Grupo 6.

Problem Set 1.

Natalia Castro Alarcon.

Victor Dulio Chique. **200319157**

Victor Ivan Sanchez. **201510287**

---

1. INTRODUCCCIÓN

En Colombia, de acuerdo con la Dirección de Impuestos y Aduanas Nacionales (DIAN) la evasión de impuestos es cercana a los $65 billones de pesos, estimaciones indican que la evasión esta cercana a 5,4 puntos del PIB, de los cuales 0,7 puntos corresponden evasión de impuestos de personas naturales. [1] Algunos trabajos han estimado el subreporte de ingresos en 1.66 veces mas que el reporte inicial. (Rocha (2014)). Por lo anterior, usando metodologías de aprendizaje estadístico, aun incipientes en economía, estimamos un modelo de predicción de salarios basado en información de la GEIH para el año 2018 en Bogotá. Este modelo predictivo es potencialmente útil para encontrar casos de fraude fiscal y adicionalemente, apoyar la toma de decisiones de política para intervenir familias vulnerables. Principalmente estimamos un modelo que explica el salario por la edad y otro que busca mirar brechas salariales entre hombres y mujeres.

Encontramos que existe una relación no lineal (cuadrática) entre el salario y la edad, y por tanto la edad tienen rendimientos marginales decrecientes sobre el salario, es decir a mayor edad mayor salario no obstante el incremento es cada vez menor hasta alcanzar un máximo. Un año mas de edad aumenta en promedio un 1.23%, sin embargo, dado que la relación es cuadrática, según la edad dicho incremento es diferente, por ejemplo un joven de 18 años, al incrementar un año su edad su salario incrementará un 3.6% por el contrario para un individuo de 60 años, incrementar en un año su edad disminuye su salario en 1.9%. Respecto a la brecha salarial encontramos diferencias significativas entre hombres y mujeres ($p\text{-value}<0.01$). Ser mujer disminuye el salario en un 4.4% cuando se realiza una regresión simple y 9.1% con variables de control, así mismo, las mujeres alcanzan su "*peak age"* mas rapido que los hombres, por tanto, el mercado laboral esta castigando con mayor severidad la edad en las mujeres que en los hombres.

2. DATOS

Usamos datos de la Gran Encuesta Integrada de Hogares GEIH en Bogotá para el año 2018, la cual contiene información sobre las condiciones de empleo de las personas, características generales de la población como sexo, edad, estado civil y nivel educativo. Contiene también información sobre fuentes de ingreso.

---

[1] Ver: https://www.larepublica.co/especiales/reforma-tributaria-2022/segun-la-dian-y-el-minhacienda-la-evasion-de-impuestos-es-cercana-a-80-billones-3422523

El proceso de adquisición de la información se realizó mediante el *scraping* de la información contenida en el sitio web https://ignaciomsarmiento.github.io/GEIH2018_sample/, sobre esta información no existe ningún tipo de restricción. El sitio web contiene información extraída y clasificada de la GEIH con variables originales y construidas que proporcionan datos relevantes para el análisis. Se realizo un web *scrapping* de los 10 bloques de información, los cuales estaban particionados en diferentes URL´s. Estos bloques de información se consolidaron para terminar con una sola base de datos integrada de la información relevante.

a. Limpieza de datos

La GEIH en lo concerniente a los procedimientos de conformación, depuración, imputación y empalme del ingreso contiene 137 variables con las características de los individuos y aproximadamente 40 variables que contienen información relacionada a ingreso, fue necesario realizar un procedimiento de limpieza de la base en el cual se restringe el universo de datos en aquellos relevantes para los objetivos del trabajo, por tanto centramos el análisis en las siguientes variables: *i) Sexo*; Variable dicótoma que toma valor 1 si es hombre y 0 si es mujer. Usamos esta variable con la finalidad de identificar brechas salariales entre hombres y mujeres. *ii) Edad;* contiene la edad en años del individuo, la muestra se analiza para aquellos mayores de 18 años. Esta variable se usa para identificar patrones predictivos entre la edad y el salario de las personas *iii) Educación*; contiene e máximo grado escolar alcanzado iv) *Tamaño empresa;* que muestra el numero de empleados de la empresa en que el individuo labora v) *Salario;* la variable de salario usada es el salario mensual por hora vi) *Relab;* que es el oficio del individuo y nos puede ayudar a entender diferencias salariales entre las ocupaciones. La variable de salario fue construida mediante la sumatoria del ingreso recibido en el ultimo mes en el empleo que tenia el individuo y los ingresos recibidos por horas extras, consideramos que estos dos conceptos definen de forma mas limpia el salario, pues en primer lugar se originan por la relación laboral y remuneran en particular el trabajo y las horas trabajadas, por tanto podrán arrojarnos estimaciones mas precisas en nuestros modelos.

b. Estadísticas descriptivas

Las estadísticas descriptivas de la Tabla 1 nos dan luces de lo que va a ser nuestro análisis posterior. Respecto al sexo, vemos que la muestra esta balanceada en la proporción de hombres y mujeres encuestados, la edad promedio esta en aproximadamente 36 años con individuos de hasta 86, recordemos que limitamos el estudio a los mayores de 18 años, que es la mayoría de edad en Colombia. En cuanto a la educación, la población Bogotana en promedio tiene un alto nivel educativo, pues los individuos promedian 6.08 de Educación en una escala que tiene su máximo en 7, por tanto la mayoría de la población encuestada tiene niveles de educación altos.

