Universidad de los Andes
Maestría Economía Aplicada: Big Data and Machine Learning for Applied Economics
Grupo: Laura Natalia Capacho (202121025), Sebastián David Beltrán (202121021) y Yurani Gonzalez (201212100)

GitHub URL: https://github.com/sbeltro/G10_PS1

Problem Set 1: Predicting Income

El objetivo es construir un modelo predictivo para el ingreso individual, a partir de la Gran Encuesta Integrada de Hogares del DANE para 2018 en la ciudad de Bogotá.

$$Ingreso = f(X) + u$$

1. Adquisición de datos

(a) El proceso de adquisición de datos está descrito en el *script* de R adjunto al documento.

(b) No hay restricciones para acceder a la información puesto que no existe ningún archivo (robots.txt) asociado a la raíz del sitio web que indique restricciones para rastrear la información de la página. Sin embargo, al momento de realizar el *web scraping* para obtener las bases de datos, nos enfrentamos a un problema debido a que la página web es dinámica y no estática. Por lo anterior, no es posible extraer las tablas de forma directa, en la medida que la información no carga inmediatamente. Las tablas realmente provienen de una promesa *json*, asociada a un link externo que es de donde pudimos extraer los datos.

(c) Para adquirir los datos seguimos el siguiente proceso:

1. Crear una lista vacía "Links" en la que se van a guardar los links de cada una de las páginas que contienen las tablas con la información
2. Crear una lista vacía "base" en la que se van a almacenar las tablas que se extraen de cada página web
3. Crear un *loop* que para i = 1, 2, …, 10, ejecuta los siguientes pasos:
   3.1. Guardar en la lista "Links" cada uno de los links de las 10 páginas web, definiéndolo como un documento de *.html*
   3.2. Descargar y preprocesar las páginas web descargadas de cada link (*read_html)*
   3.3. Convertir los nodos de tipo tabla de la página web en tablas de R (*html_table)*
   3.4. Almacenar cada una de las tablas en la lista "base"
4. Renombrar la primera columna de todas las bases, que está vacía, para evitar errores
5. Convertir cada una de las tablas en la lista "base" a un formato *tibble*
6. Hacer un *append* de las 10 bases almacenadas en la lista "base", para obtener una base de datos completa con toda la información

2. Limpieza de datos

Luego de hacer una exploración de las variables disponibles en la base de datos se hizo una selección de aquellas que consideramos más relevantes para explicar el ingreso individual de

una persona. Destacamos que en este trabajo se hace referencia al ingreso total, constituido tanto por ingresos laborales provenientes de salarios o independientes, como por ingresos de ayudas, subsidios, bonificaciones y diversas fuentes que serán detalladas más adelante.

En este sentido, los determinantes del ingreso son tanto características de la persona o su hogar, como del tipo de trabajo que desempeñe (rama de actividad económica, condiciones de formalidad, tipo de empresa, entre otros). Así como otros factores que pueden aumentar la probabilidad de recibir otros ingresos (subsidios o auxilios, pensiones, entre otros).

Variables de interés

En primer lugar, se tomó un conjunto de características sociodemográficas de los individuos: **edad, edad al cuadrado** (para capturar el efecto decreciente de la edad sobre el ingreso), **género, máximo nivel educativo, relación con el jefe del hogar** y **estrato**. Dentro de estas se hicieron algunas modificaciones y se crearon nuevas variables:

1. Creamos una variable de **educación** medida en años que construimos a partir del máximo nivel educativo alcanzado que se reportó en la encuesta. En efecto, si este nivel es ninguno se asignaron 0 años, si es preescolar 3 años, si es primaria incompleta 7 años, si es primaria completa 8 años, si es secundaria incompleta 13 años, si es secundaria completa 14 años y si es terciaria 19 años.
2. Construimos un *proxy* de **experiencia laboral**[1]. Debido a que no se tiene información reportada de los años de experiencia laboral, utilizamos el concepto de "experiencia potencial", que se crea a partir de la edad, los años de educación y los años de iniciación en el mercado laboral (Aristizábal & Ángel, 2017).
3. Generamos una variable *dummy* que caracteriza si la persona es **jefa del hogar** o no.

En segundo lugar, se consideraron un conjunto de variables correspondientes al trabajo del individuo: trabajador **asalariado o independiente, formalidad, oficio** y **tamaño de la empresa** en que labora. En este caso también se realizaron modificaciones y se crearon variables:

1. Se genera una variable *dummy* que indica si la persona trabaja en una **microempresa** o no, a partir del tamaño de la empresa, medido por el número de trabajadores. En concreto, se denotan como microempresas aquellas compañías que tengan personal no superior a 10 trabajadores (Ministerio de Comercio, Industria y Turismo [MINCIT], 2007).

Finalmente, se toman todas las variables referentes a **ingresos totales** de la persona: **ingreso laboral** (salario o independiente) y **otros ingresos**, conformados por: ingreso por intereses o dividendos, por jubilaciones o pensiones, de ayudas de hogares e instituciones, por arriendos, por horas extra, por bonificaciones, por auxilios alimentarios o de transporte, por subsidios familiares o educativos, por primas de servicios, navidad o de vacaciones, por pensiones

[1] Para las personas con educación se utiliza la siguiente aproximación de experiencia (X): Si 18 < edad < 22, X = edad – 18; si edad > 22, X = edad – educación – 6. Y para personas sin educación terciaria se aproxima como sigue: si edad > 18, X = edad – 18.

alimentarias, viáticos, accidentes o de cualquier otra fuente. En este conjunto de variables se hicieron algunas agrupaciones:

1. Se crea una variable de **ayuda de hogares** a partir de la suma de variables que denotan dinero recibido de otros hogares o personas residentes en el país y fuera del país.
2. Se crea la variable de **primas** que agrupa los ingresos por primas de servicios, navidad y de vacaciones.

