Problem Set 1: Predicting Income

Estudiantes:
Alexandra Rizo - 202210094
Héctor Ticuán - 202225884
Carlos Vergara - 201414896
Danna Bolaños - 201911675

13 de febrero de 2023

1. Introducción

Por medio de este documento, se presentan los principales resultados del Taller No. 1 frente a los datos de la Gran Encuesta Integrada de Hogares, (en adelante "GEIH"), para 2018 y su relación con el salario ("*wage*") con otras variables como el género, la edad, entre otros. Todo lo anterior, en el marco de lo solicitado en el documento de instrucciones aplicable al Taller No. 1 "*Problem Set 1: Predicting Income*".

Para estos efectos, este documento se divide en los siguientes acápites, a saber: (i). Fuente de datos; (iii)Regresión - perfil Edad y Salario ("*Age-wage profile*"); (iii). La brecha salarial con base en el género ("*Gender earnings gap*") y; (iv). Predicción frente a ganancias ("*earning*").

Los principales resultados del estudio, se resumen a continuación:

2. Fuente de datos:

Los datos utilizados para nuestro análisis representan la encuesta realizada por el DANE "GEIH"disponibles en el repositorio https://ignaciomsarmiento.github.io/GEIH2018 sample/. La cual estaba dividida en diez (10) secciones o chunks con rangos de observaciones continuos. A continuación unos breves comentarios entornor a la GEIH, el proceso de adquisición de los datos y la limpieza de los datos respectivamente.

2.1. Descripción de las fuentes de datos:

La GEIH es una encuesta sobre las condiciones de empleo de las personas. Esta herramienta estadística recopila información sobre la población, los hogares y las condiciones socioeconómicas del país. Esta encuesta se lleva a cabo de forma periódica y con ella se toman decisiones de política pública, seguimiento y evaluación de los distintos programas sociales y económicos implementados.

La GEIH incluye preguntas sobre demografía, educación, empleo, ingresos, vivienda, salud y otros temas relacionados con la calidad de vida de la población.

Esta encuesta está diseñada y elaborada por el DANE - Departamento Administrativo Nacional de Estadística. El DANE se encarga de planear, evaluar e implementa la producción y comunicación de Estadísticas en el país. A través de esta encuesta, esta organización puede conocer aspectos tales como la tasa de ocupación, las actividades que desempeñan los colombianos y su remuneración y el comportamiento del mercado laboral de las mujeres, lo jóvenes y otros grupos poblacionales.

Para este trabajo puntual, utilizaremos datos de Bogotá para el año 2018.

2.2. Adquisición de los datos:

La base de datos esta dividida en 10 chunks. Antes de iniciar debemos preparar el espacio llamando los paquetes y las librerías, seguido de ello, hacemos el scrapping de los 10 chunks; al iniciar con este proceso se presentan barreras para su abquisición dado el número de observaciones, inicialmente notamos que los chunks están separados.

Posteriormente, en el proceso de inspeción para cada apartado notamos que el proceso de abstracción de los datos debe ser realizado teniendo en cuenta el origen de los datos, dado que al abrir cada chunk nos encontramos con una pagina en blanco mientras se realiza la carga completa de los datos, la cual sera la impresión que tome r sino se realiza de la manera adecuada, para esto se realiza el proceso de scraping posterior a una inspección exhaustiva de la pagina en donde se encuentra el repositorio para su carga correcta. procedemos a unificarlos: Iniciamos dándole un nombre a la URL haciendo web scrapping para recopilar la información de forma automática de la URL entregada por las instrucciones del taller. Posteriormente ubicamos los datos de los distintos chunk en un dataframe previamente creado.

2.3. Limpieza de datos:

Como restricción del problema, solo se analizan los datos de las personas con una edad mayor o igual a 18 años y que además se encuentren empleadas. Esto reduce la muestra de 32,177 a 22,640.

Dado que existe valores con NA y no se dispone de datos para imputar estos datos, se decide no tenerlos en cuenta dado que pueden afectar de manera significativa la capacidad predictiva de nuestros modelos debido su alta presencia con un $56,31\,\%$ aproximadamente, dentro de nuestra variable dependiente wage, con esto pasamos de 22,640 datos a 9,892 lo cual reduce la muestra a un $43,69\,\%$ aproximadamente de nuestra base inicial, lo cual es una muestra aceptable para la realización de nuestros modelos.