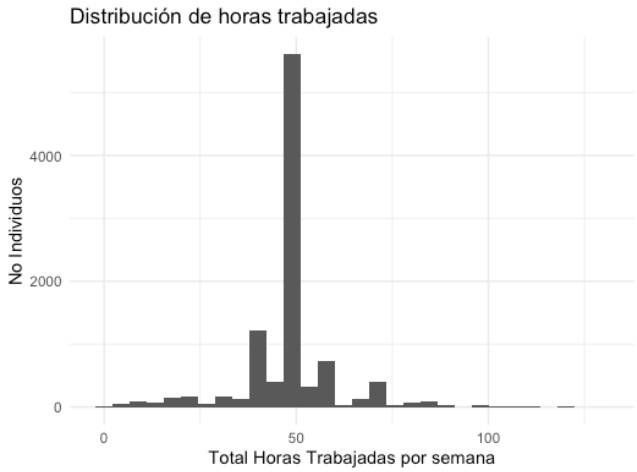
Tabla 1.

Variables incluidas en la base de Datos

Statistic	N	Mean	St. Dev.	Min	Max
sex	9,967	0.506	0.500	0	1
edad	9,967	36.078	11.942	18	86
maxEducLevel	9,967	6.084	1.112	1	7
hoursWorkUsual	9,967	48.027	12.336	1	130
salario_mensual	9,967	1,524,879.000	2,040,110.000	10,000	30,200,000
salario_mensual_hora	9,967	8,314.517	11,832.650	208.333	312,500.000
log_salario_mensual_hora	9,967	8.681	0.705	5.339	12.652

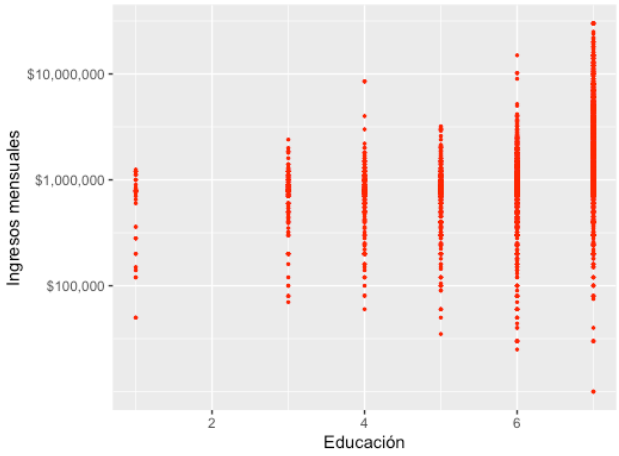
Respecto a las horas promedio trabajadas por semana coinciden con la jornada laboral legal, ver Grafico 1. Con algunos individuos (outliers) que trabajan mucho mas (130 horas) y otros mucho menos (1 hora).

Grafico 1.



Respecto a el salario mensual promedio de los individuos, este se ubico en $ 1.524.879, esto quiere decir que el salario promedio de los bogotanos para el año 2018 fue aproximadamente 1.75 veces mayor al salario mínimo mensual vigente para ese año, el cual se ubicaba en $869.453 incluyendo auxilio de transporte. No obstante, existe una variabilidad importante en los datos, pues la desviación estándar es 1.33 veces mayor que el salario promedio del grupo, esta variabilidad de los ingresos es bien ilustrada por el grafico 2.

Grafico 2. Relación de Ingresos mensuales y Educación



Nuestro principal objetivo es encontrar un modelo predictivo con el potencial de darnos valores predichos de los salarios de forma precisa y de esta manera usar el modelo para diferentes fines importantes, como encontrar casos de fraude o subreporte de ingresos, o a su vez, determinar aquellos individuos vulnerables donde se podría eventualmente focalizar la asistencia. Dentro de nuestras variables, características como la edad, el sexo y la educación han sido largamente estudiadas en economía laboral como características claves que explican el salario de los individuos y que además tienen patrones predictibles, como es el caso de la edad.

3. AGE-WAGE PROFILE

El salario es el pago regular de dinero de acuerdo a las horas trabajadas pactado mediante un contrato. La evidencia empírica sobre los efectos de la edad en el salario encuentra que, a medida que aumenta la edad, el salario también crece hasta cierta edad, a partir del cual comienza su descenso, cuya especificación corresponde al siguiente modelo:

$$log(w) = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u \qquad (1)$$

Donde: $log(w)$ es el logaritmo natural de salario por hora y Age es la edad.

En esta sección se estima esta especificación del salario. La variable dependiente corresponde al logaritmo del salario por hora, dicha transformación busca reducir el rango de la variable a una unidad más pequeña que el salario en niveles, y facilita su interpretación. La variable predictora es la edad. Los datos son obtenidos de la Gran Encuesta Integrada de Hogares (GEIH) correspondiente a Bogotá.

