con el funcionamiento subyacente del proceso de estimación de modelos mediante MCO (en el cual se busca minimizar los errores de estimación al cuadrado), lo que lo vuelve el criterio óptimo para evaluar modelos desarrollados mediante la metodología de estimación seleccionada, y; ii) porque a diferencia del ECM tradicional, el rECM mejora la interpretabilidad de los errores de estimación, gracias a que transforma las unidades al cuadrado de ECM a las mismas unidades de la variable de interés (COP).

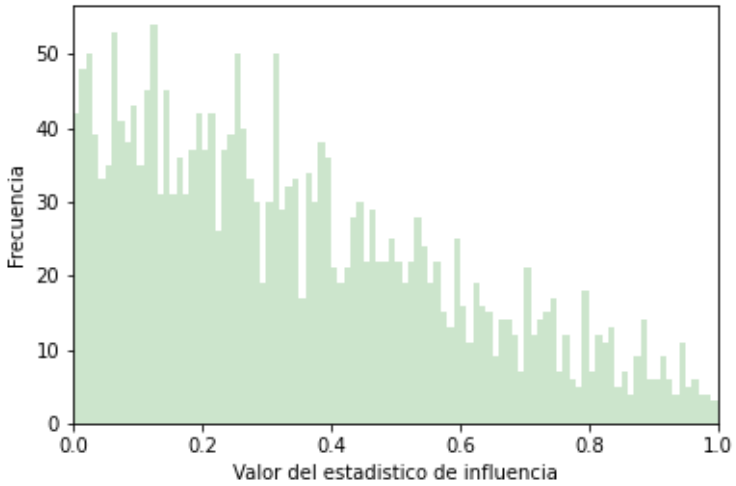
Tabla 8. Evaluación fuera de muestra de las especificaciones.

Especificación	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
rMSE	0,838	0,852	0,821	0,579	0,567	0,605	0,574	0,580	0,582	0,581

Nota: El rMSE reportado fue evaluado en la muestra de *testing* y se encuentra en unidades logarítmicas para facilitar su visualización.

Luego de realizar la evaluación por fuera de muestra, en la Tabla 8 se puede evidenciar que los dos (2) mejores modelos para predecir el ingreso laboral, (5) y (7), corresponden a estimaciones nuevas que no fueron trabajadas durante las secciones anteriores del trabajo. El modelo (5), que es el de mejor desempeño relativo, dejó de lado a la variable de género y agregó un componente lineal y uno cuadrático de la variable horas trabajadas a la semana, con la intención de capturar posibles no linealidades entre la cantidad de trabajo utilizado y su valor marginal, e incluyó interacciones entre las horas trabajadas a la semana y la edad, bajo el razonamiento de que el numero de horas que una persona puede trabajar esta relacionado con sus capacidades físicas y biológicas. El modelo (7), *second best*, manejó un razonamiento distinto al modelo (5), ya que decidió profundizar mucho más en las posibles interacciones que pueden existir entre la variable de género y otros posibles determinantes del ingreso como lo son el nivel educativo y la relación laboral, que pueden presentar una correlación más fuerte con la variable de interés, y agregó otra variable que representa la disposición a trabajar más durante la semana, lo cual puede corresponder a un buen predictor de individuos que buscan obtener mayor ingreso laboral.

Figura 3. Distribución del Estadístico de Influencia (α).



Ahora bien, con el objetivo de evaluar la relación existente entre las observaciones y los valores predichos por el modelo (5), se realizó la Figura 3, que presenta la distribución de los estadísticos de influencia calculados sobre la porción de datos de testeo. El gráfico sugiere que existe una gran concentración de observaciones para las cuales el estadístico de influencia se ubica entre cero (0) y 0,4. Lo anterior quiere decir que el grueso de la muestra aporta de manera similar a la construcción de los estimadores del modelo. Sin embargo, existe una proporción de observaciones cuyo valor del estadístico de influencia es superior a 0,6, incluso superando el umbral de 0,9, lo cual indica que un grupo de *outliers* si presenta una relevancia importante en la estimación. Por tanto, la Dirección de Impuestos y Aduanas Nacionales (DIAN) debería concentrarse en los *outliers*, pues es probable que posean características especiales en las variables utilizadas para la regresión que generen una distorsión en las estimaciones.

