Problem Set 1:

Predicting Income

Integrantes: Juan Diego Duarte[1]; Erick Julian Villabon[2]; Carlos Torres Sandoval[3]; Tania Reina[4]

1. 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. Esta brinda 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. Por lo que este estudio utiliza datos del reporte de "Medición de Pobreza Monetaria y Desigualdad" de 2018, centrándonos en la población mayor de edad en Bogotá, allí se 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. En consecuencia, estos datos podrían ayudar a reducir la brecha fiscal y adicionalmente, permitirían identificar familias e individuos en condición de vulnerabilidad que serían potencialmente beneficiarios de asistencia que focalizarán el gasto.

Con el fin de obtener una estimación 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. Donde se expone una relación positiva entre el nivel educativo de un individuo y sus ingresos futuros, así como con su experiencia laboral (Hartog, 2016). Por lo tanto, hemos decidido emplear esta teoría como base para nuestras proyecciones con el objetivo de obtener resultados coherentes y fundamentados. Una correcta identificación de los ingresos reales de la

---

[1] Código: 202011999
[2] Código: 201815677
[3] Código: 202225155
[4] Código: 202015300

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.

En este orden de ideas, se realizó una caracterización causal de algunos de los elementos que determinan el salario de las personas, posteriormente, se estableció un modelo de predicción de los salarios con el fin de que este se pueda aplicar por fuera de muestra para evitar la brecha fiscal. En primera instancia, encontramos que la relación entre la edad y el salario esta caracterizada por una función cuadrática. En esta, el salario máximo se alcanza con una edad de entre 42 y 44 años, y un año de vida adicional aumenta o reduce el salario en un (8.9 – 0.2*Edad) % el salario.

Se analizó la brecha salarial de género en la economía laboral mediante una serie de modelos de regresión. Inicialmente, se encontró que las mujeres ganaban un 14.7% menos que los hombres en un modelo simple. Sin embargo, al incorporar controles como la experiencia laboral, las horas trabajadas, el tipo de trabajo, y la edad, la brecha de género se amplió considerablemente, llegando a un 24%. Estos resultados sugieren que factores adicionales ajenos al género influyen en las disparidades salariales. Además, se aplicó la metodología FWL para mejorar la eficiencia del modelo, obteniendo un coeficiente similar. El análisis de bootstrap proporcionó un valor predicho y un intervalo de confianza, lo que respalda la evidencia de una brecha salarial significativa de género en la muestra estudiada.

El siguiente 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 caracteriza el efecto causal de la edad sobre los ingresos de las personas. La sección 4 analiza la brecha salarial de género en la economía laboral mediante una serie de modelos de regresión. Por último, la sección 5 realiza las predicciones del salario.

2. 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[5]. 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

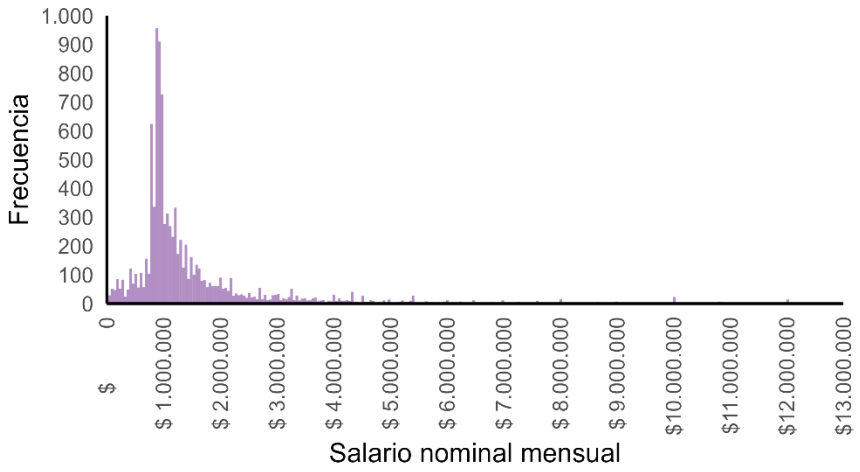
[5] Página a realizar web scraping: https://ignaciomsarmiento.github.io/GEIH2018_sample/

	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%. Al desagregar estos resultados por genero (Hombre – Mujer) podemos evidenciar una brecha en indicadores de educación y mercado laboral como el salario mensual, ingreso total y educación superior.

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.