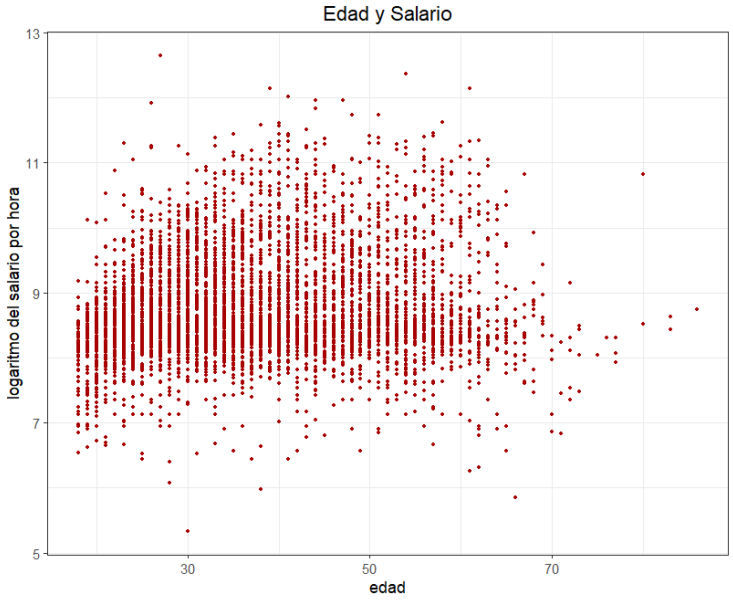
Tabla 2.

Estadisticas Descriptivas

Statistic	N	Min	Pctl(25)	Median	Mean	Pctl(75)	St. Dev.	Max
Salario	9,967	208.3	4,069.0	4,846.9	8,314.5	7,552.1	11,832.6	312,500.0
Edad	9,967	18	26	34	36.1	44	11.9	86

El salario promedio en la ciudad de Bogota asciende a $ 8,314.1 por hora, cuya variabilidad es elevada, revelada por la desviación estándar de $ 11,832.1 y los cuartiles 1 y 3, con un rango de $ 3,483.115. La edad promedio de las personas mayores a 18 años, que trabajan está en 36 años. En el Grafico 3 se observa que la amplitud de variación del salario para cada edad es importante, que está asociada con las características individuales de cada persona. También se puede notar la presencia de *outliers*.

Grafico 3. Relación entre logaritmo del salario por hora y la edad



a. *Regresión*

Tabla 3. Regresión modelo (1)

```
Resultados de la Regresion
================================================
                        Dependent variable:
                    ----------------------------
                        Logaritmo del Salario
------------------------------------------------
edad                          0.0595914***
                               (0.0035464)

edad al cuadrado              -0.0006548***
                               (0.0000441)

Constant                      7.4769350***
                               (0.0663397)

------------------------------------------------
Observations                      9,967
R2                               0.0379413
Adjusted R2                      0.0377482
Residual Std. Error      0.6920231 (df = 9964)
F Statistic          196.4784000*** (df = 2; 9964)
================================================
Note:                  *p<0.1; **p<0.05; ***p<0.01
```

b. Significancia e interpretación

Significancia: El modelo estimado muestra que la variable edad es estadísticamente significativa a un nivel de significancia del 1% (p-value<0.01) tanto la parte lineal como no lineal (cuadrático). Es decir, la estimación confirma una relación no lineal entre la edad y el salario. También, el estadístico F muestra que la especificación no lineal de la relación salario-edad es estadísticamente significativa de manera conjunta al 1% de nivel de significancia (p-value<0.01).

Interpretación: Al mostrar el modelo una relación no lineal entre la edad y el salario, para ver el efecto que tiene la edad en el salario es necesario derivar el logaritmo del salario respeto a la edad, lo cual resulta en una semi elasticidad.

$$\frac{\partial \log(w)}{\partial edad} = \beta_2 + 2\beta_3 Age$$

El cálculo de la semi elasticidad requiere utilizar el promedio de la edad de los bogotanos, que es $\overline{Age} = 36.1$ años; también puede ser evaluada para diferentes edades.

$$\frac{\partial \log(w)}{\partial edad} = 0.0595914 + 2(-0.0006548)(36.1)$$

$$\frac{\partial \log(w)}{\partial edad} = 0.0123148$$

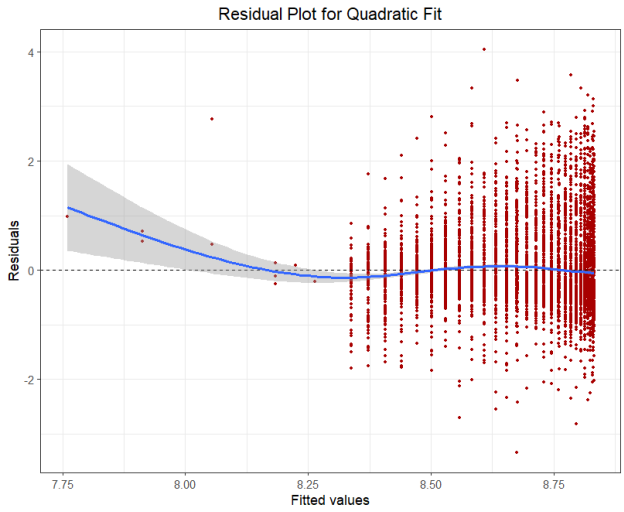
Esta semi elasticidad nos indica que una persona que tiene una edad promedio de 36.1 años, ante el incremento de un año más en su edad, su salario aumenta 1.23%. Pero esta semi elasticidad cambia según la edad. Es decir siendo joven, cuanta más edad tenga, su salario aumentará, pero cada menos hasta llegar a cierta edad (rendimientos marginales decrecientes) y comenzará a disminuir conforme se haga más adulto. Por ejemplo, para alguien que tiene 18 años, tener un año más implica un incremento en su salario de 3.6% y para otro que tiene una edad de 60 años, tener un año más se traduce en una disminución del 1.9% de su salario.