Dado que nuestra variable de respuesta son los sueldos horarios, se realiza un análisis descriptivo de la misma la cual sera entendida como los ingresos laborales asalariados nominales por hora de todos los occ, incluyendo propinas y comisiones (esta variable aparece en el dataframe como $y\_ingLab\_m\_ha$).

N	Min.	1st Qu.	Median	Mean	SD	3rd Qu.	Max.
9892	326.7	4226.5	5055,6	8822,2	12886,1	8049,5	350583,3

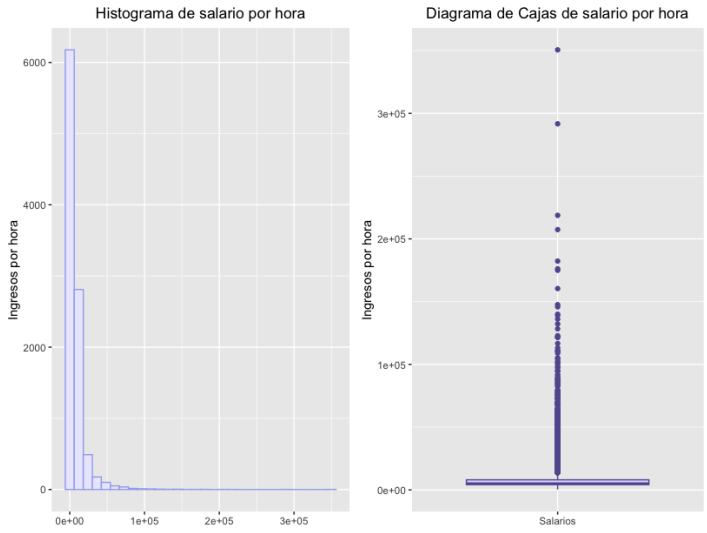

Figura 1: Histograma y diagrama de cajas para salario por hora

Se observa claramente como tiene una cantidad relevante de valores atípicos. Por lo cual se propone la siguiente alternativa:

Realizar un proceso de transformación de la variable por medio de logaritmo el cual propone una mayor estabilidad para nuestro estimador

1. Eliminar los valores atípicos que tengan un valor mayor a:

$$Valores\_atipicos = Observaciones > Q_3 + 1{,}5 * IQR \tag{1}$$

Donde Q_3 es el percentil del 75 % e IQR ese el rango intercuantil.

Al realizar esto obtenemos que 1.289 valores son atípicos y al eliminar estos valores se obtiene lo siguiente.

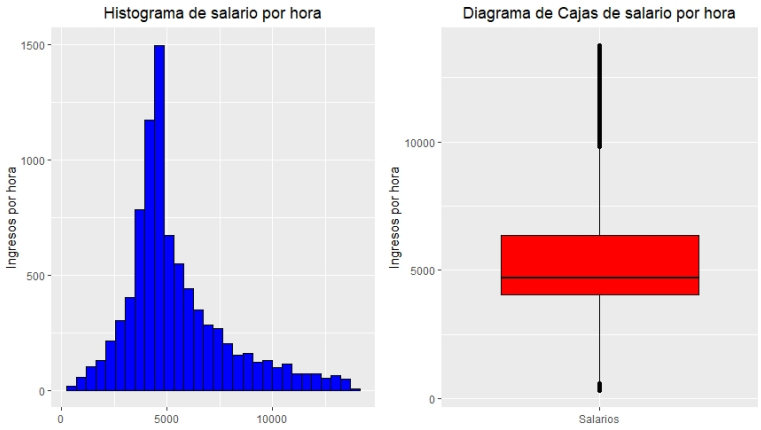

Figura 2: Histograma y diagrama de cajas para salario por hora sin atípicos

Posteriormente analizamos la transformación de la variable de respuesta por medio de logaritmo, como se observa en la figura 3.

