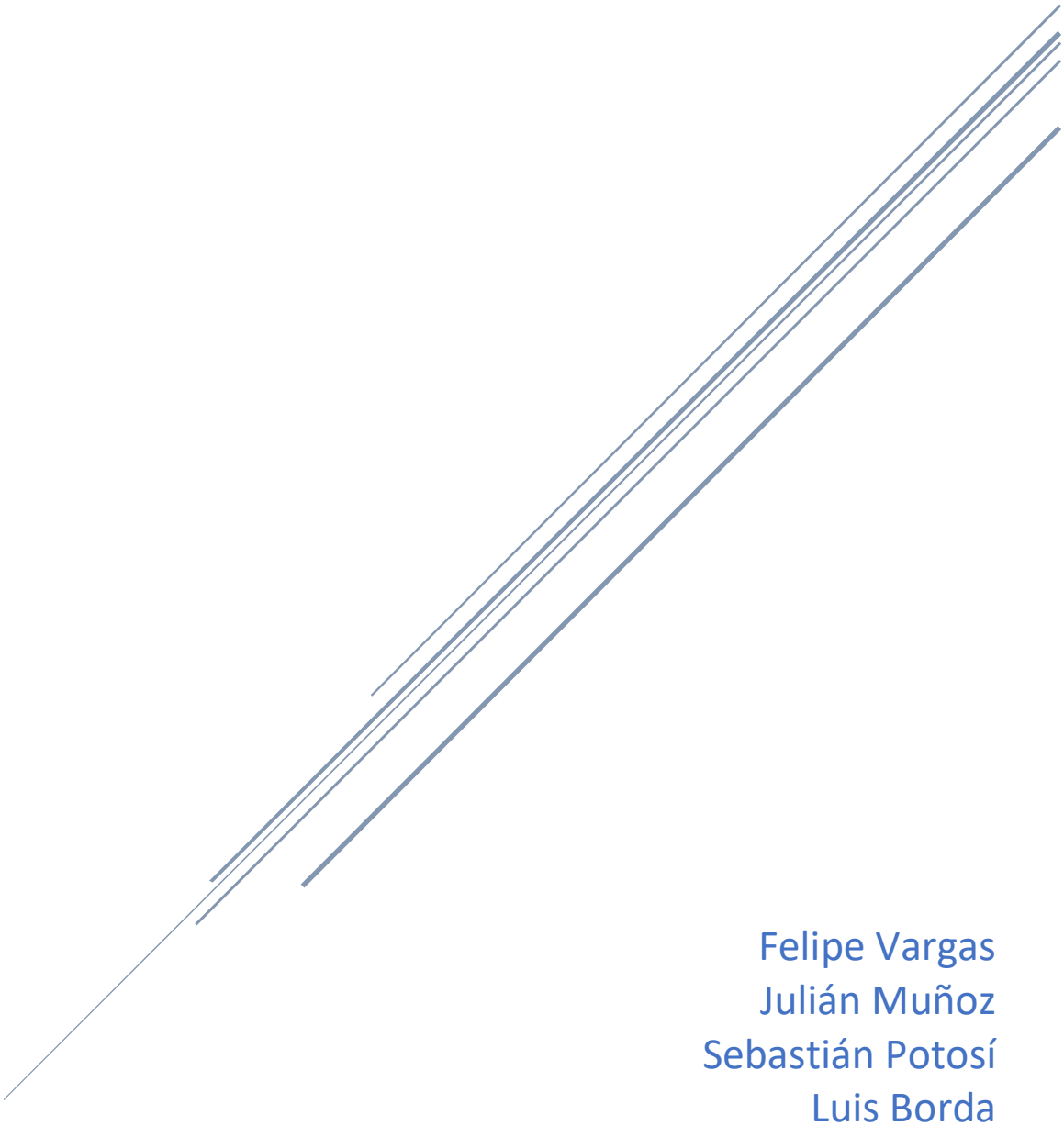
TALLER 1

BIG DATA Y MACHINE LEARNING PARA ECONOMÍA APLICADA



Felipe Vargas
Julián Muñoz
Sebastián Potosí
Luis Borda
MECA

Contenido

1 INTRODUCCIÓN

La evasión tributaria tiene grandes repercusiones sobre el desarrollo y crecimiento de los países, limitando las capacidades del gobierno como gestor de política pública. Los desafíos relacionados con la regresividad en la recaudación de impuestos aumentan la brecha de desigualdad y afectan directamente el bienestar de los ciudadanos, esto implica que, aquellos individuos cuyo nivel de ingresos es alto, su tasa efectiva para tributar disminuye, afectando la equidad social y haciendo que los individuos con menos recursos tengan cargas impositivas mayores. Esto puede resultar en una disminución de los ingresos fiscales totales del gobierno afectando significativamente al presupuesto del gobierno y dificultando la posibilidad de emplear el gasto de manera eficiente y en alguna medida austero. (Moller, 2012)

