Problem Set 1:

Predicting Income

**Integrantes:** \_\_\_\_\_\_\_\_\_\_

## **Introducción**

Un desafío para el sector público es precisar los ingresos individuales para aproximarse al recaudo de impuestos y evitar el fraude fiscal, sin embargo, la complejidad del sistema tributario y la recolección de datos hace del estudio tributario una tarea difícil dada la complejidad del sistema tributario y la disponibilidad de los datos (IRS, 2022). En Estados Unidos, las estimaciones realizadas por el Sistema de Impuestos Internos (IRS) muestran para 2017 – 2019 una brecha tributaria bruta promedio de $540 mil millones de dólares por año y una proyección de la tasa de cumplimiento voluntario del 85%. Para el caso de Colombia, tener una aproximación del ingreso resulta una importante medida para estimar la tributación regional debido a las disparidades fiscales en Colombia, donde regiones del centro del país registran menores desequilibrios en su capacidad fiscal debido a su menor brecha de tributación (Bonet y Ayala, 2016).

Dado esto, una solución al momento de obtener los datos es la Gran Encueta Integrada de Hogares (GEIH) debido a que es la principal fuente de información de las estadísticas del mercado laboral en Colombia. Brindando información de las condiciones de los hogares con respecto a sus ingresos y condiciones de vida o características poblacionales como nivel de educación, sexo, edad, entre otros.

Este estudio utiliza datos para Bogotá del reporte de la “Medición de Pobreza Monetaria y Desigualdad” de 2018, esta base recoge información de la GEIH y datos adicionales que pueden ser de utilidad al momento de realizar predicciones de los ingresos con el objetivo de identificar casos de fraude que podrían ayudar a reducir la brecha fiscal y adicionalmente, poder identificar familias e individuos en condición de vulnerabilidad que serían potencialmente beneficiarios de asistencia adicional a través de focalización del gasto.

Con el fin de obtener una estimación más precisa de los ingresos reales de la población, se han aplicado diversos modelos predictivos. En este contexto, hemos seleccionado variables de control basadas en la ecuación de Mincer. Según Mincer (1974), existe una relación positiva entre el nivel educativo de un individuo y sus ingresos futuros, así como con su experiencia laboral. Por lo tanto, hemos decidido emplear esta teoría como base para nuestras proyecciones con el objetivo de obtener resultados coherentes y fundamentados. La mejora en la aproximación de los ingresos reales de la población no solo es valiosa por sí misma, sino que también proporciona un conocimiento más profundo sobre la distribución de ingresos. Al desglosar esta distribución según variables como la edad y el género, podemos obtener información crucial para la formulación o ajuste de políticas públicas destinadas a la redistribución de recursos hacia grupos específicos de la población.

## **Datos**

Los datos de la encuesta “Medición de Pobreza Monetaria y Desigualdad 2018” es elaborada a partir de la “Gran Encuesta Integrada de Hogares” por el DANE. Usando los datos de esta encuesta se pretende predecir el nivel de ingreso laboral de ocupados mayores de 18 años radicados en la ciudad de Bogotá, con el fin de desarrollar un modelo predictivo que pueda generar alertas relevantes para las autoridades encargadas de la recaudación de impuestos, utilizando ciertos parámetros específicos.

Para recopilar los datos, realizamos web scraping seleccionando 10 archivos de la web que contienen información de la encuesta para cada mes de 2018 en la ciudad de Bogotá. Es importante destacar que la página de la cual se extrajo los datos no impone restricciones para el uso de web scraping. Inicialmente, se cuentan con 32.177 observaciones para Bogotá, pero nos enfocamos en 24.568 personas mayores de edad y 19.801 de la población ocupada. Luego, eliminamos las observaciones que tenían valores faltantes en la variable de salario nominal mensual, lo que resultó en la eliminación de 9.892 observaciones debido al alto porcentaje de datos faltantes.