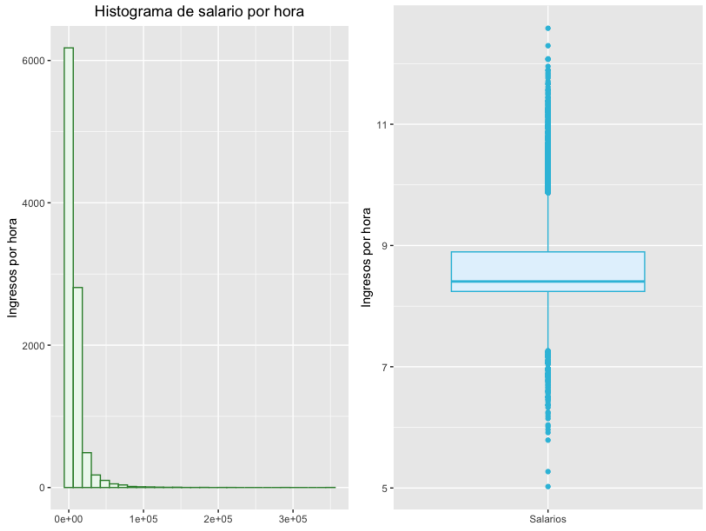

Figura 3: Histograma y diagrama de cajas para salario por hora sin atípicos

En la figura 3 se encuentra una variable de salario con menos datos atípicos para nuestro análisis

En nuestro caso tomaremos la variable $y\_ingLab\_m\_ha$ como nuestra variable dependiente w dado que esta incluye el ingreso nominal laboral de asalariados por hora y con las transformaciones previamente realizadas a la base nos permitirá un análisis efectivo de su comportamiento.

$$w = f(X) + u \tag{2}$$

Nuestro análisis indica que el comportamiento de las variables segun la edad es altamente relevante,

para nuestra data la edad donde se alcanza el salario maximo promedio son los 51 años aproximadamente, por otro lado este estadistico cambia al analizar los datos teniendo en cuenta se genero; dado que, en el cao de las mujeres esta edad se reduce a 47, pero, en el caso de los hombres se presente en 5 años menos ubicándose en la edad 42 años, lo que quiere decir que los hombres tienden a alcanzar el salario mas alto a una menor edad que las mujeres al rededor de su vida.

3. Regresión perfil Edad-Salario

Numerosos estudios de economía laboral sugieren que el perfil edad - salario del trabajador típico sigue una trayectoria predecible: "Los salarios tienden a ser bajos cuando el trabajador es joven aumentan a medida que el trabajador envejece, alcanzando un máximo en el torno a los 50 años; y la tasa salarial tiende a permanecer estable o a disminuir ligeramente después de los 50 años".

En esta sección vamos a estimar el perfil edad-salario de los individuos de esta muestra:

$$log(w) = \beta_1 + \beta_2 Edad + \beta_3 Edad^2 + u \tag{3}$$

En la figura 4 se observo que no existe correlación entre la variable salario y la edad. Por lo cual se procede a realizar un análisis descriptivo de esta.

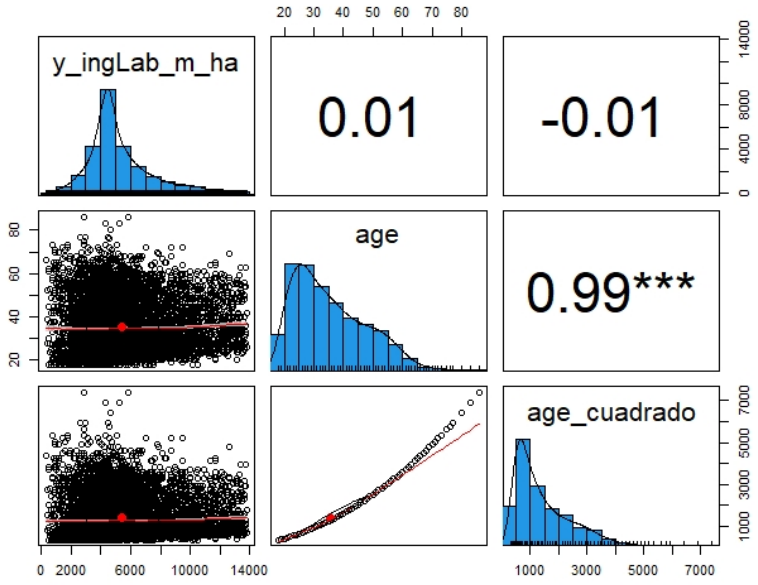

Figura 4: Correlación entre edad y salarios por hora

Si bien la edad es una variable entera para esta muestra no necesariamente es entera, por lo cual se tratara como una variable continua, dado a lo especificado en la parte de limpieza de datos la edad mínima es de 18 años y el promedio de 35.55 años, lo cual es razonable dado que la mayoría

