[18 DE SEPTIEMBRE 2023]

[Problem Set 1: Predicting Income]

Por: [Andrés Chavarro, Ricardo Silva, Federico Camacho & Juan David Vélez]

Códigos: 202011832, 201821978, 202011648 & 201923348

[Introducción]

Con la base de datos del GEIH exploramos las relaciones de la variable edad su especificación adecuada y posible concavidad, hicimos indagación en el enfoque de género todo esto robusteciendo los errores con Bootstrap y hallando intervalos de confianza para los estimadores de las regresiones principales de los enunciados. Encontramos efectivamente que hay una relación cóncava de edad y salario en donde el salario máximo se encuentra en la edad de 45, existe una brecha salarial entre hombres y mujeres, donde las mujeres ganan con un intervalo de entre 2 a 4% menos que los hombres. Intervalos que van variando dependiendo de la cantidad de controles que se incluyen en la regresión. En el modelo predictivo trabajamos basados en la ecuación de Miencer y en esta indagación teórica de las estructuras laborales, corrimos varios modelos que exploraban interacciones, formas cuadráticas y controles de variables dummys y categóricas. El Mse mas bajo que nos dio fue de 0,31 que es un 40% más eficiente que el primer modelo donde no se incluían mayores flexibilidades y complejidades. También sufrimos de overfitting, pero con prueba y error pudimos llegar al modelo final que esta especificado en el final del documento.

[Descripción de limpieza]

Para analizar los determinantes de los salarios por hora de los individuos en Colombia, se utilizó la Gran Encuesta Integrada de Hogares (GEIH). Acceder a esta fuente de datos implicó realizar un proceso de web scraping. Fue necesario crear un bucle que iterara para extraer información de cada una de las 10 subpáginas (chunks) dentro de la página principal. Posteriormente, se unió toda esta información y se agregaron las variables pertinentes, generando así un marco de datos completo. Se utilizaron comandos de HTML, y la clave residía en acceder a la red (net work) de cada página para comprender cómo estaban almacenados los datos. La consolidación de las variables resultó sencilla, ya que estas eran consistentes en todas las páginas; la diferencia radicaba en las distintas personas presentes en cada página. Luego en la selección de variables estudiamos el diccionario del GEIH y la definición de cada una de ellas. Se enfocó en variables estrechamente relacionadas con el poder adquisitivo de los hogares, buscando patrones de comportamiento en dichas variables. La base de datos contiene una amplia gama de información sobre ingresos extras percibidos por el hogar, ya sea por arriendo, remesas, subsidios o programas de ayuda. También aborda preguntas sobre el empleo, horas extras, disposición a trabajar más o menos, entre otras de interés según la GEIH.

[Estadísticas descriptivas]

Sin embargo, muchas de estas preguntas presentan una gran cantidad de datos faltantes o problemas de medición, especialmente aquellas que no tienen respuestas numéricas y son muy específicas, lo que dificulta su interpretación. Por tanto, nos centramos en aproximadamente 18 de las 180 variables disponibles, abarcando aspectos como educación, edad, género, condiciones laborales e ingresos (trabajados en forma logarítmica y por hora). Preferimos trabajar con variables que fueran dummies o que tuvieran un alto nivel de observaciones completas para asegurar la robustez de nuestros modelos.

Dado que nuestra variable de interés presentaba muchos datos faltantes, optamos por imputarlos para algunas regresiones. En particular, para los métodos FWL y Bootstrap, se imputaron los valores faltantes con la media de la distribución, que calculamos excluyendo esos valores. La media del salario por hora en logaritmo resultó ser de 8.72.

Posteriormente, restringimos la base de datos a personas mayores de 18 años y que no estuvieran desempleadas, reduciendo la cantidad de observaciones de 32,177 a 22,640.

En la sección de estadísticas descriptivas, nos interesaba entender el comportamiento de los salarios y los niveles de ingresos en esta base de datos de colombianos restringida. A continuación, encuentran dicha tabla de estadísticas descriptivas después de trabajar en la limpieza de datos. Donde se encuentran condiciones laborales como si es formal, si cotiza pensión, si está afiliado a salud. Condiciones de la firma de tamaño que nos aproxima a su productividad. Variables demográficas de ser jefe de hogar, el estrato, el genero, la edad y la ubicación geográfica. Aquí podrán observar algunas características estadísticas básicas de estas variables.