3. Perfil de salario por edad

Para entender de manera integral el comportamiento de los salarios es fundamental comprender que factores 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 (Cherrington, et al, 1979). 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 (Lazear, 1976)

El mecanismo indirecto sería el siguiente: Las personas jóvenes tienen una mayor probabilidad de poseer menos años de educación y años de experiencia; mientras más años de vida se tiene, menores son esas probabilidades (Myck, 2010). Así mismo, dentro de países intensivos en mano de obra, el mecanismo directo se puede caracterizar así: Las personas jóvenes tienen una menor capacidad de adaptación al ambiente laboral, lo cual va mejorando al pasar los años (Lee, et al, 1985). De la misma manera, mientras menos edad se tiene, más competencia laboral existe, y, por ende, en ciertos rangos de edad la productividad se mantiene alta o aumentando por efecto de mercado (Lazear, 1976). En contra posición, las personas jóvenes están asociadas a unas cualidades “blandas” (motivación, necesidad de trabajar, escalabilidad de puestos laborales, etc.) que van mejorando al pasar los años, pero que en un cierto punto empiezan a empeorar gravemente (Lee, et al, 1985).

Todos estos mecanismos asociados a la edad provocan que la literatura sea muy concisa en mostrar que en países intensivos en mano de obra 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 se puede 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, en países intensivos en mano de obra.

Este *trade-off* entre edad y salarios se puede caracterizar en un modelo cuadrático (Lazear, 1976) de la siguiente manera:

$$Log(Salario\ mensual) = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 + u$$

Aquí la edad se relaciona de manera cuadrática con el salario bajo los mecanismos que ya describimos. Los resultados de este modelo, bajo la muestra que describimos en el Punto 2 de este documento, son los siguientes:

Tabla 2. Regresión Salario y edad con errores estándar robustos

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

El efecto marginal de la edad sobre el salario mensual es de (8.9 – 0.2*Edad) puntos porcentuales por cada año adicional. En términos sencillos, cada año de vida adicional aumenta 8.9 % el salario mensual, pero este efecto marginal 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

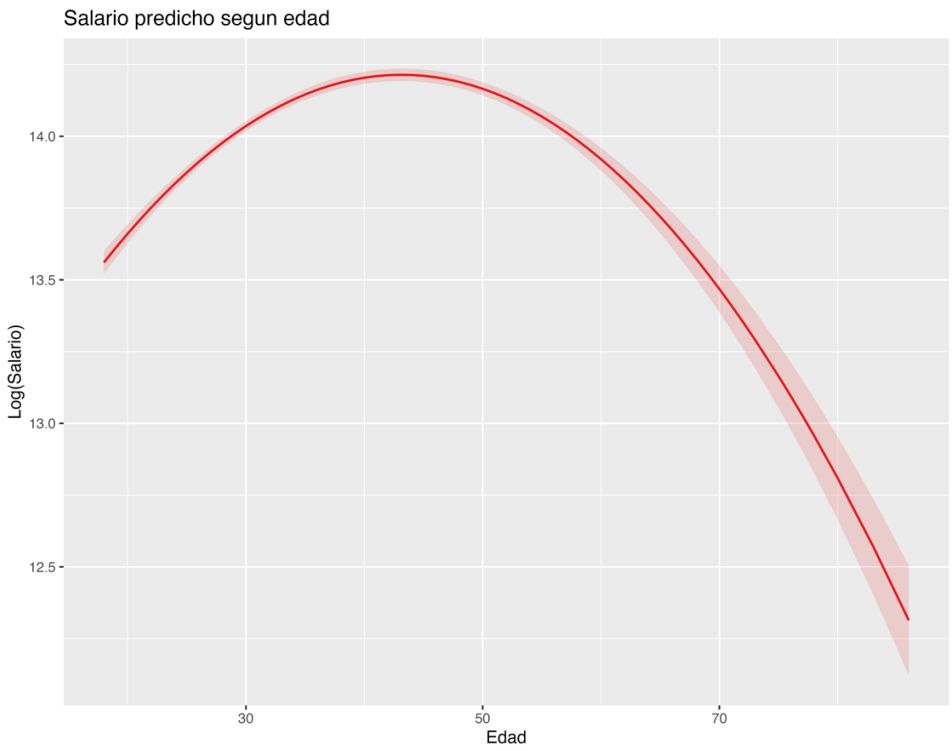
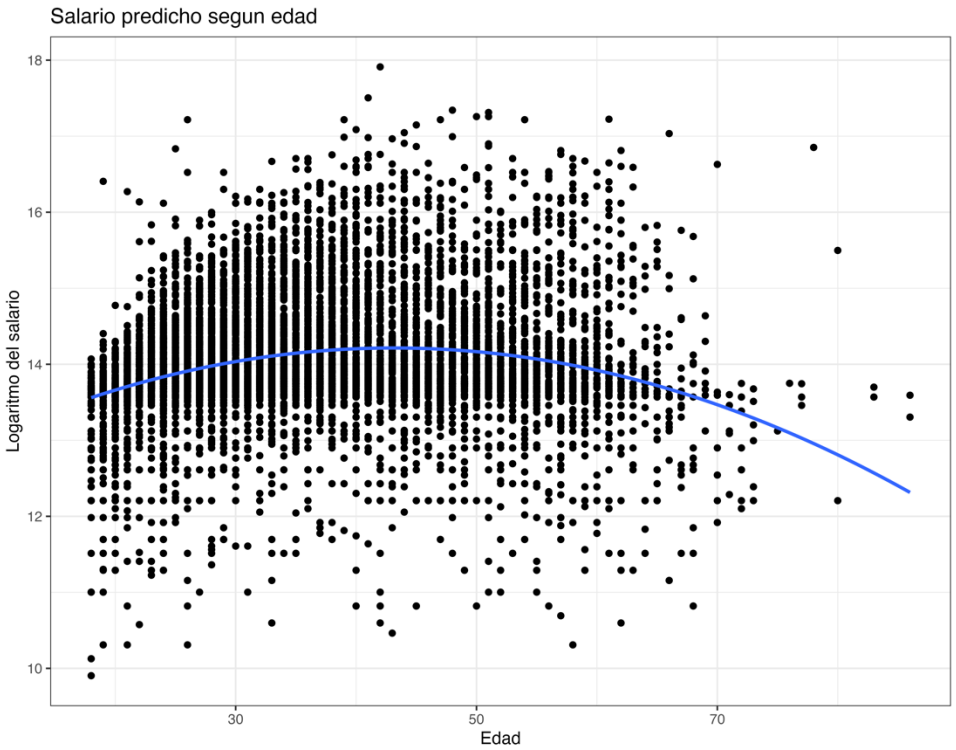


