María Carolina Gutiérrez Cód. 201922402
Sara Torres Benavides Cód. 201923245
Sofía Quiroga Robles Cód. 201913999
Luis Carlos Rodríguez Cód.

Problem Set 1: Predicting Income

I. Introducción

El enfoque de este trabajo es hacer un modelo que logre predecir el ingreso de los ciudadanos de Bogotá a partir de determinadas variables independientes. Poder predecir el ingreso de los ciudadanos ofrece varias ventajas, entre estas, permite identificar personas en situaciones vulnerables, entender que factores determinan el ingreso y cuáles de estos factores son características inherentes a la persona, como su edad o su sexo biológico. Además, un modelo de predicción de ingresos permite identificar si hay una brecha entre el salario reportado por los ciudadanos que pagan impuestos y el salario predicho (la estimación del salario verdadero). El reporte de ingresos de los ciudadanos es esencial para poder calcular el recaudo fiscal, por lo tanto, poder predecir la brecha entre el reporte y los ingresos verdaderos, es una oportunidad para mejorar la gestión del recaudo.

El trabajo consiste en cuatro fases diferentes: la adquisición y limpieza de los datos, un análisis de la relación edad-ingresos, un análisis de la brecha salarial de género y la diferencia en las edades óptimas entre los sexos y, por último, un análisis del poder predictivo de los modelos especificados. Después de llevar a cabo los diferentes análisis de regresión y de métodos de remuestreo, se encontró que la edad óptima estimada es menor a la hipótesis del enunciado (50 años) y que el salario es decreciente en la edad. Además, se encontró que existe una brecha salarial condicionada y no-condicionada entre hombres y mujeres, las mujeres reciben en promedio un salario menor al de los hombres, incluso cuando tienen características similares y se dedican a trabajos parecidos. Finalmente, se compararon modelos con diferentes niveles de complejidad y el que tuvo un mejor desempeño predictivo fue uno de los modelos más complejos, con 12 variables explicativas. Es importante tener en cuenta que los modelos complejos suelen tener buen desempeño prediciendo dentro de la muestra y por fuera de la muestra el error tiende a ser mayor.

II. Datos

2.1 Base de datos

La base de datos que se usará para ajustar el modelo es la base de la Gran Encuesta Integrada de Hogares (GEIH). Esta base contiene información de una muestra representativa de la población en edad de trabajar del país. La encuesta es llevada a cabo por el Departamento Administrativo Nacional de Estadística (DANE) cada año con el objetivo de obtener indicadores del mercado laboral colombiano (Datos Abiertos Gobierno, 2021). La población de enfoque de este trabajo son las personas ocupadas y asalariadas de Bogotá mayores de 18 años, de esta población se puede encontrar información socioeconómica relevante. Además de contener variables generalmente asociadas al salario como años de educación y edad, contiene variables del contexto social de la persona como su sexo biológico, estrato, educación de sus padres, entre otras. Se considera que esta base de datos es relevante útil para poder plantear el modelo de predicción de ingresos de los bogotanos

Adicionalmente, algunas características destacables de la GEIH, es que la obtención de los datos se realiza de forma aleatoria y por muestreo. Sin embargo, tiene la limitación de que conforma un sistema de datos del tipo sección cruzada repetida, en la cual los hogares encuestados no son los mismos a lo largo del tiempo. La base también presenta el tipo de muestra, identificando estrados y conglomeraciones, al igual que determina la probabilidad de que un hogar sea encuestado. La unidad de medida corresponde a las viviendas, hogares e individuos. La GEIH es caracterizada por la extensa cobertura del territorio colombiano. Sin embargo, excluye los departamentos de la Amazonía y Orinoquía, las cabeceras municipales que no son capitales de departamento, así como los centros

poblados y rural disperso. También se excluye la población de Providencia y el centro poblado y rural disperso de San Andrés. En este trabajo solo se usarán los datos recolectados para hogares en Bogotá.

2.2 Adquisición de datos

La base de datos a utilizar son los datos de la GEIH 2018, la cual cuenta con una totalidad de 178 variables y 32177 observaciones. La base se obtuvo en una página web propiedad del profesor de la Universidad de Los Andes Ignacio Sarmiento: https://ignaciomsarmiento.github.io/GEIH2018_sample/. Para obtener los datos de la página se usó la técnica de web scraping utilizando el paquete de "rvest" de R-estudio. Dentro de la página, los datos se encuentran divididos en 10 "chunks" o particiones con un promedio de 3218 observaciones cada una. Al momento de utilizar el programa, no se pudo realizar el webscraping utilizando el link de la página de cada chunk de datos que se quería obtener, esto se debía a que la librería "rvest" carga todos los datos que aparecen en los primeros momentos de la página, esto generaba un problema ya que las tablas tardaban varios segundos en aparecer, por lo que no se podía hacer scraping de las tablas con los datos que se descargaban. Para solucionar este problema, se utilizó el enlace que se generaba en cada página después de que se cargaran las tablas (este enlace se podía obtener de la sección "red" al analizar los elementos de la página). Luego de descargar la información de la página, se procedió a separar la tabla con las variables y observaciones de los demás datos descargados. Este proceso se repitió con los 10 chunks que conformaban toda la base de datos para al final unir todas las tablas en una sola y así tener la base completa de la GEIH 2018.

2.3 Limpieza de la base de datos

La base de datos de la GEIH 2018 obtenida en el punto anterior estaba en formato de "lista", este formato no es tan útil al momento de hacer análisis datos en el programa de R, por lo que se necesitó cambiar de formato "lista" a "dataframe", el cual es un tipo de formato especial del programa que facilita el análisis de datos ya que se comporta parecido a una matriz, por lo que se pueden tratar los datos por filas, columnas y observaciones.

Luego de obtener el dataframe, se procede con la limpieza de la base, la cual consiste en obtener solo los datos que se van a usar para ajustar el modelo de predicción. El primer paso es tratar las observaciones en blanco o "Missing values" (NA). En este caso, como la base es amplia (32177 observaciones), se puede eliminar los missing values y seguir teniendo una muestra grande (16682 observaciones). Después, como el modelo de interés es uno que predice el salario por hora, el siguiente paso es obtener las observaciones de solo las personas en edad de trabajar, es decir las personas mayores a 18 años. Usando la variable que indica si la persona es ocupada o no, se seleccionan únicamente a las personas ocupadas. Luego de esto se separan las variables que se usarán en el trabajo, que en este caso son: sexo, estrato, edad, nivel de educación máximo, tiempo trabajado en la empresa e ingresos totales.