Estadísticas Descriptivas

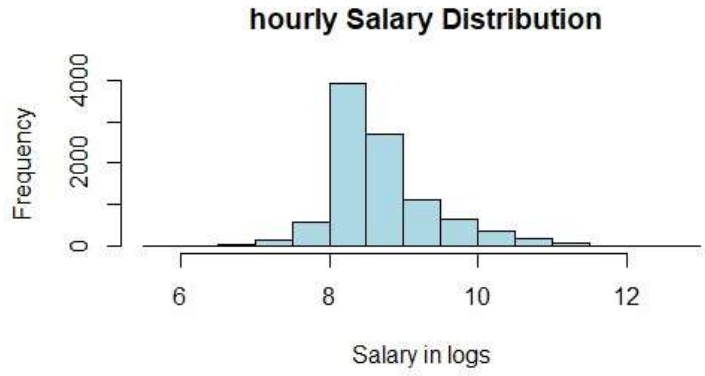
Statistic	N	Mean	St. Dev.	Min	Max
Directorio	22,640	4,656,436.00	82,372.73	4,514,331	4,804,455
Secuencia-p	22,640	1.01	0.13	1	4
Orden	22,640	1.97	1.22	1	12
Clase	22,640	1.00	0.00	1	1
Dominio	22,640	6.44	3.39	1	12
Mes	22,640	2.59	1.02	1	6
Estrato1	22,640	0.47	0.50	0	1
Sexo	22,640	43.03	17.37	18	106
Edad	9,892	8,822.23	12,886.16	326.67	350,583.30
Ingresos laborales Hora	9,892	1,745,416.00	2,403,441.00	20,000.00	60,100,000.00
Ingresos laborales Mensual	14,764	1,617,551.00	2,431,319.00	84.00	70,000,000.00
Ingreso total Mensual	14,764	8,541.87	13,866.13	0.47	350,583.30
Ingreso total Hora	22,640	0.29	0.46	0	1
College	22,640	11.00	0.00	11	11
Depto	22,640	0.00	0.00	0	0
Empleado	16,542	0.59	0.49	0	1
Formal	22,638	5.77	1.39	1	7
Educación máxima	22,640	2.25	1.74	1	9
Parentesco Jefe Hogar	16,542	63.76	89.49	0	720
Tiempo trabajando	20,874	1.38	0.76	1	3
Seguridad Salud	16,542	3.15	1.65	1	5
Tamaño Firma	16,542	47.40	15.66	1	130
Total horas trabajadas	16,542	0.45	0.50	0	1
Micro Empresa	22,640	247.28	58.48	107	808
Pesos frecuencia	10,553	1.99	0.18	1	9
Otros ingresos	16,542	1.45	0.54	1	3
Pensiones	16,542	4.93	3.34	1	9
Número personas en Firma	16,542	49.77	28.08	1	99

Gráfica 1. Estadísticas Destrictivas

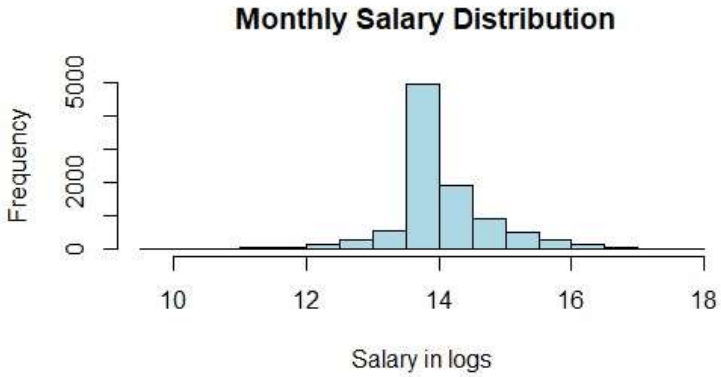
Observamos que el ingreso mensual promedio es de 1,745,416, mientras que la mediana de ingresos es de 1,032,560. Esta diferencia notable entre la mediana y el promedio sugiere una gran disparidad en los ingresos, donde salarios altos impulsan la media a niveles cercanos a más de 2 salarios mínimos del 2018, que estaban en $781,242.

Intentamos representar la distribución mediante un histograma, pero la visualización no fue óptima. Para superar esto, convertimos las variables de ingreso total mensual e ingreso por hora a logaritmo, lo que permitió una mejor observación de la distribución. En ambos casos, se observaron puntos álgidos donde se concentra la mayoría de la distribución. La distribución de ingresos mensuales demostró ser leptocúrtica, concentrándose alrededor de su media y mostrando sensibilidad a los salarios altos.

En cuanto a nuestra variable de interés, el salario por hora, casi todas las observaciones se centraron alrededor de 8 y 8.5. Esto refuerza la idea de que la imputación de datos faltantes utilizando la media de 8.7 no afecta negativamente la distribución



Gráfica 2.