Gráfico 4: Salario predicho según edad



De esta manera se e puede observar, en ambos gráficos, la relación cuadrática evidente entre la edad y el salario mensual, teniendo así una edad que maximiza el salario mensual que se encuentra entre

los 40 y 50 años. De hecho, el grafico 3 muestra el valor estimado del salario según la edad, 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 que existe una mayor variabilidad de los datos en estos puntos. Esto nos dice que lo salarios de las personas con mayor edad son menos homogéneos, en comparación a los de las personas más jóvenes. Por esta razón se usaron errores estándar robustos a la hora de la regresión correspondiente a la Tabla 2. 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 (este caracteriza la manera en que se va comportando la pendiente del efecto marginal de la variable) son significativas al 99% de confianza. A su vez, en esta regresión poseemos un R^2 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, aun así solo se está prediciendo el comportamiento del salario en un 5.8%. 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, no se puede afirmar que el modelo este haciendo un excelente trabajo a la hora de predecir los datos de la muestra. Esto último no implica que el modelo no nos sirva para encontrar causalidad y caracterizar los efectos marginales la edad, pero el modelo se queda corto a la hora de generar predicciones acertadas del salario dentro de la muestra.

El problema de caracterizar esta relación cuadrática entre el salario y la edad, en ultimas, se vuelve un problema de encontrar esa edad que maximiza el salario mensual. En este orden de ideas, la expresión que maximiza ese salario es la siguiente:

$$-\frac{\beta_2}{2 * \beta_3} = Edad^*$$

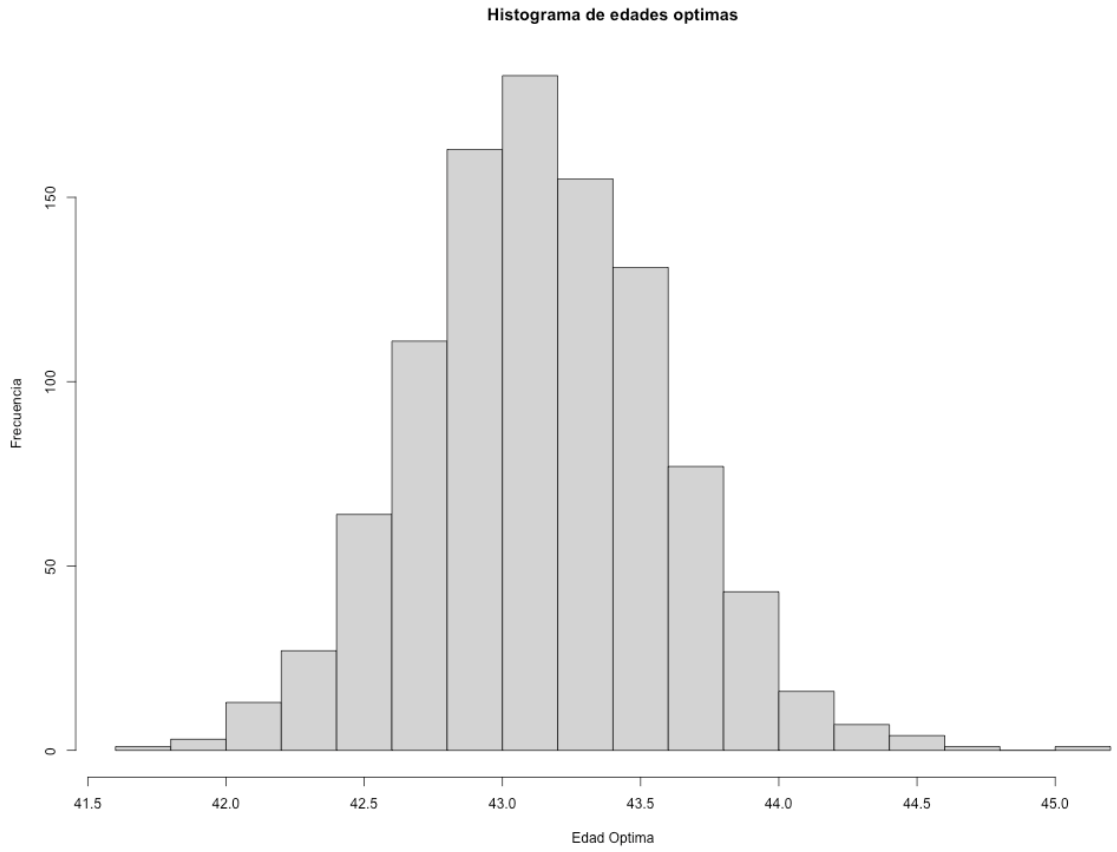
De manera tal que, a la hora de predecir este valor, por medio de la metodología *bootstrap* con 1000 iteraciones, obtenemos el siguiente valor predicho:

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*



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

4. Brecha salarial de género

La brecha salarial por género es uno de los temas fundamentales para entender en la economía laboral. En esta sección vamos a evaluar si con los datos anteriormente descritos en la sección 2 podemos encontrar evidencia de dicha brecha.

Para comenzar, vamos a estimar un modelo sencillo para explicar el salario teniendo como variable independiente únicamente el género de los trabajadores.

$$\log(Salario) = \beta_0 + \beta_1 Mujer + u$$

Para hacer esta estimación tenemos la variable de género como una dummy, cuando la variable es 1, se trata de una mujer, en caso contrario es un hombre.

Tabla 3. Regresión Salario y sexo

	Variable dependiente:
	Log(Salario mensual)
Mujer	-0.147***
	(0.015)
Constante	14.088***
	(0.011)
Observaciones	9,892
R^2	0.009
Nota:	*p<0.1; **p<0.05; ***p<0.01

En el modelo más sencillo podemos ver en la tabla 3 que cuando se trata de una mujer, el salario suele ser un 14.7% menos que cuando se trata de un hombre. Aunque se trata de un resultado significativo estadísticamente, queremos evaluar si al interactuar con distintas variables que pueden tener incidencia en el ingreso se siguen obteniendo estos resultados tan contundentes que ratifican la evidencia de la brecha de género en el salario.

Para agregar los controles a nuestro modelo, se realizó una revisión de literatura y se usó como guía un informe de la Comisión Económica Europea en la que identifican posibles causas de las diferencias en el ingreso de las mujeres.

La primera que identifican es la segregación sectorial que se relaciona con la mayor cantidad de mujeres empleadas en sectores peor remunerados que tienen a ser sistemáticamente infravalorados, para esta causa, agregamos al modelo la variable de oficio. Otro aspecto fundamental es el trabajo no remunerado, que en la mayoría de los casos recae sobre la mujer, este trabajo hace que tengan menos horas disponibles para un trabajo remunerado y puede ser significativo en la brecha total, para este aspecto agregamos la variable de horas trabajadas. Finalmente, agregamos la edad y la experiencia laboral para lograr controlar aspectos relacionados con los techos de cristal y la discriminación salarial.

$$\log(Salario) = \beta_0 + \beta_1 Mujer + \beta_2 Experiencia\ laboral + \beta_3\ Horas\ trabajadas + \beta_4\ Oficio + \beta_5\ Edad + u$$

Tabla 4. Regresión Salario y sexo con controles.