Para finalizar, se decidió realizar una estimación más robusta de los errores de predicción fuera de muestra para los dos (2) modelos con mejor capacidad predictiva, el (5) y el (7). Esto, con la intención de poder tomar una decisión final bien informada acerca de cual se debería utilizar para identificar episodios de fraude fiscal en Bogotá. La metodología utilizada para la toma de decisión corresponde a la *Leave-one-out-Cross-calidation* (LOOCV), la cual permite realizar estimaciones por fuera de muestra para cada una de las observaciones y posteriormente extrae el rMSE de cada una. Al final esto permite calcular un promedio del rMSE por fuera de muestra teniendo en cuenta todas las observaciones de la muestra utilizada.

Tabla 9. *Leave One Out Cross Validation (LOOCV).*

Especificación	(5)	(7)
rMSE	1,099	0,412

Nota: El rMSE reportado está en unidades logarítmicas para facilitar su visualización.

Dada la Tabla 9, el modelo con menor rMSE resultó ser de la especificación (7); es decir, cuando se hizo $LOOCV$ regresando todas las observaciones se encontró que este brinda el ajuste más "poderoso" tanto dentro como fuera de muestra. Lo que quiere decir en términos económicos que gran parte de la variación de los ingresos laborales está siendo explicada por el género, que debe incluirse con obligatoriedad, y las redes preexistentes del individuo. A saber, con quien está ligado en sus relaciones laborales, qué contactos tiene, entre otros.

6 Conclusiones

En síntesis, este trabajo estudió las razones subyacentes del fraude fiscal y la brecha salarial de género intríseca en el mercado laboral de Bogotá D.C. a partir de una muestra obtenida por *data-scraping* de la GEIH para el año 2018. Allí se encontró evidencia estadística para corroborar que la edad es una función cóncava respecto al ingreso por lo que este se maximizaba a la edad de 42,6 años. Este valor fluctuaba entre géneros y se acrecentaba con el pasar del tiempo, resultado consistente con la literatura relevante en la materia. Además, se mostró que, tanto condicional como no condicional en las variables de control (características laborales), las mujeres en Bogotá D.C. no poseen brecha salarial respecto a los hombres lo cual podría deberse a un eventual sesgo de selección, pero sus márgenes de distancia de ingreso laboral si crecen a lo largo del tiempo. Por último, dadas varias especificaciones para predecir el ingreso se acreditó a un modelo en particular cuya función de pérdida era la mínima relativo al resto; donde se encontró que la gran mayoría de las observaciones no son de gran influencia, pero hay individuos que requieren atención especial de la DIAN porque presumiblemente evaden impuestos.

Bibliografía

CEPAL. (2017). Progreso y evolución de la inserción de la mujer en actividades productivas y empresariales en América del Sur. *Revista CEPAL*, Número 122, pp. 417-423.

DANE. (2006). Neighborhood size selection in the k-nearest-neighbor rule using statistical confidence. *Princeton University Press*.

Javaheri, B. (2021). Knn with Examples in Python. *DOMINIO Datalab*. (Disponible en: https://www.dominodatalab.com/blog/knn-with-examples-in-python)

Mwangi, A. (2010). Implementing K Nearest Neighbors with Python. *DataDrivenInvestor*. (Disponible en: https://medium.datadriveninvestor.com/implementing-k-nearest-neighbors-with-python-1c0b7cdf85f2)

Ramírez, A., Vargas-Correa, R., García, G., & Londoño, D. (2020). Optimal selection of the number of control units in knn algorithm to estimate average treatment effects. *Cornell University*, arXiv preprint arXiv:2008.06564.

Wang, J., Neskovic, P., & Cooper, L. (2017). Mercado laboral por departamentos. *Boletín Técnico*, Volumen 39, Edición 3, pp. 417-423.