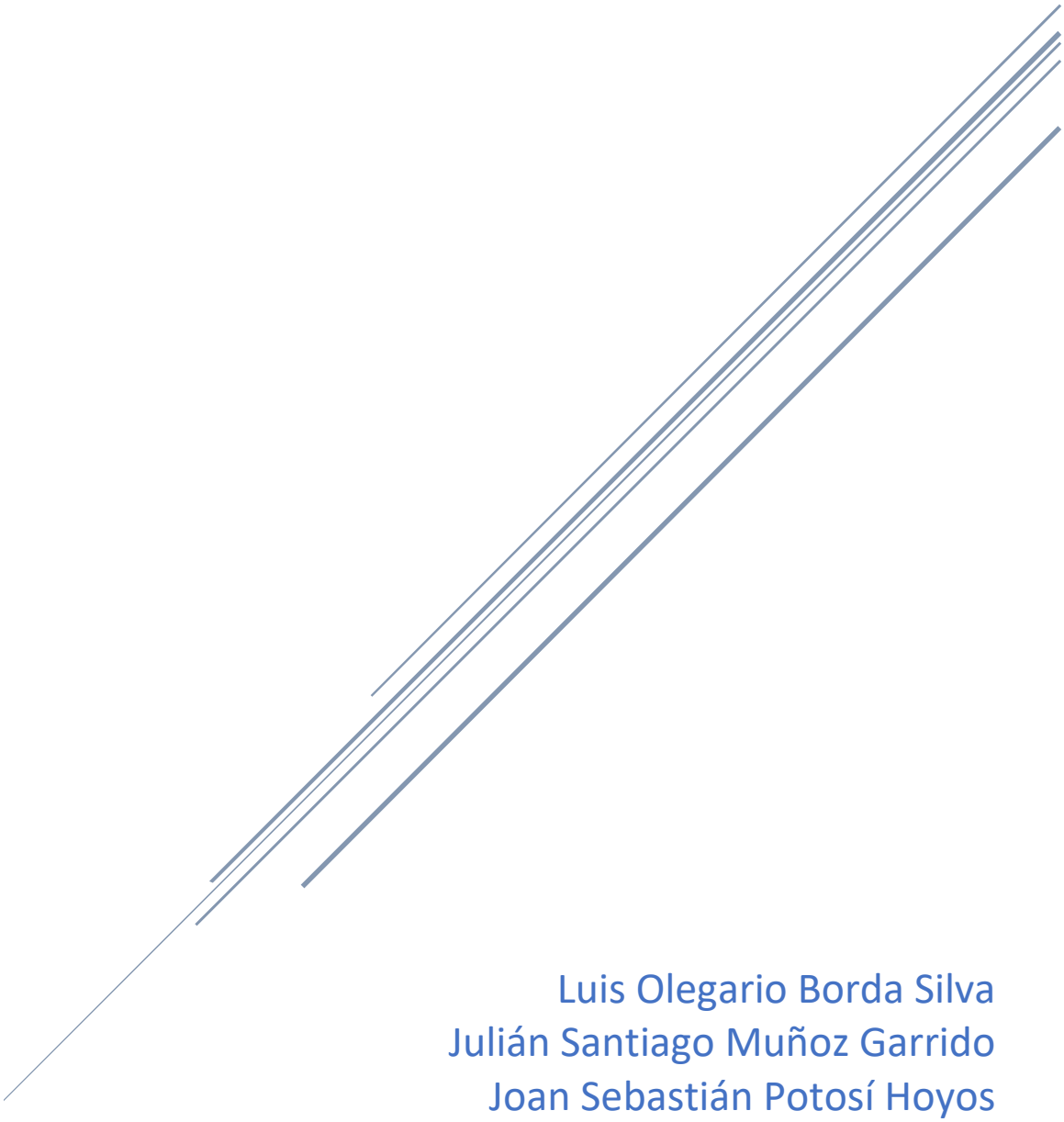
PROBLEM SET 1: PREDICTING INCOME

BIG DATA Y MACHINE LEARNING PARA ECONOMÍA APLICADA



Luis Olegario Borda Silva
Julián Santiago Muñoz Garrido
Joan Sebastián Potosí Hoyos
Juan Felipe Vargas Guachetá

