Big Data y Machine Learning
Problem Set 1

Nicolás González LLano 201813698
Paula Andrea Alarcón Chaparro 201812516
Martín Lara 201711300

Septiembre 5 de 2022

En el siguiente enlace puede encontrar el repositorio Github del Problem Set 1.

1 Datos

a) **Scrapping**

Los datos estaban originalmente en la página personal del profesor del curso dentro de 10 enlaces diferentes. Si bien hacía parte del ejercicio que fueran difíciles de extraer los diez enlaces solo diferían en un número (del uno al diez). Sin embargo, los datos no eran directamente accesibles desde la página que alojaba los datos, por ese motivo había que usar el llink específico de los datos en el html.

Una vez logramos es acceso a los datos desde R hicimos un loop que los descargara y los alojara en una sola base de datos haciendo un "rowbind" para pegarlos verticalmente.

Los datos eran de la Gran Encuesta Integrada de Hogares del DANE con una serie de modificaciones y cáculos hechos por porfesores de la Facultad de Economía. Inicialmente la base de datos tenía 178 variables y 32177 observaciones que reflejaban el estado del mercado laboral en Bogotá. Despues de generar la base de datos esta se subió al repositorio en la carpeta Stores como un archivo .Rdata y .csv para seguir trabajando en las modificaciones.

Se hicieron una serie de modificaciones a la base de datos para depurar las variables que no eran de interés y a los individuos que no se busca entren en la predicción de impuestos que se lleva a cabo de aquí en adelante.

b) **Data cleaning**

En un primer acercamiento a los datos, optamos por tener en cuenta solo aquellos datos con un bajo porcentaje de valores faltantes, esto simplemente como una primera depuración que nos permite tener mayor claridad sobre los datos que vamos a manejar. Las variables con un porcentaje de valores faltantes mayor o igual a 45%, los descartamos de primero. Luego, en un segundo proceso de depuración, estudiamos las variables restantes y escogimos aquellas que pueden explicar a grandes razgos el ingreso de una persona. Por otro lado, también buscamos variables caracteristicas de personas o de contexto de la muestra. Para comenzar, eliminamos todas aquellas personas que son desempleadas o son menores de 18 años. Para ello usamos las variables dsi y age. Por ello, como se evidencia en la tabla 1, age tiene un valor minimo de 18 y dsi (desempleado) solo tiene valores en cero, es decir que nuestra base no tiene la observación de ningún desempleado. Para corroborar que los datos tuvieran sentido, sumamos los factores de expansión de las observaciones restamtes y obtuvimos una población de 5.598.422 personas, un valor razonable frente a una ciudad con aprox. 8 millones de habitantes en 2018. Esto quiere decir que entre desempleados y menores de edad suman 2.401.578 personas. Como variable de ingreso tomamos ingtot, esto debido a que contaba con todas las observaciones, lo que nos iba a permitir contar con más datos para nuestras estimaciones, además de que el ingreso laboral es el más estudiado por la literatura y el más fiable, pues es muy dificil de esconder. Al analizar las estadisticas descriptivas de nuestras variables también es posible evidenciar como casi un 30% de la población tiene estudios de pregrado. Finalmente, aunque optamos por variables con un muy bajo porcentaje de valores faltantes, variables como horas trabajadas tienen algunos valores faltantes, por lo que decidimos imputarlos usando el metodo de imputación por valores más cercanos, los resultados después de imputación se pueden observar en la tabla 2.