c. Ajuste del modelo

El ajuste del modelo medido a partir del R-cuadrado nos dice cuánto de la varianza del logaritmo del salario es explicado por el modelo. En este caso un R-cuadrado de 0.0379413 significa que solo el 3.79% de la varianza del salario es explicado por el modelo. Este es bastante bajo y sugiere que el modelo no tiene buen ajuste. Sin embargo, esto no necesariamente es malo, pues los predictores son estadísticamente significativos y los coeficientes del modelo aun representan el cambio en el salario frente a cambios en la edad. Por lo tanto, se justifica en parte el ajuste del modelo.

Otra manera de ver es a partir del gráfico de los errores del modelo. Este grafico muestra que los errores no se alejan sistemáticamente de cero y no tienen un patrón (línea azul). En consecuencia, sugiere que la edad al cuadrado mejora el ajuste a los datos.

Grafico 3.

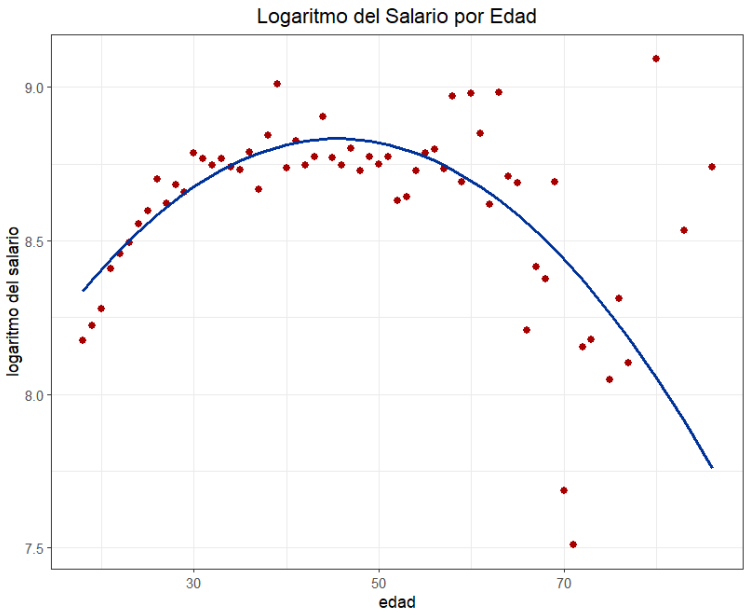


d. Edad-salario estimado, "peak age" e intervalos de confianza

La relación entre la edad y el salario es no lineal como se ve en el gráfico, en una primera etapa el salario crece conforme aumenta la edad del trabajador, hasta llegar a los 46 años (45.50549), edad a partir del cual entra a una segunda etapa, cuando su salario comienza a descender.

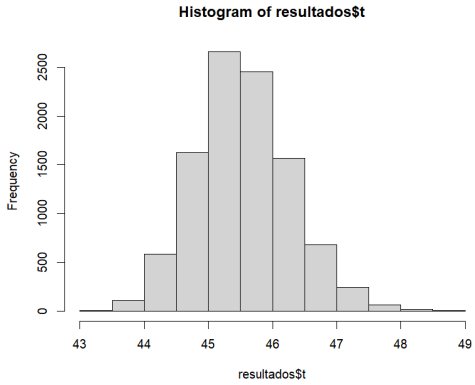
El dato 45.50549 de edad (*peak age*) se ha calculado usando Bootstrap para obtener 10000 estimaciones de la edad, basado en 10000 muestras con reemplazo sobre la muestra original. El Bootstrap genera una distribución cercana a la normal de la edad, tal como se observa en el histograma y el gráfico Cuantil-Cuantil.

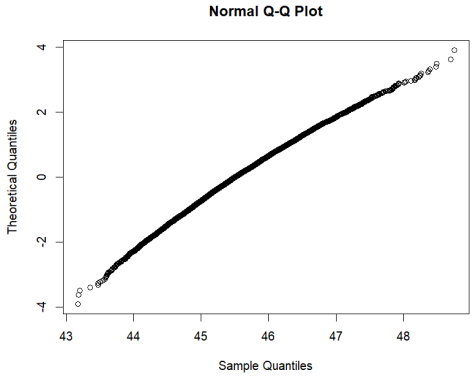
Grafico 4.



Por último, el intervalo de confianza de 95% para la edad es IC:[44.02, 46.91], en otras palabras, la edad a partir del cual el salario comienza a caer estará entre 44 y 47 años para los bogotanos.