Tratamiento de valores faltantes

En el procesamiento de los datos encontramos que hay muchas observaciones con datos faltantes (*missings)*, en este sentido, dimos un tratamiento especial a las variables correspondientes de acuerdo con ciertos criterios. En principio se dio un tratamiento a las variables existentes en la base de datos, a continuación, se detalla el procesamiento:

- La variable **oficio** es una variable categórica que denota la ocupación de la persona y toma valores de 1 a 99. En este caso los *missings* se reemplazaron por 0, que corresponde a una nueva categoría de ocupación que agrupa "otras ocupaciones".
- La variable **formal** es una variable categórica que toma el valor de 1 si se trata de un trabajador formal y 0 en otro caso. Para los *missings* en este caso se asignó el valor de la variable p6090 (¿Está afiliado, es cotizante o es beneficiario de alguna entidad de seguridad social en salud?). Lo anterior, puesto que el empleo informal se refiere a los trabajadores que, entre muchos factores, pertenecen a una empresa o desempeñan un trabajo sin contrato laboral y sin aportes a seguridad social (Departamento Administrativo Nacional de Estadística [DANE], 2009).
- La variable **iof1es** corresponde al ingreso por intereses y dividendos imputado, y para sus valores faltantes se asignó el valor de la variable iof1 que es la misma variable, pero antes de imputación.
- La variable **iof2es** corresponde al ingreso por jubilaciones y pensiones imputado, y para sus valores faltantes se asignó el valor de la variable iof2 que es la misma variable, pero antes de imputación.
- La variable **y_vivienda_m** representa la renta por vivienda, en este caso a los *missings* se les asignó el valor reportado en la variable p7500s1 (Valor asociado a ¿El mes pasado, recibió pagos por: a. arriendos de casas, apartamentos, fincas, lotes, vehículos, equipos etc.?)
- Las variables **y_horasExtras_m**, **y_bonificaciones_m**, **y_auxilioAliment_m**, **y_auxilioTransp_m**, **y_subFamiliar_m**, **y_subEducativo_m**, **y_primaServicios_m**, **y_primaNavidad_m**, **y_primaVacaciones_m**, **y_primas_m**, **y_viaticos_m**, **y_accidentes_m**, **y_total_m** son ingresos por conceptos de horas extra, bonificaciones, auxilios de alimentación y transporte, subsidios familiar y educativo, primas de servicios, navidad y vacaciones, primas totales, viáticos, accidentes e ingreso salarial y de independientes, respectivamente. Para los valores faltantes en estas variables se asignó el valor cero (0). Lo anterior, partiendo del supuesto de que, si no hay valor en estos ingresos, es porque no existió o fue igual a cero.

Posteriormente, procedimos a crear algunas variables de interés, y así mismo se hizo un tratamiento sobre sus valores faltantes.

- La variable **Micro_empresa** que creamos para denotar si la persona trabaja en una microempresa o no. En este caso, se decidió eliminar las observaciones para las cuales esta variable tenía *missings*, esto puesto que no tenemos información suficiente para asignar a microempresa o no arbitrariamente.
- En la variable **educ** que creamos y mide el nivel de educación en años, utilizamos el máximo nivel de educación, y en ese caso solamente se encontró un valor faltante, al cual se le asignó el valor de cero (0).

Estadísticas descriptivas

- La Tabla 1 presenta las principales estadísticas descriptivas de los datos. Contamos con una muestra de 16,397 observaciones para todas las variables. Encontramos que el ingreso promedio es de COP2,698,186 millones (mn), con una desviación estándar de COP5,144,721 mn y se observa un valor máximo atípico de más de COP160 mn. Del mismo modo, encontramos que el 47% de la muestra son mujeres y 48% jefes de su hogar, y en cuanto a caracteristicas laborales, 31% son trabajadores independientes, 59% son formales y 51% trabajan en microempresa.

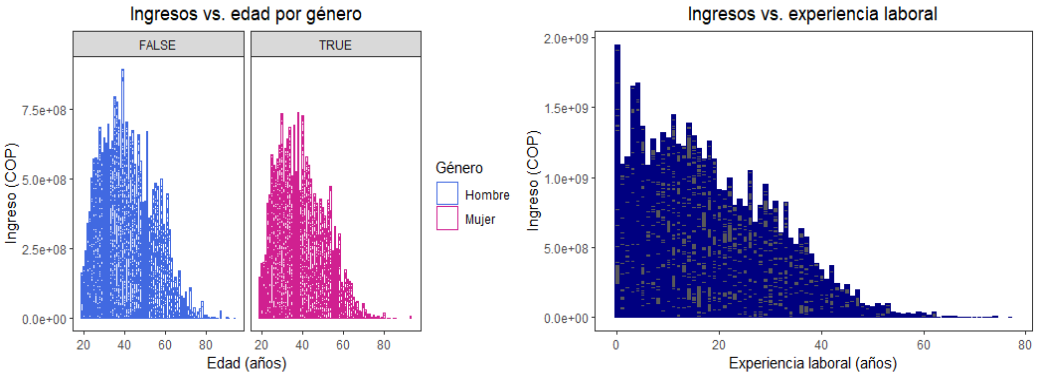
Tabla 1. Estadísticas descriptivas de las variables*

Variable	N	Media	Desv. Est.	Min	Máx
Ingreso	16,397	2,698,186	5,144,721	0	160,833,333
Edad	16,397	40	13	19	94
Educación	16,397	15	4	0	19
Experiencia	16,397	19	15	0	77
Oficio	16,397	50	28	1	99
Estrato	16,397	3	1	1	6
Ingreso laboral	16,397	1,451,278	2,359,661	0	70,000,000
Otros ingresos	16,397	1,127,691	3,674,663	0	130,000,000
Mujer (=1)	16,397	7715	47.1%	-	-
Cuenta propia (=1)	16,397	5080	31.0%	-	-
Formal (=1)	16,397	9676	59.0%	-	-
Jefe hogar (=1)	16,397	7872	48.0%	-	-
Microempresa (=1)	16,397	8443	51.5%	-	-

*Para las variables categóricas la media corresponde al número de personas en la muestra que toman el valor de 1 en cada variable. La desviación estándar indica la proporción sobre el total.

Luego de armar la base de datos y analizar las estadísticas descriptivas principales, se procedió a explorar cómo se caracterizan los individuos en la muestra utilizada. En la Gráfica 1 se observa que los hombres tienden a tener ingresos más altos que las mujeres a una edad más temprana. En ambos casos, el ingreso tiene un declive con el aumento de edad en alrededor de los 46 años. Además, se demuestra un comportamiento similar para la experiencia laboral, donde a mayor número de años el ingreso aumenta y luego comienza a decrecer después de los 20 años.

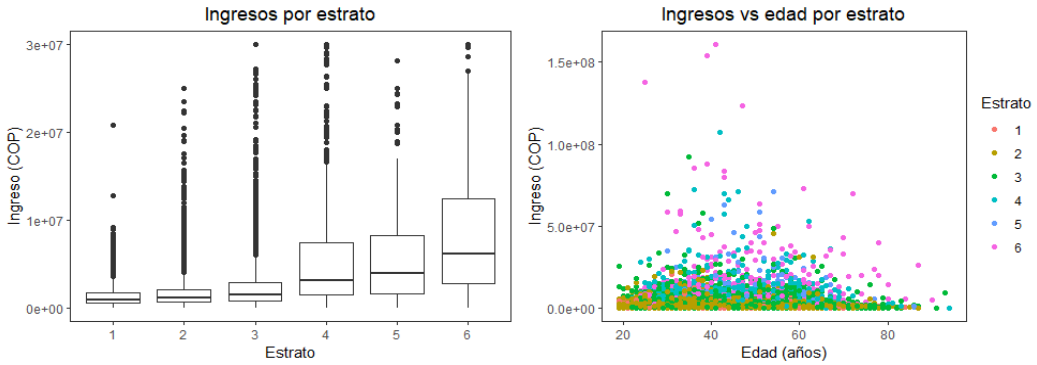
Gráfica 1. Ingresos vs. edad y experiencia laboral



Fuente: GEIH 2018, construcción propia.