La evasión tributaria se ve fuertemente afectada por la informalidad y el desempleo, la informalidad laboral es la proporción de ocupados que reciben ingresos al margen del control tributario y bajo condiciones que no están dentro del aval institucional. El trabajador informal, normalmente, no paga sus cargas correspondientes a la obra social y a la jubilación, además, no dispone de seguro médico (entre otras cosas)**.** El problema de la informalidad laboral en Colombia repercute significativamente en la seguridad social, especialmente porque los trabajadores informales no tienen la posibilidad de realizar las cotizaciones que constituyen una fuente de financiación en el sistema de pensiones.

Así mismo, la ausencia de seguridad social es una característica que define a las personas ocupadas dentro del sector informal, en este sentido, se puede hablar de una relación inversa entre subsidios o ayudas por parte del Estado y el incentivo de participar en el mercado laboral, donde a mayor ayuda del gobierno la probabilidad de pertenecer a la informalidad es mayor. Estas ayudas pueden generar incentivos perversos sobre la decisión de participar en el mercado laboral por parte de los integrantes de un hogar beneficiario del programa, generando mayor comodidad en pertenecer al régimen subsidiado que al contributivo y a su vez, propiciando a que las personas prefieran trabajos informales (Farné, David, & Paola, 2016), hacienda el problema de la evasión tributaria cada vez más insostenible.

Según la teoría convencional de ingreso de los hogares, la entrega de una transferencia produce un puro efecto ingreso al igual que por alquileres, intereses, sueldos y salarios, entre otros. Por lo que, "si la entrega del subsidio está sujeta a los (bajos) ingresos del hogar, puede que algunos miembros de la familia decidan trabajar menos, o simplemente no trabajar, para no exponerse a ser reportados como receptores de remuneraciones por encima de los umbrales máximos establecidos para ser beneficiarios" (Moffit, 1996). Atendiendo a estas consideraciones, se entiende que los subsidios pueden desestimular la permanencia en el mercado del trabajo o la formalización en el mismo.

En retrospectiva, es imperativo el estudio de ingresos individuales en el contexto de una reforma estructural para la evasión de impuestos, considerando las diversas condiciones salariales de la población mayor a 18 años, empleando métodos técnicos que ayuden a predecir el nivel de ingresos asertivamente. Por lo que es de interés en este trabajo de

investigación, estimar un modelo econométrico a partir de los datos obtenidos en la Gran Encuesta Integrada de Hogares para 2018 realizada por el Departamento Administrativo Nacional de Estadística (DANE), que permita estimar correctamente el ingreso de los individuos de la muestra, considerando aspectos como su género, edad, etc. El trabajo de investigación tendrá cinco secciones, de la cual la primera fue esta introducción, en la segunda sección se explicará el proceso de adquisición, limpieza y descripción de los datos y las variables utilizadas. La tercera establece un modelo perfil Edad-Salario en el que a partir de una transformación aritmética básica de la variable "Edad" (elevando a la segunda potencia) se podrá analizar si el salario de una persona aumenta de manera proporcional a medida que su edad también incrementa, se utilizará *bootstrap* y *fit-model* para mostrar los resultados. La cuarta analiza la brecha salarial de género. La quinta y última parte, se realizarán predicciones en los ingresos por modelo, las conclusiones y limitaciones de esta investigación.

2 DATOS

En esta sección se realizará una descripción de los datos utilizados para el análisis y su finalidad. Posteriormente, se mostrará el proceso realizado para obtener dichos datos, su proceso de limpieza y procesamiento. Por último, el lector podrá encontrar un reporte de estadísticas descriptivas de las variables que se utilizarán para el desarrollo del *set*.

Descripción de los datos

La información utilizada proviene del informe de Medición de Pobreza Monetaria y Desigualdad del año 2018, realizado por el DANE de la Gran Encuesta Integrada de Hogares (GEIH). Esta encuesta proporciona información estadística sobre el tamaño y estructura de la fuerza de trabajo (empleo, desempleo y población fuera de la fuerza de trabajo), los ingresos laborales y no laborales de los hogares, la pobreza monetaria y la pobreza monetaria extrema de la población residente en el país (DANE, 2023).