	Variable dependiente:		
	Log (Salario mensual)		Residuales Salario
	(1)	(2)	(3)
Mujer	-0.147***	-0.240***	
	(0.015)	(0.013)	
Experiencia laboral		0.002***	
		(0.0001)	
Horas trabajadas		0.015***	
		(0.001)	
Oficio		-0.012***	
		(0.0002)	
Edad		0.003***	
		(0.001)	
residualesMujer			-0.240***
			(0.013)
Constante	14.088***	13.754***	0.000
	(0.011)	(0.036)	(0.006)
Observaciones	9,892	9,892	9,892
R^2	0.009	0.307	0.032
Note:	*p<0.1; **p<0.05; ***p<0.01		

En la tabla 4 se muestran los resultados del modelo con los controles anteriormente explicados. Se puede ver que respecto a la primera estimación se obtiene un mayor impacto de la diferencia de género en el salario, obteniendo un 24% menos de salario respecto a los hombres. Aunque todos nuestros controles son estadísticamente significativos, es de resaltar que a medida que al agregar controles, se tiene un R^2 más grande lo que indica que nuestro modelo es más preciso explicando con los datos de nuestra muestra, pero que puede hacerse menos preciso cuando intentemos hacer predicción con datos fuera de ella.

En la columna (3) podemos ver los resultados de la estimación utilizando FWL. En este caso podemos ver que obtenemos un coeficiente igual al de usar todos los controles, pero un R menor que nos va a permitir tener mejores predicciones fuera de muestra haciendo nuestro modelo más eficiente.

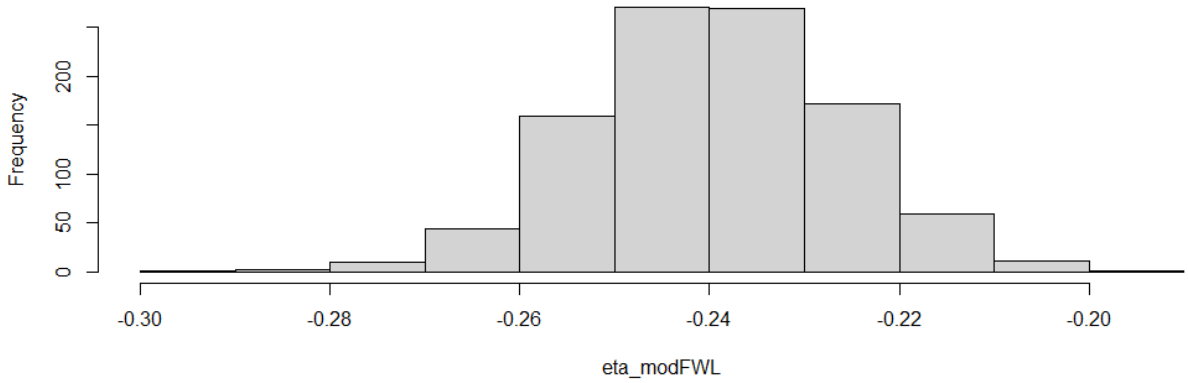
A la hora de predecir este valor, por medio de la metodología *bootstrap* con 1000 iteraciones, obtenemos el siguiente valor predicho:

Tabla 4. Intervalo de confianza

Intervalo de confianza	**Valor**
Límite inferior	-0.2148
Valor predicho	-0.2394
Límite superior	-0.2663
Error estándar del estadístico	0.0133

Con la siguiente distribución del estadístico:

Gráfico 5: Distribución de los valores estimados por *bootstrap*



5. Predicción de salario

Con el propósito de evaluar la capacidad predictiva de los modelos anteriores, así como los nuevos modelos propuestos en esta sección, se implementa un enfoque que garantiza la replicabilidad de los resultados. En este sentido, se inicia seleccionando una semilla que asegura la reproducibilidad de las estimaciones. A continuación, la muestra de datos se divide aleatoriamente en dos partes: el 70% de los datos constituyen la muestra de entrenamiento, mientras que el 30% restante conforma la muestra de prueba. Esta división se realiza con la finalidad de utilizar la muestra de entrenamiento para desarrollar y ajustar los modelos, y posteriormente, la muestra de prueba se emplea para evaluar su capacidad predictiva