Maestría en Economía Aplicada

Contenido

1. Introducción

La evasión tributaria, es uno de los principales fenómenos sociales que actualmente encienden las alarmas del Estado colombiano, en particular de la administración tributaria dado que, no solo representa un canal por el cual se fugan recursos públicos, sino que además refleja la poca eficiencia institucional en los programas de fiscalización. Por esta y otras razones, combatir el fraude fiscal se ha convertido en uno de los principales propósitos del gobierno nacional y una apuesta permanente de toda reforma tributaria (Hoyos, 2021)

En consecuencia, diferentes autores se han dedicado al estudio de las causas de la evasión fiscal. Por ejemplo, Moller plantea que los desafíos relacionados con la regresividad en la recaudación de impuestos aumentan la brecha de desigualdad y afectan directamente el bienestar de los ciudadanos, lo cual afecta la equidad social y hace 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 su presupuesto. (Moller, 2012)

Otros autores coinciden en que la evasión tributaria está fuertemente influenciada por la informalidad y el desempleo. 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.

Ahora bien, el alto nivel de afectación recaudatoria que se ha alcanzado en Colombia por cuenta de la evasión tributaria ha sido de tal escala que, según reportes de la DIAN, se estima que la evasión fiscal estaría quitando al Estado cerca de $65 billones anuales, es decir 5.4 puntos del PIB (DIAN, 2022). Cabe resaltar que entre los tributos que más se pierden recursos por evasión están el impuesto a la renta y el IVA.

Bajo este contexto, el desarrollo de análisis precisos sobre los ingresos de las personas es crucial, ya que permite detectar casos de fraude fiscal y diseñar herramientas de política pública para mitigar sus impactos en la administración tributaria. En este orden de ideas, en el presente problem set, se estructura un modelo de predicción del ingreso de los hogares en Colombia para el año 2018, utilizando como fuente de información la Gran Encuesta Integrada de Hogares (GEIH). Para realizar este ejercicio se aplicaron los conceptos aprendidos en el curso "Big data y Machine Learning para economía aplicada" de la Universidad de los Andes.

El documento inicia con corta introducción, posteriormente en la segunda sección, se hace una descripción de los datos utilizados incluyendo información relevante sobre la GEIH, el proceso de obtención y limpieza de los datos y algunas estadísticas descriptivas de las variables utilizadas. Luego, en la sección 3 se presenta un análisis de la relación edad salario, sobre el entendido que los salarios tienden a ser bajos cuando el trabajador es joven e incrementan a medida que envejece. En la sección 4, se hace un análisis de la brecha salarial por género, presentando las estimaciones obtenidas y discutiendo los resultados a la luz de los modelos planteados. Finalmente, en la sección 5, se hace una predicción de las ganancias, haciendo uso de 5 modelos

con diferentes niveles de complejidad y se compara su desempeño predictivo en terminos del MSE.

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*.

2.1. 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.

2.2. 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.

2.3. 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:

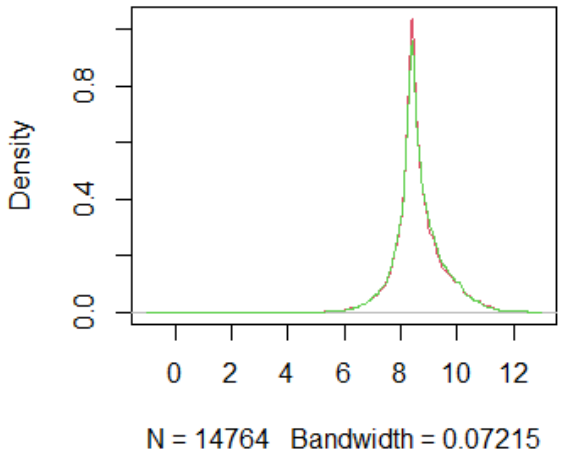
Tabla 1: Variables seleccionadas y valores missing

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.