Las temáticas por las cuales se indagan en la GEIH permiten caracterizar a la población según sexo, edad, parentesco con el jefe del hogar, nivel educativo, afiliación al sistema de seguridad social en salud, grupos poblacionales como etnias, campesinos, LGBT o con algún tipo de discapacidad, otras formas de trabajo como producción de bienes y servicios para autoconsumo, trabajo en formación y voluntariado, entre otras.

Actualmente, la GEIH cuenta con una muestra anual aproximada de 315.000 hogares a nivel nacional, lo que hace que sea la encuesta de mayor cobertura a nivel nacional. De modo que permite obtener indicadores confiables y series continuas para analizar la fuerza de trabajo del país y los principales indicadores del mercado laboral, considerados como información fundamental para la toma de decisiones de política pública.

Respecto al diseño estadístico, la GEIH tiene cobertura nacional con diferentes niveles de desagregación temporal y geográfica: total nacional, total de cabeceras de ciudades (con o sin áreas metropolitanas), grandes agrupaciones (cabeceras, centros poblados y rural disperso) y

departamentos. Además, tiene desagregación anual, semestral, trimestral y mensual. Por último, su unidad de observación al igual que su unidad de análisis son las viviendas, los hogares y las personas.

Ahora bien, para este caso particular, el análisis se centra en las personas empleadas mayores de 18 años que viven en Bogotá. En consecuencia, la base de datos a utilizar contiene información para 16.542 registros.

Proceso de adquisición de los datos

Teniendo en cuenta que para este problem set la base de datos a utilizar se encuentra almacenada en una página web, se hizo necesario efectuar un scrape del website. Para esto, en primer lugar, se exploró la URL referida para así identificar de qué forma estaba almacenada la información en el website. Al realizar este ejercicio se pudo observar que la base de datos a utilizar se encuentra dividida en 10 tablas diferentes.

Posteriormente, se identificó la dirección URL a utilizar. Para esto, se inspeccionó el website basado en un enfoque de ensayo y error para así identificar las tablas a extraer. Una vez hecho esto, la principal restricción a la que nos enfrentamos para consolidar la base es que los datos estaban en tablas distintas. Por esta razón, fue necesario hacer una iteración para pegar cada tabla y así consolidar la base final.

Limpieza de los datos

Por último, se filtró la base de datos para quedar con las observaciones de personas mayores de 18 años y que se encuentran ocupadas. El resultado es una base de datos con 14.763 observaciones y 178 variables, ahora para poder seleccionar dentro de este global reducido por los respectivos filtros, se realizó un conteo de los datos faltantes en la base de datos. Se mantienen las variables que tienen el 15% de datos faltantes o menos, para que los modelos estimados no pierdan poder de predicción, de modo que se pasaron de 178 variables a 65, sobre estas ultimas es que hicimos la selección de las variables de interés.

En consecuencia, seleccionamos las variables que a partir de la teoría económica y de los datos que se tenían al alcance resultaban significativos, llegando a las siguientes variables:

- **Nivel educativo (maxEducLevel):** Becker en su libro *"Education and Earnings" (1975)* sostuvo que la educación es una inversión en capital humano y que las personas que invierten en su educación tienden a ganar salarios más altos. La educación formal, como títulos universitarios o posgrados, puede mejorar las perspectivas salariales.
- **Tiempo de Ocio (P6240):** En el artículo *"A Theory of the Allocation of Time" (1965)* En, Becker examina cómo las personas toman decisiones sobre cómo asignar su tiempo entre el trabajo, el ocio y otras actividades, lo que tiene implicaciones para los ingresos laborales.
- **Experiencia laboral (P6426):** Becker menciona en su artículo *"Investment in Human Capital: A Theoretical Analysis" (1962)* que el que desarrolla que la experiencia laboral, vista como capital humano afectan los ingresos laborales a lo largo del tiempo.
- **Sexo- Edad- Estrato:** Becker también abordó la discriminación laboral y cómo factores como el género, la raza o la etnia pueden influir en los salarios. Sus investigaciones contribuyeron a la comprensión de la discriminación en el mercado laboral y cómo puede afectar la distribución de los ingresos.

También vimos pertinente crear una nueva variable con las siguientes características:

- **Rezago Educativo (maxEducLevel*age):** Es la multiplicación entre la edad y el nivel educativo y es relevante ya que las personas con rezago educativo (Educación no acorde a la edad) pueden tener menos habilidades y conocimientos que las personas con mayor nivel educativo. Esto puede limitar sus oportunidades de empleo y reducir su capacidad para desempeñarse en trabajos que requieren habilidades específicas. Como resultado, es posible que reciban salarios más bajos.