N	Min.	1st Qu.	Median	Mean	SD	3rd Qu.	Max.
8603	18.00	26	33	35.55	12.1	44	86

Cuadro 1: Estadísticos de la edad

de las personas en edad productiva son menores a 50 años con base en la hipótesis. Tenemos una muestra de edad sesgada hacia la izquierda como se observa en la figura 5 y con valores atípicos menores a los de salaria, por lo cual no se realiza ninguna limpieza, ni transformación en este momento.

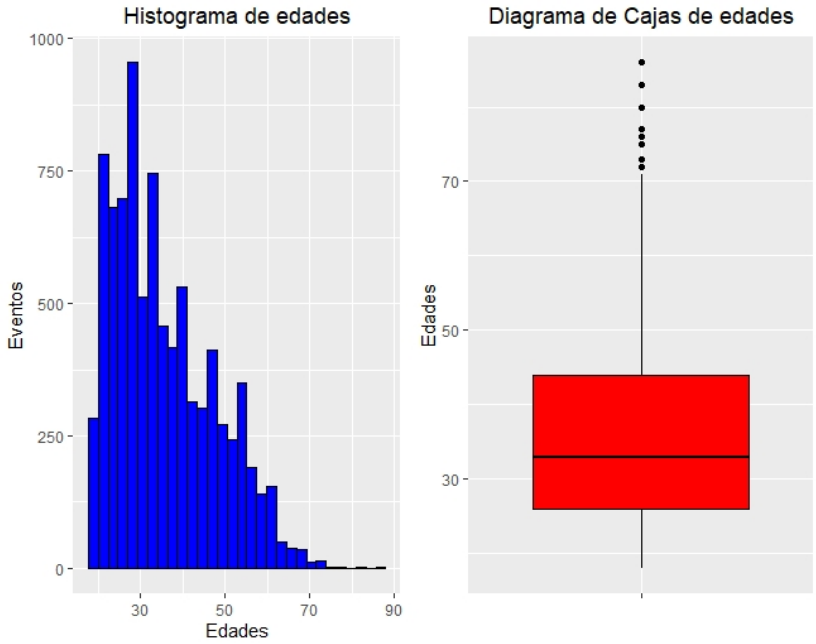

Figura 5: Histograma y diagrama de cajas para edad

Con los datos de la muestra establecidos, se procede a regresar la regresión lineal por medio de la herramienta RSTUDIO y la función $lm()$, como se menciono en la ecuación 2, la edad y la edad al cuadrado serán las únicas variables predictoras utilizadas en este modelo. Para este primer modelo, se utilizan dos base de información, la una con la data completa y la otra eliminando los datos atipicos, como se explico en la sección anteriores, esto tiene como finalidad revisar el impacto de estos a nuestro modelo.

Cuadro 2: Tabla de regresión de edad con y sin datos atipicos

	Variable de respuesta:	
	Salario por hora	
	sin atipicos	con atipicos
Edad	179.766***	646.432***
	(12.859)	(64.616)
Edad cuadrado	−2.233***	−6.230***
	(0.160)	(0.799)
Constante	2,196.908***	−5,520.979***
	(238.952)	(1,214.216)
Observations	8,603	9,892
R^2	0.022	0.025
Adjusted R^2	0.022	0.025
Residual Std. Error	2,383.871 (df = 8600)	12,722.090 (df = 9889)
F Statistic	98.173*** (df = 2; 8600)	129.382*** (df = 2; 9889)
Note:		*p<0.1; **p<0.05; ***p<0.01

Como se mencionó al inicio de esta sección, los salarios tienden a subir con la edad y llega a un punto máximo, antes de decrecer. Esto se observa por medio de los coeficientes de la edad y la edad al cuadrado. Los primeros años dado que el coeficiente de edad es más grande que el coeficientes de la edad al cuadrado el salario aumenta, pero dado que la variable edad al cuadrado esta al cuadrado llegara un punto en el cual empieza a disminuir, este punto es nuestro edad pico y tiene el mejor salario por hora.