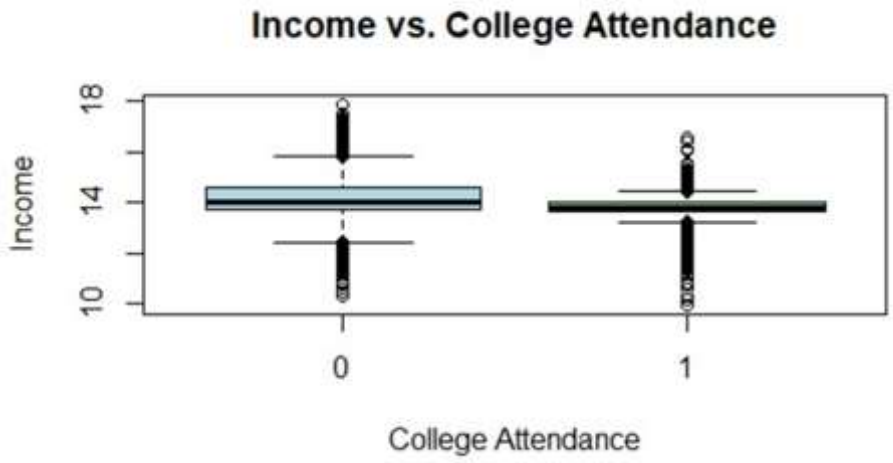


Gráfica 3.

Posteriormente, exploramos la interacción entre nuestras variables de interés, inspirados en la teoría económica que sugiere un impacto positivo en los ingresos familiares al recibir educación secundaria, brindando acceso a mejores oportunidades laborales. Buscamos medir si la educación superior recompensa a los estudiantes en el mercado laboral. Utilizamos un diagrama de caja y bigotes para comparar los percentiles entre grupos y observar sus distribuciones.

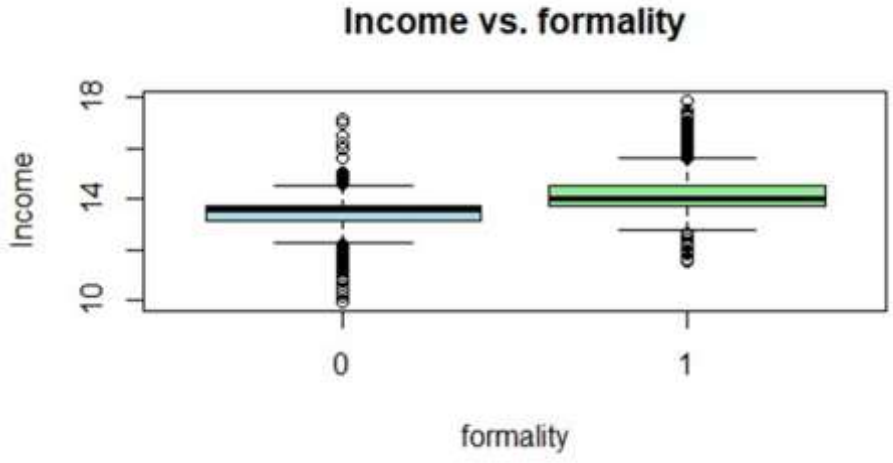
Al analizar la hipótesis relacionada con la educación en esta sección de estadísticas descriptivas, encontramos una situación intrigante. No se confirmó que las personas que asisten a instituciones de educación terciaria tengan un salario mediano más alto que aquellos que no asisten. Sin embargo, observamos una hiper concentración en la caja, donde el 50% de la distribución de ingresos se sitúa en el salario logarítmico mensual de 14. Esto indica una menor variabilidad en los ingresos y una mayor probabilidad de estar en este rango. Por otro lado, las personas que no asisten tienen una mediana ligeramente mayor pero una variabilidad de ingresos mucho más alta. Entre el primer y tercer cuartil, la variación en el nivel logarítmico es significativamente mayor, oscilando entre 10 y 16. Esta variabilidad es notablemente diferente a la desviación del grupo educado, que en estos mismos cuartiles apenas varía y se mantiene cercana a la mediana.

Interpretamos estos resultados considerando que la educación terciaria no garantiza salarios mensuales más altos. Sin embargo, sí parece homogeneizar los salarios dentro de un rango, reduciendo la incertidumbre en el pago. Es decir, brinda cierta tranquilidad al asegurar ingresos dentro de rangos más predecibles.



Gráfica 4.