La Gráfica 2 muestra el comportamiento del ingreso por estrato, este tiene una media similar entre estratos 1-2, y a partir del estrato 3 comienza a aumentar el ingreso promedio al igual que la varianza, indicando observaciones atípicas. Asimismo, a pesar de la edad, el estrato es factor determinante, estratos más altos reciben más ingresos incluso con la misma edad.

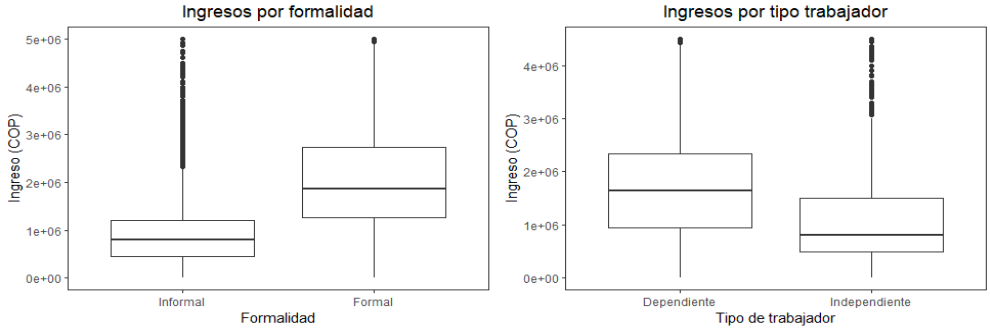
Gráfica 2. Ingresos por estrato socioeconómico



Fuente: GEIH 2018, construcción propia.

Igualmente, la formalidad y tipo de trabajador demuestran ser factores importantes en la varianza de ingresos, es decir, aquellos trabajadores con un trabajo formal o que son dependientes, tienen mayores ingresos de aquellos que no lo son (ver Gráfica 3).

Gráfica 3. Ingresos por tipo de trabajo (formalidad)



Fuente: GEIH 2018, construcción propia.

3. Perfil de edad-ingreso *(Age-earnings profile)*

Como se enunció anteriormente, en este documento se hace referencia al ingreso total de los trabajadores. De esta forma, se analiza un ingreso que está constituido tanto por ingresos laborales provenientes de salarios o independientes, como por ingresos de ayudas, subsidios y otras fuentes.

En este sentido, la variable de **Ingreso** que se utilizará para las estimaciones se construyó sumando: **ingreso laboral** (salario o independiente) y **otros ingresos**, conformados por: ingreso por intereses o dividendos, por jubilaciones o pensiones, de ayudas de hogares e instituciones, por arriendos, por horas extra, por bonificaciones, por auxilios alimentarios o de transporte, por subsidios familiares o educativos, por primas de servicios, navidad o de vacaciones, por pensiones alimentarias, viáticos, accidentes o de cualquier otra fuente.

La selección de las variables se fundamenta en que nuestro interés está en aproximar el ingreso total del individuo, proveniente de cualquier fuente, tanto laboral como no laboral. Con esto, las variables seleccionadas agrupan todas las posibles fuentes de ingreso monetario de la persona, dentro de las variables disponibles en la base.

Basados en esta aproximación, se realiza la estimación del siguiente modelo de perfil edad-ingreso, y los resultados se observan en la Tabla 2:

$$Ingreso = B_1 + B_2 edad + B_3 edad^2 + u$$

Tabla 2. Modelo: perfil edad-ingreso

	Ingreso
	(1)
Intercepto	-1,317,388.5***
	(355023.24)
edad	190,025.0***
	(17584.57)
edad2	-2,008.7***
	(202.56)
R^2	0.008
Adj. R^2	0.008
Estad. F	67.670***
N	16,397

****p < 0.001; **p < 0.01; *p < 0.05*

Fuente: GEIH 2018, Cálculos propios.

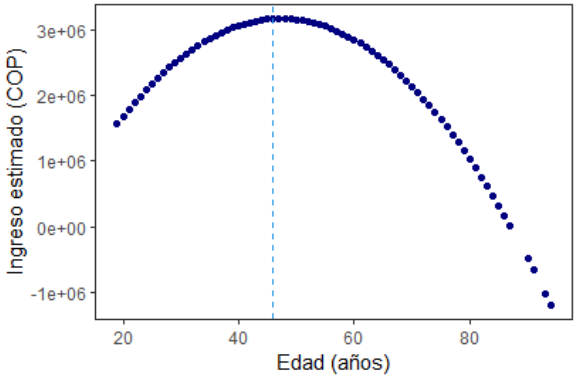
Al analizar el ajuste del modelo vemos que tanto el $\mathbf{R^2}$ como el $\mathbf{R^2}$ ajustado son de 0.008, es decir, las variaciones en las variables explicativas explican en tan solo 0.8% las variaciones en el ingreso del individuo. Lo anterior, indica que esta estimación no permite ajustar de manera sobresaliente la variable de interés, y esto podría deberse a que existen muchos otros factores no incluidos en el modelo que pueden estar explicando el ingreso.

De otro lado, en cuanto al **estadístico F**, que indica significancia conjunta de las variables independientes, podemos afirmar que existe suficiente evidencia estadística para rechazar la hipótesis nula, con lo cual, las variables incluidas son, en conjunto, significativas para explicar el ingreso. Asimismo, en los resultados de la estimación se observa que individualmente las variables de edad y edad al cuadrado resultan significativas al 1% como determinantes del ingreso.

Como se evidencia en la Gráfica 4, la relación entre el ingreso y la edad no es lineal, con lo cual, sabemos que se tiene un impacto positivo sobre el ingreso con cada año adicional, hasta un punto, en el cual cada año adicional de edad empiezan a impactar negativamente. Procedemos a encontrar ese "pico de edad" que sugiere la ecuación estimada, utilizando el método de remuestreo *bootstrap* para calcular los errores estándar del modelo y construir así los intervalos de confianza para la edad.

