Profesor: Ignacio Sarmiento-Barbieri

Problem Set 1: Predicting Income

Big Data y Machine Learning para Economía Aplicada, 2023-1

Integrantes: Jorge Rodríguez[1]. Iván Velásquez[2], Santiago González Cepeda[3] , María José Colmenares Wolff[4]

1. Introducción

Tanto la complejidad del sistema tributario como los datos disponibles, son variables que hacen del estudio tributario algo no tan sencillo. De lo anterior, se deriva lo que se conoce como la brecha fiscal, lo cual es la diferencia entre las estimaciones de los ingresos fiscales frente a lo que realmente es el recaudo real que es pagado de manera oportuna y voluntaria. En el caso de Estados Unidos la brecha fiscal estimada fue de USD $456 mil millones para los años 2014 – 2016, por otro lado la tasa de cumplimiento voluntario que mide el cumplimiento relativo entre el estimado y el real se ubicó en 85 % según el Servicio de Impuestos Internos (IRS por sus siglas en ingles) para los mismos años. Es posible, que este fenómeno se replique en Colombia, ya que ha sido caracterizada por tener una estructura tributaria antitécnica, compleja, donde hay una combinación de gravámenes relativamente elevados con un cúmulo de exenciones Clavijo (2011) y puede haber una cabida para que se presente una subdeclaración de ingresos por parte de las personas, lo cual tendría un efecto negativo en el recaudo tributario.

Para tener una mejor aproximación de los ingresos reales de la población, se utilizarán diferentes modelos donde la edad y el género hacen parte de la forma como se quiere hacer una predicción de dichos ingresos. Esto tendrá diferentes beneficios que se derivan directamente de un mayor poder de tributación a ciertos individuos, lo que implica un mayor recaudo, el cual se utiliza para inversión en infraestructura, salud, educación, entre otros. En ese sentido, los controles escogidos para las estimaciones están sustentados en la ecuación minceriana de ingresos, toda vez que sugiere ciertas variables que pueden ayudar a predecir los ingresos laborales de las personas. Según Mincer (1974) existía una relación positiva entre la escolaridad de un individuo y sus ingresos posteriores, asimismo con la experiencia. Por lo cual, se considera utilizar esta teoría para así poder generar unos resultados razonables. Además, tener una mejor aproximación a los ingresos reales de la población, permitirá tener un mejor conocimiento de la distribución de ingresos y, al desagregarlo por variables como la edad y el género, pueden ser de gran utilidad para la construcción o ajuste de políticas públicas que tengan como estrategia la redistribución a ciertos grupos poblacionales.

De esta manera, la Gran Encuesta Integrada de Hogares (GEIH) es una herramienta muy útil ya que brinda información sobre las condiciones de empleo del hogar, es decir, si las personas trabajan o no, cuál es su trabajo, cuántos son sus ingresos, si tienen seguridad social y salud y si están en búsqueda de empleo. Además de unas características generales de la población como sexo, edad, nivel de educación, estado civil, etc. DANE (2022). Igualmente, este será un enfoque que solo tomará datos de Bogotá dejando de lado el resto de información a nivel nacional, tanto de cabeceras, como regional, departamental y las capitales de los departamentos.

En este estudio, únicamente se utilizará una derivación de la GEIH que fue utilizada por el Departamento Administrativo Nacional de Estadística para construir medidas de pobreza monetaria y desigualdad. Esta derivación agrupa las variables que se quieren estudiar en este problema en particular y descarta otras tantas variables que están incluidas en la GEIH que no son necesarias para el objetivo del ejercicio. Como grupo demográfico, solo se utilizarán los datos de individuos residentes en Bogotá. Como resultado, esta predicción de ingresos lo que busca es que se puedan conocer los fraudes que se presentan y de igual forma, poder

[1]Cod. Uniandes: 201715669
[2]Cod. Uniandes: 201114762
[3]Cod. Uniandes: 201719971
[4]Cod. Uniandes: 201811929

identificar familias e individuos vulnerables que serian potenciales beneficiarios de políticas que focalicen el gasto.

