Problem Set 1:

Predicting Income

**Integrantes:** Juan Diego Duarte[[1]](#footnote-1); Erick Julian Villabon[[2]](#footnote-2) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

## **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.

Falta un resumen de nuestros resultados o lo que estamos hallando

El resto de este trabajo se desarrolla de la siguiente manera: la sección 2 proporciona información sobre la recopilación, tratamiento y descripción de los datos. Posteriormente, la sección 3

## **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.

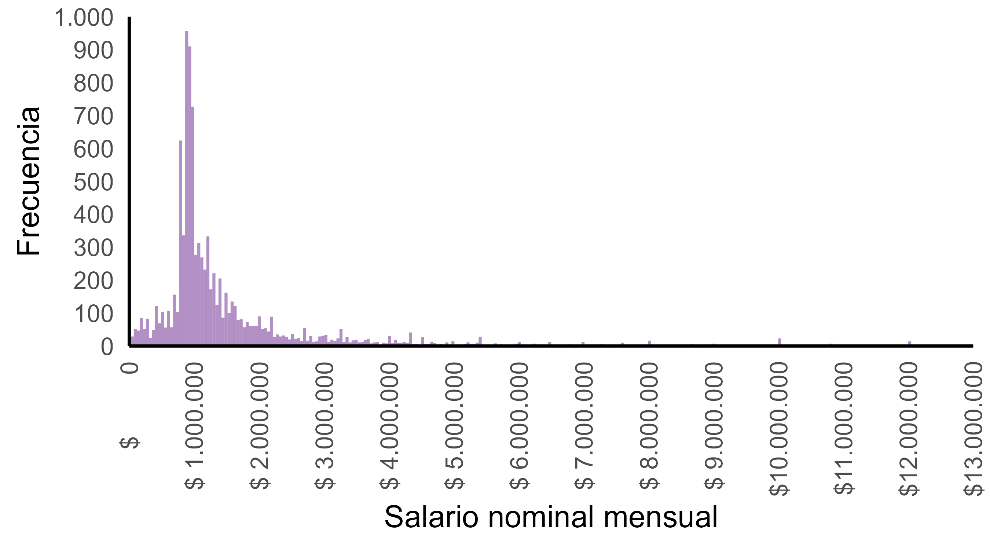
Tabla 1. Estadísticas Descriptivas – Colombia 2018

|  |  |  |
| --- | --- | --- |
|  | **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**

Teniendo esto en cuenta, para entender de manera integral el comportamiento de los salarios es fundamental comprender que factores que los afectan y de qué manera lo hacen. En este orden de ideas, según la literatura uno de los elementos más importantes para la determinación de los salarios y la productividad laboral, es la edad. En términos generales, la literatura determina que los años de una persona capturan de manera directa e indirecta factores que están directamente asociados con la productividad laboral.

El mecanismo indirecto sería el siguiente: Las personas jóvenes tienen una mayor probabilidad de poseer menos años de educación y años de experiencia, y mientras más años de vida se tiene, menores son esas probabilidades. Así mismo, el mecanismo directo se puede caracterizar así: Las personas más jóvenes tienen una mayor capacidad de adaptación al ambiente laboral, lo cual aumenta su productividad. De la misma manera, mientras menos edad se tiene, más competencia laboral existe, lo cual en ultimas aumenta la productividad por efecto de mercado. A lo cual se le suma que las personas jóvenes están asociadas a un capital humano mucho más actualizado, y unas cualidades “blandas” (motivación, necesidad de trabajar, escalabilidad de puestos, etc.) especiales que aumentan la productividad.

Todos estos mecanismos asociados a la edad provocan que la literatura sea muy concisa en mostrar que hay un claro *trade-off* cuando hablamos del efecto de la edad sobre el salario. Básicamente, con una mayor edad se van ganando años de educación y años de experiencia, pero a su vez se van perdiendo otras cualidades y se van ganando limitaciones que reducen la productividad. Por lo que la literatura es muy clara al determinar que durante la juventud y la adultez los salarios aumentan a medida que van pasando los años, pero se llega a un punto en que los salarios empiezan a decrecer al aumentar la edad.