2.4 Variables

A continuación, se presentan las primeras variables seleccionadas para la investigación junto con las estadísticas descriptivas y una breve descripción de su significado. De la muestra anteriormente seleccionada, se cuenta con un total de 16.542 datos.

Tabla 1: Estadísticas descriptivas

Variable	Media	Desv. Est.	Mín.	Pctl. 25	Pctl. 75	Máx.
sexo	0.53	0.499	0	0	1	1
estrato	2.551	1.011	1	2	3	6

edad	39.436	13.483	18	28	50	94
max_educativo	5.001	1.099	1	4	6	9
tiempo_trabajo	63.758	89.488	0	7	84	720
salud	1.089	0.339	1	1	1	9
salario	1.769.378,995	2.675.627.739	0	8e+05	1.723.158.25	85.833.333.33

En primer lugar, la variable *sex* hace referencia al género del individuo seleccionado de la muestra. Inicialmente, es una variable dicótoma que toma el valor de 1 cuando es hombre y 0 si es mujer. Como indica la media, el 53% de la muestra tomada en consideración son hombres y el restante (47%) son mujeres. Esta variable es de gran importancia para tomar en cuenta al momento de estimar el modelo de predicción porque en el mundo laboral, es común que por diversos factores las mujeres tengan un salario diferente, normalmente inferior, al de los hombres. En particular, por sesgos de género, a las mujeres se les ofrece menor salario a comparación de los hombres al igual que se autoseleccionan a trabajar en labores estereotipadas como por ejemplo enfermería. Al colocar esta variable, es posible capturar la diferencia existente entre los salarios de hombres y mujeres y predecir el salario para cada uno de estos dos perfiles con mayor precisión.

En segundo lugar, la variable *estrato* es una variable categórica que indica el estrato de la persona entrevistada. Esta variable toma los valores entre 1-6. El promedio de la variable (2.5) indica que un colombiano promedio se encuentra entre los estrados 2 y 3. Sin embargo, esta variable es altamente dispersa, pues el 75% de la muestra se encuentra entre el estrato 1 y 4 (este cálculo corresponde a sumar y restar dos veces la desviación estándar al promedio de la variable). Adicionalmente, es importante mencionar que, como lo indican los percentiles, únicamente el 25% de la muestra pertenece a los estratos 1 y 2, mientras que el 75% de la muestra se encuentra hasta el estrato 3, donde el restante 25% de la muestra componen los estratos 4, 5 y 6. Esta alta variabilidad da indicios de los desigual que es la población colombiana, donde la mayoría perteneces a la clase media o baja, mientras que los más ricos (clasificados como aquellos que ganan más de 10.000.000 COP) se encuentran en los estratos más altos pero que apenas representan el 25% de la muestra. Teniendo en cuenta esta alta variabilidad y la desigualdad de ingresos, es necesario tomar en cuenta esta variable como predictora en el modelo, pues influye en gran medida en los niveles de salario obtenidos. En parte, esta relación se puede dar porque las personas más ricas puedes tener mayor acceso a privilegios, como educación. Por lo tanto, esta variable permite aislar el efecto que tienen los distintos estratos con el salario, ya que, en el caso colombiano, mientras más alto el estrato, es común que el salario sea mayor.

En tercer lugar, la variable *edad* señala la edad de la persona al momento de responder la encuesta. La media de edad de la muestra es de 39.43 años. Esto significa que la media de edad de todas las personas que están ocupadas es de alrededor de 39 años. Este valor es un buen reflejo de la media que deberían tener las personas en edad de trabajar teniendo en cuenta 2 factores, la edad mínima para de 18 años, y la edad mínima para pensionarse que es un promedio de 60 años entre hombres y mujeres (hombres 62 y mujeres 57). Si la población fuera homogénea entre las edades, el promedio de edad de trabajo debería ser de 38.75 años, valor que es parecido a la media de edad que tenemos. La edad puede ser considerada como una aproximación de la experiencia, y experiencia se puede traducir como una mayor capacidad de hacer el trabajo, por lo que los salarios normalmente son mayores mientras más edad se tenga, pero esto no es constante, ya que se puede llegar a un punto donde la experiencia ya no afecte a los salarios. Esto se puede deber a que mientras más viejo se es, la productividad ya no es la misma que la de una persona joven, y aunque se sepa cómo hacer el trabajo de mejor forma, siempre esta mejora tiene un límite, el cual puede ser alcanzado por una persona más joven con una cantidad decente de experiencia. Por lo que es difícil que el salario crezca por términos de productividad. En otras palabras, es común

que la experiencia tenga pendiente positiva y rendimientos decrecientes. Por lo anterior, la edad es un acercamiento de la experiencia que tiene la persona y sus capacidades actuales para realizar el trabajo.

Por otro lado, la variable *max_educativo* es una variable categórica que indica el nivel máximo de educación que la persona entrevistada tiene y lo califica entre una variable del 1 al 7, el número 9 es un indicados si la persona no respondió. En promedio, la muestra tiene un nivel educativo secundaria incompleto (6to-10mo), identificado como categoría 5. Es altamente preocupante, pues este resultado indica que un colombiano promedio, tomado de la muestra, no alanza a terminar el colegio. Esto da indicios de la alta probabilidad de incurrir en deserción escolar. En parte, en Colombia ocurre debido a las presiones familiares, donde no se impulsa la educación sino el trabajo para aportar ingresos y alimentos al hogar. De hecho, alrededor del 75% de los encuestados en 2018 se encontraban con bajos niveles de educación, entre ningún tipo de estudios como apenas acabando el colegio. Note que la varianza en esta variable es relativamente alta, pues con una desviación estándar, una persona puede pasar de un nivel educativo de secundaria incompleto a uno incompleto, y de una incompleto a uno de apenas cruzar quinto de primaria. Es preocupante pensar que la mayoría de la muestra no haya accedido a educación universitaria; esto implica que no existe una buena formación de capital humano ni tampoco una buena provisión del servicio por parte del estado. Así, se eligió esta variable porque la educación es uno de los factores más importantes que definen el salario de las personas pues el nivel educativo es una muestra del nivel de conocimientos y preparación que tienen las personas, mientras mayor nivel de educación, se espera que sean capaces de desarrollar tareas y trabajos más complejos, por lo que sus salarios cambiarían en función de la dificultad y requisitos del trabajo.