2. Datos

Los datos de la encuesta "Medición de Pobreza Monetaria y Desigualdad 2018" creados a partir de la GEIH de 2018, es usada principalmente por el DANE para realizar las estimaciones del Índice de Pobreza Multidimensional (IPM), sin embargo, múltiples estudios usan estos datos para evaluar modelos de desarrollo rural (Serrano-Malaver et al. (2018)), evaluar los niveles de educación y género a partir de la descomposición del IPM (Cortes Sabogal et al. (2016)) y otros estudios relacionados con tributación y pobreza (García Ocampo (2018)). Usando los datos de esta encuesta se pretende predecir el nivel de ingreso (laboral) de los individuos de la ciudad de Bogotá ocupados mayores de 18 años, con el fin de contar con un modelo predictivo que permita bajo ciertos parámetros tener una medida que funcione para crear las alertas necesarias en los entes encargados de la recolección tributaria. No se hace uso de la variable de ingreso total, dado que, por definición en esta encuesta, dicha variable tiene información del valor aproximado de ingresos especies que en la practica puede generar ruido a la hora de tener modelos predictivos, ya que es muy arbitrario el valor que toman.

La obtención de los datos fue a partir de web scraping, donde se seleccionan 10 archivos de la web que componen toda la información de la encuesta para todos los meses del 2018 para la ciudad de Bogotá, sin embargo los datos originales se encuentran en la página oficial del DANE, dicha página no tiene limitaciones para el uso de web scraping [5]. La base originalmente cuenta con 42,069 observaciones para la ciudad de Bogotá de las cuales 34,460 son personas mayores de edad y 16,542 pertenecen a la población ocupada, de esta muestra de individuos se eliminan aquellas observaciones que presentan valores perdidos en la variable de salario nominal mensual (6,650 observaciones) dado que no se utilizara ninguna técnica de imputación dado el alto porcentaje de valores perdidos. Para el tratamiento de valores atípicos se realiza una técnica llamada "Winsorizar" que consiste en retener las observaciones atípicas, haciendo que estas tomen valores particulares dependiendo del percentil que se elija, para este caso los valores atípicos se internalizan en la muestra del tal manera que el 1 % de la muestra toman el valor del percentil 1, y el 1 % más alto de la distribución de los salarios toma el valor del percentil 99. Finalmente, la muestra objeto de estudio cuenta con 9,892 observaciones.

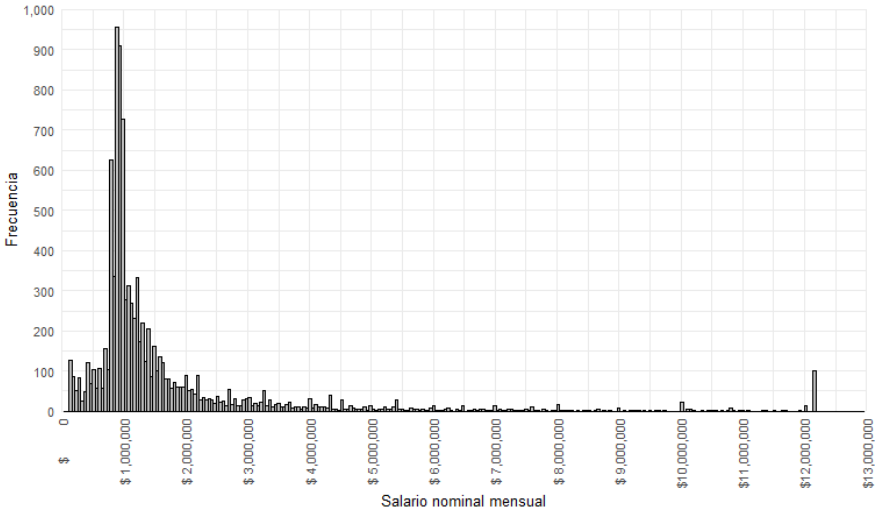
Cuadro 1: Estadísticas Descriptivas - Colombia 2018 por género