Table 1: Estadísticas descriptivas de las variables antes de imputar

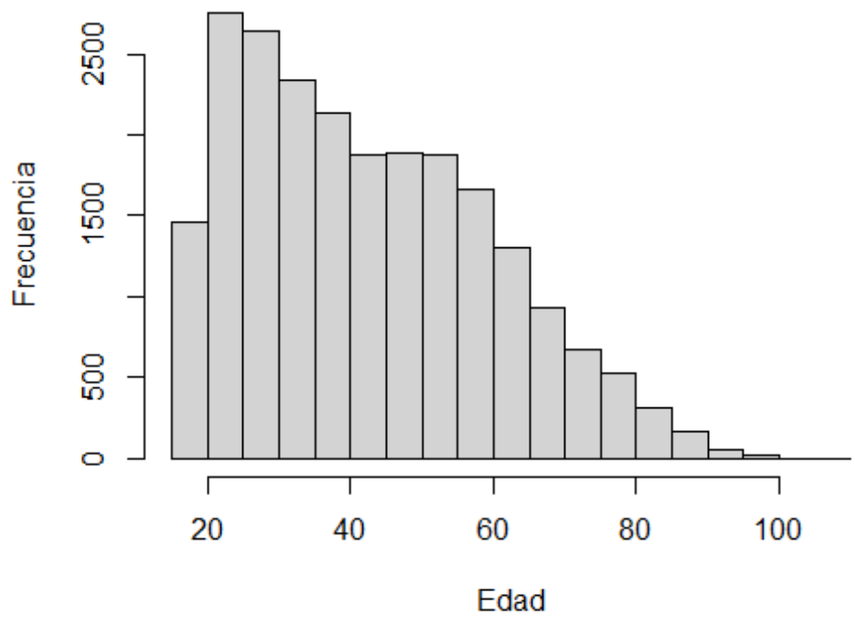
Statistic	N	Mean	St. Dev.	Min	Max
age	22,640	43.033	17.368	18	106
college	22,640	0.294	0.456	0	1
cotPension	16,542	1.455	0.542	1	3
cuentaPropia	22,640	0.226	0.418	0	1
dsi	22,640	0.000	0.000	0	0
estrato1	22,640	2.590	1.025	1	6
fex_c	22,640	247.280	58.488	106.613	808.241
hoursWorkUsual	16,542	47.008	15.543	1	130
ingtot	22,640	1,464,343.000	2,448,759.000	0.000	85,833,333.000
ingtotob	22,640	1,298,605.000	2,323,974.000	0.000	85,833,333.000
maxEducLevel	22,638	5.766	1.388	1	7
oficio	16,542	49.773	28.080	1	99
p6426	16,542	63.758	89.488	0	720
p7500s1a1	22,640	73,912.650	439,987.800	0	20,000,000
p7500s2a1	22,640	121,959.400	832,352.000	0	80,000,000
p7510s5a1	22,640	90,921.440	3,875,585.000	0	400,000,000
sex	22,640	0.469	0.499	0	1
totalHoursWorked	16,542	47.403	15.662	1	130

La segunda tabla de estadisticas descriptivas, nos permite evidenciar que los valores imputados no cambiaron de manera importante en su valor de la media ni en la desviación estándar.

Table 2: Estadísticas descriptivas de las variables después de imputar

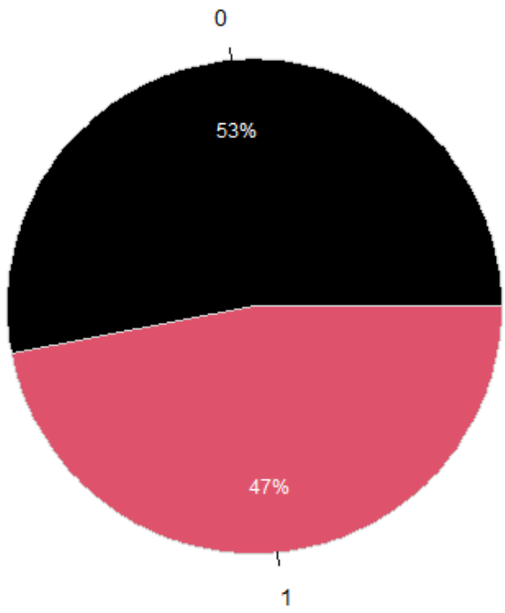
Statistic	N	Mean	St. Dev.	Min	Max
age	22,640	43.033	17.368	18	106
college	22,640	0.294	0.456	0	1
cotPension	22,640	1.444	0.530	1	3
cuentaPropia	22,640	0.226	0.418	0	1
dsi	22,640	0.000	0.000	0	0
estrato1	22,640	2.590	1.025	1	6
fex_c	22,640	247.280	58.488	106.613	808.241
hoursWorkUsual	22,640	49.572	14.970	1	130
ingtot	22,640	1,464,343.000	2,448,759.000	0.000	85,833,333.000
ingtotob	22,640	1,298,605.000	2,323,974.000	0.000	85,833,333.000
maxEducLevel	22,638	5.766	1.388	1	7
oficio	16,542	49.773	28.080	1	99
p6426	22,640	60.652	80.057	0	720
p7500s1a1	22,640	73,912.650	439,987.800	0	20,000,000
p7500s2a1	22,640	121,959.400	832,352.000	0	80,000,000
p7510s5a1	22,640	90,921.440	3,875,585.000	0	400,000,000
sex	22,640	0.469	0.499	0	1
totalHoursWorked	22,640	48.609	14.044	1	130
cotPension_imp	22,640	0.269	0.444	0	1
hoursWorkUsual_imp	22,640	0.269	0.444	0	1
p6426_imp	22,640	0.269	0.444	0	1
totalHoursWorked_imp	22,640	0.269	0.444	0	1