2.4. Análisis descriptivo de los datos

Para el desarrollo del problem set, se utilizaron las variables estrato, sexo, edad, actividad principal, experiencia medida en meses, 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: Estadísticas 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	Experiencia	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	▁▁▇▁▁
y_total_m_ha	Salario por hora	1778	8541,87	13866,1	0,5	4837,5	350583,3	▇▁▁▁▁

Fuente: Elaboración propia, 2023

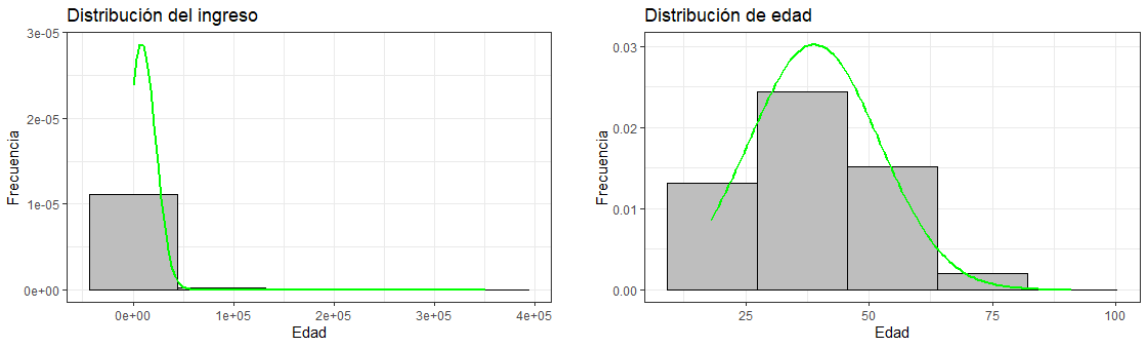
Iniciando con nuestra variable a predecir, se observa que el ingreso por hora tiene una media $8.541 pesos, con un valor mínimo de 0,5 pesos y un valor máximo de $350.583. El histograma preliminar de la tabla muestra que los ingresos se concentran en los rangos de ingresos más bajos. Por su parte, la variable sexo tiene una proporción muy similar entre hombres y mujeres lo cual es consistente con la distribución de la población por género en el país.

La edad de los individuos se concentra en los rangos etarios más jovenes, su media es 39 años, el valor mínimo es 18 años y el máximo de 94. La variable experiencia, tiene una media de 39 meses, con un mínimo de 0 y un valor máximo de 720 meses. Respecto al estrato, su moda es 2 y de acuerdo al histograma preliminar se identifica que su distribución tiene un sesgo hacia a la izquierda es decir, la muestra se concentra en los estratos más bajos. Finalmente, la variable ocu muestra que la totalidad de la muestra se encuentra ocupada.

Lo anterior es consistente al graficar la distribución de las variables ingreso y edad. En la figura 2, Panel A, se observa que los ingresos se concentran en los valores más bajos de la distribución. Por su parte, el panel B muestra que la edad de las personas encuestadas se concentra entre los 25 y 50 años aproximadamente.

Figura 2: Distribución del ingreso y edad

Panel A **Panel B**

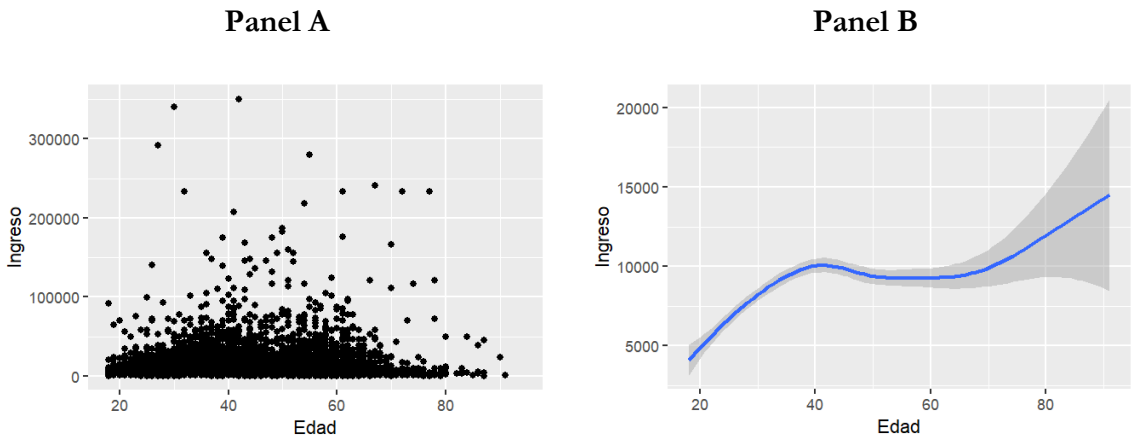


Fuente: Elaboración propia

Relación entre el ingreso y las variables seleccionadas para el análisis.