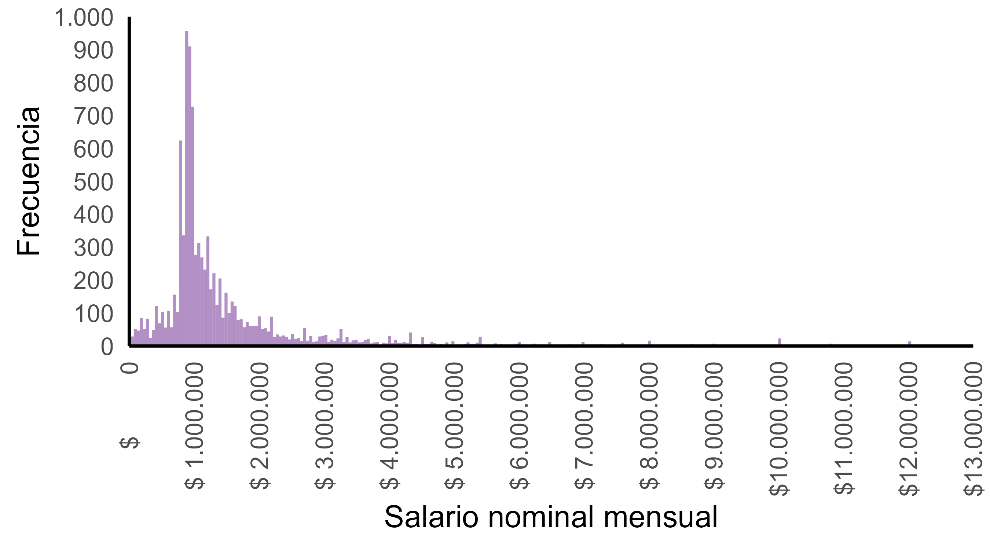
Tabla 1. Estadísticas Descriptivas – Colombia 2018

|  |  |  |
| --- | --- | --- |
|  | **Todos** | |
|  | Media | Desviación Estándar |
| **Individuo** |  |  |
| Edad | 36,24 | 12,02 |
| Amo casa | 0,030 | 0,171 |
| **Hogar** |  |  |
| Núm. hijos hogar | 0,235 | 0,424 |
| Estrato | 2,509 | 0,975 |
| **Educación** |  |  |
| Estudiante | 0,010 | 0,101 |
| Primaria | 0,005 | 0,067 |
| Secundaria | 0,095 | 0,293 |
| Media | 0,346 | 0,476 |
| Superior | 0,453 | 0,498 |
| **Mercado Laboral** |  |  |
| Salario mensual | 1.745.416 | 2.403.441 |
| Ingreso total | 1.872.592 | 2.509.096 |
| Experiencia trabajo actual | 49,73 | 73,20 |
| Horas trabajadas en la semana | 48,02 | 12,15 |
| Informalidad | 0,233 | 0,422 |
| **Observaciones** | 9.832 | |

Fuente: Cálculos propios usando Medición de Pobreza Monetaria y Desigualdad 2018 del DANE

La tabla 1 resume los estadísticos descriptivos de la muestra para el total de los datos que se utilizaron en la estimación de los modelos. En la muestra podemos encontrar 9.832 personas con un promedio de edad de 36,2 años, donde el promedio de los hogares cuenta reportan un nivel socioeconómico –estrato– de 2,51. En cuanto a los niveles educativos en Bogotá la educación superior predomina con el 45,3%. El salario mensual promedio es de $1.745.416 para el año 2018, sin embargo, el ingreso total es mayor dado que incluye otras fuentes de ingresos estando en promedio son de $1.872.592; las horas que trabajan a la semana en promedio es de 48,02. Por último, los niveles de informalidad llegan a ser del 23,3%.

Gráfico 1: Distribución de los salarios nominales



La figura 1 muestra la distribución de los salarios mensuales para la muestra, en la cual se puede evidenciar una concentración de los salarios entre $0 y $2.000.000, teniendo así un sesgo hacia la izquierda en la distribución de los salarios.

## **Age-wage profile**

Tabla 2. Regresión Salario y edad

|  |  |
| --- | --- |
|  | |
|  | *Variable dependiente:* |
|  |  |
|  | Log(Salario mensual) |
|  | |
| edad | 0.089\*\*\* |
|  | (0.004) |
|  |  |
| edad^2 | -0.001\*\*\* |
|  | (0.00005) |
|  |  |
| Constante | 12.289\*\*\* |
|  | (0.071) |
|  |  |
|  | |
| Observaciones | 9,892 |
| R2 | 0.058 |
|  | |
| *Nota:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |

Tabla 4. Intervalo de confianza de la edad Optima

|  |  |
| --- | --- |
| **Intervalo de confianza** | **Valor** |
| Límite inferior | 42.45 |
| Valor predicho | 43.14 |
| Límite superior | 43.89 |

Tabla 3. Regresión Salario y edad con errores estándar robustos

|  |  |
| --- | --- |
|  | |
|  | *Variable dependiente:* |
|  |  |
|  | Log(Salario mensual) |
|  | |
| edad | 0.089\*\*\* |
|  | (0.004) |
|  |  |
| edad^2 | -0.001\*\*\* |
|  | (0.0001) |
|  |  |
| Constante | 12.289\*\*\* |
|  | (0.079) |
|  |  |
|  | |
| Observaciones | 9,892 |
| R2 | 0.058 |
|  | |
| *Nota:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 |