Figure 1: **Histograma de la edad**



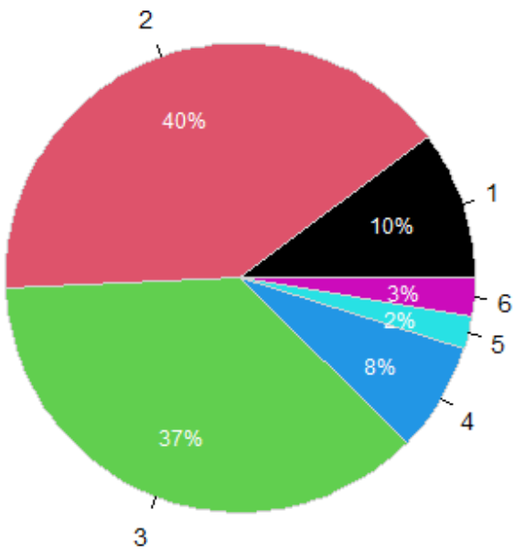
La figura 1, muestra la distribución de personas por edad. Aquí es notable la presencia de únicamente mayores de edad. Por otro lado, la edad con mayor frecuencia es entre los 20 y los 30 años.

Figure 2: **Porcentaje de la población por sexo (hombre=1; mujer=0)**



La figura 2, nos permite identificar que de nuestra población objetivo el 47% son hombres y el 53% son mujeres.

Figure 3: **Porcentaje de personas por estrato social**



Finalmente, la figura 3, nos muestra la distribución en porcentaje de la población entre estratos sociales y como era de esperarse, el grueso se encuentra entre los estratos 2 y 3, mientras que el estrato 5 y 6 suman un porcentaje del 5

2 Modelo de Ingresos

La literatura presenta que existe un Age-earnings profile que sugiere que los ingresos de las personas jóvenes tienden a ser bajos y a medida que aumenta la edad aumentan los ingresos hasta que llegan a un máximo, de ahí en adelante es posible que se mantengan o disminuyan. En esta sección queremos evaluar cuál sería ese pico de edad promedio para la población en Bogotá y el ingreso percibido. Para ello, tenemos el siguiente modelo:

$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u$$

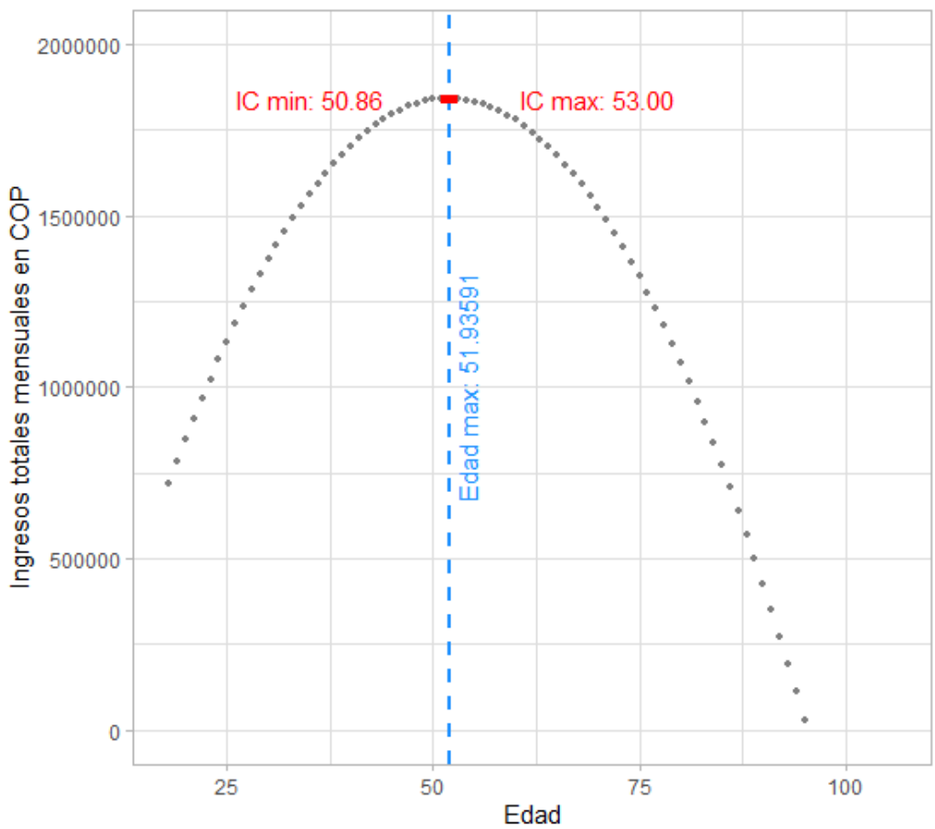
En el cual medimos el ingreso con la variable ingtot porque contaba con todas las observaciones y nos permite tener más datos para nuestras estimaciones y por tanto mayor robustez. De esta forma se estima el modelo con la regresión de la Tabla 3, en la cual el efecto de la edad en los ingresos se mide con la elasticidad que es 0,0047. Por consiguiente, la