También, analizamos la interacción entre la formalidad laboral y los salarios. Nuestra expectativa era que las personas que trabajan en empleos formales tengan ingresos más altos en comparación con aquellas que trabajan en empleos informales. Este hallazgo se confirmó al comparar los diagramas de caja y bigotes. En este análisis, encontramos que la mediana y todos los rangos cuartiles son más altos para aquellos en empleos formales. Observamos una concentración de datos en el bigote superior, indicando que este sector acumula a muchas personas con salarios atípicamente altos en comparación con la distribución. Esto contrasta con el sector informal, donde hay una representación excesiva en el bigote inferior, indicando que jala los cuartiles hacia abajo y muestra salarios incluso por debajo del salario mínimo.



Gráfica 5.

[Relación "*Wage-Age*"]

$$log(w) = \beta 1 + \beta 2Age + \beta 3Age2 + u$$

Al estimar la regresión que incluye únicamente las variables de edad y edad al cuadrado, nuestro objetivo es validar la hipótesis relacionada con la forma funcional y especificación adecuada del modelo. Si el estimador (beta) asociado al término que contiene la edad al cuadrado es negativo, esto indica una relación cóncava entre el ingreso y la edad. Este hallazgo respalda la hipótesis de que el ingreso por edad alcanza un punto máximo y, a partir de ahí, disminuye.

Al correr el modelo, observamos un primer estimador con un signo positivo, lo que sugiere un efecto de aumento del salario por hora en función de la edad. Específicamente, un incremento de un año en la edad conlleva un aumento positivo del 6,702% en el salario por hora en términos logarítmicos. También podemos interpretar que hay aumentos salariales cercanos al 70% por cada década adicional de edad.

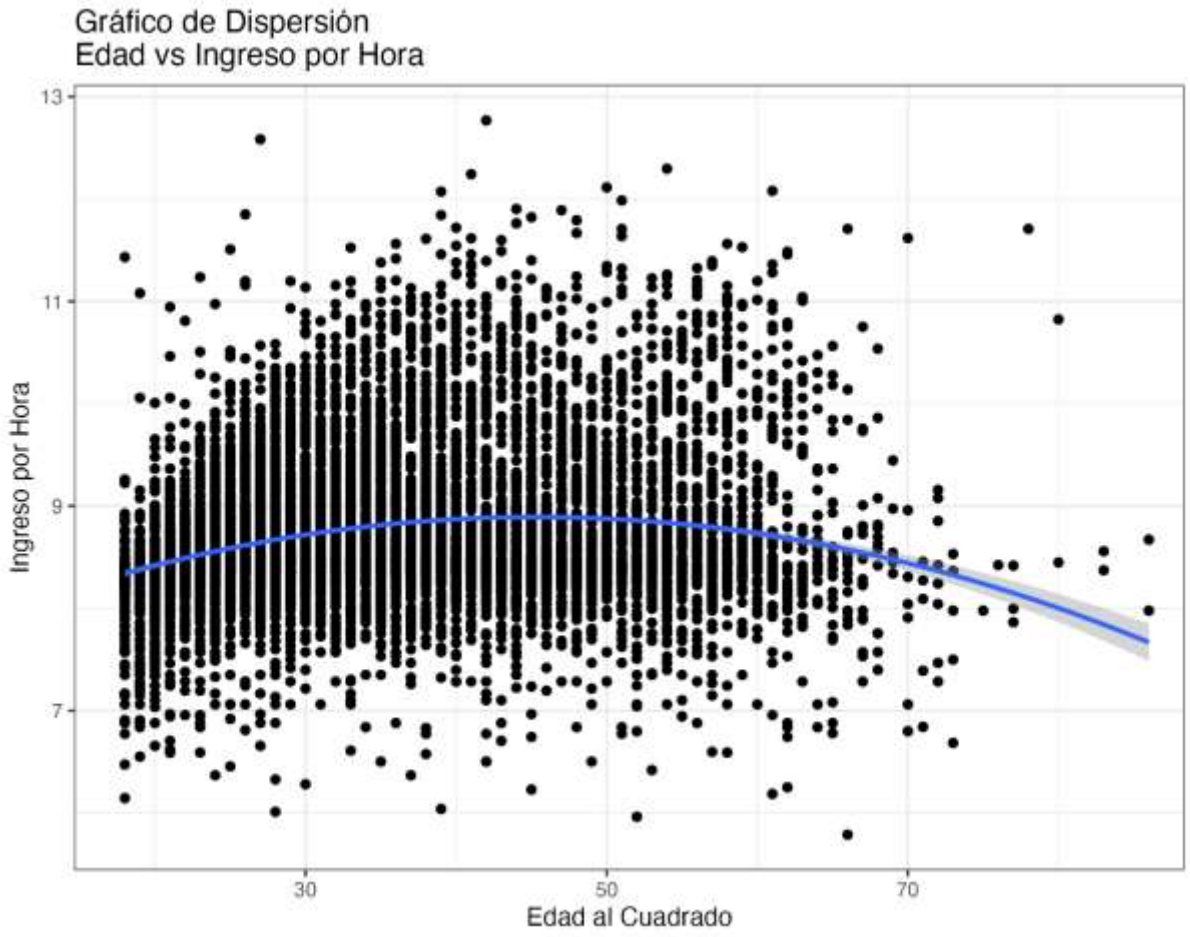
Por otro lado, el signo negativo del segundo estimador, que explora la relación cuadrática, nos indica la presencia de concavidad en la relación. El valor de este estimador es -0,0007394, ilustrando que, al ser la relación cuadrática y penalizar los valores altos de edad, cada incremento en edad se traduce en aumentos salariales menores debido a esta penalización inherente a la naturaleza cuadrática de la relación.