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = base_nueva, statistic = eta_reglsalario_mes_hora_f
n,
    R = 10000)


Bootstrap Statistics :
    original     bias    std. error
t1* 45.50549 0.03994752   0.7383577



BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = resultados, type = c("norm", "basic"))

Intervals :
Level      Normal              Basic
95%   (44.02, 46.91 )   (43.90, 46.82 )
Calculations and Intervals on Original Scale
```





4. THE GENDER EARNINGS GAP

4.1 The Gender Earning Gap: Con el fin de estudiar la posible brecha salarial entre hombres y mujeres se realizaron las siguientes estimaciones:

4.1.1. Brecha por género con modelo lineal simple: Se utilizó una regresión lineal para estudiar si la variable dicotómica *Female* (Female=1 si es mujer) explica diferencias en el salario que gana una persona por hora. El modelo utilizado fue el siguiente:

$$\log(w) = \beta_1 + \beta_1 Female + u$$

Aunque el predictor explica solamente el 0.1% de la variación del logaritmo del salario por hora, la variable *Female* es significativa. Ser mujer disminuye el logaritmo del salario en 4.4%. El intervalo de confianza es (-0.072, -0.0163) por lo cual el descuento por ser mujer es significativamente diferente de cero. El resultado confirma que sí existe una brecha salarial por género.

Tabla 4.

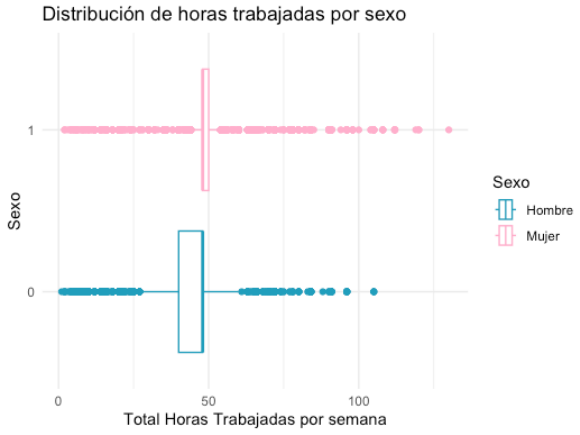
Regresión Log Salario - Female

	Dependent variable:
	Logaritmo del Salario
female	-0.044***
	(0.014)
Constant	11.469***
	(0.010)
Observations	10,046
R2	0.001
Adjusted R2	0.001
Residual Std. Error	0.709 (df = 10044)
F Statistic	9.689*** (df = 1; 10044)
Note:	*p<0.1; **p<0.05; ***p<0.01

	2.5 %	97.5 %
(Intercept)	11.44958060	11.48858050
female	-0.07179884	-0.01631226

Adicionalmente, el grafico 5 muestra las horas trabajadas por sexo. Se observa que las mujeres reportan mayor cantidad de horas respecto a los hombres. Esto complementa la evidencia de la existencia de una brecha salarial y podría poner de manifiesto condiciones desfavorables en el sentido de que las mujeres en promedio dedican más horas de trabajo y en promedio menos salario.

Grafico 5.



4.1.2. Equal pay for equal work: Existen diferentes variables que influyen en el salario que recibe una persona y aunque cada empleador ofrece un salario diferente, este no debería variar por el hecho de ser mujer u hombre. Las personas que realizan el mismo tipo de trabajo deberían recibir el mismo salario. Para estudiar si la brecha que se encontró en el apartado anterior se mantiene cuando se controla por tipo de trabajo y características similares del empleado se utilizó un modelo lineal en el que además de la variable *Female*, se incluyeron: *relab*=tipo o relación de trabajo, *maxEduLevel*=máximo nivel educativo, *tam_empresa*=tamaño de la empresa, *edad* y *edad al cuadrado*.

$$\log(w) = \beta_1 + \beta_2 Female + +\beta_3 relab + \beta_4 maxEdulevel + \beta_5 Edad + \beta_6 Edad\_sqr + \beta_7 tam\_empresa + u$$

Los siguientes son los resultados de la regresión utilizando MCO.

Tabla 5.

Regresión Log Salario - X2

	Dependent variable:
	Logaritmo del Salario
female	-0.091***
	(0.012)
relab	0.121***
	(0.012)
maxEducLevel	0.244***
	(0.006)
edad	0.046***
	(0.003)
edad_sqr	-0.0004***
	(0.00004)
tam_empresa	0.072***
	(0.002)
Constant	8.317***
	(0.066)
Observations	10,046
R2	0.315
Adjusted R2	0.315
Residual Std. Error	0.587 (df = 10039)
F Statistic	769.496*** (df = 6; 10039)
Note:	*p<0.1; **p<0.05; ***p<0.01

Este modelo explica el 31.5% de la variación del logaritmo del salario. En este modelo *Female* continúa siendo significativa y afecta negativamente al logaritmo salario en un porcentaje mayor: 9.1%. Es decir que este modelo predice una brecha por genero aún más grande.

La regresión también se llevó a cabo siguiendo los pasos del teorema Frish-Waugh-Lovell (FWL). En este caso se utilizará *Female* como X_1 y *relab, maxLevelEduc, edad, edad_sqr y tam_empresa* como X_2. El teorema especifica tres pasos:

- Paso 1: Se realiza la regresión de la variable dependiente, en este caso log(w), respecto a X_2 y se obtienen los residuales. De esta manera se obtiene la parte de *log(w)* que no explican las variables X_2.

$$\log(w) = \beta_1 + \beta_2 relab + \beta_3 maxEdulevel + \beta_4 Edad + \beta_5 Edad\_sqr + \beta_6 tam\_empresa + u$$

- Paso 2: Se realiza la regresión de X_1 respecto a X_2 y se obtienen los residuales. De esta manera se obtiene la parte de X_1 que no explican las variables X_2.

$$Female = \beta_1 + \beta_2 relab + \beta_3 maxEdulevel + \beta_4 Edad + \beta_5 Edad\_sqr + \beta_6 tam\_empresa + u$$

- Paso 3: Se realiza la regresión de los residuales del paso 1 respecto a los residuales del paso 2. De esta manera se obtiene la parte del logaritmo del salario que sólo explica X_1. El Teorema predice que el coeficiente de MCO de *Female* debe ser igual al coeficiente del paso 3. A continuación se muestran los resultados obtenidos por MCO y por FWL:

Tabla 6.

	Dependent variable:	
	resid_reg_3	log_salario_mensual_hora
	(1)	(2)
resid_reg_4	-0.091***	
	(0.012)	
female		-0.091***
		(0.012)
relab		0.121***
		(0.012)
maxEducLevel		0.244***
		(0.006)
edad		0.046***
		(0.003)
edad_sqr		-0.0004***
		(0.00004)
tam_empresa		0.072***
		(0.002)
Constant	-0.000	8.317***
	(0.006)	(0.066)
Observations	10,046	10,046
R2	0.006	0.315
Adjusted R2	0.005	0.315
Residual Std. Error	0.587 (df = 10044)	0.587 (df = 10039)
F Statistic	56.200*** (df = 1; 10044)	769.496*** (df = 6; 10039)
Note:		*p<0.1; **p<0.05; ***p<0.01

4.1.3. FWL con Bootstrap: Se estimaron los coeficientes de FWL utilizando *Bootstrap* que estima la distribución de los errores utilizando submuestras con reemplazo. *Bootstrap* entonces captura mejor la varianza de los errores. Esto es importante porque si se reportan errores pequeños cuando verdaderamente son más grandes, el coeficiente puede parecer significativo cuando verdaderamente no lo es. En este caso la varianza de los errores es más grande que la estimada por MCO y FWL pero el t-value continúa siendo mayor a 1.96.

Tabla 7.

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = base_nueva, statistic = fwl_bootstrap_log_salario,
    R = 1000)


Bootstrap Statistics :
         original       bias    std. error
t1* -1.857553e-17 0.0001443976 0.005803229
t2* -9.058492e-02 0.0001367564 0.017352631
```