En primer lugar, se encuentra que la edad*, es decir, el punto de inflexión en la edad es 47, con lo cual, cada año adicional aumenta el ingreso, y esto ocurre hasta los 47 años, edad a partir de la cual cada año adicional de edad empieza a reducir el ingreso de la persona. Lo anterior, se puede confirmar visualmente en la Gráfica 8. Por otro lado, se encuentra un error estándar de 0.953, y se utiliza la fórmula estándar[2] para construir un intervalo de confianza de la media con un nivel de significancia del 5%. Se encuentra para la media de la edad un $IC\ (95\%) = [45\ ,49]$, que indica que con un 95% de probabilidad podemos afirmar que el verdadero valor del pico de edad promedio de todos los individuos se encuentra entre 45 y 49 años.

Gráfica 4. Ingreso estimado vs. edad



Fuente: GEIH 2018, construcción propia.

Con el fin de poder comparar el modelo de perfil de edad-ingreso aquí desarrollado con los modelos que se estimarán más adelante, se decide hacer una transformación de la variable de interés, ingreso, y se toma el logaritmo de este. Del mismo modo, la transformación se hace con el fin de capturar de forma más precisa el crecimiento del ingreso en términos relativos, en la medida que el logaritmo permite eliminar el efecto de las unidades de la variable sobre

[2] La fórmula para IC de media en muestras grandes es: $IC = [\overline{X} - 1.96 * SE\ ; \overline{X} + 1.96 * SE\ ]$. El valor de 1.96 proviene de la distribución normal estándar, donde 1.96 es el valor crítico asociado al grado de confianza de 95%.

los coeficientes. Con esto, se busca aportar estabilidad en los regresores, y reducir las observaciones atípicas (Wooldridge, 2010). De este modo, se realiza la estimación del modelo (1), y en términos de la significancia de las variables y el ajuste del modelo se obtienen los mismos resultados.

$$(1)\ \log(Ingreso) = B_1 + B_2 edad + B_3 edad^2 + u$$

De la estimación de ambos modelos se puede ver en la Gráfica del Anexo 1 que existe una varianza importante entre los datos estimados y observados, lo que nos indica que este modelo no permite un muy buen ajuste.

4. La brecha de ingresos

Se estima la brecha de ingresos incondicional utilizando el siguiente modelo:

$$(2)\ \log(Ingreso) = B_1 + B_2 mujer + u$$

Tabla 3. Modelo: brecha de ingresos incondicional

	L_Ingreso
	(2)
Intercepto	14.166***
	(0.022)
mujer (=1)	-0.338***
	(0.032)
R^2	0.007
Adj. R^2	0.007
Estad. F	112.866***
N	16,397

****p < 0.001; **p < 0.01; *p < 0.05*

Fuente: GEIH 2018, Cálculos propios.

El coeficiente B_2 acompaña a la variable mujer, que indica el género de la persona y que toma el valor de 1 cuando el individuo es mujer y 0 cuando es hombre. En este modelo, el parámetro captura el efecto que tiene el género sobre los ingresos. Habiendo realizado la estimación, obtuvimos un B_2 de -0.338, significativo a un nivel de significancia 1%. De este modo, tenemos suficiente información estadística para afirmar que, en promedio, las mujeres tienen un ingreso 33.8% inferior que los hombres, manteniendo todo lo demás constante.

Por otra parte, el R^2 indica que el ajuste del modelo no es bueno, ya que las variaciones en el género explican únicamente el 0.7% de variaciones en ingreso. El resultado se entiende en la medida que únicamente se está explicando el ingreso teniendo en cuenta el género, omitiendo muchas otras variables que son relevantes.

Posteriormente, se estima el perfil de edad-ingreso por género utilizando los modelos (2.m) y (2.h), donde el primero realiza una estimación únicamente considerando mujeres, y el segundo considerando solamente hombres.

(2.m) $Ingreso = B_1 + B_2 edad + B_3 edad^2 + u$

(2.h) $Ingreso = B_1 + B_2 edad + B_3 edad^2 + u$

Tabla 4. Modelo: perfil edad-ingreso por género

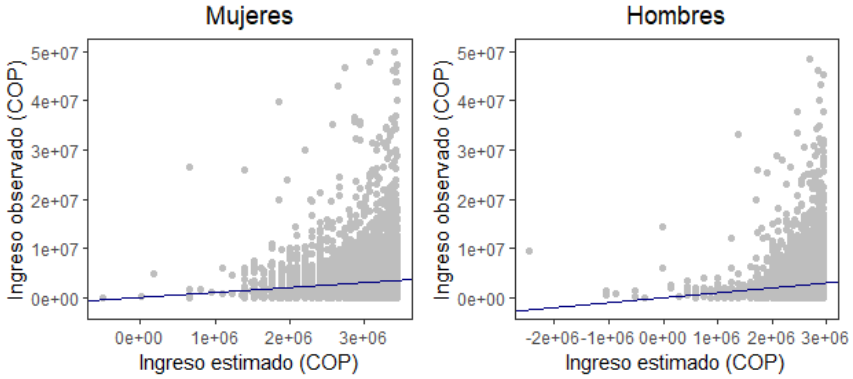
	Ingreso	
	(2.m)	**(2.h)**
Intercepto	-1,022,470.3*	-1,829,025.9***
	(490,531.9)	(508,230.0)
edad	185,345.0***	209,167.2***
	(24,503.4)	(25,036.1)
edad2	-2,158.8***	-2,075.8***
	(285.97)	(285.60)
R^2	0.007	0.012
Adj. R^2	0.007	0.011
Estad. F	28.759***	50.822***
N	7,715	8,682

****p < 0.001; **p < 0.01; *p < 0.05*

Fuente: GEIH 2018, Cálculos propios.

Después de estimar el modelo de ingresos-edad diferenciando por género, podemos concluir que, en la ciudad de Bogotá, el impacto de la edad sobre el ingreso es diferente para hombres y mujeres. En concreto, podemos ver que el intercepto de las mujeres es mayor, sin embargo, la pendiente para el caso de los hombres es mayor (Gráfica 5). Intuitivamente, esto indica que las mujeres tienen un ingreso más alto si consideramos que tienen justamente 18 años, no obstante, cada año adicional, implica mayores ingresos para hombres que para las mujeres.

Gráfica 5. Estimación modelos: perfil edad-ingreso por género



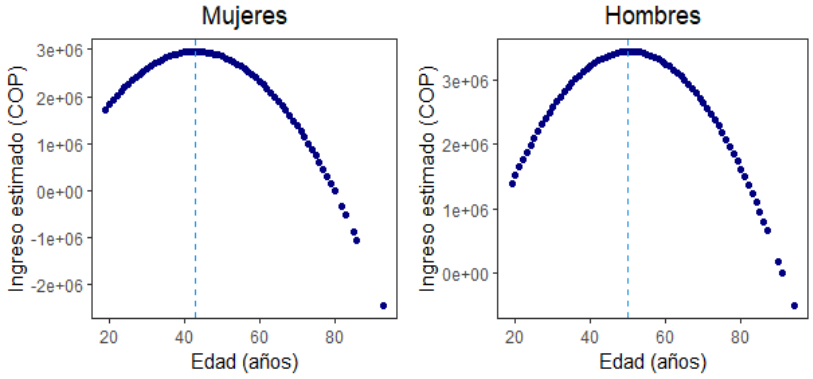
Fuente: GEIH 2018, construcción propia.