Para ambos grados polinomiales de edad, la regresión exhibe valores t elevados y una significancia estadística notable, ya que los coeficientes (betas) son significativos al 99%. Esta alta significancia respalda la validez de la interpretación, confirmando una relación cóncava y la disminución de ingresos después de cierta edad. Específicamente, al derivar el modelo de regresión original y establecer la derivada en 0, identificamos que el pico de salario ocurre a los 45 años de edad. Este cálculo se realiza derivando sobre la variable edad y luego igualando a cero y despejando, esto da $(-\frac{\beta_1}{2\beta_2})$

	Dependent variable:
	Logaritmo de salario por hora
Edad	0.07^{***}
	(0.004)
Edad Cuadrada	-0.001^{***}
	(0.0000)
Constant	7.37^{***}
	(0.07)
Observations	9,892
R^2	0.04
Adjusted R^2	0.04
Residual Std. Error	0.71 (df = 9889)
F Statistic	228.44^{***} (df = 2; 9889)
Note:	*p<0.1; **p<0.05; ***p<0.01

Regresión 1.

Posteriormente, para visualizar de manera gráfica la relación entre ingresos y edad, creamos un diagrama de dispersión. En este gráfico, se aprecia una sutil indicación de una relación cóncava, reflejada en un patrón que se asemeja a una "u" invertida. Por ejemplo, tanto las personas de 25 años como las de 75 años muestran observaciones salariales cercanas a 7 en valor logarítmico, lo cual sugiere que la distribución de ingresos desciende después de cierto punto, porque datos de edad alejados vuelven a encontrase. Desde alrededor de los 25 años hasta antes de los 50 años, la relación con la edad tiende a ser ascendente, indicando que a medida que las personas tienen más años, también tienen ingresos mayores. No obstante, esta relación no parece mantenerse a medida que la edad avanza, si no revertirse y que los ingresos luego a medida de una edad son más bajos.



Gráfica 6.

[Relación: *The gender earning GAP*]

	Dependent variable:	
	log_salario_hora	
	Modelo Incondicional	Modelo Condicional
	(1)	(2)
Mujer	-0.045***	-0.020**
	(0.015)	(0.009)
Observations	9,892	16,542
R^2	0.001	0.327
Adjusted R^2	0.001	0.324
Residual Std. Error	0.727 (df = 9890)	0.463 (df = 16450)
F Statistic	9.317*** (df = 1; 9890)	87.960*** (df = 91; 16450)
Note:		*p<0.1; **p<0.05; ***p<0.01

Anteriormente se aborda la brecha salarial que existe entre hombres y mujeres. Donde la imagen usual es que efectivamente existe tal brecha, y que además tiene una magnitud importante. Así pues, realizaremos una comparación de los salarios sin controles, otras estimaciones con controles, con el objetivo de clarificar si esta brecha se genera a raíz de la diferenciación entre trabajadores, y cómo última cuestión compararemos la edad "peak" de los hombres con el de las mujeres.