Respecto a la significación se encuentra que ambas variables son significativas para el modelo, pero se tiene un valor de R^2 bastante bajo. Destaca que el valor de R^2 es mayor con los valores atípicos, esto nos dice que posiblemente si bien existe valores atípicos, estos no son datos incorrectos y que ayudan al modelo, para un mejor análisis al respecto, en el punto 5 se verifican los modelos.

En las figuras 7 y 3 observamos los gráficos de perfil estimado.

Nota: Para una mejor visualización se acoto el eje y a 25 000, dado que existe un valor atípico que distorsiona la gráfica.

Como se hablo en el inició del documento, los salarios tienden a aumentar con los años hasta llegar a un punto máximo, luego decae, según la teoría a partir de los 50 años, en nuestro modelo sin datos atípicos es al rededor de los 40 años y con valores atípicos en 45 años, es importante mencionar que usando únicamente una variable los sesgos son altos, más aun cuando sabemos que se dispone de muchas más variables para mejorar el modelo.

Adicionalmente usamos bootstrap, dado que puede ser una técnica útil para tratar con outliers y graficamos con los respectivos intervalos de confianza.

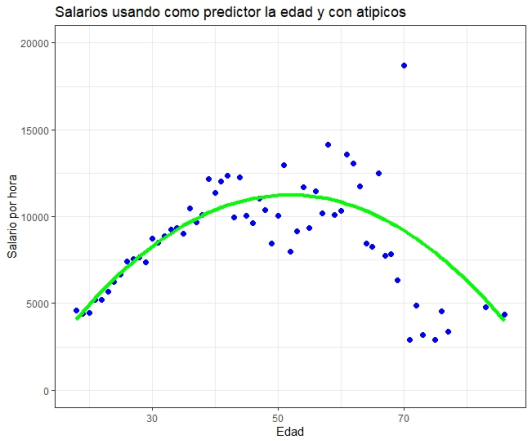

Figura 6: Con atipicos

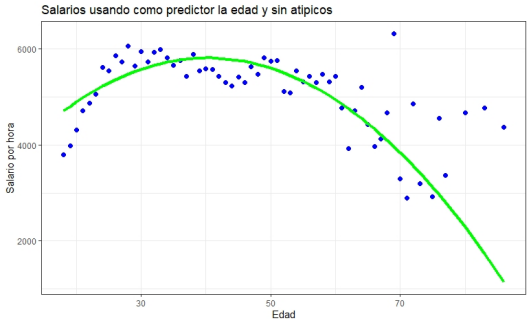

Figura 7: Sin atipicos

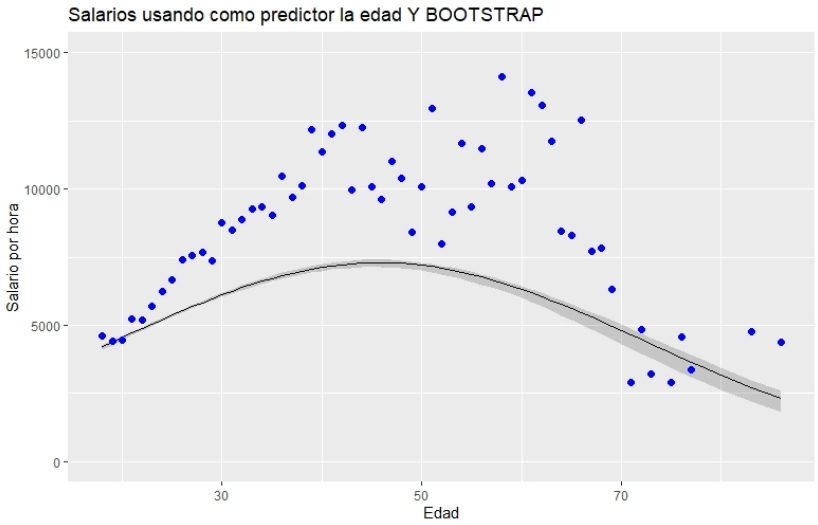

Figura 8: Perfil de predicciones usando BOOSTRAP

Finalmente para el modelo usando BOOSTRAP, se calcula la edad pico con el salario más alto y sus respectivos intervalos de confianza. La edad fue 45 años con un salario de 7.275 COP e intervalos de confianza de [7.125, 7.428] y un nivel de , respecto a los datos sin BOOSTRAP