Igualmente, la variable *tiempo_trabajo* registra el tiempo que lleva la persona trabajando en el mismo puesto medido en meses. El promedio de la muestra es de 64 meses. La muestra tiene tiempos de trabajo dispersos, el mínimo siendo de un solo mes y el máximo de alrededor de 60 años. Esta variable fue seleccionada porque el tiempo en el mismo puesto de trabajo, puede indicar la permanencia que tienden a tener las personas ocupadas en un solo puesto. Un salario bajo junto a un tiempo de trabajo alto indica condiciones que no le permiten a una persona ascender a un mejor puesto de trabajo.

Por otro lado, la variable s*alud* indica si la persona está afiliada o cotiza a un servicio de seguridad social. Es una variable binaria que toma el valor de 1 sí está afiliada y 0 de lo contrario. El 9 indica que la persona no contestó la pregunta. Esta variable permite tener una idea de la formalidad del trabajo de la persona, ya que todos los empleados con contrato formal en Colombia deben cotizar salud. Las personas de la muestra están afiliadas a un servicio de seguridad social, ya que el trabajo se concentra en la muestra de ocupados. Por lo tanto, no se usó la variable en los modelos para predecir el salario ya que era redundante.

Finalmente, se hace referencia a la variable del *salario*, no como predictora pero sí como medida de caracterización de la variable con interés de predecir. Esta variable se definió como la suma de los ingresos totales para cada una de las personas encuestadas. Se considera que esta es una medida para aproximar los salarios de un individuo, pues la mayor proporción de los ingresos de un hogar colombiano provienen en del salario por el trabajo. Además, la muestra tomada en consideración fue filtrada por aquellos individuos que mencionaban estar ocupados, por lo cual se asume implícitamente que trabajan y la mayor parte de los ingresos provienen de este.

Para un individuo promedio de la muestra, el salario esperado es de 1.769.378 COP. Representa más que el salario mínimo. Sin embargo, es importante mencionar que incluso hay personas que no ganan ningún salario (monto equivalente a 0), y también existe una persona que gana casi 50.000.000 COP más que el promedio. Además, el ingreso es altamente volátil en la muestra a una desviación estándar, una persona puede ganar alrededor de 4 millones o alrededor de cero pesos. Es importante mencionar la forma en la que se distribuye el salario en la muestra. El 25% de la muestra gana alrededor de

800.000COP, algo inferior al salario mínimo del momento. Aún más preocupante termina siendo que el 75% de la muestra gana alrededor de 1.723.000COP, muy cercano al salario mínimo. Mientras que, el restante percentil más alto de la distribución, ganan sumas de dinero entre los 2.000.000COP hasta los 85.833.333 COP. Nuevamente, este resultado da evidencia de la alta desigualdad económica y de oportunidades de la muestra y del país per se, pues los más ricos reciben ingresos mucho mayores a los que gana un ciudadano promedio. Además, los más ricos terminan siendo alrededor del 1% de la población en Colombia.

Ante la alta volatilidad del comportamiento del salario en Colombia, deben ser tomados en consideración varios predictores como los previamente mencionados para predecir de forma óptima, según las características de las personas, el salario esperado estimado.

III. Perfil Edad-Ingreso

Para probar la hipótesis de que los salarios se maximizan, para una persona promedio, cuando se tiene alrededor de 50 años de edad, se estimó la ecuación (2) pero para datos de Colombia, específicamente de Bogotá. Al momento de estimar el perfil de un colombiano de la muestra seleccionada de la GEIH, se obtiene el siguiente resultado:

Tabla 2.1. Regresión entre logaritmo del ingreso total y la edad

	(1) Logaritmo del ingreso
Edad	0.058***
	(0.003)
Edad^2	-0.001***
	(0.00003)
Constante	12.805***
	(0.059)
Observaciones	16,277
R-cuadrado	0.024
R cuadrado ajustado	0.024
F- estadístico	197.768

Errores estándar en paréntesis
*** p<0.01, ** p<0.05, * p<0.1

De lo anterior, el modelo estimado puede escribirse de la siguiente forma:

$$log(w) = 12.805 + 0.058 \cdot Edad \ - \ 0.001 \cdot Edad^2$$

En primera instancia, resulta pertinente mencionar que el modelo estimado presenta estimadores con los signos esperados, significancia individual al igual que significancia global. Note que entre mayor edad tenga un individuo, mayor será su ganancia esperada, pues implícitamente se captura el efecto de la experiencia laboral e incluso nivel de escolaridad. Esto se refleja con el signo positivo del coeficiente

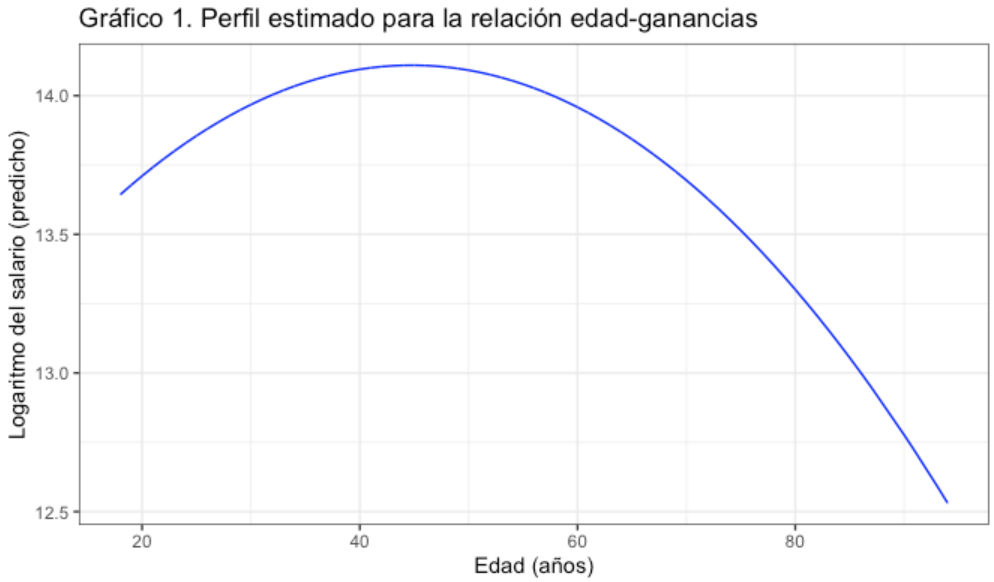
de edad. Sin embargo, esta relación positiva, no perdura para toda la vida, pues las personas de mayor edad tienden a ganar menos dinero bien sea por haber sido reemplazados en el trabajo o porque simplemente llegaron a la edad de jubilación alrededor de los 60 años. Esta relación cóncava de la función se refleja con el signo negativo del coeficiente de la edad al cuadrado. Respecto a la significancia global, es posible afirmar con un 99% de confianza que las variables edad y edad al cuadrado explican la variación en el comportamiento del logaritmo natural del ingreso. Sin embargo, al notar el bajo valor del R^2, equivalente a 2%, indica que las dos variables independientes únicamente alcanzan a explicar la variación en el logaritmo natural del ingreso en un 2%, mientras lo demás se encuentra en el error. En cuanto a este resultado, no resulta sorprendente el bajo valor del R^2, pues el modelo (2) deja por fuera variables independientes importantes, como por ejemplo el nivel educativo, años de experiencia laboral, entre otros. Esto lleva a que el modelo presente el problema de variables omitidas y que los estimadores sean inconsistentes, sesgados e ineficientes. Esto implica que los estimadores encontrados estén subestimando o sobreestimando el efecto verdadero que tiene la edad sobre el salario. Sin embargo, el caso sencillo es una forma de validar a simple vista el cumplimiento de la hipótesis de la maximización del salario a cierta edad.