	Todos		**Mujer**		**Hombre**		
	Media	Desv. Estándar	Media	Desv. Estándar	Media	Desv. Estándar	Diff[1]
A. Individuo							
Edad	36,24	12,02	36,58	11,90	35,90	12,14	***
Prop. Amo(a) de casa	0,030	0,171	0,055	0,229	0,005	0,072	***
B. Hogar							
Prop. Hijos en el hogar	0,235	0,424	0,232	0,422	0,238	0,426	
Estrato	2,51	0,98	2,58	1,00	2,44	0,95	***
C. Educación[2]							
Estudiantes	0,010	0,101	0,011	0,103	0,010	0,099	
Primaria	0,005	0,067	0,005	0,074	0,004	0,060	
Secundaria	0,095	0,293	0,081	0,273	0,109	0,312	***
Media	0,346	0,476	0,320	0,467	0,371	0,483	***
Terciaria	0,453	0,498	0,503	0,500	0,403	0,491	***
D. Mercado Laboral[3]							
Salario mensual	1.682.266	1.876.130	1.606.083	1.837.522	1.757.622	1.910.751	***
Ingreso Total	1.872.592	2.509.096	1.790.266	2.309.982	1.954.025	2.689.353	***
Experiencia trab. Actual	49,73	73,20	49,45	73,71	50,01	72,71	
Horas trabajas en la semana	48,02	12,15	45,86	11,68	50,16	12,24	***
Prop. Informalidad	0,233	0,422	0,248	0,432	0,217	0,412	***
Observaciones	9.892		4.919		4.973		

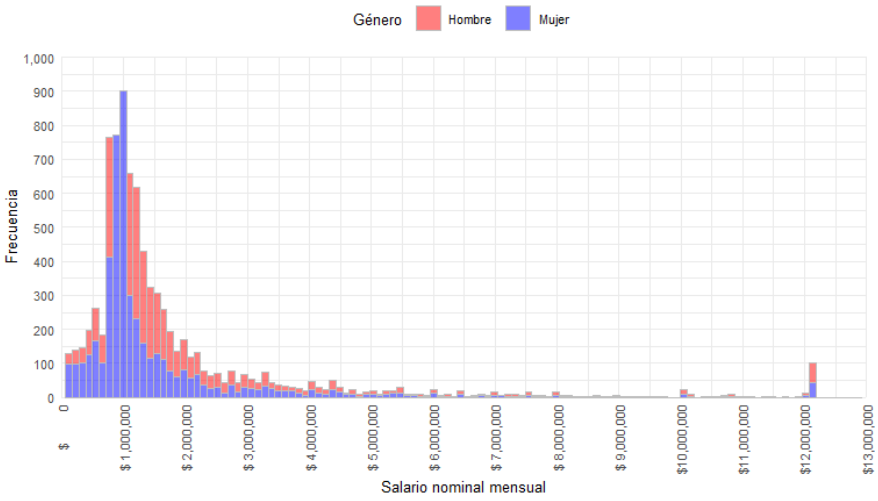
Notas: Cálculos propios usando datos DANE Medición de Pobreza Monetaria y Desigualdad 2018. [1] Indica el nivel de significancia estadística de la diferencia de medias entre hombres y mujeres bajo la $Ho : \mu(Hombre) - \mu(Mujer) = 0$. [2] Los valores representados en esta sección corresponden a la proporción de personas que esta dentro de cada categoría.[3] La experiencia en el trabajo actual esta medida en meses. La unidad monetaria de las variables referentes al ingreso son pesos colombianos nominales. *** p<0.01, ** p<0.05, * p<0.1

[5]Esto se verifica a través de https://microdatos.dane.gov.co/index.php/catalog/608/get$_{m}icrodata/robots.txt$

En el cuadro 1, resume los estadísticos descriptivos de la muestra para el total y por género de los datos que se utilizara en la estimación de los modelos. En la muestra hay 4,919 mujeres y 4,973 hombres. La edad promedio de la población es de 36.2 años y las mujeres tienen una edad promedio mayor que los hombres; existe una mayor proporción de mujeres que son amas de casa que hombres. El nivel socioeconómico o estrato es en promedio de 2.5. En cuanto a los niveles educativos reportados se encuentra que existe una mayor proporción de mujeres con algún tipo de educación superior que hombres siendo 50.3 % y 40.3 % respectivamente. Un dato relevante es que para Bogotá la educación superior es la categoría predominantes en términos educativos. El salario mensual promedio es de 1,682,266 siendo este superior al salario mínimo del año 2018 (781,242), existe una diferencia estadísticamente significativa al 1 % entre hombres y mujeres en términos salariales, esta brecha a persistido a través de los años (2008-2019) en las diferentes ciudades tal como lo muestra Florez, Melo-Becerra, and Posada (2021). El ingreso total es mayor dado que incluye otras fuentes de ingreso no laborales. La experiencia promedio de los individuos es de aproximadamente 4 años en el trabajo que reportan actualmente, y las horas que trabajan a la semana son en promedio de 48.02, destacando que el numero de horas trabajadas es menor en las mujeres. Finalmente, la proporción de informalidad es más alta en las mujeres que en los hombres. Para todas las variables reportadas, en la última columna del cuadro 1, se reporta si existen diferencias promedio estadísticamente significativas entre hombres y mujeres.