En primer lugar, al analizar la relación entre las variables ingreso y edad, el panel A de la figura 3 en principio no evidencia una tendencia clara pero si permite identificar casos atipicos para ciertos grupos etarios. Por su parte, el panel B muestra que hay un incremento del ingreso pronunciado en edades entre los 20 y 40 años, y a partir de este hay un punto de inflexión. El ingreso vuelve a tener una relación positiva en edades mayores a los 70 años.

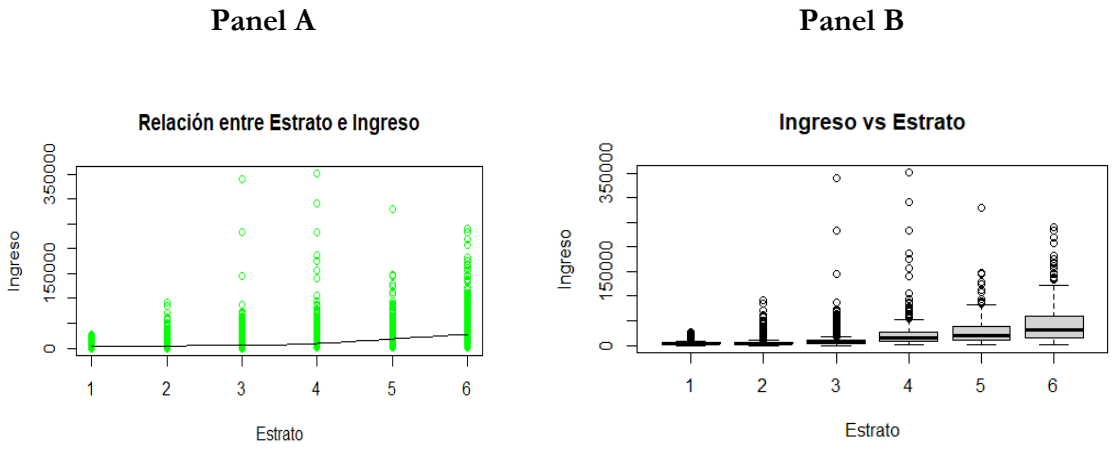
Figura 3: Relación ingreso edad



Fuente: Elaboración propia

Por otro lado, la figura 4 muestra la relación entre estrato e ingreso. En el panel A es posible observar que las personas en estratos 3 y 4 alcanzan los ingresos más altos en esta muestra. En el panel B se observa además que la media del ingreso incrementa a medida que el estrato socioeconómico es más alto. Con base en estas apreciaciones, a priori es posible inferir que existe una relación positiva entre el ingreso y el estrato.