Comparación de coeficientes de *Female* y errores:

Tabla 8.

	MCO	FWL	BOOTSTRAP
Coeficiente Female	-0.091	-0.091	-0.0905
Errores	0.012	0.012	0.017

La brecha por género entonces es significativa.

4.1.4 Predicted Wage: A continuación se muestran dos gráficos de edad-salario. El primero toma en cuenta todo el rango de edad de la muestra. Se puede apreciar que existe una alta variabilidad después de los 50 años de edad que puede ser el motivo por el que se genera un alto *peak age* de 59 años. En el gráfico de la derecha se restringe la edad para personas entre 18 y 55 años. En ese rango el *peak age* es de 48.9 años con un CI (46.04, 51.62). El CI se calculó utilizando Bootstrap. Se utilizó el modelo lineal:

$$\log(w) = \beta_1 + \beta_2 Female + +\beta_3 relab + \beta_4 maxEdulevel + \beta_5 Edad + \beta_6 Edad\_sqr + \beta_7 tam\_empresa + u$$

Tabla 9.

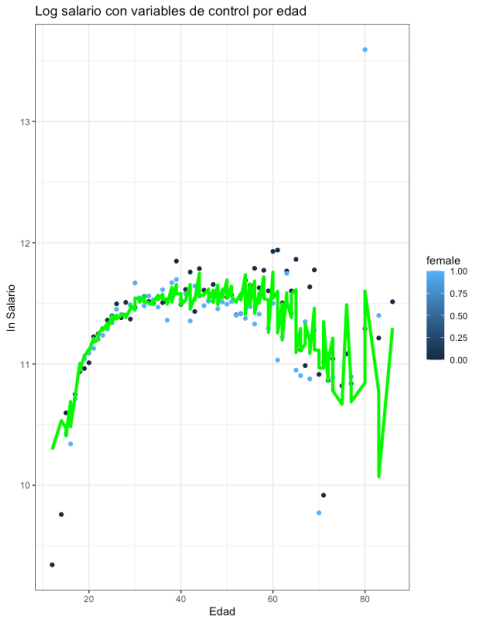
```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = base_nueva, statistic = eta_mod2_fn, R = 1000)


Bootstrap Statistics :
    original    bias    std. error
t1* 48.93428 0.1060933   1.424415
```

Tabla 10.