estimación arroja que en promedio el cambio en un año de edad tiene efectos positivos en 0,0047 unidades. La estimación cuenta con 22.640 observaciones y un R^2 muy cercano a 0 de 0,020 que indica que la estimación no es la más adecuada.

Table 3: Regresión de Earnings

	Dependent variable:
	ingtot
age	98,332.390***
	(4,768.292)
age2	−942.654***
	(49.055)
Constant	−737,212.000***
	(104,879.000)
Observations	22,640
R^2	0.020
Adjusted R^2	0.020
Residual Std. Error	2,424,762.000 (df = 22637)
F Statistic	226.152*** (df = 2; 22637)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 4: **Valores predichos de la regresión: Ingresos según la edad**



Ahora bien, en este modelo queremos evaluar cuáles serían los picos de edad en los que los individuos alcanzan su máximo ingreso. La figura 4 se construyó a partir de la regresión presentada anteriormente y los intervalos de confianza se hallaron con la metodología bootstrap. En este sentido, el gráfico presenta que, en promedio, las personas de Bogotá llegan a su ingreso máximo a los 52 años aproximadamente y en promedio tienen un ingreso de 1.700.000 en el pico, con un intervalo de confianza entre 50,86 y 53 años con $\alpha = 5\%$. Igualmente, con los valores predichos el gráfico presenta que en promedio los ingresos tienden a decrecer después del peak-age.