Figura 4: Relación ingreso estrato

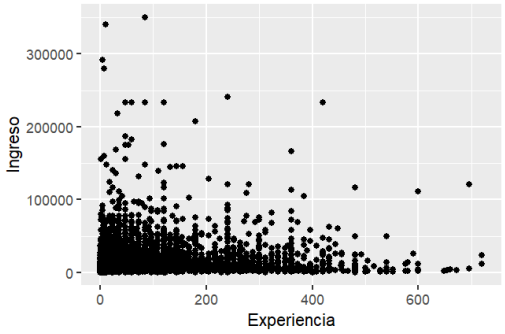


Fuente: Elaboración propia

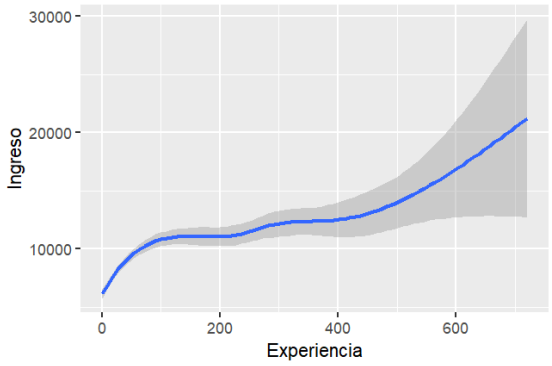
Finalmente, la figura 5 muestra la relación entre la experiencia y el ingreso. El Panel A, muestra que existen algunos casos atipicos para ciertos rangos de experiencia, y a partir del panel B es posible inferir que existe una relación positiva entre la experiencia y el ingreso.

Figura 5: Relación ingreso experiencia

Panel A



Panel B



Fuente: Elaboración propia

3. Perfil edad - salario

En economía, existen diferentes estudios que evidencian que los salarios tienden a ser bajos cuando las personas son más jóvenes y aumentan a medida que la persona envejece, alcanzando su punto máximo alrededor de los 50 años y tiende a permanecer estable o disminuir ligeramente después de los 50 años lo cual es consistente con los resultados expuestos en la figura 3. Ahora bien, es este punto, se estima el perfil edad salario a partir del siguiente modelo:

$$\log(w) = \beta_1 + \beta_2 Age + \beta_3 Age^2$$

Los resultados de esta estimación se muestran en la tabla 4.

Tabla 3: Modelo salario 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

Fuente: Elaboración propia

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

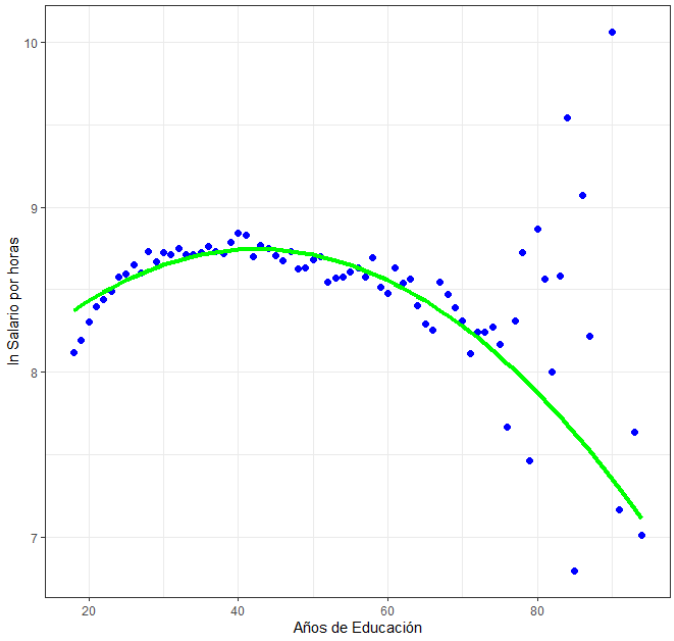
El modelo de regresión tiene un R cuadrado de 0.003, lo que indica que solo el 0.3% de la variabilidad de la variable dependiente es explicada por el modelo. Este valor es muy bajo y sugiere que el modelo no se ajusta bien a los datos. El estadístico F de 48.9 indica que el modelo es significativo, es 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 está 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.

Figura 6: Relación entre años de educación y salario por horas



Fuente: Elaboración propia

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 calculó 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 por género

En este punto se hace un análisis de los ingresos teniendo en cuenta brechas por género.

4.1. Estimación de la brecha salarial

Para el desarrollo de este punto se estimó el siguiente modelo

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

Los resultados de las estimaciones se muestran en la tabla 4.

Tabla 4: Modelo brecha salarial género

	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) 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.