Este *trade-off* entre edad y salarios se puede caracterizar en un modelo cuadrático de la siguiente manera:

De manera tal que el salario se relaciona de manera cuadrática con el salario, como ya describimos. Los resultados de este modelo sobre la muestra que describimos en el Punto 2 de este documento se pueden presentar de la siguiente manera:

Tabla 2. Regresión Salario y edad con errores estándar robustos

|  |  |
| --- | --- |
|  | |
|  | *Variable dependiente:* |
|  |  |
|  | Log(Salario mensual) |
|  | |
| Edad | 0.089\*\*\* |
|  | (0.004) |
|  |  |
| Edad2 | -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 |

De manera tal que el efecto marginal de la edad sobre el salario mensual es de (8.9 – 0.2\*Edad) puntos porcentuales por cada año de edad adicional. En términos sencillos, cada año de vida adicional aumenta 8.9 % el salario mensual, pero este efecto se va reduciendo a medida que aumenta la edad de la persona, hasta que llega un punto en el que tener más edad implica una reducción del salario. Para ilustrar mejor este efecto, es importante plantear los siguientes ejemplos: Si una persona pasa de 0 años de vida a 1 año, entonces su salario se mejoró en un 8,9%; pero, si una persona pasa de tener 20 años a tener 21 años, entonces su salario en promedio se habrá mejorado en un 4.9%; y si una persona pasa de tener 50 años a 51 años, entonces su salario en promedio se reducirá en 1.1%. Esta relación se puede ver ilustrada claramente por los siguientes gráficos:

Gráfico 3: Intervalos de confianza del salario predicho según edad

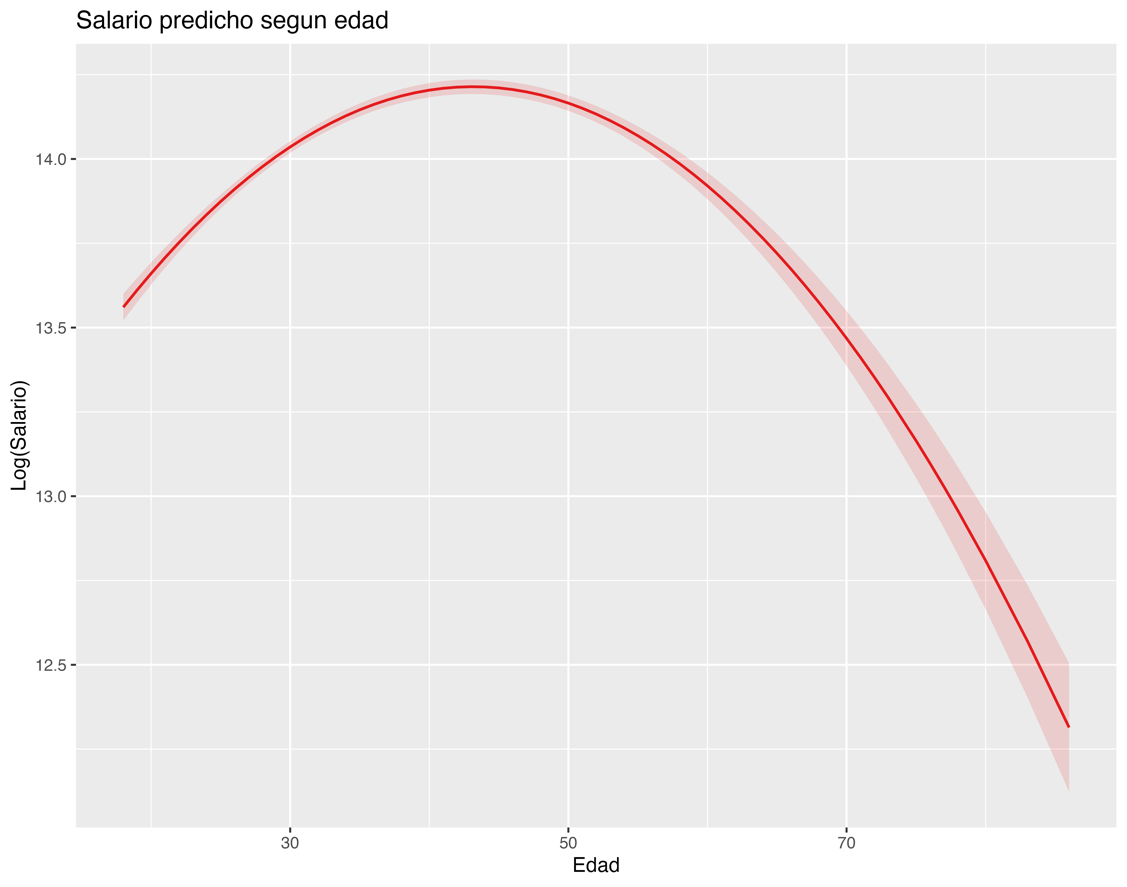
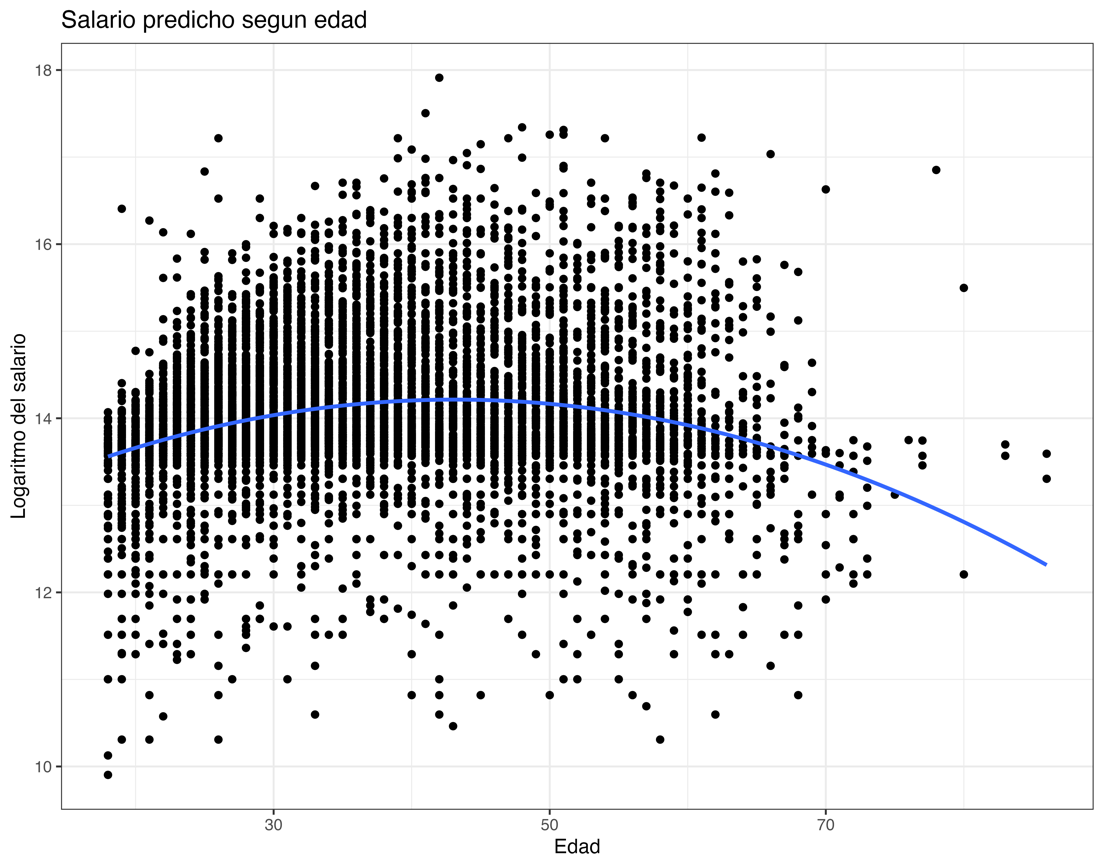
****

Gráfico 4: Salario predicho según edad

****

En este caso, se puede observar en ambos gráficos la relación cuadrática evidente entre la edad y el salario mensual, llegando a un punto entre 40 y 50 años que maximiza en el salario mensual. Sumado a esto, el grafico 3 se compone del valor estimado del efecto marginal de la edad para cada edad determinada, junto a su intervalo de confianza; para lo cual podemos observar que a medida que va aumentando la edad, hay un mayor error estándar, gracias a la mayor variabilidad de los datos en estos puntos. El grafico 4 ilustra la relación entre ambas variables y muestra la distribución de los datos de la muestra.