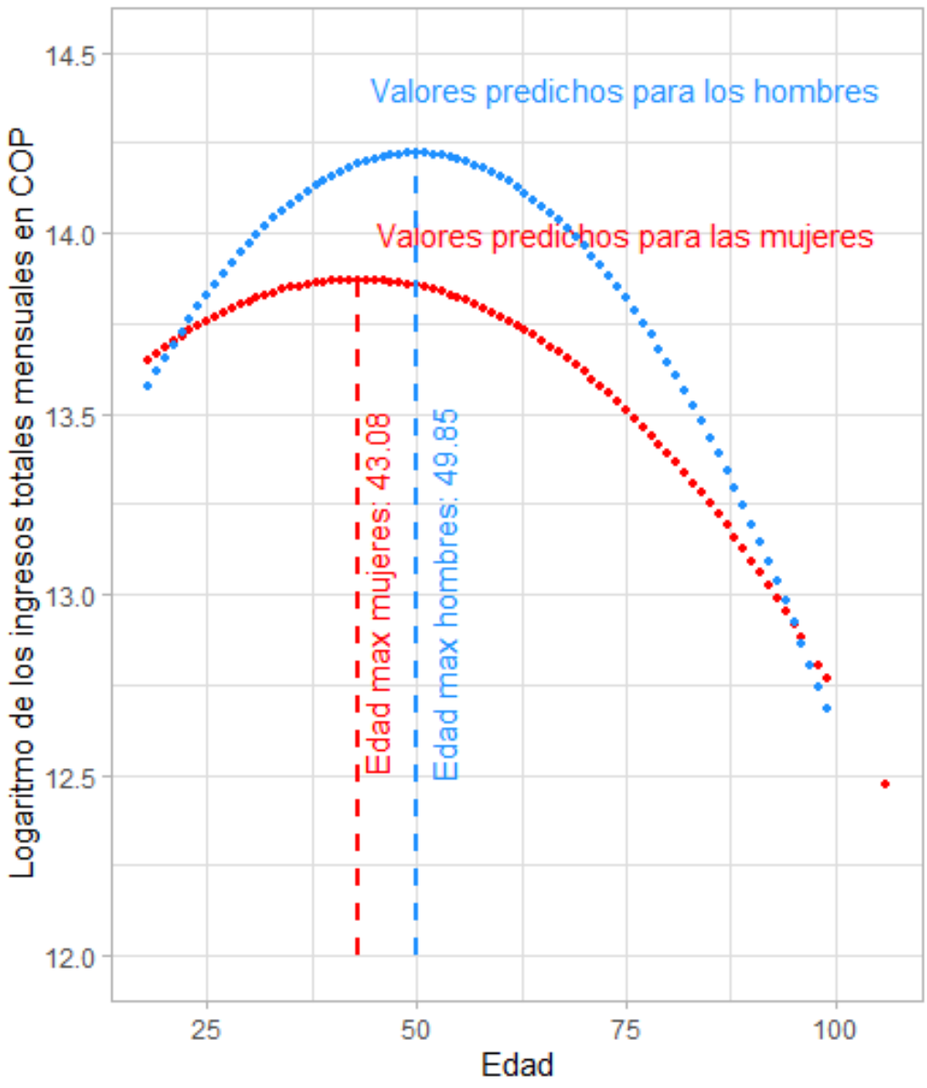
3 Brecha de Género

La brecha de ingresos entre hombres y mujeres ha sido una de las grandes preocupaciones de los encargados de las políticas públicas en el país, pues es cierto que las mujeres perciben menores ingresos que los hombres por factores como la carga de labores de cuidado no remunerado, el acceso a empleos en edad fértil, el nivel de educación, entre otros. Según ONU Mujeres (2020) las labores del hogar juegan un papel fundamental en el desarrollo educativo y laboral de las mujeres, ya que se presentan trade-offs entre el trabajo doméstico y el trabajo remunerado que afectan la trayectoria laboral. A pesar de que su participación es relevante en la productividad del país, la mitad de las mujeres está fuera de la fuerza laboral o tiene empleos en condiciones desfavorables, lo cual tiene fuertes repercusiones en la economía y bienestar de Colombia ONU Mujeres (2020). En este capítulo se desarrollarán diferentes factores que pueden explicar la brecha de género en cuanto a ingresos en Bogotá, en especial se evaluará la incidencia del nivel

educativo en los ingresos diferenciado por género.

En la misma línea de la sección anterior, realizaremos las estimaciones con la variable de ingresos totales como medida del ingreso justificando que esta variable toma el total de los ingresos obtenidos y es la que más datos reporta de las demás variables que miden ingreso. Asimismo, evaluaremos si existen diferencias entre los picos de edad, mencionados anteriormente, y sus respectivos intervalos de confianza con $\alpha = 5\%$ para grupos diferenciados entre hombres y mujeres con sus respectivos intervalos de confianza que hallamos con la metodología bootstrap. En la figura 5 se observa que los hombres llegan a sus ingresos máximos aproximadamente a los 50 años, con intervalos de confianza entre 48,71 y 50,98 mientras que las mujeres lo alcanzan a los 43 años, con intervalos de confianza entre 41,37 y 44,78 años (más estrechos que el de los hombres). En este sentido, hay evidencia suficiente para decir que existe una brecha de género en los picos de edad, en donde el hombre alcanza su máximo ingreso a una mayor edad y percibiendo mayores ingresos que la mujer, generando así mayores ganancias intertemporales para él.

Figure 5: Valores predichos: Ingresos según la edad y sexo



Ahora bien, con el fin de evaluar la brecha de género en los ingresos iniciaremos haciendo una primera estimación simple para ver el efecto de ser mujer sobre los ingresos de una persona, para esto se presenta la regresión de la Tabla 4 que cuenta con 19.239 observaciones, en donde la variable dependiente es el logaritmo natural de los ingresos totales y la variable independiente es una dicótoma que toma el valor de 1 si la persona es mujer y 0 de lo contrario. De acuerdo con los resultados de la Tabla 4, se observa que con una significancia del 1% hay un efecto negativo de 24% sobre el ingreso total por ser mujer y por tanto se puede concluir que existe una brecha de género en el modelo no condicionado. Cabe destacar que el R^2 de esta regresión es muy bajo y cercano a 0 (0.016) y por lo tanto no especifica adecuadamente el modelo.

$$log(Income) = \beta_1 + \beta_2 Female + u$$

Table 4: Regresión de Female

	Dependent variable:
	lningtot
fem	-0.241^{***}
	(0.014)
Constant	14.014^{***}
	(0.010)
Observations	19,239
R^2	0.016
Adjusted R^2	0.016
Residual Std. Error	0.949 (df = 19237)
F Statistic	309.842^{***} (df = 1; 19237)
Note:	*p<0.1; **p<0.05; ***p<0.01