Los nuevos modelos estimados son:

- Modelo 3

$$\log(Salario) = \beta_0 + \beta_1 mujer + \beta_2 \exp trab\ actual + \beta_3 horas\ trab\ usual + \beta_4 oficio + \beta_5 edad + u$$

- Modelo 4

$$\log(Salario) = \beta_0 + \beta_1 mujer + \beta_2 hijo\ hogar + \beta_3 horas\ trab\ usual + \beta_4 oficio + \beta_5 \exp trab\ actual + u$$

- Modelo 5

$$\log(Salario) = \beta_0 + \beta_1 edad + \beta_2 edad^2 + \beta_3 mujer + \beta_4 hijos\_hogar + \beta_5 horas\_trab\_usual + \beta_6 \exp trab\ actual + \beta_7 informal + u$$

- Modelo 6

$$\log(Salario) = \beta_0 + \beta_1 edad + \beta_2 edad^2 + \beta_3 mujer + \beta_4 hijos\_hogar + \beta_5 horas\_trab\_usual + \beta_6 \exp\ \_trab\_actual^2 + \beta_7 Oficio + \beta_8 informal + \beta_9 Estrato + u$$

- Modelo 7

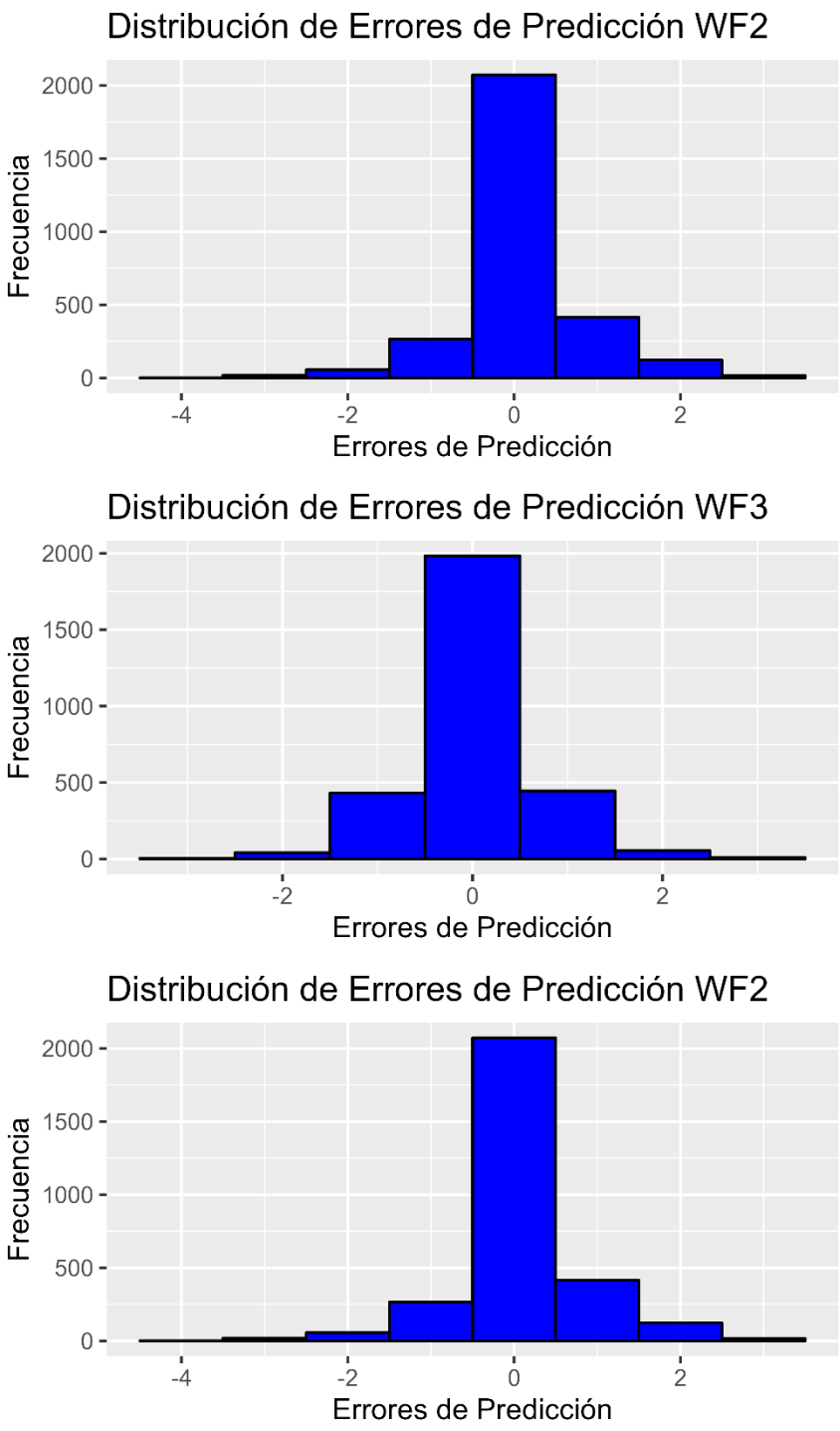
$$\log(Salario) = \beta_0 + \beta_1 secundaria + \beta_2 media + \beta_3 superior + \beta_4 \text{exp\_trab\_actual} + \beta_5 \exp\ \_trab\_actual^2 + u$$