Finalmente se empleó para variable dependiente en cada uno de los modelos que se implementaran más adelante el salario por horas, como la siguiente variable:

- **Salario por hora (y_total_m_ha):** Seleccionada debido que es nuestra variable de interés y esta variable capta el salario más ingresos de labor como independiente, lo cual es equivalente a la nominal mensual por horas.

Tras seleccionar estas variables con la muestra ya reducida a la población de interés aún teníamos dentro de estas ocho variables valores faltantes, con la siguiente estructura:

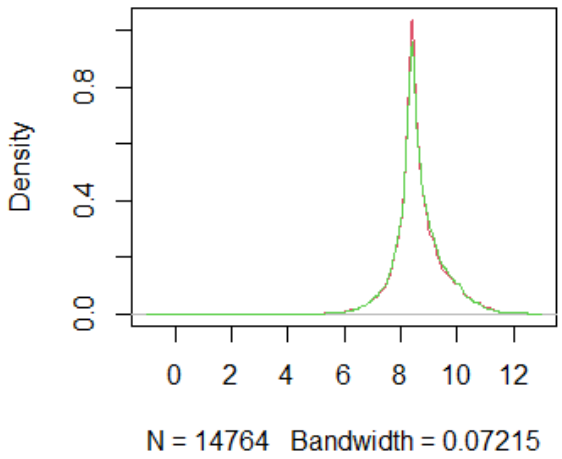
Tabla 1: Variables seleccionadas y valores missing

Variable	Naturaleza	Vacios	% Vacios
estrato1	Estrato socieconomico	0	0,0%
sex	Sexo	0	0,0%
age	Edad	0	0,0%
p6240	Repartir el tiempo libre	0	0,0%
p6426	Experiencia laboral	0	0,0%
maxEducLevel	Años de Educación	1	0,0%
ocu	Ocupado	0	0,0%
maxEducLevel*Edad	Rezago educativo	1	0,0%
y_total_m_ha	Salario + Independiente - nominal mensual	1778	10,7%

Fuente: Elaboración propia

Donde particularmente es de interés poder poblar los valores de la variable de salario por hora (y_total_m_ha) ya que cuenta con un porcentaje de valores faltantes de casi el 11%, para poblar esta variable se empleó el método de regresión estocástica; este procedimiento comienza por determinar la intercepción, pendiente y varianza residual en el modelo lineal. A continuación, estima el valor predicho para cada dato que falta y añade un componente aleatorio basado en el residuo a la predicción, este método mantiene los coeficientes de regresión intactos y también conserva la correlación entre las distintas variables. Tras realizar este proceso de imputación, la siguiente gráfica permite observar que tan bien quedo la imputación:

Figura 1: Salario por hora



Fuente: Elaboración propia

Donde la línea roja es con los datos imputados y la línea verde con los datos faltantes, se observa un buen ajuste y se puede continuar con la variable imputada.

Análisis descriptivo de los datos

Para el desarrollo del problem set, se utilizarán las variables estrato, sexo, edad, actividad en que ocupó la mayor parte de su tiempo, antigüedad laboral (tiempo que lleva trabajando con su actual empleador), nivel de educación más alto alcanzado, si la persona es ocupada y el salario por hora. La siguiente tabla muestra las estadísticas descriptivas de las variables mencionadas:

Tabla 2: 2Estadísticas descripticas de las variables utilizadas

Nombre variable	Descripción	Missing	Media	Desviación estándar	P0	P50	P100	Histograma
estrato1	Estrato	0	2,55	1,0	1	2	6	▇▂▁▁
sex	Sexo	0	0,53	0,5	0	1	1	▇▁▁▁▇
age	Edad	0	39,44	13,5	18	38	94	▆▇▂▁
p6240	Actividad principal	0	1,55	1,4	1	1	6	▇▁▁▁▁
p6426	Antigüedad laboral	0	63,76	89,5	0	24	720	▇▁▁▁▁
maxEducLevel	Nivel de educación más alto alcanzado	1	5,95	1,2	1	6	7	▁▁▁▁▇
ocu	Toma el valor de 1 si la persona esta ocupada	0	1,00	0,0	1	1	1	▁▁▇▁▁
dsi	Toma el valor de 1 si la persona esta desempleada	0	0,00	0,0	0	0	0	▁▁▇▁▁
y_total_m_ha	Salario por hora	1778	8541,87	13866,1	0,5	4837,5	350583,3	▇▁▁▁▁