De manera paralela podemos observar que tanto el efecto marginal lineal, como el efecto marginal del factor cuadrático (como se va comportando la pendiente del efecto de la variable) son significativas al 99% de confianza. A su vez, en esta regresión poseemos un R2 de 0.058, que se traduce en que nuestro modelo está prediciendo el comportamiento del salario mensual en un 5.8%. Aunque con el método de OLS se están generando estimadores insesgados, y por ende, se está maximizando la predicción dentro de la muestra, esta solo esta prediciendo un 5.8% del salario mensual. De hecho, en este caso tenemos un RMSE de 0.7431, que quiere decir que en promedio el logaritmo del salario predicho se diferencia del real en un 0.7431 dentro de la muestra; esto último puede ser considerado como un valor grande si se tiene en cuenta las escalas bajo las que se mueven los Log(Salario Mensual) en el Grafico 3. En términos generales, a priori no se puede afirmar que el modelo este haciendo un excelente trabajo a la hora de predecir los datos de la muestra. Cabe destacar que esto no implica que el modelo no nos sirva para encontrar causalidad y efectos marginales, porque como ya mostramos anteriormente estos son relevantes estadísticamente hablando.

Para caracterizar de la mejor forma posible esta relación cuadrática del salario y la edad, se vuelve fundamental entender ese punto que maximiza el salario mensual. En este orden de ideas, la expresión que maximiza ese salario es la siguiente:

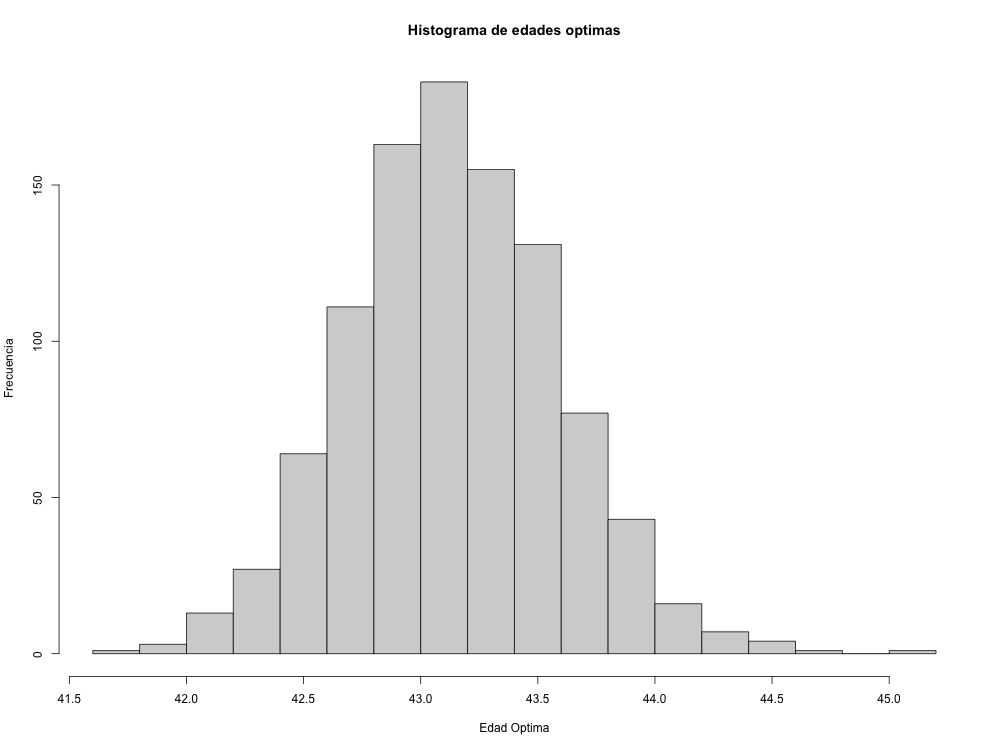
De manera tal que a la hora de predecir este valor, por medio de la metodología *bootstrap* con 1000 iteraciones, obtenemos el siguiente valor predicho:

Tabla 3. Intervalo de confianza de la Edad Optima

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

Con la siguiente distribución del estadístico de Edad Optima:

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

****

Por lo que podemos concluir que la edad que maximiza el salario mensual esta entre 42.35 años y 43.89 años, con un 95% de confiabilidad.

## **…**

**Referencias**

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

1. Código: [↑](#footnote-ref-1)
2. Código: [↑](#footnote-ref-2)