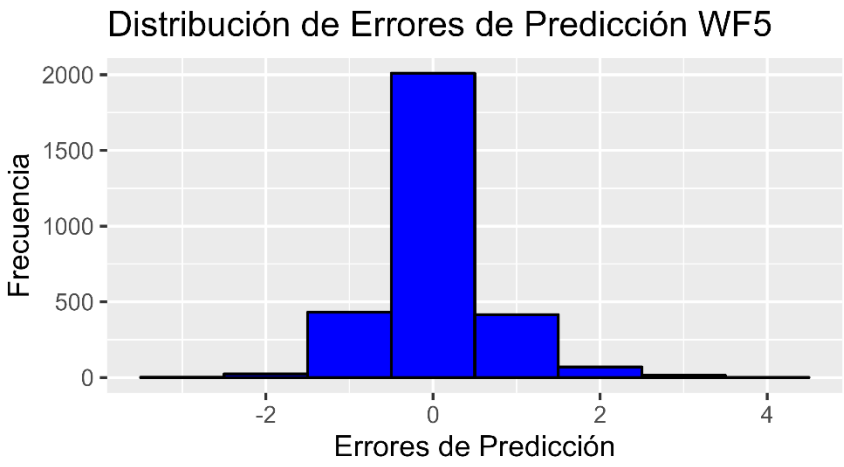
- Modelo 8

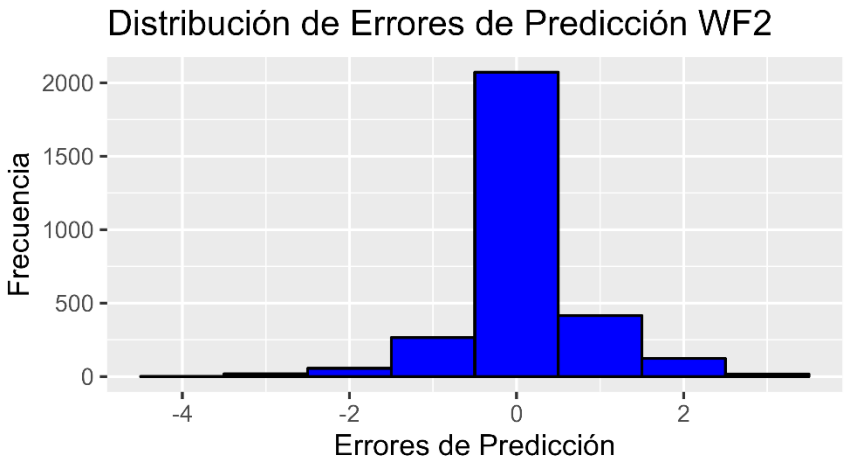
$$\log(Salario) = \beta_0 + \beta_1 edad + \beta_2 edad^2 + \beta_3 mujer + \beta_4 hijos\ hogar + \beta_5 mujer * hijos\ hogar + \beta_6 mujer * amo\ casa + \beta_7 \exp\ \_trab\_actual^2 + \beta_8 Oficio + \beta_9 informal + u$$