4. Regresión frente a la brecha salariar entre géneros

En el marco del ejercicio, se planteo que uno de los principales aspectos a ser analizados por los encargados de formular e implementar proyectos, programas o políticas en materia salarial, se encuentra lo relativo a la brecha salarial entre géneros. En efecto, y respecto al caso colombiano, se han adelantado diversos estudios en relación con este punto, logrando concluir inclusive, que la existencia de brechas entre los géneros hace parte de una serie de violencias hacia el género femenino desde una aproximación histórica, económica e inclusive social[1]
En este caso, se presentará una aproximación a la brecha salarial entre géneros, a partir de los datos de la GEIH de 2018 con fundamento en el siguiente modelo:

$$log(w) = \beta_1 + \beta_2 Femenino(Female) + u \tag{4}$$

Para iniciar, realizamos un análisis descriptivo de la variable sexo según los datos de la GEIH para 2018:

N	Hombre	Mujer
9891	4919	4972

Cuadro 3: Estadísticos del género (sexo)

De la Tabla No. 3 se puede apreciar que, entre 9891 observaciones, 4919 corresponden a personas que se identificaron como hombre y 4972 a personas que se identificaron como mujer. Al respecto, vale la pena aclarar que, según los documentos técnicos de la GEIH no se contempló otra posibilidad para la persona encuestada de definir su identidad sexual como hombre o mujer.

Con este marco, se procede a realizar la regresión incondicionada y condicionada entre las variables sexo y la logarítmica del ingreso laboral. En el marco de la condicionada, donde se aplicó el teorma FWL y FLW con Bootstap se obtuvieron los siguientes resultados, según consta en la siguiente cuadro donde se evalúa el poder predictivo de nuestras variables de tal forma que se plantea una comparativa entre los distintos modelos previamente creados junto con sus coeficientes, el grado de significancia de sus variables, su respectivo $R\hat{2}$, $R\hat{2}$ ajustado, el error residual de la desviación estándar; los cuales representan estadísticos descriptivos de las capacidades predicativas de nuestros modelos. Espacio para Tabla 4.

De la tabla anterior podemos inferir que las variables age, dummy sex (0 = female, 1 = Male), maxEducLevel (maximo nivel de educación alcanzado), oficio(tipo de oficio que desempeña dentro de la empresa), estrato1(estrato de energía) y age_cuadrado, presentan una alta significancia dentro del modelo no condicionado como dentro de los modelos condicionados con un p value 0 01.

Por otro lado el $R\hat{2}$ de nuestros modelos nos indica que se explican en una mayor cantidad aquellos con mayor cantidad de predictores como los osn el modelo 2 y 3 los cuales explican en un 43 % aproximadamente el modelo, mientras que el modelo mas simple con una variable independiente (sex), solo lo explica en un 0,09 %.

Durante el proceso de evaluación de variables para incluir en los modelos condicionados se tienen en cuenta aquellas que nos puedan llegar a mostrar un entorno de control dentro de las condiciones laborales, lo que quiere decir que se incluyen dado que presentan características especificas sobre las condiciones y capacidades laborales, como lo son el tipo de oficio (oficio), la edad (age), el estrato (estrato1) y el máximo nivel de educación alcanzado (maxEducLevel) y la variable previamente analizada en el modelo 1 que representa los generos (sex).

[1] Ver: Oscar Hernán Cerquera, et al. *La brecha salarial por género en Colombia y en el Departamento de Caldas* Rev. Ánfora, Vol. 27, No. 48, (2020), pág. 113-136 y; Adriana Sabogal. *Brecha salarial entre hombres y mujeres u ciclo económico en Colombia.* Revista coyuntura Económica: Investigación económica y social, vo. 31, No. 1, (2012), pág. 53 y 91, entre otros