(a) Todos



(b) Por género

Figura 1: Distribución del salario nominal mensual

En la Figura 1, se muestra la distribución de los salarios nominales mensuales para la muestra, es evidente que la distribución esta sesgada hacia la izquierda, y en su gran mayoria los salarios se centran entre 0 y 1,000,000. En cuanto a la distribución por sexo, los hombres tienen la distribución salarial más hacia la derecha.

3. Perfil de salario por edad

Lo primero para realizar el ejercicio de hacer la predicción de los salarios, es empezar por estudiar cómo diferentes variables tienen influencia en este. Para esta sección, evaluaremos la evolución del salario a través del tiempo, más específicamente en el tiempo de vida de las personas. Esto es relevante para entender la forma en que diferentes grupos de edad reciben ingresos, y uno de los objetivos de esta sección es encontrar la edad en donde, en promedio, una persona alcanzaría sus ingresos máximos.

Para esto, se realizó un modelo de regresión sencillo, que sigue la siguiente ecuación:

$$ln(salario) = \beta_0 + \beta_1 Edad + \beta_2 Edad^2 + u$$

En la tabla 1 se presentan los resultados de esta estimación.

Cuadro 2: Estimación Salario - Edad

	Dependent variable:
	log_salario_m
Edad	0.087***
	(0.004)
Edad^2	-0.001***
	(0.00004)
Constante	12.335***
	(0.068)
Observations	9,892
R^2	0.060
Adjusted R^2	0.060
Residual Std. Error	0.711
F Statistic	315.885***
Note:	*p<0.1; **p<0.05; ***p<0.01

• An interpretation of the coefficients and it's significance. • A discussion of the model's in sample fit.
• A plot of the estimated age-earnings profile implied by the above equation.Including a discussion of the "peak age" with it's respective confidence intervals. (Note: Use bootstrap to construct the confidence intervals.)

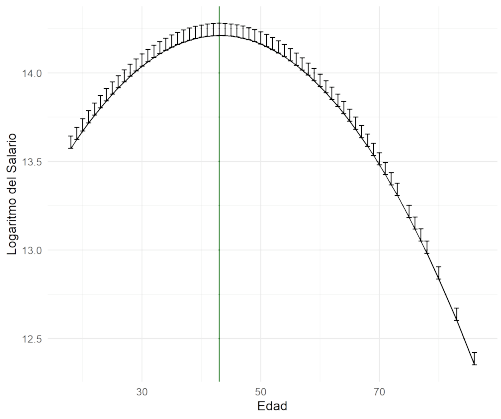


Figura 2: Todos

4. Brecha salarial de género

5. Predicción de Salarios

Para evaluar el poder de predicción de los modelos anteriores y de los que se proponen en esta sección, primero se establece una semilla que permite la replicabilidad de los modelos estimados, y en este caso la muestra se divide aleatoriamente en dos partes, una parte correspondiente al 70 % y otra al 30 %, esto estableciendo que el 70 % es la muestra de entrenamiento y el 30 % es la muestra de testeo. Dado que los modelos usan la variable Oficio como predictor y que esta tiene grandes bondades al reconocer las diferencias salariales entre las distintas ocupaciones, el muestreo aleatorio se realizó de manera estratificada con el fin de garantizar que en las dos muestras existan individuos de cada ocupación y de esta manera poder estimar los parámetros de manera completa.