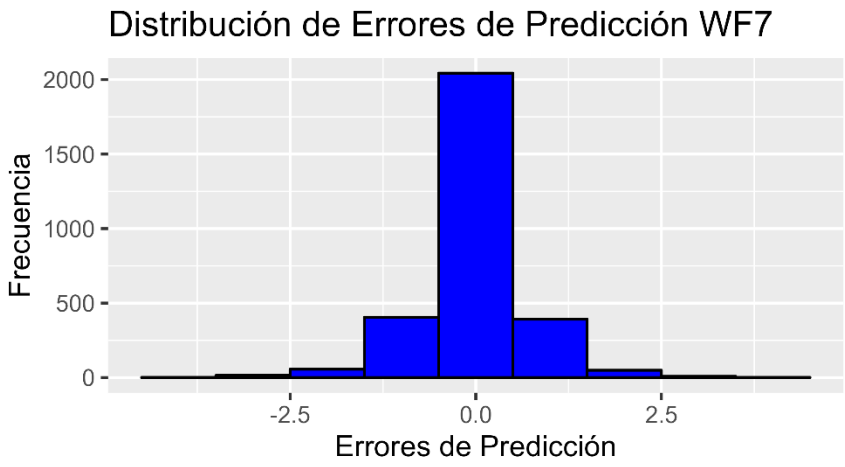
En este caso los nuevos modelos reconocen la no linealidad entre las variables y su relación con el salario, las experiencias en el actual empleo reportado y las horas de trabajo a la semana representan rendimientos decrecientes. La interacción entre mujer e hijos hogar y mujer, amo de casa exhiben una relación decreciente frente al salario como se esperaría. Los siguientes gráficos presentan la distribución de los MSE de los modelos estimados.

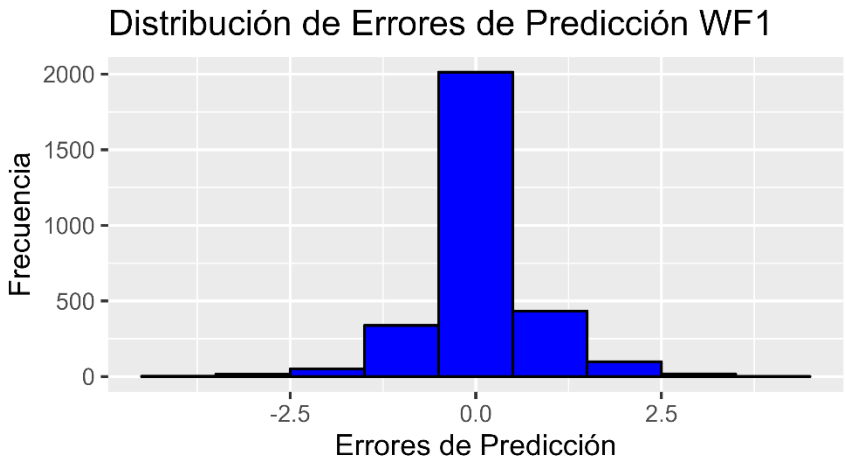
Gráfico 6: Distribución de los MSE




Distribución de Errores de Predicción WF5
Frecuencia
2000
1500
1000
500
0
-2
0
2
4
Errores de Predicción


Distribución de Errores de Predicción WF2
Frecuencia
2000
1500
1000
500
0
-4
-2
0
2
Errores de Predicción





Los gráficos muestran que a medida que a medida que sube la complejidad del modelo el MSE va disminuyendo, sin embargo, en nuestros modelos no se está realizando ningún sobreajuste al momento de aumentar la complejidad del modelo, esto sustentado bajo la teoría de económica al momento de incluir las variables. En este sentido, el modelo que mejor predice por fuera de muestra es el modelo 8. Aunque claramente este es el modelo que más complejidad tiene, este a su vez, es el que mejor predice por fuera de muestra.

GitHub

Mediante el siguiente enlace:

https://github.com/Erick-Villabon/Problem_Set_1

Referencias

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

Lazear, E. (1976). Age, Experience, and Wage Growth. The American Economic Review, 66(4), 548–558. http://www.jstor.org/stable/1806695

Cherrington, D. J., Condie, S. J., & England, J. L. (1979). Age and work values . Academy of Management Journal, 22, 617-627 .

Myck, M. (2010). Wages and Ageing: Is There Evidence for the 'Inverse-U' Profile?. Oxford Bulletin of Economics and Stadistics.

Lee, R., & Wilbur, E. R. (1985). Age, Education, Job Tenure, Salary, Job Characteristics, and Job Satisfaction: A Multivariate Analysis. Human Relations, 38(8), 781–791. https://doi.org/10.1177/001872678503800806

Halog, J., & Gerritsen, S. (2016). Mincer Earnings Functions for the Netherlands 1962–2012. *De Economist*, 164, 235-253.