La relación positiva entre ingreso y edad es coherente con la teoría económica, sin embargo, destacamos que la relación se mantiene hasta un punto en el que el ingreso empieza a disminuir con cada año adicional. A este punto se le conoce como pico implícito de la edad y lo calculamos usando la metodología de *Bootstrap* con la que, mediante un remuestreo, calculamos los errores estándar del modelo, y obtenemos el intervalo de confianza[2].

De acuerdo con la estimación, en el caso de las mujeres, la edad en la cual el ingreso comienza a decrecer es de 43 años, con un error estándar de 0.867, y un $IC\ (95\%) = [41\ ,45]$. Con esto podemos afirmar que, el pico de inflexión de la edad estará en este intervalo para con un 95% de probabilidad. En el caso de los hombres, vemos la edad en la cual el ingreso cae con un año adicional es a los 50 años, con un error estándar de 1.911, con el cual encontramos un $IC\ (95\%) = [47\ ,54]$. De modo que, el valor verdadero del pico de edad promedio para los hombres se encuentra entre los 47 y los 54 años de edad, el 95% de las veces.

Si analizamos los intervalos de confianza construidos para ambos géneros, podemos ver que no se traslapan, es decir, no tienen edades en común. Esto nos muestra que incluso los hombres más jóvenes que empiezan a experimentar una caída en su ingreso, a los 47 años, son mayores que las mujeres que más se demoran en tener este cambio de la relación edad-ingreso, a los 45 años (Gráfica 6).

Gráfica 6. Ingreso estimado vs. edad



Fuente: GEIH 2018, construcción propia.

Equal pay for equal work?

(a) Comúnmente se habla de que en el mercado laboral debería existir un "salario igual para trabajos iguales", apelando a que, si se consideran empleados con caracteristicas similares tanto a nivel del individuo como de su trabajo, no deberían existir brechas de género en sus ingresos. Para analizar esta concepción estimamos el modelo (3.1) y (3.2), que además de tener en cuenta el género como determinante del ingreso, incluye algunos controles por las caracteristicas del trabajo:

$$(3.1)\ \log(Ingreso) = B_1 + B_2 mujer + B_3 cuentaPropia + B_4 formal + B_5 Micro\_empresa + u$$

$$(3.2)\ \log(Ingreso) = B_1 + B_2 mujer + B_3 cuentaPropia + B_4 formal + B_5 Micro\_empresa + B_6 oficio + u$$

(b) Se repite las estimaciones anteriores utilizando el teorema de Frisch Waugh Lovell (FWL), donde se añadió una variable (Ingreso_out) para omitir los *outliers* que fueran menores a cuatro veces el intercuantil del ingreso. Se encuentra que el coeficiente asociado a la variable de género (mujer) se reduce para ambos modelos (3.1) y (3.2), con lo cual,

podríamos afirmar que la estimación por MCO podría estar sesgada por valores atípicos (*outliers)* y sobreestima el efecto de ser mujer sobre el ingreso.

(c) El coeficiente B_2 que acompaña la variable mujer indica cual es el efecto de ser mujer sobre el ingreso del individuo.

Tabla 5. Modelos: brecha de ingresos condicional

	L_Ingreso		**L_Salario**	
	(3.1)	**(3.2)**	**(3.3)**	**(3.4)**
Intercepto	13.63***	14.05***	12.58***	12.21***
	(0.05)	(0.06)	(0.11)	(0.13)
Mujer (=1)	-0.32***	-0.41***	-0.06	0.02
	(0.03)	(0.03)	(0.07)	(0.07)
cuentaPropia (=1)	0.31***	0.29***	-0.22**	-0.20*
	(0.04)	(0.04)	(0.08)	(0.08)
Formal (=1)	1.28***	1.20***	0.77***	0.84***
	(0.04)	(0.04)	(0.09)	(0.09)
Micro_empresa (=1)	-0.64***	-0.59***	-1.08***	-1.12***
	(0.04)	(0.04)	(0.09)	(0.09)
oficio		-0.01***		0.01***
		(0.00)		(0.00)
R^2	0.165	0.173	0.042	0.043
Adj. R^2	0.165	0.173	0.041	0.043
Estad. F	812.349***	687.662***	177.949***	147.047***
N	16,397	16,397	16,397	16,397

****p < 0.001; **p < 0.01; *p < 0.05*

Fuente: GEIH 2018, Cálculos propios.

El modelo (3.1) demuestra que existe una brecha de ingresos por género, incluso cuando se controla por el tipo de trabajador (dependiente o independiente) con la variable cuentaPropia, el tipo de trabajo (formal o informal), y el tamaño de la empresa en que trabaja el individuo (Micro_empresa). En concreto, como se observa en la Tabla 5, el modelo (3.1) señala que ser mujer tiene un impacto negativo, pues reduce el ingreso de la persona en cerca de 32,3%. De igual forma, en el modelo (3.2), cuando se incluye una variable que indica el tipo de ocupación del individuo (oficio), la brecha no solo se mantiene, sino que se acentúa, y en este caso ser mujer reduce 41% los ingresos.

En ambos modelos las variables de control son significativas estadísticamente al 1%, de igual forma, el estadístico F señala que en conjunto las variables seleccionadas son estadísticamente significativas y por ende explican bien la variable de ingreso. Entre tanto, el R^2 indica que la varianza del ingreso solo es explicada en alrededor del 17% por las variables independientes seleccionadas.

Ahora bien, dado que esperamos que al controlar por las características del empleado y de su trabajo la brecha se reduzca, y los hallazgos difieren de esto, decidimos estimar nuevamente

los modelos anteriores, pero esta vez considerando como variable dependiente, el ingreso laboral (salario o ingreso de independientes) únicamente, puesto que el ingreso anteriormente considerado incluye además de ingreso laboral, otros ingresos. Los modelos estimados son:

$$(3.3)\ \log(Ing\_laboral) = B_1 + B_2 mujer + B_3 cuentaPropia + B_4 formal + B_5 Micro\_empresa + u$$

$$(3.4)\ \log(Ing\_laboral) = B_1 + B_2 mujer + B_3 cuentaPropia + B_4 formal + B_5 Micro\_empresa + B_6 oficio + u$$

Cuando se considera ingreso laboral los hallazgos antes de controlar por el tipo de ocupación son iguales, ser mujer reduce el ingreso, en este caso en 6%. Sin embargo, cuando se controla por esta característica (oficio), encontramos que ahora la brecha anterior no solo desaparece, sino que se invierte y ahora ser mujer tiene un impacto positivo en el ingreso en comparación con ser hombre. En este caso, el coeficiente asociado indica que ser mujer incrementa los ingresos en aproximadamente 1,5% (Tabla 5).