Para revisar la brecha de género en personas con características similares en educación y, para ajustar el modelo inicial, realizamos una regresión condicionada con diferentes variables de control como hoursWorkUsual, cuentaPropia, age y age2 que representan las horas trabajadas en promedio de una semana, si la persona es independiente, la edad y la edad al cuadrado respectivamente. En este caso, para medir la educación utilizamos la variable categórica maxEducLevel que mide el máximo nivel educativo del individuo como se presenta a continuación:

Table 5: Descripción de la variable maxEducLevel

Valor	Etiqueta
1	Ninguno
2	Prescolar
3	Primaria incompleta (1-4)
4	Primaria completa (5)
5	Secundaria incompleta (6-10)
6	Secundaria completa (11)
7	Terciaria
9	N/A

Sin embargo, si se realiza la regresión con la variable maxEducLevel los resultados estarían mal estimados, pues el programa tomaría estos valores como si terminar primaria fuera exactamente el doble de importante que terminar preescolar. Para resolver este problema dividimos las categorías de maxEducLevel en diferentes variables dummy donde educx toma el valor de 1 si la categoría es x, por ejemplo para representar que el máximo nivel educativo es primaria se crea la variable dummy educ4 que toma el valor de 1 si maxEducLevel es 4 y 0 de lo contrario.

En este sentido, para evaluar la brecha de género en los ingresos por nivel educativo decidimos realizar dos modelos: el primero es un modelo no restringido que contiene la variable fem y sus respectivas interacciones con el nivel de educación (Tabla 6, modelo 1) y el segundo modelo restringido en donde suponemos que fem toma el valor de 0 (Tabla 6, modelo 2).

Modelo no restringido:

$$log(Income) = \beta_1 + \delta_1 fem + educx\beta_2 + educxfem\delta_2 + Z\beta_3 + u$$

Modelo restringido:

$$log(Income) = \beta_1 + educx\beta_2 + Z\beta_3 + u$$

Donde el logaritmo natural el ingreso es la variable dependiente, fem es una dicótoma que toma el valor de 1 si la persona es mujer, educx es la matriz que representa las dummys del máximo nivel educativo alcanzado, educxfem es la matriz que representa las interacciones de los niveles de educación con fem y Z es la matriz de hoursWorkUsual, cuentaPropia, age y age^2.

Table 6: Regresión de Female

	Dependent variable:	
	lningtot	
	(1)	(2)
fem	−0.159***	
	(0.017)	
educ3	0.265***	0.248***
	(0.075)	(0.072)
educ3fem	−0.066	
	(0.056)	
educ4	0.460***	0.407***
	(0.071)	(0.069)
educ4fem	−0.152***	
	(0.041)	
educ5	0.551***	0.471***
	(0.070)	(0.069)
educ5fem	−0.224***	
	(0.038)	
educ6	0.699***	0.680***
	(0.068)	(0.068)
educ6fem	−0.064**	
	(0.026)	
educ7	1.358***	1.362***
	(0.067)	(0.068)
hoursWorkUsual	0.011***	0.012***
	(0.0004)	(0.0004)
cuentaPropia	−0.418***	−0.405***
	(0.013)	(0.013)
age	0.053***	0.050***
	(0.002)	(0.003)
age2	−0.0004***	−0.0004***
	(0.00003)	(0.00003)
Constant	11.465***	11.373***
	(0.083)	(0.084)
Observations	16,276	16,276
R^2	0.342	0.325
Adjusted R^2	0.341	0.325
Residual Std. Error	11.241 (df = 16261)	11.381 (df = 16266)
F Statistic	603.298*** (df = 14; 16261)	870.878*** (df = 9; 16266)
Note:		*p<0.1; **p<0.05; ***p<0.01

Las regresiones de la tabla 6 arrojan resultados intuitivos con la literatura. En primer lugar, en la primera estimación del modelo no restringido, que cuenta con 16.276 observaciones y tiene un R^2 de 0.342 mayor que el de la regresión simple, las variables que tienen el efecto de los diferentes años de educación para las mujeres son negativos con una confianza del 99%, excepto en el caso de primaria incompleta que no es significatia y en educación secundaria que es significativa al 0.05%.Por otro lado, aquellas variables que no tienen interacción con la dummy fem (sin contar los controles) tienen coeficientes positivos y significativos al 1%. Estos resultados son intuitivos, pues se espera que los años de educación de una persona tengan impactos positivos en sus ingresos, mientras que para las mujeres puede existir una brecha por factores como el trade off entre el trabajo del cuidado y el trabajo remunerado. En este caso, el efecto estimado del nivel máximo en educación de las mujeres se ve reflejado por la diferencia entre los coeficientes de educx y educxfem. Así, el efecto de que una mujer haya completado la secundaria (educ6) impacta en el ingreso en 63% mientras que para el hombre es de 69%.