En cuanto a la interpretación del coeficiente de edad, es posible afirmar con una confianza incluso del 99%, que el estimador es estadísticamente diferente de cero. En parte puede deberse por realizar la estimación con gran número de observaciones (muestra de 16,277), implicando que el error estándar del coeficiente sea bastante inferior al estadístico, aumentando la probabilidad de cometer error tipo II. La interpretación económica indica que por un año más de edad del individuo, este tendrá un aumento en las ganancias de 5.8% en promedio. Este efecto es equivalente a un aumento en el logaritmo del salario del 0.45% como porcentaje de la media. Es posible argumentar que este efecto es relativamente bajo respecto a la media del logaritmo de los ingresos pues, en promedio, al incrementar la edad de una persona el salario (en logaritmo) esperado aumenta tan solo un 0.42% a comparación de la media.

En cuanto al coeficiente de edad al cuadrado, este es significativo también al 1%. Nuevamente, la alta significancia puede deberse al gran número de observaciones utilizadas para calcular los errores estándar. Ahora bien, más que una interpretación económica, el análisis del coeficiente se centra en el signo que indica el estimador. Estadísticamente es diferente de cero, y para efectos de este ejercicio, el estimador resultó siendo negativo, implicando que la función de maximización de los ingresos conforme la edad es estrictamente cóncava, por lo cual es posible calcular la edad a la cual se maximizan los ingresos a lo largo de la vida en promedio de un individuo. De igual forma, esto es coherente con la hipótesis a probar.

Un punto a analizar es también el intercepto estimado. Es importante mencionar que se tomó en cuenta las personas mayores a 18 años al igual que no se tomaron en cuenta las observaciones que indicaban salarios o ganancias nulas, pues como para este ejercicio se requería hacer la trasformación monotónica del logaritmo, se requería que este estuviera definido. En este sentido, la ecuación estimada no pasa por el origen. Con esto en mente, la interpretación del intercepto carece de sentido económico.

Ahora bien, para tener referencia sobre la tendencia de la estimación que maximiza el logaritmo del salario según la edad, a continuación, se presenta la gráfica que ilustra el comportamiento:



Como es posible observar, la gráfica indica una función cóncava para la relación del logaritmo de los ingresos en función de la edad. Esta tendencia invertida se debe a que, como se explicó anteriormente, la edad de una persona refleja la experiencia laboral e incluso reputación, por lo cual entre los primeros años (18-44) la tendencia es positiva pero luego decrece, pues, por ejemplo, el costo de entrenar a personas mayores es cada vez más grande al igual que durante la vejez, se depende más de una jubilación (proporción del salario) que del salario *per se*. Al analizar esta relación, es posible apoyar la hipótesis propuesta en el enunciado, en donde el perfil promedio de un individuo trabajador indica que los salarios tienden a ser menores cuando se está joven, pero van aumentando a medida que aumenta la edad; llega a un máximo alrededor de los 50, aunque para estos datos parece ser más alrededor de los 55 años, y luego la tendencia parece decrecer después de los 50.

Note que los individuos más jóvenes tomados en consideración (18 años), en promedio ganan alrededor del salario mínimo en Colombia (900.000 y 1.000.000 COP). Usualmente, estos individuos se caracterizan por ser recién graduados y sin experiencia, por lo cual no se les reconoce una gran suma de dinero. Sin embargo, a medida que va incrementando la edad, el salario también presenta una tendencia creciente hasta ubicarse en un máximo. Para la muestra tomada en consideración, el salario máximo recibido se maximiza cuando el individuo tiene 44.7 años, o mejor aproximado a 44 años de edad, pues el individuo sigue estando en sus 44 y no 45 años. Con este dato, el salario máximo alcanzado durante toda la vida corresponde a un valor esperado de 1.342.105 COP. A comparación del dato de la media para los 44 años, 1.409.041 COP, es menor en 4.75%. Esto puede deberse a la gran variabilidad entre los individuos tomados en consideración de la muestra, donde los más ricos a la edad de 44 años, direccionan un alto promedio del salario.

Luego, pasados los 44 años, tal como ilustra la gráfica, se percibe una caída en el logaritmo del salario esperado, llegando incluso a valores menores de lo que ganaban los individuos en su juventud. Por ejemplo, a los 80 ganan alrededor de 568.070 COP en promedio, mucho menos de lo que ganaban recién entraban a la oferta laboral de 18 años (900.000 COP). Nuevamente, esto se debe en parte a que las personas mayores de 62 ya no pertenecen a la oferta laboral por la legislación colombiana, pero hay unos pocos, grandes eminencias, que de igual forma siguen trabajando. También, hay varias personas, por no decir la mayoría de los colombianos, que no logran acceder a un fondo de pensiones, por lo cual

los ingresos disminuyen incluso a cero y quienes acceden al beneficio de la pensión, claramente reciben una proporción del ingreso menor.

Por otro lado, es importante centrar la discusión sobre la edad a la cual se maximiza en ingreso. El modelo predice que esta edad corresponde a los 44 años de un individuo promedio de la muestra, y comparando con la edad de 50 que propone en enunciado, son menores los años necesarios para alcanzar el máximo salario a lo largo de la vida.