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = resultados, type = c("norm", "basic"))

Intervals :
Level      Normal              Basic
95%   (46.04, 51.62 )   (45.57, 51.20 )
Calculations and Intervals on Original Scale
```

Grafico 6. Salario estimado para Mujeres

Rango de edad completo



Acotada a 55 año máximo

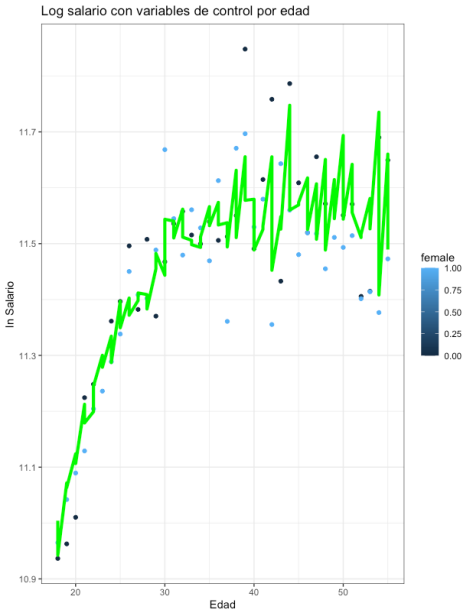
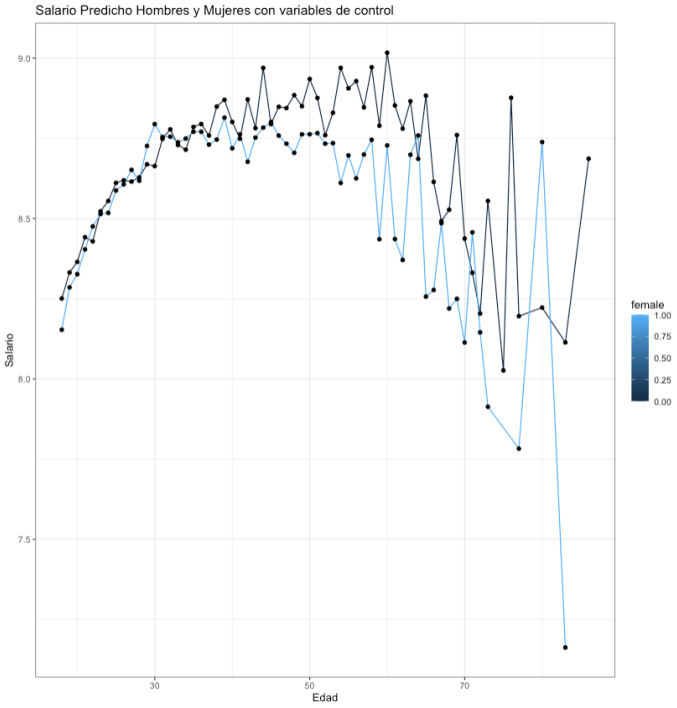


Grafico 7. Salario predicho Hombres y Mujeres



5. PREDICTING EARNINGS

En esta sección se especifican nuevos modelos con la finalidad de encontrar el que tiene mejor poder predictivo.

5.1 Division de la muestra: Aquí se dividide la muestra en dos, 70% para entrenamiento y 30% para prueba. Con el primero obtenemos los estimadores que mejor se ajustan, con el modelo estimado predecimos y evaluamos el mejor modelo con el más bajo error de predicción fuera de la muestra.

5.2 Reporte y comparación: Se contaba con tres modelos previamente estimados de edad - salario, sexo - salario, y sexo – salario con controles. Adicional a estos se estiman 7 modelos más con relaciones no lineales de tipo polinómicas, de raíz cuadrada y de interacción entre los predictores, de los mimos se calculan los MSE para evaluar su desempeño predictivo mediante el error de predicción (MSE) fuera de la muestra (prueba). Los resultados se muestran en la siguiente tabla.

Tabla 11.

	model	MSE
1	model1	0.4852
2	model2	0.5015
3	model3	0.3400
4	model4	0.4167
5	model5	0.3390
6	model6	0.3375
7	model7	0.3365
8	model8	0.3195
9	model9	0.3369
10	model10	0.3361

5.3 Resultados y discusión:

i. De los tres primeros modelos en las dos secciones anteriores, aquel que tiene un mejor desempeño fuera de la muestra es el salario explicado por la no linealidad de la edad, sexo, relación de trabajo, máximo nivel educativo y tamaño de la empresa, cuyo MSE es 0.34.

   Al compararlos con los modelos adicionales, el **modelo 8** que establece relaciones no lineales de tipo raíz cuadrada y su interacción con los predictores y entre los otros predictores, es el que muestra el menor error predictivo fuera de la muestra con un **MSE de 0.3195**. Así, este modelo es superior a los otros nueve, pues cuando complejizamos más los modelos con relaciones polinómicas de grado cinco y ocho, como en los modelos 9 y 10, el error predictivo comienza a aumentar.

   Las estimaciones muestran que mientras más complejidad haya en el modelo el performance mejora y a partir del modelo 9 el performance es menor, medido con el MSE.

ii. El modelo con el menor error predictivo es el modelo 8, cuya especificación es siguiente:

Tabla 12

Regresion con mejor error predictivo (model8)

	Dependent variable:
	Logaritmo del Salario
edad	0.132***
	(0.041)
sqrt(edad)	-1.838***
	(0.526)
female	-0.512
	(0.524)
relab	-0.995**
	(0.503)
tam_empresa	-0.324***
	(0.095)
maxEducLevel	-1.113***

	(0.244)
edad:female	-0.003
	(0.014)
sqrt(edad):female	0.034
	(0.174)
edad:maxEducLevel	-0.025***
	(0.006)
sqrt(edad):maxEducLevel	0.340***
	(0.078)
edad:relab	-0.025*
	(0.013)
sqrt(edad):relab	0.339**
	(0.165)