En segundo lugar, la estimación del modelo restringido arroja coeficientes positivos y significativos al 1% para las variables de educación, los cuales, siguiendo con el mensaje anterior, son intuitivos pues a mayor educación se esperan mejoras en el ingreso. Este modelo tiene 16.276 observaciones al igual que el anterior y un R^2 de 0.325 que es cercano pero menor que el del modelo sin restricción. Ambos tienen un mejor ajuste que el modelo simple presentado al inicio de esta sección.

Ahora bien, a continuación se presenta la prueba de hipótesis sobre los efectos de la educación entre hombres y mujeres, en la cual si el ingreso promedio fuera idéntico para ambos grupos no deberían haber cambios de intercepto ni pendiente entre modelos.

$H_0 : \delta_1 = \delta_2 = 0$ (Los retornos a la educación son iguales para ambos grupos)

H_a : Alguno difiere de 0 (Existe una brecha de género)

$$F = \frac{(SCE_r - SCE_{nr})/q}{SCE_{nr}/n - k - 1}$$

Donde q es el número de restricciones, n el número de observaciones y k el número de variables independientes

en el modelo sin restricción.

$$F = \frac{(13.151 - 12.772,05)/q}{12.772,05/16.276 - 14 - 1} = 485,89 > F(5, 16261) = 2,21$$

Con base en los resultados, es posible concluir que se rechaza la hipótesis nula y por tanto el ingreso medio difiere entre hombres y mujeres con el mismo nivel educativo.

Ahora bien, para realizar la estimación del modelo no rstringido, también se puede usar la metodología FWL con la cual pretendemos remover un parámetro ruidoso de la regresión y calcular el vector de residuales sin el elemento ruidoso en todas las observaciones, para así hallar los estimadores.

En nuestro caso, como se presenta en la tabla 7, la regresión por la metodología FWL (modelo 1) tiene los mismos coeficientes o muy cercanos a la regresión sin restricción que habíamos calculado inicialmente. Sin embargo, el ajuste del modelo inicial es mejor que por la metodología FWL, pues el R^2 del primero es 0,342 y del segundo 0,296. La igualdad en los coeficientes de ambas regresiones se da principalmente por la relación que genera la matriz aniquiladora entre los residuos y los verdaderos errores del modelo.

Table 7: Comparación entre FWL y "Long Reg"

	Dependent variable:	
	res_lningtot	lningtot
	(1)	(2)
res_fem	−0.159***	
	(0.017)	
res_educ3	0.264***	
	(0.075)	
res_educ3fem	−0.066	
	(0.056)	
res_educ4	0.460***	
	(0.071)	
res_educ4fem	−0.152***	
	(0.041)	
res_educ5	0.551***	
	(0.070)	
res_educ5fem	−0.224***	
	(0.038)	
res_educ6	0.699***	
	(0.068)	
res_educ6fem	−0.064**	
	(0.026)	
res_educ7	1.358***	
	(0.067)	
res_hoursWorkUsual	0.011***	
	(0.0004)	
res_age	0.053***	
	(0.002)	
res_age2	−0.0004***	
	(0.00003)	
fem		−0.159***
		(0.017)
educ3		0.265***
		(0.075)
educ3fem		−0.066
		(0.056)
educ4		0.460***
		(0.071)
educ4fem		−0.152***
		(0.041)
educ5		0.551***
		(0.070)
educ5fem		−0.224***
		(0.038)
educ6		0.699***
		(0.068)
educ6fem		−0.064**
		(0.026)
educ7		1.358***
		(0.067)
hoursWorkUsual		0.011***
		(0.0004)
cuentaPropia		−0.418***
		(0.013)
age		0.053***
		(0.002)
age2		−0.0004***
		(0.00003)
Constant	0.005	11.465***
	(0.006)	(0.083)
Observations	16,276	16,276
R^2	0.296	0.342
Adjusted R^2	0.296	0.341
Residual Std. Error	11.241 (df = 16262)	11.241 (df = 16261)
F Statistic	526.390*** (df = 13; 16262)	603.298*** (df = 14; 16261)
Note:		*p<0.1; **p<0.05; ***p<0.01