A pesar de antes haber realizado la inferencia estadística con los errores estándar de la regresión, es importante realizar inferencia estadística con la desviación estándar calculada mediante Bootstrap, los cuales están sujetos a heteroscedasticidad. Al realizar este ejercicio, se encuentra una desviación estándar de 0,499 para un re-muestreo con reemplazo de 1000 observaciones; así, el respectivo intervalo de confianza corresponde a: [-0.92, 1.03] para el estimador de la edad.

En primera instancia, es importante mencionar que los errores encontrados con Bootstrap son bastante mayores a los encontrados con la regresión, de hecho, son mayores en una magnitud de 0.496. En parte, esto se debe a que primero, los errores con Bootstrap se realizaron para pseudo muestras de tamaño 1000, mientras que en la regresión de la Tabla 2.1, se utilizaron 16,277 observaciones. En segunda medida, también se debe al bajo ajuste del modelo en términos del R^2 (2.4%). Según los intervalos de confianza calculados con Bootstrap, el estimador de la edad no es estadísticamente significativo, pues los intervalos contienen el 0. Esto indica que, en muestras repetidas, el 95% de las veces el parámetro está contenido en el intervalo, por lo cual no existe evidencia estadística para afirmar que el coeficiente de la edad es estadísticamente significativo utilizando Bootstrap.
Este resultado es importante para poner sobre la mesa la discusión de la guerra de las estrellas y de la significancia. Claramente, con muestras muy grandes, se tiende a rechazar la hipótesis nula, indicando que hay significancia en los coeficientes, sin embargo, esto no necesariamente implica un mayor efecto. En este caso, el re-muestreo de las 1000 observaciones, indica que la significancia es efecto del tamaño de la muestra más que el efecto verdadero de la edad sobre las ganancias del logaritmo del salario.

IV. Brecha salarial de género

La brecha salarial de género es una de las preocupaciones principales de la economía con enfoque de género. En este trabajo se hará un análisis de brecha salarial por sexo biológico usando los datos de la población ocupada de Bogotá. Para analizar la brecha salarial no-condicionada, se generó una variable binaria “female” que toma el valor de 1 si la persona es de sexo biológico femenino y el valor 0 de lo contrario.

El modelo de brecha salarial no-condicionada es:

$$logw_i = \beta_o + \beta_1 female_i + u$$

Donde la variable dependiente es el logaritmo del salario y la variable independiente es la variable dummy “female”.

Un slogan común es “Equal pay for equal work”, dando a entender que cuando el trabajo es el mismo, no debería existir brecha en el salario. Teniendo esto en cuenta, se plantea un modelo de brecha salarial condicionada, incluyendo variables para controlar por características similares de trabajo. Entre las variables de control se incluyó la variable de nivel máximo de educación que la persona ha recibido, suponiendo que las personas que han alcanzado un nivel igual de educación tienen trabajos con

características similares y, por lo tanto, su remuneración es parecida. También se incluyó el tiempo que lleva en su trabajo actual medido en meses, suponiendo que hay sectores de trabajo en los que estar en un mismo lugar de trabajo por más tiempo es más común que en otros. Además, se incluyeron las variables edad y edad al cuadrado (edadsq), considerando que las personas en grupos de edad más cercanos se dedican ocupan puestos de trabajo parecidos, por ejemplo: las personas en la primera mitad de sus veintes se dedican a trabajos para principiantes, mientras personas en la segunda mitad de sus cuarentas, pueden estar ocupando puestos de trabajo directivos. Por último, se incluyó el estrato como variable de control, ya que las personas en el mismo estrato socioeconómico pueden estar dedicadas a actividades parecidas y tener características socioeconómicas similares.

El modelo de brecha condicionada es:

$$logw_i = \beta_o + \beta_1 female_i + \sum_{j=2}^{6} \beta_{ij} \, \mathrm{X}_{ij} + u$$

Donde $\mathrm{X}_1 = \{maxEducativo, tiempoTrabajo, estrato, edad, edadsq\}$ es la matriz de controles y $\mathrm{X}_2 = \{\text{female}\}$ es la variable de interés.

A partir del slogan "Equal Pay for Equal Work" se esperaría que la brecha salarial condicionada fuera nula o lo más cercana a 0, es decir, para trabajadores con las mismas características y trabajos parecidos, el salario debería ser igual, sin importar el sexo biológico. Al llevar a cabo ambas estimaciones se encontraron los coeficientes estimados para cada variable

Tabla 4.1 Modelos brecha de género condicionada y no-condicionada

	Brecha no condicionada	Brecha condicionada
(Intercepto)	6.05 ***	6.06 ***
	(0.01)	(0.01)
female	***-0.16 ******	***-0.19 ******
	(0.01)	**(0.01)**
maxEducativo		0.14 ***
		(0.01)
tiempo_trabajo		0.05 ***
		(0.01)
estrato		0.14 ***
		(0.01)
edad		0.47 ***
		(0.04)
edadsq		-0.47 ***
		(0.04)
Observaciones	16542	16542
R2	0.01	0.10

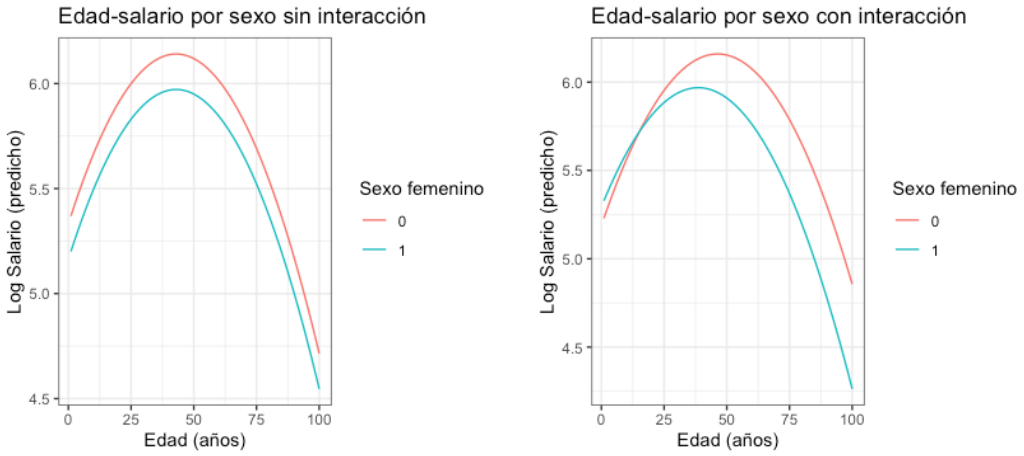
All continuous predictors are mean-centered and
scaled by 1 standard deviation.
*** p < 0.001; ** p < 0.01; * p < 0.05