Como una primera aproximación ante el problema se realizó la siguiente regresión $\log(salario) = \beta_0 + \beta_1 Female + u$; en esta podemos ver que se da un coeficiente de – 0,004, pensando en que nuestra variable base es hombre tenemos que en promedio las mujeres están recibiendo 4% menos salario que los hombres. Ahora bien, uno de los argumentos más comunes en contra de la brecha salarial es que los trabajadores, en muchos casos, no son comparables, ya que hay diferencias en motivaciones, en niveles de productividad entre los individuos y teniendo en mente esto decidimos correr la siguiente regresión:

$$\log(salario) = \beta_0 + \beta_1 female + \beta_2 Sizefirm + \beta_3 P6870 + \beta_4 Oficio + \beta_5 College + \beta_6 Cotpension + \epsilon$$

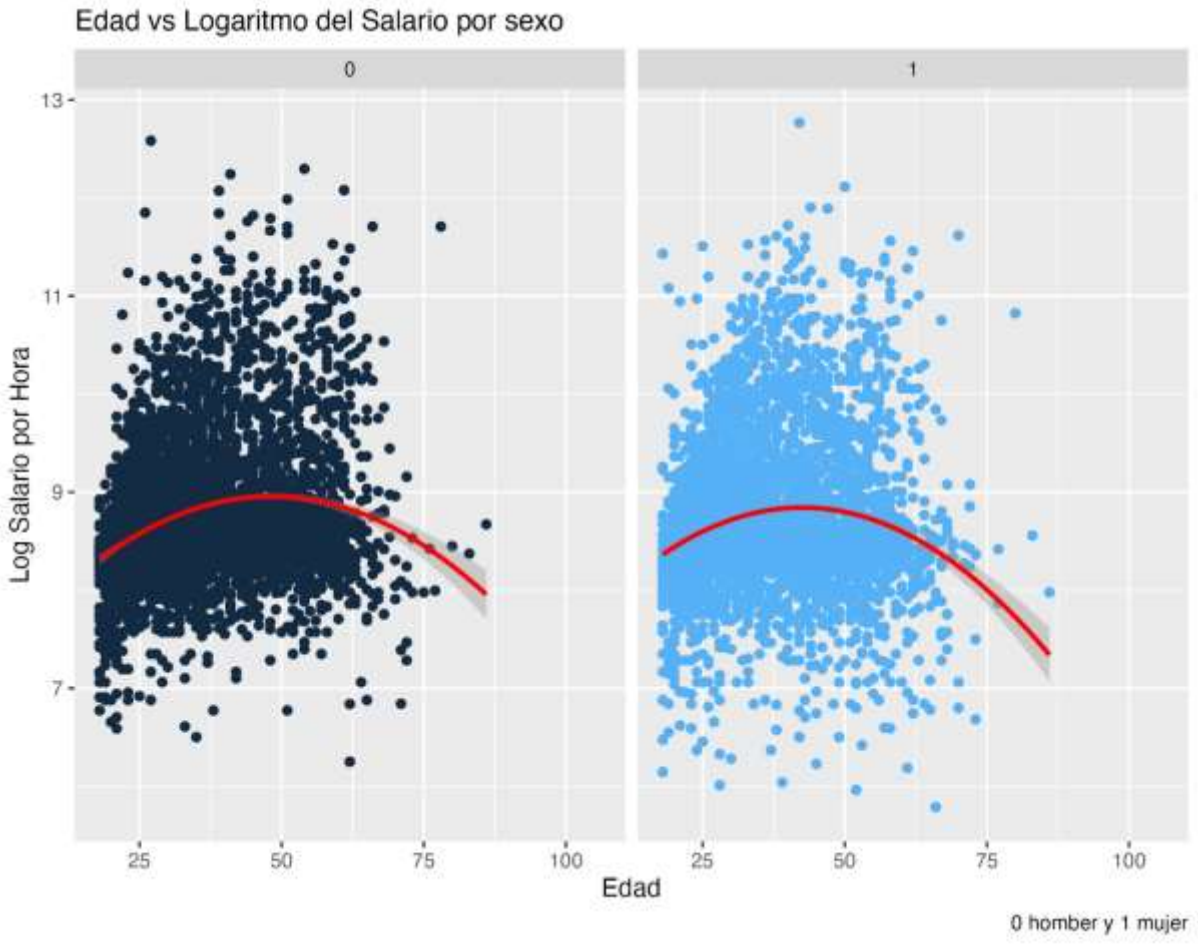
Donde cada una de las variables de control tienen un sentido especifico

- Sizefirm: Este es el tamaño de la firma donde la persona trabaja, resulta clave puesto que a mayor tamaño de firma la probabilidad de recibir mejores salarios es más alta, por lo que es importante controlar.
- P6870: Esta variable es el número de personas que hay en la firma, esto puede ser importante ya que entre más grande la firma, lo salarios pueden tender a ser más alto, ya que son más competitivas, más productivas, etc.
- Oficio: Esta variable es categórica que genera la diferenciación entre profesiones de las personas, esto evidentemente afecta el salario, ya que existen profesiones que son mejor pagas, y más solicitados en el mercado laboral.
- College: Esta es una variable dummy, que toma el valor de 1 si fue a la universidad, y toma valor de cero de lo contrario. Esta variable resulta evidentemente vital, ya que la diferencia salarial de una persona que va a la universidad y aquella persona que no va es bastante grande.
- Cotpensión: Esta variable indica si la persona cotiza, no cotiza o si ya está pensionada, es vital ya que una persona que cotiza a pensión está en un trabajo regulado, lo que indica que los salarios son más altos.