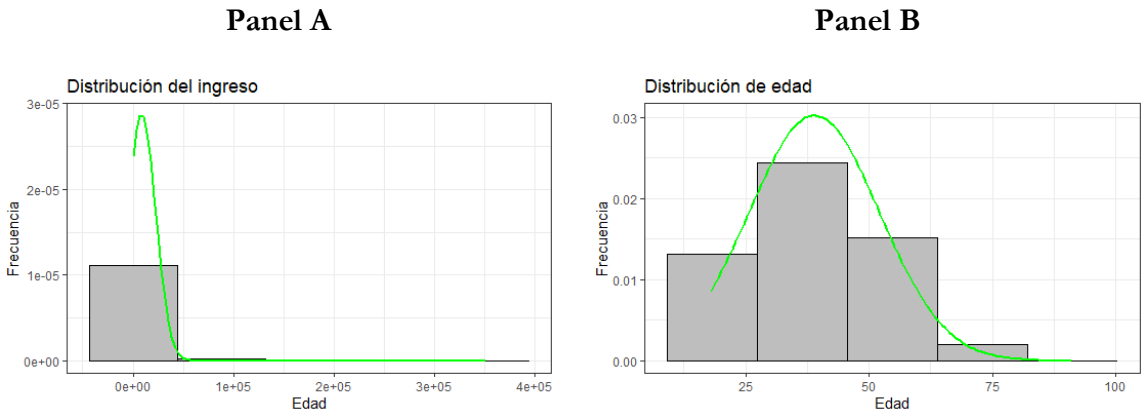
Fuente: Elaboración propia, 2023

Iniciando con nuestra variable a predecir, el ingreso tiene una media de 8541, con un valor mínimo de 0,5 y un valor máximo de 350.583. El histograma preliminar de la tabla muestra que los ingresos se concentran en los rangos de ingresos más bajos. Respecto al estrato, el valor mínimo es 1 y su valor máximo es 6, el histograma preliminar muestra una concentración en los

estratos más bajos. La edad promedio de los encuestados es 39 años, con un valor mínimo de 18 años y un valor máximo de 94.

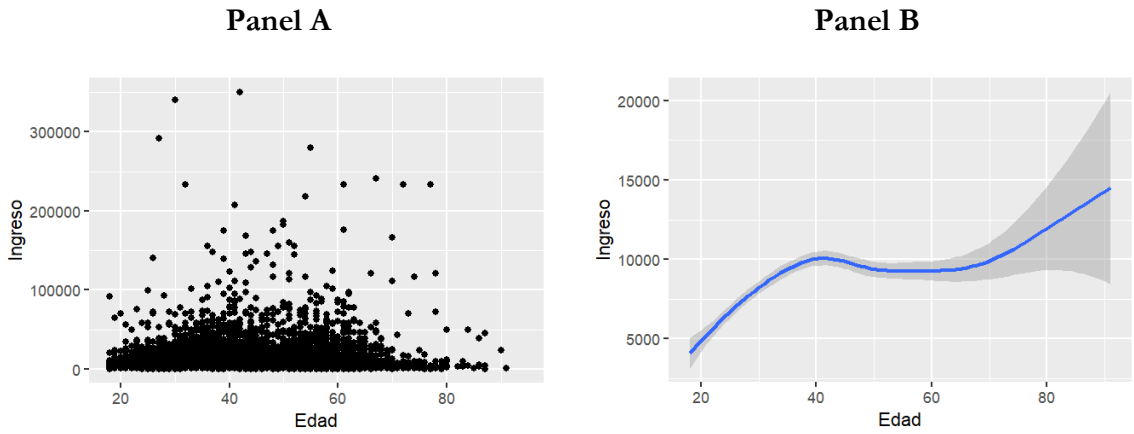
Ahora bien, estos datos son consistentes con