Tabla 6. Modelos: brecha de ingresos condicional (Teorema FWL)

	L_Ingreso	**residuo**	**L_Ingreso**	**residuo 2**
	(3.1)*		**(3.2)***	
Intercepto	13.51***	-0.00	13.69***	-0.00
	(0.04)	(0.01)	(0.06)	(0.01)
Mujer (=1)	-0.32***		-0.36***	
	(0.03)		(0.03)	
Ingreso_out	1.98***		1.91***	
	(0.06)		(0.06)	
cuentaPropia (=1)	0.32***		0.31***	
	(0.04)		(0.04)	
Formal (=1)	1.17***		1.14***	
	(0.04)		(0.04)	
Micro_empresa	-0.53***		-0.51***	
(=1)	(0.04)		(0.04)	
residuo_mujer		-0.32***		
		(0.03)		
oficio			-0.00***	
			(0.00)	
residuo_mujer2				-0.36***
				(0.03)
R^2	0.221	0.008	0.222	0.009
Adj. R^2	0.220	0.008	0.222	0.009
Estad. F	928.074***	127.625***	780.044***	151.687***
N	16,397	16,397	16,397	16,397

****p < 0.001; **p < 0.01; *p < 0.05*

Fuente: GEIH 2018, Cálculos propios.

Por otro lado, en la estimación con el Teorema de FWL encontramos que ser mujer en esta estimación reduce el ingreso en 31,9% sin controlar por el tipo de oficio, y 35,8% después de controlar por esta característica (Tabla 6). Del mismo modo, cuando se realiza la estimación únicamente considerando ingreso laboral, se obtiene una reducción en los coeficientes (Anexo 2). Antes de controlar por el tipo de ocupación, ser mujer tiene un impacto negativo de 3,7% en el salario, y luego de incluir esta variable, la brecha cambia y ahora ser mujer impacta positivamente el salario del individuo pues lo incrementa 1,46%.

La reversión de la brecha cuando se considera solamente el ingreso laboral puede deberse a que la brecha inicial de ingreso que afecta negativamente a las mujeres se debe más a un problema de selección y no a uno de discriminación, como lo explican Di Paola y Berges (2000). Puntualmente, se ha encontrado que la causa de las diferencias salariales entre hombres y mujeres se debe en gran medida a un problema de selección. Las mujeres suelen aceptar trabajos con una menor retribución monetaria, pero con otro tipo de compensaciones que los hacen más "agradables", lo cual hace que cuando se estima el modelo controlando por el tipo de ocupación, es decir, cuando se comparan dos con las mismas características de empleo y lo único que varía es su género, la brecha que tradicionalmente se espera desaparece.

5. Predicción de ingresos

(a) En busca de evaluar el poder predictivo de los modelos, se divide la muestra en dos submuestras: una de entrenamiento (70%) y una de prueba (30%).

i. En primer lugar, como punto de referencia, se estima un modelo que solo incluye la constante (0).

$$(0)\ \log(Ingreso) = B_1 + u$$

ii. En segundo lugar, utilizando ahora las dos submuestras, se realiza la estimación de los modelos previos (1), (2), (3.1) y (3.2):

$$(1)\quad \log(Ingreso) = B_1 + B_2 edad + B_3 edad^2 + u$$

$$(2)\quad \log(Ingreso) = B_1 + B_2 mujer + u$$

$$(3.1)\quad \log(Ingreso) = B_1 + B_2 mujer + B_3 cuentaPropia + B_4 formal + B_5 Micro\_empresa + u$$

$$(3.2)\quad \log(Ingreso) = B_1 + B_2 mujer + B_3 cuentaPropia + B_4 formal + B_5 Micro\_empresa + B_6 oficio + u$$

iii. En tercer lugar, se exploran algunas transformaciones de las variables independientes, de manera que se pueda controlar la estimación por otras características del individuo o de su trabajo. Para establecer la forma funcional de estos modelos, incluimos características que consideramos relevantes para explicar el ingreso. Por ejemplo, agregamos variables que indican si la persona es trabajador dependiente (asalariado), o si, por el contrario, trabaja por cuenta propia.

Consideramos importante hacer esta distinción ya que ambos tipos de trabajadores pueden tener diferencias significativas en cuanto a beneficios extrasalariales, obligaciones y capital. Así mismo, se ha encontrado que existen diferencias en cuánto impacta cada año adicional de experiencia o edad dependiente del tipo de trabajados, por lo cual, también decidimos incluir interacciones entre estas variables y cuentaPropia (Guataquí et al., 2009).

$$(4)\quad \log(Income) = B_1 + B_2 edad + B_3 edad^2 + B_4 educ + u$$

$$(5)\quad \log(Income) = B_1 + B_2 edad + B_3 edad^2 + B_4 educ + B_5 mujer + B_6 educ * mujer + u$$

$$(6)\quad \log(Income) = B_1 + B_2 edad + B_3 edad^2 + B_4 educ + B_5 cuentaPropia + u$$

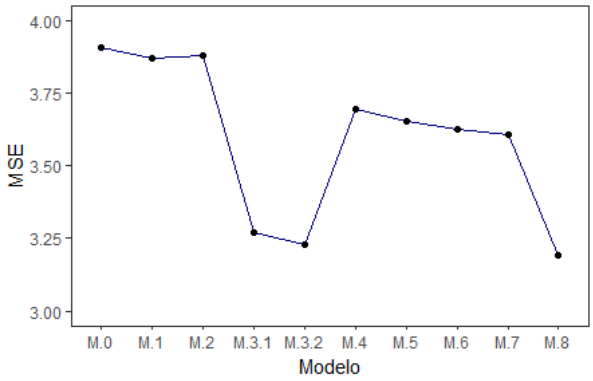
$$(7)\quad \log(Income) = B_1 + B_2 edad + B_3 edad^2 + B_4 educ + B_5 cuentaPropia + B_6 cuentaPropia * edad + B_7 cuentaPropia * edad^2 + B_8 cuentaPropia * educ + u$$

$$(8)\quad \log(Income) = B_1 + B_2 edad + B_3 edad^2 + B_4 educ + B_5 mujer + B_6 educ * mujer + B_7 cuentaPropia + B_8 cuentaPropia * edad + B_9 cuentaPropia * edad^2 + B_{10} cuentaPropia * educ + B_{11} formal + B_{12} oficio + B_{13} Micro\_empresa + u$$

iv. Después de realizar las estimaciones, se calculó el Error cuadrático medio (*MSE* en inglés) para cada modelo (Anexo 3), desde el modelo de referencia, hasta los modelos propuestos anteriormente. En la Gráfica 7 podemos observar que a medida que incluimos variables que capturan características del trabajo (modelos (3.1) y (3.2)), el *MSE* disminuye y el ajuste del modelo mejora.