Cuando realizamos la regresión evidenciamos que había un sesgo por variable omitida en la regresión sin controles. Podemos ver que esta vez se da que las mujeres, en promedio, ganan –1.98% menos en cuanto al salario que los hombres. Para mejorar la robustes de las estimaciones decidimos estimar con Bootstrap, de forma de que exista un coeficiente de –1.98%, lo que esta reafirma lo ya mencionado anteriormente, con una magnitud diferente. En cuanto al R2-ajustado- podemos mencionar que, en el incondicional, es de 0,001, y en el condicional 0,324, por lo que vemos una mejoría de la capacidad de predicción al utilizar los controles;

intuitivamente, esto se debe a que cuando existen controles, estos son capaces de predecir a su vez el cambio en salario.

En lo que respecta a las disparidades salariales y a las especificaciones particulares de los modelos de ingresos para hombres y mujeres, se observa que en ambos casos la relación entre el salario por hora y la edad muestra una curvatura cóncava, independientemente del género. Se nota que en cierto punto de la edad se alcanza un pico máximo de ingresos y posteriormente estos comienzan a disminuir. Esta tabla permite discernir la rapidez con la que se llega a dicho punto álgido y cómo descienden los ingresos después de alcanzar ese máximo. Se puede apreciar que, en el caso de las mujeres, la pendiente descendente es más pronunciada y marcada, ya que las mujeres mayores terminan con salarios más bajos. No obstante, es notable la gran similitud en cuanto al rango de distribución salarial y las observaciones que se encuentran en un intervalo similar, que son concordantes a lo que se ha encontrado a lo largo del documento donde la brecha salarial de genero no es tan alta ´para ser observada gráficamentee. Los salarios son comparables y la curva de valores predichos dentro de la muestra muestra esa concavidad en ambos casos, ajustándose adecuadamente a los datos. El peak de mujeres parece ser un poco más acelerado, y esta en los 30,68 años que difiere del pico de los hombres y justamente reafirma ese movimiento mas rapido de la curva predicha. Este tipo de resultados visuales subraya la importancia de que los modelos de regresión consideren esta forma particular de los datos en su estructura e incluyan interacciones cuadráticas y exploren hasta cubicas.



Gráfica 7.

[Predicción]

Para desarrollar un modelo predictivo del ingreso por hora de las personas, nos basamos en la ecuación de Mincer y en la literatura de economía laboral. Esta literatura ha intentado explicar cuáles son las variables y factores que inciden de manera significativa en los ingresos de los hogares. Se ha sugerido que los determinantes clave son la educación, la edad y la experiencia, buscando analizar los efectos marginales de estas variables en los ingresos. Se han propuesto diversas aproximaciones econométricas al problema, y en este estudio, nos enfocamos en aprovechar este aporte teórico.

Construimos cinco modelos con distintos niveles de complejidad, explorando diferentes grados polinomiales de las variables de edad, experiencia y años de educación (formación académica). La variable de años de experiencia en el mercado laboral no estaba directamente disponible en la encuesta, por lo que la calculamos restando la edad de la persona a los años máximos de educación alcanzada. Además, ajustamos este cálculo para las personas mayores de 65 años, donde la resta de edad y educación solo se contabiliza hasta esa edad. Partimos del supuesto de que después de los 65 años, las personas se suelen pensionar, evitando así observaciones de experiencia demasiado elevadas para personas de edad avanzada.