Figura 2: 2 Distribución del ingreso y distribución de la edad



Fuente: Elaboración propia

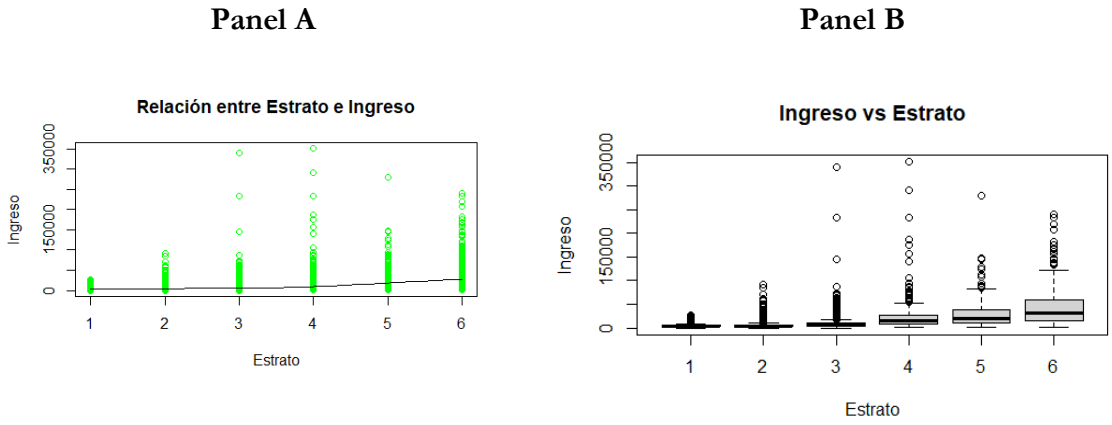
Figura 3: Distribución del ingreso y distribución de la edad



Fuente: Elaboración propia

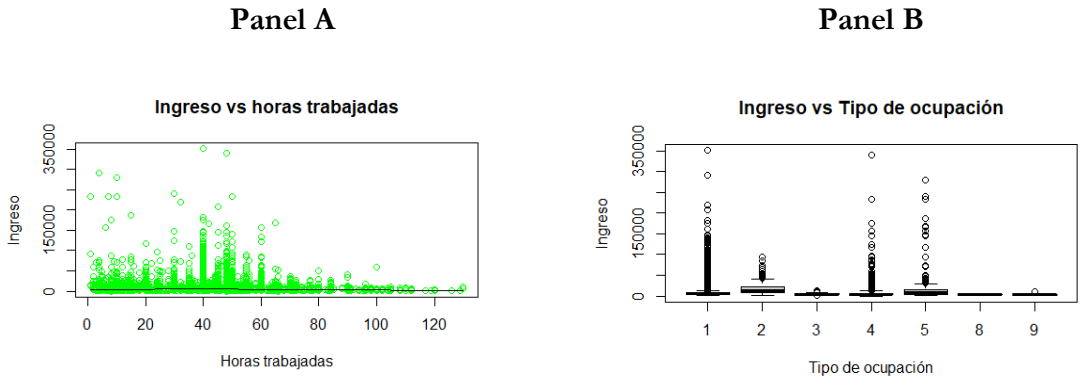
Estrato

Figura 4: Relación ingreso estrato



Fuente: Elaboración propia

Figura 5: Relación ingreso horas trabajas y relación ingreso tipo de ocupación



3 PERFIL EDAD-SALARIO

Tabla 3:

Tabla de Regresión

```
                        Modelo Salario depende de Edad
================================================
Dependent variable:
---------------------------
log_w
------------------------------------------------
age                              0.053***
(0.003)

age_2                            -0.001***
(0.00003)

Constant                         7.629***
(0.055)

------------------------------------------------
Observations                      16,542
R2                                 0.022
Adjusted R2                        0.022
Residual Std. Error          0.818 (df = 16539)
F Statistic             189.686*** (df = 2; 16539)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

Interpretación de los coeficientes

La regresión tiene como variable dependiente el logaritmo del salario por horas y dos variables independientes: la edad en años y la edad en años al cuadrado. Los coeficientes de estas variables son 0.053 y -0.001, respectivamente. Obteniendo las siguientes interpretaciones:

El coeficiente de la edad en años (0.053) indica que, manteniendo constante la edad al cuadrado, un aumento de un año en la edad se asocia con un aumento del 5.3% en el salario por hora. Esto sugiere que la edad puede ser un factor importante en la determinación del salario.

El coeficiente de la edad al cuadrado (-0.001) indica que, manteniendo constante la edad en años, un aumento de un año al cuadrado se asocia con una disminución del 0.1% en el salario por hora. Esto sugiere que la relación entre la edad y el salario no es lineal, sino que puede tener una forma curvilínea. El signo negativo del coeficiente de la edad al cuadrado indica que la relación entre la edad y el salario es cóncava hacia abajo. Es decir, a medida que la edad aumenta, el efecto positivo de la edad en el salario disminuye.

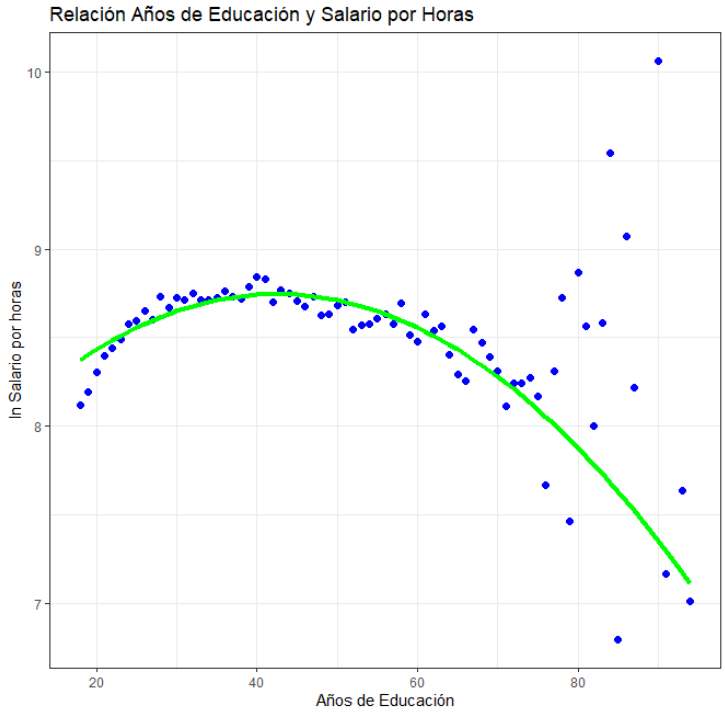
El coeficiente de determinación (R2) de la regresión puede proporcionar información sobre la bondad de ajuste del modelo, para este caso en particular es de 0.022, esto indica que las variables independientes explican muy poco de la variabilidad en la variable dependiente. Cabe mencionar finalmente, que ambas variables resultaron significativas estadísticamente y por ende es válido y útil analizar los resultados obtenidos.

Ajuste del modelo

El modelo de regresión tiene un R cuadrado de 0.003, lo que indica que solo el 0.3% de la variabilidad de la variable dependiente es explicada por el modelo. Este valor es muy bajo y sugiere que el modelo no se ajusta bien a los datos. El estadístico F de 48.9 indica que el modelo es significativo, es decir, que al menos una de las variables independientes tiene un efecto significativo en la variable dependiente. Sin embargo, esto no significa que el modelo sea útil o que se ajuste bien a los datos. El Residual Std. Error de 0.826 indica que la desviación estándar de los residuos es de 0.826 unidades de la variable dependiente. Un valor bajo de Residual Std. Error indica que el modelo se ajusta bien a los datos, pero en este caso, el valor es relativamente alto en comparación con la media de la variable dependiente. En resumen, el modelo es significativo, pero no se ajusta bien a los datos y tiene una capacidad limitada para explicar la variabilidad de la variable dependiente.

Gráfico del perfil estimado

En el siguiente gráfico se puede observar el perfil estimado del logaritmo natural del salario por horas junto con los años de educación, tal como se mencionaba anteriormente la regresión toma una forma cóncava, se evidencia de manera que hay un punto de inflexión el cual esta en 42 años, este es el máximo de la función y por ende es el punto en el cual empieza a disminuir la relación entre salarios y edad. También es importante mencionar que gráficamente pareciese que la regresión pierde poder de ajuste a medida que la edad aumenta, posteriormente de pasar l0s 75 años lo puntos del gráfico de dispersión toman valores muy extremos y no hay un claro comportamiento del salario que pueda ajustarse a la regresión.



Es importante mencionar también cual fue la metodología mediante la cual se calculó el valor máximo y cuál es su respectivo intervalo de confianza, realizando el siguiente procedimiento:

- A través de un Bootstrap se calculo el estimador y el error estándar, obteniendo, con la semilla dispuesta un error estándar muy pequeño, cuyo valor es de:

$$ee = 0.001731169$$