A partir de los resultados de ambas regresiones, se encontró que la brecha salarial de género existe cuando se condiciona en otras variables y cuando no. La brecha salarial no condicionada es de 0.16, es decir, la diferencia del salario entre el sexo femenino y masculino es del 16% del salario promedio. Para la estimación de la brecha condicionada se encontró un resultado contrario a lo que supone el slogan "Equal Pay for Equal Work", cuando se controla por características de trabajo y trabajador, la brecha salarial es mayor. Según la estimación, en promedio el salario de una persona de sexo femenino es 19% menor que el salario de una persona de sexo masculino. Entonces, cuando se compara entre personas de características similares en términos sociales y de la actividad económica que realizan, la brecha salarial es casi igual al 20%.

Usando el teorema FWL, se encontraron los mismos coeficientes para las variables de control (X_1), cuando se regresan en la variable dummy *female* (X_2) y luego se regresan los residuos de regresar el logaritmo del salario en X_2. Con el método de remuestreo "Bootstrap", se estimó la misma brecha salarial en promedio que con el método de FWL. Sin embargo, los errores estándar son diferentes, el de Bootstrap es menor que el de FWL. Esto se explica porque con Bootstrap no se asumen errores homoscedásticos y son una estimación de la desviación estándar de la distribución de la muestra, a partir de esto, los errores estimados son menores que los errores de FWL que sí se asumen como homoscedásticos por el método de Mínimos Cuadrados Ordinarios.

Por otro lado, se llevó a cabo un análisis de edad-ingreso con enfoque de sexo biológico para encontrar la edad óptima de los individuos y si esta es diferente por sexo. Al realizar la estimación de edad óptima se encontraron dos edades óptimas diferentes entre sexo masculino y femenino. Para el sexo femenino se encontró una edad óptima de alrededor de 43 años, mientras que para el sexo masculino se encontró una edad óptima de 53 años. Realizando el análisis gráfico a partir de un modelo de predicción, se puede estimar que las personas de sexo femenino alcanzan su salario máximo a una edad menor. Infortunadamente, esto no es porque se demoren menos en alcanzar el mismo salario máximo que las personas de sexo masculino, se estima que sus ingresos empiezan a decrecer con la edad a una edad menor que los de sexo masculino. Para el análisis gráfico y las predicciones se plantearon dos modelos: uno con interacción entre la edad y sexo femenino (2) y otro sin la interacción (1). Se planteó el modelo con interacción para estimar el efecto de un cambio marginal en la edad (aumento en un año) en el salario promedio de una persona de sexo femenino.

$$\log(w) = \beta_0 + \beta_1 female_i + \beta_2 edad_i + \beta_3 edad^2{}_i + u \quad (1)$$
$$\log(w) = \beta_0 + \beta_1 female_i * edad_i + \beta_3 edad^2{}_i + u \quad (2)$$



En ambos casos, el salario máximo que alcanza una persona de sexo femenino es menor que el de una persona de sexo masculino. En el caso de la interacción es evidente que mientras los ingresos de una persona de sexo masculino empiezan a ser decrecientes respecto a la edad después de los 50 años, los de una persona de sexo femenino empiezan a decrecer alrededor de los 40 años.

Teniendo en cuenta los resultados anteriores, es evidente que, al estimar los salarios de las personas, para aquellas de sexo femenino el salario es menor, incluso cuando las características del trabajo y la persona son similares. La brecha salarial se evidencia incluso en términos de la edad óptima. La brecha puede tener dos componentes: un problema de discriminación y otro de selección. Suponiendo que la mayoría de las personas de sexo femenino de la muestra son mujeres cis-género y la mayoría de las personas de sexo masculino son hombres cis-género, se puede considerar la estimación de la brecha como una brecha salarial por género. Por un lado, la discriminación laboral al sexo femenino puede darse de forma vertical, los puestos de trabajo con salarios más altos se concentran los hombres y por lo tanto la brecha salarial no condicionada afecta negativamente el salario de las mujeres. Por otro lado, la discriminación laboral se puede dar de forma horizontal, en el mismo puesto de trabajo, las mujeres reciben un salario menor que el del hombre. Así, la brecha salarial condicionada también afecta a las mujeres negativamente y es mayor, porque en las variables de control las personas no difieren mucho, entonces es más evidente que el sexo biológico genera variación en el salario.

La discriminación puede ser estructural, generalmente las mujeres tienen una carga mayor de trabajo de cuidado y doméstico, necesitan mayor flexibilidad en horas y ausencia, a veces deben interrumpir su trayectoria laboral para dedicarse completamente al trabajo de cuidado y esto disminuye los retornos de la experiencia que podrían tener (Goldin, 2014). Sin embargo, la discriminación también puede ser estadística, en ausencia de información completa, los empleadores agrupan a todas las mujeres por su sexo y toman decisiones con base en sus creencias sesgadas sobre el grupo (Eswaran, 2014). La discriminación, tanto estadística como estructural, puede moldear las creencias de las mujeres y generar un problema de selección. Las mujeres se autoseleccionan para estudiar carreras que se adhieren más a los roles de género, dejando de considerar carreras con retornos más altos por la discriminación que creen que pueden enfrentar (STEM Women, 2022). El problema de selección está conectado a la discriminación desde la niñez, si las mujeres están acostumbradas a que el trabajo de cuidado recae en ellas desde una edad temprana, es posible que empiecen a buscar oportunidades laborales que permita la flexibilidad que requieren para llevar a cabo el trabajo doméstico (Blau y Winkler, 2021). Por lo tanto, para acatar el problema de brecha salarial se requieren esfuerzos en términos de flexibilidad laboral, modificación de incentivos de empleadores y empleados, economía del cuidado y roles de género.

V. Predicciones de ingreso

A partir de los resultados e inferencia estadística realizada anteriormente, se busca evaluar el poder predictivo de las especificaciones del logaritmo del salario con respecto a la edad y el género de los bogotanos. Para esto, se divide la base de datos original entre el grupo de entrenamiento con 11348 observaciones, y un grupo de prueba con 4929 observaciones, respectivamente de la muestra seleccionada de la GEIH.