Cuadro 4: Tabla de regresión entre sexo y salario

	Dependent variable:		
	log(y_ingLab_m_ha)		
	(1)	(2)	(3)
age		0.0580***	0.0565***
		(0.0028)	(0.0027)
sex	0.0446***	0.2072***	0.2111***
	(0.0146)	(0.0114)	(0.0113)
maxEducLevel		0.1706***	0.1804***
		(0.0062)	(0.0062)
oficio		−0.0067***	−0.0066***
		(0.0002)	(0.0002)
estrato1		0.2362***	0.2265***
		(0.0063)	(0.0063)
age_cuadrado		−0.0006***	−0.0006***
		(0.00003)	(0.00003)
Constant	8.7022***	6.0562***	6.0407***
	(0.0104)	(0.0696)	(0.0679)
Observations	9,892	9,891	9,891
R^2	0.0009	0.4278	0.4250
Adjusted R^2	0.0008	0.4274	0.4246
Residual Std. Error	0.7273 (df = 9890)	0.5506 (df = 9884)	0.5465 (df = 9884)
F Statistic	9.3166*** (df = 1; 989)	1,231.5750*** (df = 6; 988)	1,217.5860*** (df = 6; 988)
Note:			*p<0.1; **p<0.05; ***p<0.01

Adicionalmente, se construye el gráfico relacionado con las edades pico en relación al salario y el género, a saber:

5. Predicción de salario

En esta subsección se evalúa los modelos creados anteriormente. Inicialmente los tenemos 5 modelos.

- Modelo solo utilizando como variable predictora edad (edad y edad al cuadrado).
- Modelo solo utilizando como variable predictora el sexo.
- Modelo usando la edad, sexo y el nivel máximo de estudio.
- Modelo completo, usando las variables edad, sexo, el grado de escolaridad, las horas trabajadas en la semana, el tamaño de la firma donde labora y el máximo nivel de educación aprobado

Para hacer el análisis de los modelos, dividimos la muestra en train con el 70 % de la muestra y el restante para test, esto nos ayuda a que los modelos no se sobrejusten a la muestra y poder ser evaluados con la muestra de test. Nota: Para que sea reutilizable se agrega la semilla

Cuadro 5: Tabla de regresión entre sexo y salario - Condicionada y no Condicionada

	Dependent variable:		
	log(y_ingLab_m_ha)		
	(1 - Incondicionado)	(2 - Condicionado FWL)	(3 -ondicionado FWL Bootstra
Género	0.0447355***	0.1203609***	0.1376570***
	(0.0146269)	(0.0125933)	(0.0125680)
Edad		0.0617382***	0.0606845***
		(0.0031687)	(0.0032494)
Edad al cuadrado		−0.0005532***	−0.0005285***
		(0.0000393)	(0.0000405)
Max. Nivel educativo		0.3284640***	0.3416336***
		(0.0060116)	(0.0061131)
Constante	8.7021840***	5.2303250***	5.1375930***
	(0.0103705)	(0.0713727)	(0.0719087)
Observaciones	9,891	9,891	9,891
R^2	0.0009450	0.2671003	0.2789951
Adjusted R^2	0.0008440	0.2668038	0.2787033
Residual Std. Error	0.7273377 (df = 9889)	0.6230595 (df = 9886)	0.6219855 (df = 9886)
F Statistic	9.3540480*** (df = 1; 9889)	900.7214000*** (df = 4; 9886)	956.3544000*** (df = 4; 9886
Note:			*p<0.1; **p<0.05; ***p<0.