4.2. Equal Pay for Equal Work?

Se incorporaron las variables de control "máximo nivel educativo del individuo" y el "tiempo que lleva el individuo en el trabajo" para estimar el modelo Frisch-Waugh-Lovell (FWL) y se obtuvieron los siguientes resultados:

Tabla 5: Modelo con controles

	Modelo FWL	
	log_w	log_w_resid
	(1)	(2)
mujer	-0.1157863***	
	(0.0115074)	
Max_educ	0.2959297***	
	(0.0047014)	
Tiempo_empresa	0.0015599***	
	(0.0000644)	
mujer_resid_f		-0.1157863***
		(0.0115067)
Constant	6.8226340***	0.0000000
	(0.0295029)	(0.0057236)
Observations	16,542	16,542
R2	0.2079797	0.0060845
Adjusted R2	0.2078360	0.0060244
Residual Std. Error	0.7361937 (df = 16538)	0.7361492 (df = 16540)
F Statistic	1,447.5920000*** (df = 3; 16538)	101.2543000*** (df = 1; 16540)
Note:		*p<0.1; **p<0.05; ***p<0.01

Fuente: Elaboración propia

Se cumple el teorema Frisch-Waugh-Lovell, toda vez que los coeficientes de los *bethas* estimados son iguales:

$$\hat{\beta} mujer = \widehat{\beta^*} mujer\_resid\_f$$

$$-0.1157863 = -0.1157863$$

FWL con *Bootstrap*

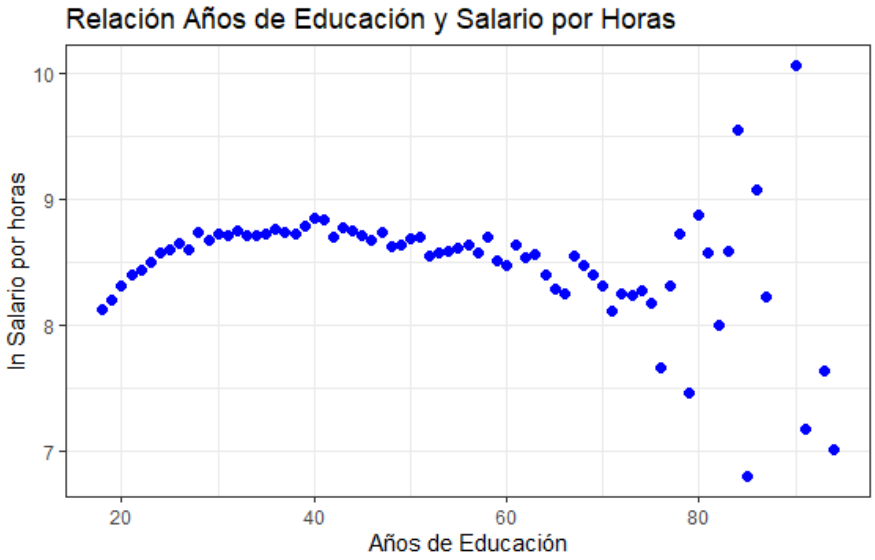
Por medio de *Bootstrap* se estimaron los errores estándar de los coeficientes a partir del ajuste de regresión lineal, siendo este:

$$ee_{FWL} = 0.01201014$$

En retrospectiva, los resultados muestran que, manteniendo constante los demás controles, si el individuo es mujer el salario por hora disminuye en 11% en comparación con los hombres. Al incluir los controles en nuestro modelo, se sigue evidenciando la brecha salarial entre hombres y mujeres. Aunado a esto, el estimador es estadísticamente significativo a los niveles de confianza convencionales y son válidos para interpretación.

Finalmente, el R^2 es de 0.6% 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.

Por último, se muestra el plot edad-salario predicho:



5. Predicción de las ganancias

En las secciones anteriores, se estimaron diferentes modelos teniendo en cuenta la inferencia. Ahora bien, en esta subsección, se evaluará el poder predictivo de estas especificaciones. Para esto se siguió el siguiente procedimiento:

i. 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.
ii. 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:

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

5.1. Desempeño predictivo de los modelos

Una vez planteados los modelos, se realizaron las estimaciones y se calculó su MSE respectivo, obteniendo los siguientes resultados:

Tabla 6: MSE para los modelos planteados

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

Fuente: Elaboración propia

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.

