Taller 1 – Big data and machine learning
Julián Andrés Camargo - 201823119
Federico Ramírez - 201914038
Maria Camila Pinillos – 201913864

Link Repositorio: https://github.com/federicorr/Taller-1---BDML

Taller 1 BDMLE

Este trabajo parte de el constante subregistro de ingresos que existe en EEUU, el cual contribuye a que solo el 83.6% de los impuestos sean pagados voluntariamente y a tiempo. Por esto, a lo largo de este trabajo se buscará producir un modelo de ingreso individual que permita atacar este sub reporte. Para esto, se utilizará la Gran Encuesta Integrada de Hogares, base de datos del Departamento Administrativo Nacional de Estadística (DANE) y el Departamento Nacional de Planeación (DNP), la cual recoge datos socioeconómicos de la población colombiana, en esta podemos encontrar datos como edad, sexo, salario y demás variables que permiten categorizar a la población.

Esta información se extrajo del siguiente link, sin embargo, no se pudo hacer *scrapping* directamente de esta, por lo cual, tocó recurrir al link de donde la página sacaba los datos. Ya con este se pudo armar toda la base de datos, para todos los grupos de datos.

Originalmente, la base contaba con 32177 observaciones provenientes de la ciudad de Bogotá, sin embargo, como se quiere mirar el ingreso es mejor restringir la base a mayores de edad, lo que deja un total de 24568 observaciones y un total de 177 variables que contienen diferentes características socioeconómicas de los individuos.

Teniendo en cuenta el objetivo de modelar el ingreso, es necesario realizar una buena selección de la variable que represente esta característica en la base. Varias de las 177 variables representan el ingreso de los individuos en diferentes dimensiones, sin embargo, en el trabajo se eligió utilizar la variable y_total_m como medida del ingreso ya que recoge tanto el ingreso salarial como el ingreso independiente de las personas. Esto permite aproximar de una mejor manera los

ingresos de las personas y abarcar grupos de interés como lo son los informales o los independientes. En este punto hubo una discusión acerca de si era mejor elegir esta variable o ingtot que representa los ingresos totales de las personas. Sin embargo, se eligió y_total_m por la imposibilidad de distinguir entre NAs y 0 en la variable ingtot.

Ya teniendo la variable dependiente que nos mide el ingreso se empezó a indagar por otras relaciones entre las variables. Primero, se consideró la opción de utilizar la variable y_salary_m ya que permitía tener ver el salario percibido por las personas, además de las propinas y demás ingresos que tuvieran. Pero en esa situación solo estaríamos teniendo en cuenta a las personas, dependientes y con empleo. Dejando por fuera a un grupo importante de personas que no cuentan con un salario base pero que si cuentan con más ingresos. Al final, aunque consideramos que la variable y_salary_m si hubiera podido ser útil, era más significativa continuar trabajando con la variable y_total_m. Segundo, se investigó la relación entre la edad y los ingresos, basándose en la afirmación de que los ingresos tienden a ser bajos cuando la persona es joven, y a través de los años de trabajo, también aumenta el salario percibido. Salario que se espera lleva a su máximo a los 50 años e luego tienden a mantenerse o declinarse por el resto de la vida.

Ingreso-Edad

Dependent variable:	
	Ingreso
edad	128,184.800***
	(10,744.940)
edad^2	-1,276.714***
	(132.932)
Constante	-1,227,844.000***
	(201,912.000)
Observaciones	9,892
R2	0.031
Adjusted R2	0.031
Residual Std. Error	2,115,556.000 (df = 9889)
F Statistic	157.436*** (df = 2; 9889)

==========
Note: *p<0.1; **p<0.05; ***p<0.01

Para esto se estimó la regresión de ingresos contra edad y edad al cuadrado. Como el termino se incluye al cuadrado el efecto marginal de la edad contiene al termino de la edad. Por lo tanto, la interpretación en este caso debe hacerse partiendo los términos y mirando sus símbolos, como se puede ver en el término edad, efectivamente el que la edad aumente en 1 esta correlacionado con que los individuos ganen más en promedio, sin embargo, el termino de edad al cuadrado nos indica que esta función es cóncava lo cual indica que el ingreso efectivamente va aumentando, pero cada vez en menor cuantía hasta que llega a un máximo. Ambos términos son significativos al 1% lo que quiere decir que los cambios en estas variables son relevantes para entender los cambios en la variable dependiente. La regresión se estima con 9892 observaciones y tiene un r2 bajo.

Lo anterior nos confirma que la idea de que la relación ingresos y edad tiene el comportamiento descrito el cual señala que empieza siendo bajo, luego aumenta hasta que llega a un máximo y luego decrece.

También se exploró la brecha de ingresos por género, esta enfocada en entender si el hecho de ser mujer afectaba los ingresos de la persona. Para esto se estimo una regresión teniendo al logaritmo de una variable ingreso como variable dependiente y a la variable female como una variable dummy dicotoma que toma el valor de 1 cuando el individuo es mujer y 0 cuando es hombre.

Los resultados de esta regresión muestran que la variable que indica el sexo femenino es significativa al 1% y es negativa, lo que quiere decir que el hecho de que una persona sea mujer disminuye en 23.8% el ingreso de una mujer frente al ingreso de un hombre. Y esta brecha de género que se sigue presentando, se da incluso entre hombres y mujeres que desempeñan el mismo trabajo en el mismo cargo.

Además, considerando que para este punto debíamos trabajar tanto con ingresos como con salario a través de los años y género. Se dio la discusión sobre si era coherente trabajar con una

variable para entender los ingresos, y una variable para entender si este aumentaba o no a lo largo de los años. Volvió a surgir la discusión sobre si trabajar con y_total_m y y_salary_m seria coherente con lo que esperamos. Y la respuesta fue negativa, al final decidimos volver a la pregunta anterior y desarrollar la investigación teniendo como base solo y_total_m que es más significativa, de manera que hubiera más coherencia en el desarrollo de todo este trabajo. a

Ingreso-Femenino

	Dependent variable:
	log(ingreso)
Femenino	-0.238***
	(0.015)
Constante	13.981***
	(0.010)
Observaciones	14,764
R2	0.018
Adjusted R2	0.017
Residual Std. Error	0.889 (df = 14762)
F Statistic	263.841*** (df = 1; 14762)
Note:	*p<0.1; **p<0.05; ***p<0.01