No obstante, aumentar la complejidad mediante la inclusión de interacciones y variables no lineales, no necesariamente mejora el ajuste (modelos (4), (5), (6) y (7)). Finalmente, en el modelo (8), en el cual incluimos tanto los controles propuestos anteriormente como características del trabajo, es donde encontramos el mejor ajuste, al tener el menor Error cuadrático medio.

Gráfica 7. Error cuadrático medio



Fuente: GEIH 2018, construcción propia.

La tabla 7 presenta los resultados de la estimación del modelo con el mejor ajuste, el modelo (8). En primer lugar, vemos que el R^2 es de 0.189, que indica que el 18.9% de la variación en el ingreso se explica por variaciones de las covariables del modelo. Por otro lado, vemos que el p-valor asociado a la prueba de significancia conjunta rechaza la hipótesis nula de no significancia, por lo cual, podemos afirmar con un nivel de confianza del 99% que las variables son significativas en conjunto.

A nivel individual, todas las variables incluidas en el modelo son significativas para explicar el ingreso al 5%, además de tener el signo esperado. Los resultados indican que los años de educación, el ser trabajador independiente y la edad (hasta el punto de inflexión) tienen un impacto positivo en el ingreso; mientras que ser mujer, trabajar en una microempresa y la edad (después del pico) afectan negativamente el ingreso.

Por último, resaltamos el hecho de que un año de educación adicional tiene un impacto en el ingreso superior para las mujeres que, para los hombres, y menor para personas independientes que para los asalariados.

Tabla 7. Modelo (8)

	L_Ingreso
	(8)
Intercepto	11.48***
	(0.24)
edad	0.07***
	(0.01)
edad2	-0.00***
	(0.00)
educ	0.06***
	(0.01)
Mujer (=1)	-0.95***
	(0.14)
cuentaPropia (=1)	1.83***
	(0.39)
Formal (=1)	1.11***
	(0.05)
oficio	-0.00***
	(0.00)
Micro_empresa (=1)	-0.57***
	(0.05)
educ : mujer (=1)	0.04***
	(0.01)
edad : cuentaPropia (=1)	-0.03*
	(0.02)
edad2 : cuentaPropia	0.00
(=1)	(0.00)
educ _ cuentaPropia (=1)	-0.06***
	(0.01)

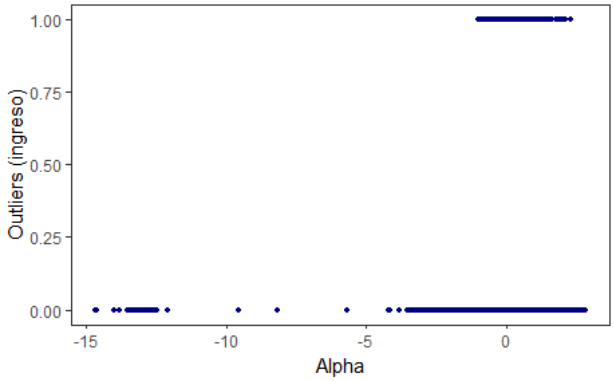
R^2	0.189
Adj. R^2	0.188
Estad. F	222.627***
N	11,478

****p < 0.001; **p < 0.01; *p < 0.05*

Fuente: GEIH 2018, Cálculos propios.

v. La estadística de *leverage* (α) nos indica esos *outliers* u observaciones que tienen una gran influencia en la estimación. Luego de calcular esta medida para observación en la muestra de prueba se encuentra a aquellos que tienen ingresos muy altos (valores atípicos) y tienen un (α) elevado (Gráfica 8), estas observaciones demuestran que están muy por encima del ingreso promedio y podrían estar sesgando el modelo. No obstante, cuando se revisa a profundidad esas observaciones con ingresos muy altos, su Hj no está teniendo influencia en el (α), sino su Uj (peso de los betas estimados). Además, el ingreso estimado está muy por debajo del ingreso reportado por lo que pueda que el modelo no este ajustado bien. Sin embargo, la DIAN puede utilizar estos puntos leverage para inspeccionar si los individuos están declarando el monto de renta correcto o están evadiendo y creando paraísos fiscales.

Gráfica 8. *Outliers* vs estadístico *leverage*



Fuente: GEIH 2018, construcción propia.

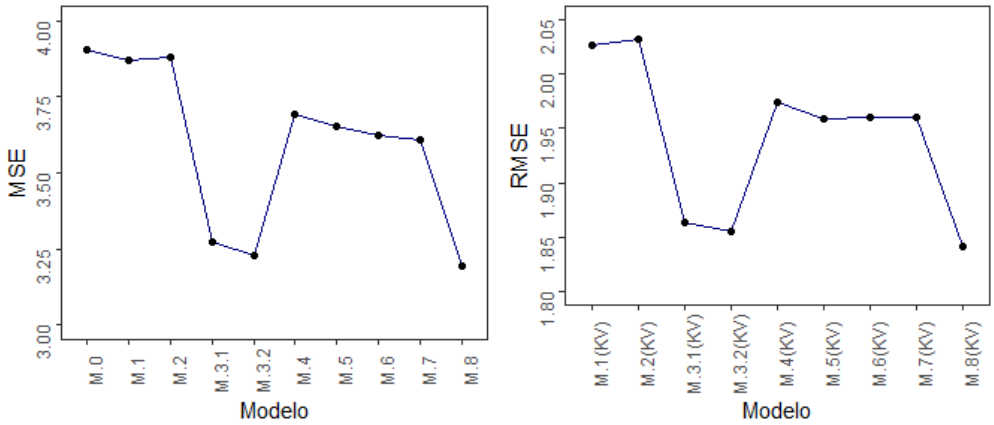
(b) Se realizó una cálculo de todos los modelos estimados hasta el momento utilizando ahora la técnica de validación cruzada en k-iteraciones (*K-fold cross validation* en inglés), para evaluar las estimaciones mediante el entrenamiento de cada modelo en un subconjunto de datos y su posterior evaluación en el subconjunto complementario de los datos.