4 Predicción de las Ganancias

Para medir el poder predictivo de los modelos especificados en los ejercicios anteriores ẏ compararlos con otros modelos más complejos empezamos el proceso dividiendo la base datos en bases de entrenamiento y prueba (70 y 30 porciento respectivamente). La idea es estimar los modelos propuestos en la base de entrenamiento y despues probar su ajuste en la base de prueba, evaluando

la capacidad de predicción fuera de muestra. Los modelos propuestos son:

1. $$Earnings = \beta_1 + \beta_2 Female + u$$

2. $$Earnings = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u$$

3. $$Earnings = \beta_1 + \beta_2 Age + \beta_3 College * TotalHoursWorked + \beta_3 Sex + u$$

4. $$Earnings = \beta_1 + \beta_2 UsualHoursWorked * Estrato + \beta_3 CotizaPension + u$$

5. $$Earnings = \beta_1 + \beta_2 UsualHoursWorked * Estrato + \beta_3 CotizaPension + \beta_4 TiempoEnEmpresa + u$$

6. $$Earnings = \beta_1 + \beta_2 ACollege + \beta_3 IngresoArriendoDePropiedadRaiz + \beta_4 IngresoPensiones + \beta_5 UsualHoursWorked * Estrato + u$$

7. $$Earnings = \beta_1 + \beta_2 AUsualHoursWorked * Estrato + \beta_3 CotizaPension + \beta_4 Age + \beta_5 Age^2 + \beta_6 Female + u$$

Como se puede ver, a partir del tercer modelo propuesto la complejidad de estos es creciente, lo anterior con el objeivo de evaluar si un modelo más complejo tiene mayor capacidad predictiva. Al medir el ajuste de los modelos en la base de prueba se evaluó el MAE y RMSE en cada una de las especificaciones anteriores para comparar su capacidad de predecir fuera de muestra.
Los resultados de la evaluación fuera de muestra de los modelos fue:

Table 8: Metricas de predición

	MAE	RMSE
1	$1,199,117$	$2,445,614$
2	$1,212,594$	$2,449,919$
3	$1,178,385$	$2,392,165$
4	$1,433,286$	$2,621,856$
5	$1,448,691$	$2,635,193$
6	$1,437,881$	$2,683,967$
7	$1,486,297$	$2,649,963$

Tanto el MAE como el RMSE son métricas de error que comparan la diferencia entre el valor predicho y el realizado, sus formulas son:

$$RMSE = \sqrt{\frac{1}{n}\sum e_n^2}$$

$$MAE = \frac{1}{n}\sum |e_n|$$

El motivo de elección de estas métricas es que ambas calculan el error de predicción de una forma simple. Sin embargo, el RMSE es más sensible a valores atípicos que el MAE. Dada la heterogeneidad de los datos es que este es en todos los casos mayor.

El error de predicción promedio de los modelos es en todos los casos mayor a un millón de pesos. Lo anterior se debe a las diferencias existentes entre los ingresos de las personas y los múltiples factores no observables que influyen en estos.

En cada ambas métricas propuestas los modelos 1 y 3 fueron los de mejor predicción. Este resultado es consistente con la teoría del Trade-off Sesgo-Varianza dado que los modelos más complejos, con mejor ajuste dentro de muestra y menor sesgo, son significativamente peores prediciendo fuera de muestra. Es interesante ver que el modelo 3, más complejo que el 2 tuvo mejor predicción.

Para los dos modelos de mejor predicción se lleva a cabo validación cruzada dejando uno afuera (Leave one out cross validation), proceso en el cuál el se estima el modelo N veces ($N = nmerodeobservaciones$) dejando una observación fuera de la estimación y probandola en la observación no incluida. Los resultados del LOOCV se presentan en la siguiente tabla:

Table 9: Metricas de predicción LOOCV

	RMSE	MAE
1	$2,435,456$	$1,183,779$
3	$2,412,498$	$1,181,094$

La predicción LOOCV del modelo tres es mejor que la del modelo uno, manteniendo lo visto para la predicción de los datos de la base de entrenamiento. Esto indica que si bien la simplicidad de los modelos propuestos en los puntos anteriores ofrece una buena predicción fuera de muestra los modelos seguían siendo demasiado simples a la hora de predecir fuera de muestra. El hecho de que un modelo ligéramente más complejo tuviera una mejor predicción que el más simple de todos implica que los modelos propuestos en los puntos anteriores estaban sacrificando mucho sesgo a cambio de ganancias en varianza.

References

ONU Mujeres, D. y. C. (Septiembre 2020). *Mujeres y hombres: brechas de género en Colombia.* Biblioteca Cámara de Comercio Bogotá.