Y ahora se obtuvo la edad pico con la derivada de la regresión, igualando a cero y despejando, para llegar a la siguiente expresión:

$$Edad_{Maxima} = \frac{-\beta_1}{\beta_2 * 2}$$

Reemplazando con los valores obtenidos en la regresión previa se obtiene una edad de:

$$Edad_{Maxima} = \frac{-0.053}{(-0.001) * 2}$$

$$Edad_{Maxima} = 42.517$$

Y con esta edad y el error estándar se puede calcular el intervalo de confianza, con la siguiente formula:

$$IC = Edad_{Maxima} \pm (ee * 1.96)$$

$$IC_{li} = 42.5 - (0.0017 * 1.96) = 42.513$$

$$IC_{ls} = 42.5 + (0.0017 * 1.96) = 42.52$$

Por tal motivo, con un nivel de confianza del 95% podríamos afirmar que la edad que es un punto máximo esta entre:

$$IC = (42.513, 42.52)$$

4 BRECHA SALARIAL DE GÉNERO

a) Estimación de la brecha salarial

$$\lg(w) = \beta_1 + \beta_2 Female + u$$

Modelo Salario de acuerdo con el sexo

	Dependent variable:
	log_w
mujer	-0.090***
	(0.013)
Constant	8.670***
	(0.009)
Observations	16,542
R2	0.003
Adjusted R2	0.003
Residual Std. Error	0.826 (df = 16540)
F Statistic	48.947*** (df = 1; 16540)

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

Al analizar los resultados obtenidos previamente podemos concluir que debido a que el coeficiente de la variable independiente (-0.090) indica que, manteniendo constante la edad, ser mujer se asocia con una disminución del 9% en el salario por hora en comparación con los hombres. Esto sugiere que existe una brecha salarial entre hombres y mujeres, adicionalmente, el estimador es significativo estadísticamente y por tal motivo sus resultados son interpretables a cualquier nivel de significancia.

Finalmente, el R2 es un valor muy bajo, de tan solo el 0.3% por tal motivo para analizar la brecha salarial entre hombres y mujeres y llegar a conclusiones relevantes es necesario incluir más variables de control que permitan que la condición de ser mujer explique en un mayor porcentaje las diferencias de salario existentes.

b) FWL y FWL con Boostrap

5 PREDICCIÓN DE LAS GANANCIAS POR MODELO

Comparar los modelos vistos en clase

a) Dividir la muestra en dos

Para la división de la muestra, se asignó el 70% de la muestra como entrenamiento y el 30% para testeo.

b) RMSE vs Otros

Se realizaron seis modelos para evaluar su valor del RMSE, variando en cada uno las interacciones entre las variables seleccionadas, realizando los siguientes modelos:

Modelo Edad:

$$Ln(w) = Edad_i + Edad_i^2 + u_i$$

Modelo Mujer:

$$Ln(w) = Mujer_i + u_i$$

Modelo Especificación 1:

$$Ln(w) = Estrato_i + Mujer_i + Edad_i + ExperienciaLaboral_i + TiempoOcio_i + Edad_i^2 + u_i$$

Modelo Especificación 2:

$$Ln(w) = Estrato_i + Mujer_i + Edad_i + ExperienciaLaboral_i + TiempoOcio_i + u_i$$

Modelo Especificación 3:

$$Ln(w) = Mujer_i * Edad_i^2 + ExperienciaLaboral_i^2 * Estrato_i + TiempoOcio_i^2 + u_i$$

Modelo Especificación 4:

$$Ln(w) = Mujer_i + Edad_i^2 + ExperienciaLaboral_i^2 * TiempoOcio_i^3 + Estrato_i * NivelEducativo_i + u_i$$

Modelo Especificación 5:

$$Ln(w) = Mujer_i * TiempoOcio_i^3 + Edad_i + Edad_i^2 * ExperienciaLaboral_i^2 * NivelEducativo_i + Estrato_i + u_i$$

Y al realizar estas regresiones se obtuvieron los siguientes resultados:

Finalmente, tras la ejecución de estas regresiones, se calculó su MSE y se obtuvieron los siguientes resultados:

Modelo	Valor MSE
Modelo 5	0,5076
Modelo 4	0,5099
Modelo 2	0,5212
Modelo 1	0,5318
Modelo 3	0,6202
Modelo Edad	0,6899
Modelo Mujer	0,6990

Como se observa los dos modelos con mayor complejidad e interacciones no lineales son aquellos que obtuvieron un error medio cuadrático menor, por ende, podrían ser los modelos que tentativamente podrían emplearse para entrenar la muestra.

c) Resultados

Dentro

d) LOOCV

Repositorio GirHub: https://github.com/Luis-Borda/PS_Repo_Taller1_G10/commit/6ec3c5f0f10d3dcc7cf316e89fe039a0c93c40a0

6 Referencias

DANE. (2023). *Metodología general Gran Encuesta Integrada de Hogares GEIH*. Bogotá.

Farné, S., David, R., & Paola, R. (2016). *Impacto de los subsidios estatales en Colombia*. Bogotá. Obtenido de https://www.uexternado.edu.co/wp-content/uploads/2017/01/CUADERNO_17-2.pdf

Moffit, R. (1996). *Incentive Effects of the U.S. Welfare System: A Review*. Journal of Economic Literature .

Moller, C. L. (2012). *¿Por qué Colombia necesita un sistema tributario más progresivo?* Banco Mundial. Obtenido de https://www.bancomundial.org/es/news/opinion/2012/12/17/why-colombia-needs-a-more-progressive-tax-system