Después de ejecutar el procedimiento para cada uno de los modelos, calculamos su Error cuadrático medio (Anexo 4). Encontramos que el modelo (8) presentó el menor valor de *MSE*, con lo cual, este continúa siendo el que tiene el mejor ajuste dentro de los modelos estimados. De modo que, tanto en la estimación con el enfoque de validación como en la estimación por validación cruzada en k-iteraciones, este sigue siendo el mejor modelo.

Como se evidencia en la Gráfica 9, en términos relativos la calidad del ajuste de los diferentes modelos tiene el mismo comportamiento desde las dos aproximaciones. Destacamos que el

ajuste mejora significativamente utilizando *K-fold CV*, donde el *MSE* es menor en comparación con el otro enfoque y los resultados son más robustos.

Gráfica 9. Error cuadrático medio



**MSE corresponde al error cuadrático medio con el enfoque de validación y RMSE con el enfoque de validación cruzada*

Fuente: GEIH 2018, construcción propia.

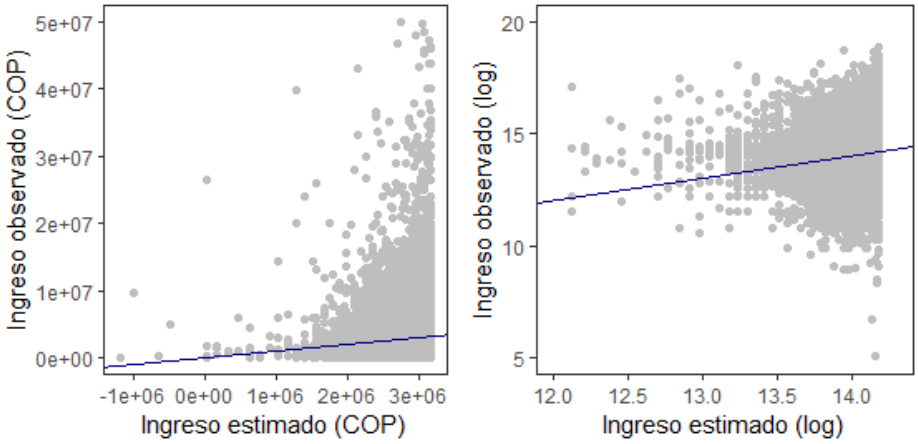
(c)

i. La creación y procesamiento del *loop* se encuentran en el código de R.

ii. El estadístico de *leverage* es relativamente menor que el LOOCV, demostrando que el modelo en general hace un buen ajuste. Inclusive, cuando se hace una comparación entre el promedio de los (Hj) y la condición teórica de (2p/n), donde "p" es el número de parámetros y "n" las observaciones, para determinar si Hj es alto y los *outliers* tienen un peso, se observa que no es el caso. El promedio de los Hj es 0.0008 y la condición es igual 0.0014 concluyendo que Hj es menor a la condición. Por lo tanto, no tendría mucho sentido ajustar el modelo y volver a aplicar el método LOOCV, que además tiene un costo computacional alto.

Anexos

Anexo 1. Estimación modelos: perfil edad-ingreso



Fuente: GEIH 2018, Cálculos propios.

Anexo 2. Modelos: brecha de salarios condicional (Teorema FWL)

	L_Salario	residuo	L_ Salario	residuo2
	(3.3)*		**(3.4)***	
Intercepto	12.41***	0.00	11.53***	-0.00
	(0.10)	(0.03)	(0.13)	(0.03)
Mujer (=1)	-0.04		0.15*	
	(0.07)		(0.07)	
y_total_m_out	2.97***		3.35***	
	(0.13)		(0.13)	
cuentaPropia (=1)	-0.31***		-0.27**	
	(0.08)		(0.08)	
Formal (=1)	0.55***		0.69***	
	(0.09)		(0.09)	
Micro_empresa (=1)	-0.89***		-0.97***	
	(0.09)		(0.09)	
residuo_mujer		-0.04		
		(0.07)		
oficio			0.01***	
			(0.00)	
residuo_mujer2				0.15*
				(0.07)
R^2	0.071	0.000	0.078	0.000
Adj. R^2	0.071	0.000	0.077	0.000
Estad. F	250.031***	0.319	230.346***	4.608**
N	16,397	16,397	16,397	16,397

****p < 0.001; **p < 0.01; *p < 0.05*

Fuente: GEIH 2018, Cálculos propios.

Anexo 3. Error cuadrático medio

Modelo	***MSE***
(0)	3.908
(1)	3.870
(2)	3.879
(3.1)	3.271
(3.2)	3.227
(4)	3.692
(5)	3.653
(6)	3.625
(7)	3.608
(8)	3.191

Fuente: GEIH 2018, Cálculos propios.

Anexo 4. Error cuadrático medio (K-partes VC)

Modelo	***MSE***
(1) VC	2.026
(2) VC	2.032
(3.1) VC	1.863
(3.2) VC	1.854
(4) VC	1.974
(5) VC	1.958
(6) VC	1.959
(7) VC	1.959
(8) VC	1.841

Fuente: GEIH 2018, Cálculos propios.

Referencias

Arango, L., Obando, N. & Posada, C. (2010). *Sensibilidad de los salarios al desempleo regional en Colombia: nuevas estimaciones de la curva de salarios*. Borradores de economía, 590. Banco de la República de Colombia.

Aristizábal, T. & López, E. (2017). *Efectos de los aumentos en la escolaridad en el mercado laboral colombiano entre 2008 y 2016.* Ecos de Economía: A Latin American Journal of Applied Economics, 21(44).

Departamento Administrativo Nacional de Estadística (2009). *Metodología informalidad Gran Encuesta Integrada de Hogares – GEIH*. Dirección de Metodología y Producción Estadística – DIMPE.

Di Paola, R., & Berges, M. (2000). Sesgo de selección y estimación de la brecha por género para Mar del Plata. In XXXV Reunión Anual de La Asociación Argentina de Economía Política.

Guataquí, J., García, A. & Rodríguez, M. (2009). *Estimaciones de los determinantes de los ingresos laborales en Colombia con consideraciones diferenciales para asalariados y cuenta propia.* Serie de documentos de trabajo, 70. Facultad de economía. Universidad del Rosario.

Ministerio de Comercio, Industria y Turismo (2007). *Guía de contratación pública para micro y pequeñas empresas – MIPYME.* Ministerio de Comercio, Industria y Turismo.

Rúa, M. & Rivera, E. (2019). *Análisis regional de los efectos de la educación y la experiencia en los salarios de Colombia*. [Trabajo de grado]. Universidad EAFIT.

Wooldridge, J. (2010). *Introducción a la econometría: un enfoque moderno*. Cengage Learning.