Los nuevos modelos estimados en esta sección son:

- Modelo 6:

$$Ln(Salario_i) = \beta_0 + \beta_1 Mujer_i + \beta_2 Mujer_i \cdot Edad_i + \beta_3 Mujer_i \cdot Edad_i^2 + \beta_4 Edad_i + \beta_5 Edad_i^2 + \beta_6 Superior_i$$
$$+\beta_7 Horas\_Trabajo_i + \beta_8 Informal_i + \beta_9 Media_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{Oficio=w\}}] + \epsilon_i \quad (6)$$

- Modelo 7:

$$Ln(Salario_i) = \beta_0 + \beta_1 Mujer_i + \beta_2 Mujer_i \cdot Edad_i + \beta_3 Mujer_i \cdot Edad_i^2 + \beta_4 Edad_i + \beta_5 Edad_i^2 + \beta_6 Superior_i$$
$$+\beta_7 Horas\_Trabajo_i + \beta_8 Informal_i + \beta_9 Media_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{Oficio=w\}}] + \beta_{10} Experiencia_i + \epsilon_i \quad (7)$$

- Modelo 8:

$$Ln(Salario_i) = \beta_0 + \beta_1 Mujer_i + \beta_2 Mujer_i \cdot Edad_i + \beta_3 Mujer_i \cdot Edad_i^2 + \beta_4 Edad_i + \beta_5 Edad_i^2 + \beta_6 Superior_i$$
$$+\beta_7 Horas\_Trabajo_i + \beta_8 Informal_i + \beta_9 Media_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{Oficio=w\}}] + \beta_{10} Experiencia_i$$
$$+ \sum_{r=1}^{6} \alpha_r [\mathbb{1}_{\{Estrato=r\}}] + \epsilon_i \quad (8)$$

- Modelo 9:

$$Ln(Salario_i) = \beta_0 + \beta_1 Mujer_i + \beta_2 Mujer_i \cdot Edad_i + \beta_3 Mujer_i \cdot Edad_i^2 + \beta_4 Edad_i + \beta_5 Edad_i^2 + \beta_6 Superior_i$$
$$+\beta_7 Horas\_Trabajo_i + \beta_8 Informal_i + \beta_9 Media_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{Oficio=w\}}] + \beta_{10} Experiencia_i$$
$$+ \sum_{r=1}^{6} \alpha_r [\mathbb{1}_{\{Estrato=r\}}] + \beta_{11} Experiencia_i^2 + \epsilon_i \quad (9)$$

- Modelo 10:

$$Ln(Salario_i) = \beta_0 + \beta_1 Mujer_i + \beta_2 Mujer_i \cdot Edad_i + \beta_3 Mujer_i \cdot Edad_i^2 + \beta_4 Edad_i + \beta_5 Edad_i^2 + \beta_6 Superior_i$$
$$+\beta_7 Horas\_Trabajo_i + \beta_8 Informal_i + \beta_9 Media_i + \sum_{w=1}^{99} \alpha_w [\mathbb{1}_{\{Oficio=w\}}] + \beta_{10} Experiencia_i$$

$$+\sum_{r=1}^{6}\alpha_r[\mathbb{1}_{\{Estrato=r\}}]+\beta_{11}Experiencia_i^2+\beta_{12}Horas\_Trabajo_i^2+\epsilon_i \quad (10)$$

En este caso los nuevos modelos reconocen las no linealidades entre las distintas variables y su relación con el salario, como se muestra en el anexo XX, la experiencia en el actual empleo reportado y las horas trabajadas a la semana representan rendimientos marginales decrecientes frente al salario, algo que va muy acorde con la teoría económica.

En el cuadro 3, se presentan los MSE de los 10 modelos estimados, la columna uno muestra el MSE en la muestra de entrenamiento y en la columna dos se muestra el MSE en la muestra de testeo. Se evidencia que el MSE aumenta muy poco en la muestra de testeo con respecto a la de entrenamiento. El MSE va bajando a medida que aumentamos la complejidad del modelo, sin embargo, no se está realizando ningún sobre ajuste al aumentar la complejidad, ya que se tiene en cuenta la teoría económica a la hora de incluir predictores. El mejor modelo es el número 10, puesto que presenta el menor MSE tanto en entrenamiento como en testeo. Los parámetros de este modelo están contenidos en el anexo XX en la columna 10, y los valores que toman tienen sentido económico puesto que reconoce que las relaciones cuadráticas de las variables que se eligieron para el ajuste presentan rendimientos marginales decrecientes.