Gráfico 4: Intervalos de confianza del salario predicho según edad

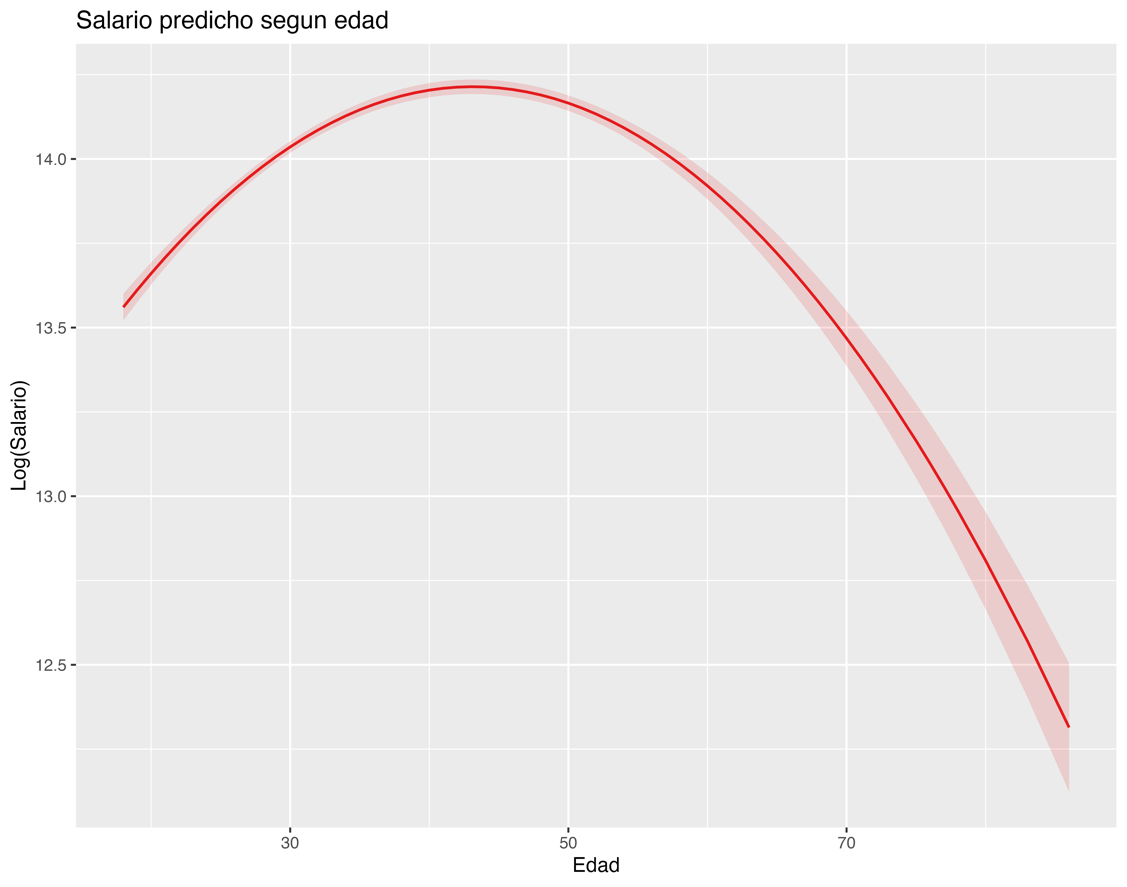
****

Gráfico 3: Salario predicho según edad

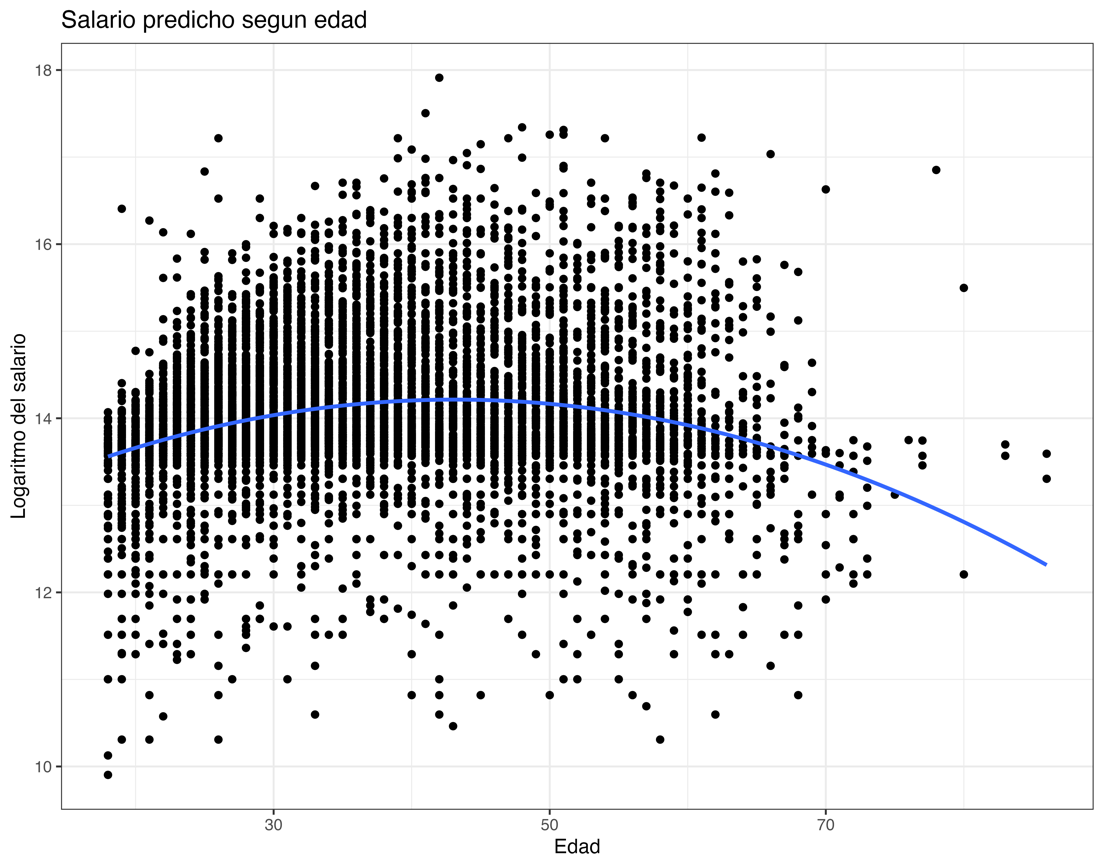
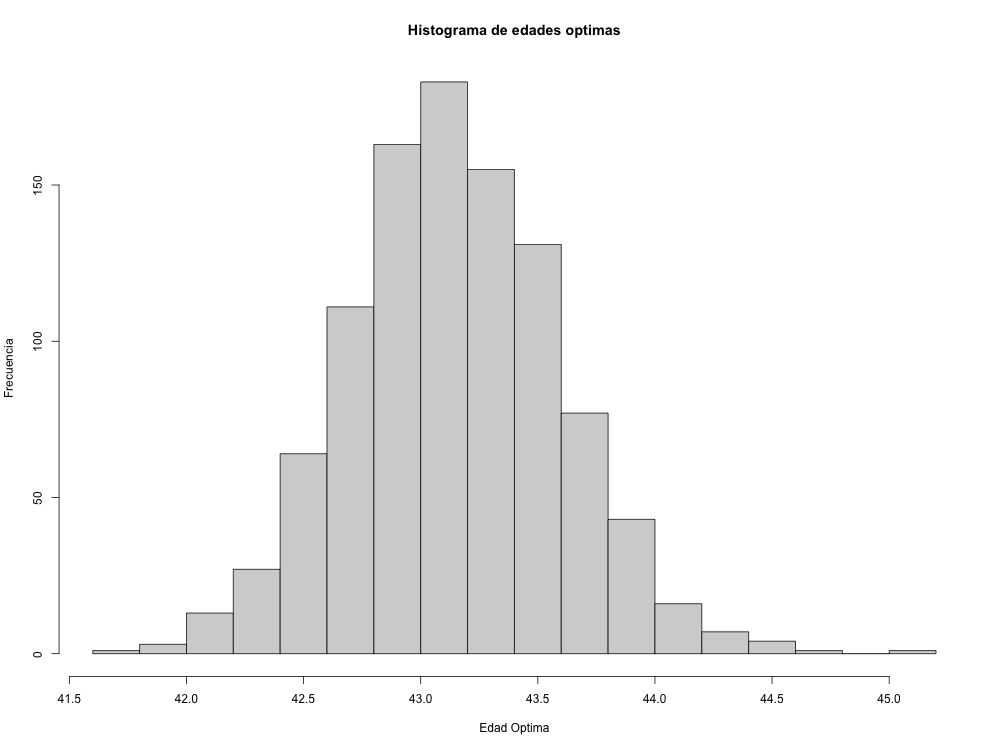
****

Gráfico 4: Distribución de los valores estimados por *bootstrap*

****

**Referencias**

Bonet-Morón, J., y Ayala-García, J. (2016). La brecha territorial en Colombia. *Documento de trabajo sobre Economía Regioanl,* (235).