De esta forma, para determinar el mejor modelo predictivo, se busca determinar aquel modelo con el menor error predictivo. Es así, como a partir de los tres modelos planteados anteriormente, se plantean dos nuevos modelos más complejos para determinar el grado de complejidad óptimo del modelo, y realizar la mejor predicción fuera de muestra. De esta forma, los modelos utilizados para comparar los errores predictivos son:

$$\log(w) = \beta_o + \beta_1 edad + \beta_2 edad^2 + u \qquad (1)$$

$$\log(w) = \beta_o + \beta_1\, edad + \beta_2\, edad^2 + \beta_3\, tiempoTrabajo + \beta_4\, tiempoTrabajo^2 + \beta_5\, edad \cdot tiempoTrabajo + \beta_6 edad \cdot tiempoTrabajo^2 + \beta_7 maxEducativo + \beta_8 maxEducativo^2 + \beta_9\, edad \cdot maxEducativo + \beta_{10} edad \cdot maxEducativo^2 + \beta_{11}\, sexo + \beta_{12}\, edad \cdot sexo + \beta_{13}\, estrato + u \quad (2)$$

$$\log(w) = \beta_o + \beta_1\, sexo + u \quad (3)$$

$$\log(w) = \beta_o + \beta_1\, sexo + \beta_2\, maxEducativo + \beta_3\, tiempoTrabajo + \beta_4\, estrato + \beta_5\, edad + \beta_6\, edad^2 + u \quad (4)$$

$$\log(w) = \beta_o + \beta_1\, sexo + \beta_2\, tiempoTrabajo + \beta_3\, tiempoTrabajo^2 + \beta_4\, sexo \cdot tiempoTrabajo + \beta_5\, sexo \cdot tiempoTrabajo^2 + \beta_6\, maxEducativo + \beta_7\, maxEducativo^2 + \beta_8\, sexo \cdot maxEducativo + \beta_9\, sexo \cdot maxEducativo^2 + \beta_{10}\, edad + \beta_{11}\, edad^2 + \beta_{12}\, sexo \cdot edad + \beta_{13}\, estrato + u \quad (5)$$

Para el caso específico de ambos modelos complejos, se decide agregar la variable de tiempoTrabajo, la cual representa el tiempo actual que lleva el individuo trabajando en esa organización, debido principalmente a que se espera que un trabajador, al aumentar su tiempo de trabajo, se vuelva experto en este, logrando procesos óptimos y eficaces. De esta forma, puede llegar a aumentar su productividad y obtener mayores logros, escalando dentro de la jerarquía de la empresa, y a su vez, obtener un mayor salario (WorkMeter, 2022). Sin embargo, se incluye también esta variable al cuadrado, dado que se considera que el trayecto laboral puede detenerse en determinado puesto de trabajo, dado por cuestiones de experiencia, habilidades y conocimientos requeridos, así como de disponibilidad laboral para ascensos. Por lo cual, el salario dejaría de crecer como lo esperado en un inicio, mostrando así un comportamiento cóncavo.

Por otra parte, en ambos modelos se decide agregar la variable de maxEducativo, siendo el nivel máximo de educación recibido por una persona. En este caso, se considera esta variable relevante dado que, bajo un nivel de estudios superior, un individuo tiene mayores posibilidades de encontrar un puesto de trabajo con un mayor salario o de mejorar su salario actual, al estar mejor capacitado y con ciertas habilidades entrenadas (La República S.A.S., 2020). Sin embargo, también se agrega esta variable al cuadrado, dado que algunas personas pueden llegar a estar sobrecalificadas para un puesto laboral, por lo cual, su salario no presentará cambios. Asimismo, se agrega la variable del estrato dado que Bogotá está marcada, en cierto grado, por las clases sociales. Con lo cual, dependiendo del estrato de una persona se espera que tenga cierta capacidad adquisitiva, lo cual implica un estimado de su salario, para cubrir sus necesidades básicas y así lograr determina si necesita apoyo económico por parte del estado. Asimismo, se incluye la interacción entre la edad y el sexo, con el interés de observar el efecto marginal de la edad sobre el salario en base de una persona con sexo masculino.

De otro modo, dentro del segundo modelo, se agrega la interacción entre la edad y el tiempo del trabajo, para determinar si existe algún efecto conjunto entre estas variables sobre el salario. Esto dado que, ante una menor edad, los individuos sienten mayores oportunidades para explorar y cambiar de trabajo sin tantas preocupaciones, por lo que presentan mayores niveles de rotación laboral. De esta forma, las personas de menor edad tienen menos niveles de tiempo de trabajo dentro de una empresa, lo cual afecta sus niveles de salario en ciertos periodos de tiempo. Asimismo, se agrega la interacción entre la edad y el nivel máximo de educación recibido, para evaluar el posible efecto de estas variables sobre el salario. En este caso, se considera que las personas de mayor edad cuentan con un mayor nivel educativo, y por ende con un posible mayor nivel de salario.

Por otro lado, dentro del quinto modelo, se añade la interacción entre el sexo y en tiempo de trabajo, con el fin de observar el efecto marginal del tiempo de trabajo sobre el salario a partir de una persona con sexo masculino. Asimismo, se incluye la interacción entre el sexo y el nivel máximo de educación recibido, para capturar el posible efecto marginal del nivel máximo de educación sobre el salario, en base del sexo masculino. Para ambos casos de las interacciones anteriores, se consideran relevantes dado lo que se ha mencionado anteriormente, por la discriminación estructural sobre los roles de género dentro de la sociedad. Los cuales afectan las dinámicas, tiempo de trabajo, y niveles educativos de las mujeres, generando un impacto negativo sobre sus niveles de salario.