5.2. Discusión de los resultados

De acuerdo con los MSE obtenidos se pueden llegar a las siguientes conclusiones:

- Modelo 5 (MSE = 0.5076): Este modelo tiene el MSE más bajo de todos los modelos que se están comparando. Esto indica que, en promedio, las predicciones de este modelo tienen el error cuadrado medio más bajo en comparación con los valores reales. En general, un MSE bajo es deseable, ya que significa que el modelo tiene un buen ajuste a los datos.
- Modelo 4 (MSE = 0.5099): Este modelo tiene un MSE muy cercano al del Modelo 5, lo que sugiere un rendimiento similar en términos de precisión de predicción. Es posible que ambos modelos estén ofreciendo un rendimiento bastante consistente y confiable.
- Modelo 2 (MSE = 0.5212): El Modelo 2 tiene un MSE un poco más alto que los dos primeros modelos. Esto indica que, en promedio, las predicciones de este modelo tienen errores ligeramente mayores en comparación con los modelos 4 y 5, pero aún así es un valor razonablemente bajo.
- Modelo 1 (MSE = 0.5318): El Modelo 1 tiene un MSE un poco más alto que el Modelo 2, lo que sugiere que, en promedio, las predicciones de este modelo tienen errores ligeramente mayores que las del Modelo 2.
- Modelo 3 (MSE = 0.6202): El Modelo 3 tiene un MSE significativamente más alto que los modelos anteriores. Esto indica que las predicciones de este modelo tienden a tener errores cuadrados más grandes en comparación con los otros modelos, lo que sugiere un rendimiento inferior en términos de precisión de predicción.
- Modelo Edad (MSE = 0.6899): El "Modelo Edad" tiene un MSE aún más alto que el Modelo 3, lo que indica un rendimiento relativamente pobre en términos de ajuste a los datos.

- Modelo Mujer (MSE = 0.6990): El "Modelo Mujer" tiene el MSE más alto de todos los modelos evaluados, lo que sugiere que este modelo tiene el peor rendimiento en términos de precisión de predicción.

En resumen, los modelos con MSE más bajo (Modelo 5 y Modelo 4) son los que ofrecen el mejor rendimiento predictivo, mientras que los modelos con MSE más alto (Modelo Mujer y Modelo Edad) tienen un rendimiento inferior. Los valores intermedios de MSE (Modelo 1, Modelo 2 y Modelo 3) tienen un rendimiento en algún lugar entre estos extremos.

De modo que el mejor modelo es el quinto modelo, el de complejidad más alta, las interacciones que pareciesen aportar fue la interacción triple entre edad, experiencia laboral y nivel educativo, esto es una combinación a su vez con el rezago educativo y la experiencia laboral, que es una variable similar al capital humano, ya que comprende estas tres dimensiones y como la teoría menciona, es una variable que aporta mucho para la definición del salario, por tal motivo, consideramos que esta interacción a la que llamamos capital humano tuvo un impacto muy positivo en la definición del salario.

5.3. LOOCV

Se seleccionaron el modelo cinco y el modelo cuatro, los cuales tuvieron el menor error predictivo y con estos se hizo el modelo LOOCV, el cual si recordamos es una extensión del modelo K-fold validation cross con la particularidad que la partición se hace uno a uno de los individuos de la muestra, al realizar este proceso con el modelo 5, el modelo con el menor MSE, obteniendo los siguientes resultados:

$$Promedio_{LOOCV-Modelo5} = 0.5187$$

Ahora, se realizó el mismo procedimiento, pero con el modelo 4, el otro modelo con el que se obtuvo el MSE más bajo, con el siguiente resultado:

$$Promedio_{LOOCV-Modelo4} = 0.5376$$

Con estos resultados obtenidos, se puede observar pequeñas variaciones, en ambos casos mayores con el modelo LOOCV, esto sugiere que, como la tasa de error o la precisión, un valor de 0.518 y 0.537 significa que, en promedio, el modelo clasificó incorrectamente aproximadamente el 51.8% y 53.7% de los ejemplos en el conjunto de validación.

Anexos

A este documento se anexa el repositorio GitHub el cual contiene la documentación de la base de datos, los scripts y los outputs correspondientes de cada punto. El repositorio se encuentra en el siguiente link:

https://github.com/Luis-Borda/PS_Repo_Taller1_G10.git

Bibliografía

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

Hoyos, V. (2021). *Las causas de la evasión tributaria en Colombia*. Bogotá: Editorial Universidad Externado.

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