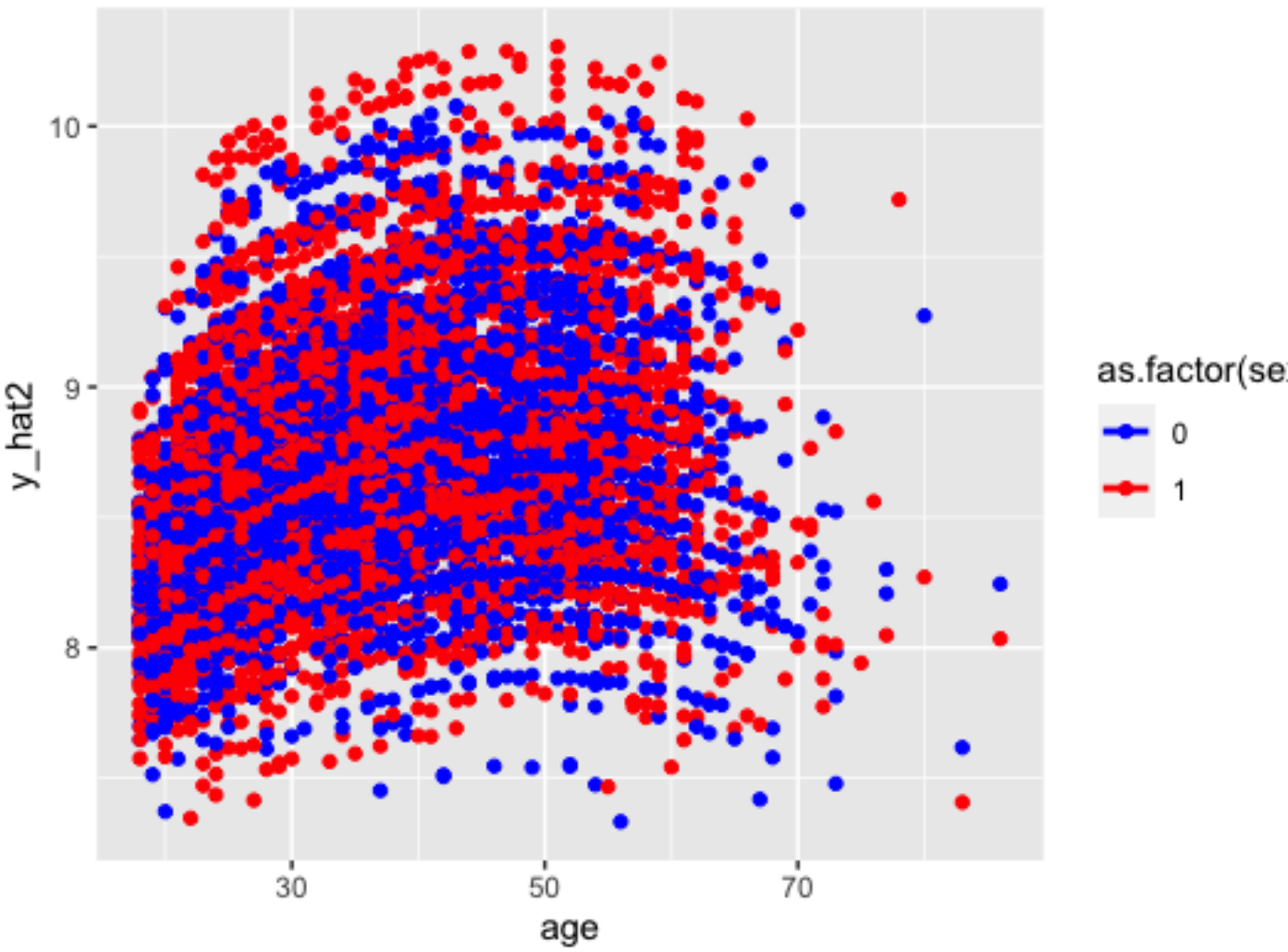

Figura 9: Gráfico de puntos frente a la brecha salaria y el género

set.seed(1121).

Las métricas utilizadas para el análisis de estos modelos fueron:

- MSE: Por ser la metrica más utilizada en Machine Learning.
- R2 Score: Calcule la puntuación de regresión R-cuadrada.

Se obtuvo como resultado:

Modelo	Muestra	R2_Score	RMSE
2*Modelo completo	TRAIN	-182506487	9860.792
	TEST	-179002437	9660.120
2*Modelo con edad	TRAIN	-74709888	6309.017
	TEST	-76043357	6296.275
2*Modelo con sexo	TRAIN	-71587508	6175.772
	TEST	-73174900	6176.382
2*Modelo completo	TRAIN	-100886747	7331.447
	TEST	-100940129	7254.124

Cuadro 6: Análisis de modelos

Los dos modelos para aplicar LOOCV son los modelos de edad y sexo, dado que tiene el menor RMSE. Al realizar este modelo se obtiene que el modelo solo con edad tiene 0.5063916 y el modelo con sexo tiene 0.5291272. Al ver que se tiene valores altos se decide utilizar la metodología LOOCV para el modelo completo para lo cual tenemos un resultado de 0.3524122 por lo cual el modelo completo tiene mejor rendimiento a modelos desconocidos.