Cuadro 3: MSE Entrenamiento y testeo

Modelo	MSE	
	Entrenamiento	Testeo
1	0,5349	0,5445
2	0,5051	0,5070
3	0,5300	0,5396
4	0,2213	0,2362
5	0,2213	0,2361
6	0,2197	0,2348
7	0,2128	0,2245
8	0,1873	0,1963
9	0,1869	0,1956
10	0,1770	0,1848

Notas: Cálculos propios usando datos DANE Medición de Pobreza Monetaria y Desigualdad 2018. La muestra de entrenamiento es el 70 % de la muestra total elegida aleatoriamente estratificando por la variable oficio.

En la Figura XX, se muestra la di

Referencias

CLAVIJO, S. (2011): «Estructura fiscal de Colombia y ajustes requeridos (2010-2020),» .

CORTES SABOGAL, M. Á. ET AL. (2016): «Nivel de educación y género: análisis de su impacto en la pobreza a partir de la descomposición del índice de pobreza monetaria en Colombia 2010-2015,» .

DANE (2022): «Gran Encuesta Integrada de Hogares,» Tech. rep., Departamento Administrativo Nacional de Estadística - DANE.

Florez, L. A., L. A. Melo-Becerra, and C. E. Posada (2021): «Estimating the reservation wage across city groups in Colombia: A stochastic frontier approach,» *Borradores de Economía; No. 1163*.

García Ocampo, Tania Lorena y Castellanos Sabogal, Y. T. (2018): «Tributación y pobreza en Colombia: un análisis desde la evolución del impuesto de renta y el índice de pobreza monetaria,» *Revista Activos*, 16, 79–98.

Mincer, J. (1974): «Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2.» .

Serrano-Malaver, M. A. et al. (2018): «Modelos de desarrollo rural e incidencia en medición de pobreza monetaria: Una aproximación econométrica desde el modelo Logit para el sector rural año 2017,» .

Cuadro 4: Modelos de predicción - Entrenamiento

	Variable dependiente: $Log(Salario)$				
	Modelo (6)	Modelo (7)	Modelo (8)	Modelo (9)	Modelo (10)
Mujer	−0.032	−0.073	−0.074	−0.055	0.003
	(0.109)	(0.107)	(0.101)	(0.101)	(0.098)
Edad	0.035***	0.034***	0.036***	0.033***	0.033***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Edad^2	−0.0003***	−0.0003***	−0.0004***	−0.0003***	−0.0004***
	(0.0001)	(0.00005)	(0.00005)	(0.00005)	(0.00005)
Educación superior	0.392***	0.375***	0.308***	0.309***	0.306***
	(0.021)	(0.020)	(0.020)	(0.020)	(0.019)
Horas trab. a la semana	0.013***	0.012***	0.012***	0.012***	0.041***
	(0.001)	(0.0005)	(0.0005)	(0.0005)	(0.002)
Informal	−0.456***	−0.426***	−0.418***	−0.414***	−0.335***
	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)
Educación media	0.123***	0.110***	0.096***	0.097***	0.087***
	(0.018)	(0.017)	(0.016)	(0.016)	(0.016)
Experiencia trab. actual		0.001***	0.001***	0.002***	0.002***
		(0.0001)	(0.0001)	(0.0002)	(0.0002)
Experiencia trab. actual^2				−0.00000***	−0.00000***
				(0.00000)	(0.00000)
Horas trab. a la semana^2					−0.0003***
					(0.00001)
Mujer·Edad	−0.002	0.001	0.002	0.001	−0.003
	(0.006)	(0.006)	(0.005)	(0.005)	(0.005)
Mujer·Edad^2	0.00001	−0.00004	−0.0001	−0.00004	0.00000
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Constante	13.614***	13.592***	13.351***	13.394***	12.730***
	(0.188)	(0.185)	(0.174)	(0.174)	(0.173)
Efectos fijos de oficio	Si	Si	Si	Si	Si
Efectos fijos de estrato	No	No	Si	Si	Si
R^2	0.589	0.602	0.650	0.651	0.669
R^2 - Ajustado	0.584	0.597	0.645	0.646	0.665
Observaciones	6,925	6,925	6,925	6,925	6,925

Notas: Cálculos propios usando datos DANE Medición de Pobreza Monetaria y Desigualdad 2018. *** p<0.01, ** p<0.05, * p<0.1