edad:tam_empresa	0.002
	(0.003)
sqrt(edad):tam_empresa	0.006
	(0.031)
female:relab	0.040
	(0.046)
female:maxEducLevel	0.047***
	(0.015)
female:tam_empresa	0.004
	(0.006)
maxEducLevel:tam_empresa	0.042***
	(0.003)
maxEducLevel:relab	-0.028**
	(0.013)
relab:tam_empresa	0.033***
	(0.006)
Constant	14.473***
	(1.643)
Observations	6,977
R2	0.356
Adjusted R2	0.354
Residual Std. Error	0.566 (df = 6956)
F Statistic	192.082*** (df = 20; 6956)
Note:	*p<0.1; **p<0.05; ***p<0.01

La mayoría de predictores considerados son estadísticamente significativos, con excepción de si es mujer y su interacción con la edad que no son estadísticamente significativas, al igual que la edad con el tamaño de la empresa. Por su parte, el ajuste del modelo, medido por el R-cuadrado mejora a 0.356, lo que significa que las variables independientes, expresadas en no linealidades e interacciones, explican en un 35.6% el comportamiento del logaritmo del salario. Además, predice mejor fuera de la muestra por presentar el MSE más bajo frente a los otros nueve modelos.

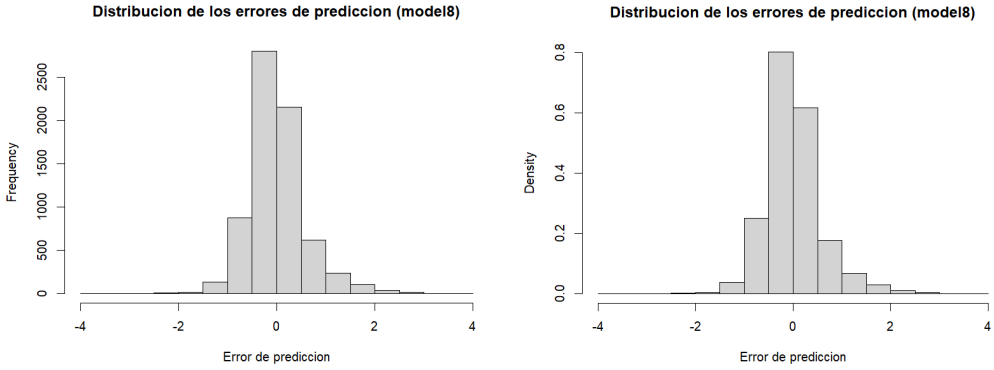
iii. Distribución de los errores de predicción del mejor modelo

El histograma muestra que el modelo para la mayor parte de la muestra de prueba predice muy buen, pues la mayoría de los datos se concentran alrededor de cero, es decir considerando determinadas variables que caracterizan a cada persona, el ingreso

estimado por el modelo es similar al salario reportado. Sin embargo, la distribución presenta colas, lo que significa en este caso, es que el salario estimado es muy diferente del salario reportado (dato real).

La cola izquierda de la distribución refleja a aquellas personas que manifiestan percibir salarios por debajo de lo que en realidad podrían tener dadas sus características (salario estimado por el modelo), sugiriendo un comportamiento cuyo incentivo, entre otros, es evadir el pago de impuestos, lo que a la autoridad tributaria debería llamarle la atención y observar con cuidado. En cuanto a la cola derecha de la distribución, están en ella declaraciones de salarios por encima de lo que sus características individuales sugieren de su salario, ello revela que el salario potencial es menor al reportado, dicha información es importante para identificar a individuos con características que requieren apoyo mediante intervenciones de políticas públicas que les permitan superar su condición de vulnerabilidad por salarios potenciales bajos.

Grafico 8.



5.4 LOOCV

LOOCV divide la muestra en *n folds*. Es decir que toma *n-1* observaciones como entrenamiento y la observación que queda por fuera la utiliza para testeo. En este caso tenemos 9967 observaciones. Para realizar este ejercicio tomamos 1000 observaciones – 1000 folds y los modelos con los mejores resultados de los apartados anteriores: modelos 7 y 8.

	Modelo 7	Modelo 8
MSE promedio de los 1000 folds	0.3452879	0.3215373

Para ambos modelos la variabilidad es mayor. Sin embargo, si se compara con los resultados de los MSE obtenidos anteriormente el modelo 8 continúa siendo el mejor modelo predictivo.