Posteriormente, incluimos controles que la misma literatura ha sugerido. Aunque no teníamos acceso a datos de raza o nacionalidad, que podrían ser de gran interés, sí pudimos acceder a información relevante relacionada con la especificidad del trabajo y las condiciones bajo las cuales se ejerce, a través de la gran encuesta integrada de hogares (GEIH). Utilizamos controles que caracterizaban distintos tipos de labores. Entre estos, contamos con una variable categórica que clasifica a las empresas según su tamaño. Consideramos que esto es relevante porque el tamaño de una firma suele estar asociado con su productividad y su capacidad para generar salarios competitivos. También incluimos variables binarias que indican si la persona reside en una región urbana o rural, así como si está afiliada al sistema de salud, lo cual también está relacionado con las condiciones laborales de la persona. Otra variable binaria indica si la persona pertenece al sector formal o informal, lo cual es particularmente significativo en la economía colombiana, donde para datos de 2023 el 57% del empleo se encuentra en la informalidad a nivel nacional y puede empeorar en rural disperso donde llega hasta el 80% (Dane,2023). Aunque inicialmente pensamos en incluir efectos fijos por departamento, resultó innecesario debido a que todas las personas encuestadas provienen de Bogotá. Esto nos permitió contar con una base homogénea en términos territoriales para nuestro análisis.

Después de seleccionar cuidadosamente estas variables, procedimos a construir modelos con diferentes niveles de complejidad. Comenzamos con un modelo inicial que intentaba predecir el salario por hora basándose en la edad, edad al cuadrado y edad al cubo. Este modelo arrojó el MSE más alto, con un valor de 0,496. Posteriormente, incrementamos la complejidad del modelo.

El segundo modelo incorporó las variables más relevantes: edad, años de educación máximos alcanzados y una variable que exploraba los pesos de la frecuencia. En este segundo modelo, se exploraron varios grados polinómicos para la variable de edad, especialmente 8, y también se probaron 3 grados para la variable de educación máxima alcanzada. Incluso se llevaron a cabo interacciones entre estas variables de edad y educación utilizando diferentes grados polinómicos.

Los modelos subsiguientes comenzaron a incorporar el cálculo de la experiencia laboral, siguiendo las sugerencias de la ecuación de Mincer (CITA). También incluyeron variables sobre la formalidad laboral, la cotización a pensiones, y características de las firmas en las que trabajaban, tales como su tamaño. En algunos modelos, se incluyeron todas las variables de control, mientras que en otros solo se consideró el tamaño de la firma o únicamente 2 controles adicionales junto a los tres principales (edad, años de educación, experiencia laboral).

Finalmente, decidimos que la variable de edad debería tener 3 grados polinómicos, la educación 2 grados y la experiencia laboral 3 grados. Así logramos identificar el modelo con el menor MSE.

$$log_{salario_{hora}} = \beta 0 + \beta 1 \cdot agei + \beta 2 \cdot agei2 + \beta 3 \cdot agei3 + \beta 4 \cdot maxEducLeveli + \beta 5 \cdot maxEducLeveli2 + \beta 6 \cdot experiencia_{ajustada} + \beta 7 \cdot experiencia_{ajustada} + \beta 8 \cdot experiencia_{ajustada} + \beta 9 maxEducLeveli * experiencia_{ajustada} + \beta 10 maxEducLeveli2 * experiencia_{ajustada} + \beta 11 \cdot formal + \beta 12 \cdot p6620 \beta 9 \cdot sizeFirm + \beta 13 \cdot regSalud + \beta 14 \cdot formal + \beta 15 \cdot p6620$$

El MSE de este modelo es de 0,314, lo que representa una notable reducción de más del 40% en comparación con el modelo inicial que no incluía interacciones. El segundo modelo con el MSE más bajo se acerca mucho en valor, registrando un MSE de 0,3192. Es importante percibir esta medida de MSE como la distancia entre la predicción realizada y el valor promedio muestral del salario por hora, expresado en logaritmo. Esta consideración es relevante en la interpretación, dado que la variable dependiente está en un nivel logarítmico, permitiendo medir el nivel predictivo de nuestro modelo.

MODELO	MSE
1	0.3372970
2	0.3372970
3	0.3633487
4	0.3156009
5	0.3132020

Se encontraron momentos en los que, a pesar de incluir más variables de control de manera lógica, un aumento excesivo en el grado polinómico de las tres variables principales, como usar grados 4 o 5, resultaba en un empeoramiento en la eficiencia del modelo. Esto podría interpretarse como un sobreajuste (overfitting) en la base de entrenamiento, lo que se traduce en una baja eficiencia al aplicarse en la base de prueba.

Se observó que al incluir la variable X y manejar grados polinómicos menores a 4, así como considerar la interacción entre años de educación y experiencia laboral, se lograba una disminución en el MSE y una mejora en la eficiencia del modelo

Referencias

Dane, 2023. Ocupación informal Trimestre móvil mayo -julio 2023. https://www.dane.gov.co/files/operaciones/GEIH/bol-GEIHEISS-may-jul2023.pdf

Repositorio en Github:

https://github.com/Randresil/BigData-MachineLearning202320