Ahora bien, para todos los modelos descritos anteriormente, se evalúa el desempeño predictivo en los datos de prueba. De esta forma, se utilizan los coeficientes estimados con los datos de entrenamiento para usarlos como predictores en los datos de prueba. Al llevar a cabo este proceso, se obtienen los siguientes resultados:

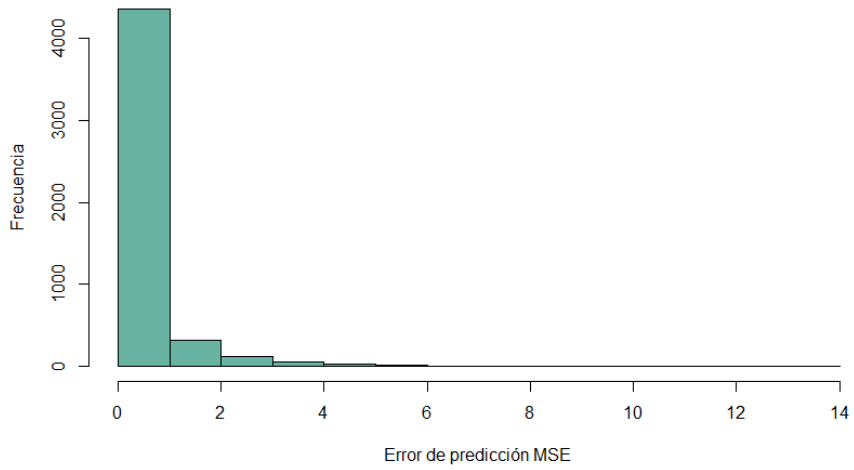
Tabla 5.1 Desempeño Predictivo

Modelos	**MSE**
Modelo 1	0.738
Modelo 2	0.464
Modelo 3	0.747
Modelo 4	0.478
Modelo 5	0.467

En este caso, es importante resaltar que aquellos modelos con un mejor desempeño predictivo son los modelos propuestos con una especificación no lineal y un mayor grado de complejidad. De esta forma, estos modelos con un mayor grado de complejidad logran tener una mejor predicción fuera de muestra, sin caer dentro de un modelo sobre ajustado. Dado lo anterior, el modelo con un mejor desempeño predictivo es el modelo 2, con el mínimo error cuadrático medio de 0.464. En este caso, el modelo 2 presenta el valor de MSE más cercano a cero, lo cual da indicios sobre la mejora en la precisión de predicción del modelo 2 frente a los demás modelos. Sin embargo, es relevante mencionar que en el caso de modelos excesivamente complejos se sobre ajusta la muestra de entrenamiento, llevando el MSE a cero y logrando una mala predicción fuera de la muestra.

A partir de lo anterior, dentro de la distribución de los errores de predicción para le modelo 2, se encuentra que estos tienen una distribución similar a la distribución F. En este caso, la mayoría se los errores de predicción se encuentran entre 0 y 2. Sin embargo, se puede evidenciar en el gráfico de distribución la existencia de algunos outliers en el modelo. En la cual, dada la mayor longitud de la cola hacia la derecha, el modelo 2 de predicción utilizado está subestimado el salario de los bogotanos. De esta forma, es posible plantear que el modelo refleja que en la realidad las personas están declarando ante la la DIAN un menor nivel de ingresos para disminuir sus niveles de impuestos. De esta forma, es necesario proponer a la DIAN una mayor revisión y control de los ingresos reportados por los bogotanos para eliminar el fraude de impuestos. Por otra parte, para asegurar la existencia de los outliers dentro del modelo, se puede pensar en la implementación de otro modelo que contenga nuevas interacciones y diferentes variables, el cual presente un MSE menor de 0.464 para asegurar una mejor predicción fuera de muestra y la existencia de dichos outliers en los errores de predicción.

Gráfico 5.1. Distribución errores de predicción



Finalmente, en base al modelo 2 y modelo 5 con el mejor desempeño predictivo, se calculan los errores predictivos usando validación cruzada con el método LOOCV. En este caso, se encuentra que para el modelo 2, el MSE es de 0.489086. De esta forma, la estimación del modelo mediante validación simple y validación cruzada LOOCV son valores muy similares, con única diferencia de 0.025, siendo en este caso mayor el valor de la validación cruzada. Asimismo, en el caso del modelo 5, el MSE es de 0.491966, siendo este valor muy similar al error de validación simple. La diferencia entre el MSE por validación cruzada LOOCV y validación simple es de 0.024, donde es mayor la estimación del error por LOOCV.

Tabla 5.2 Validación Cruzada LOOCV

Modelo	RMSE	R cuadrado	MAE
Modelo 2	0.699347	0.3649708	0.5001509
Modelo 5	0.7014031	0.3612303	0.5040376

De esta forma, es importante tener en cuenta que la estimación del error por medio de validación simple es muy variable, y depende que aquellas observaciones que se incluyen dentro del modelo de entrenamiento. Sin embargo, al estimar el error del modelo mediante LOOCV, y utilizar todo el conjunto de datos menos una observación dentro del entrenamiento, se logra reducir la variabilidad del modelo. Por otra parte, otra de las ventajas de la estimación mediante LOOCV es que permit identificar el grado de flexibilidad, de los grados del polonio, para obtener el mejor modelo de predicción. A partir de lo anterior, el mejor modelo encontrado para predecir el salario de los colombianos es el modelo 2 dado que presenta el menor error de predicción, en base a las dos validaciones presentadas, lo cual implica una mejor predicción del modelo fuera de muestra.

VI. Anexos

Enlace al repositorio de GitHub: https://github.com/SofiaQuiroga/Repositorio_Taller1_BDML.git

Enlace a la base de datos:
https://ignaciomsarmiento.github.io/GEIH2018s://ignaciomsarmiento.github.io/GEIH2018

Referencias bibliográficas

Blau, F.D Winkler, A.E. (2021) The Economics of Women, Men, and Work (Chapter 16, pp 425-455) *Oxford University Press* https://books.google.com.co/books?id=yfSszgEACAAJ

Datos Abiertos (2021) Gran Encuesta Integradora de Hogares (GEIH) https://www.datos.gov.co/Estadsticas-Nacionales/Gran-Encuesta-Integrada-de-Hogares-GEIH/mcpt-3dws

Editorial La República S.A.S. (2020, 5 febrero). *Los profesionales ganan 71% más que las personas que cuentan con básica primaria*. Diario La República. https://www.larepublica.co/economia/los-profesionales-ganan-71-mas-que-personas-con-personas-con-basica-primaria-2960985

ESWARAN, M. (2014). *Why Gender Matters in Economics*. Princeton University Press. https://doi.org/10.2307/j.ctvvh853j

Goldin, Claudia. 2014. "A Grand Gender Convergence: Its Last Chapter." *American Economic Review*, 104 (4): 1091-1119.DOI: 10.1257/aer.104.4.1091

STEM Women. 2022 "Women in STEM Statistics – A general outlook for female students" https://www.stemwomen.com/women-in-stem-percentages-of-women-in-stem-statistics#:~:text=Overall%2C%20the%20percentage%20of%20female,with%20women%20making%20up%2024%25.

WorkMeter. (2022). *Beneficios por antigüedad en la empresa*. https://www.workmeter.com/blog/beneficios-antiguedad-empresa/