Big Data & Machine Learning (MECA 4107)

Problem Set 1

María Camila Cely Moreno, Sara Ospina Giraldo

Junio 27 de 2022

Repositorio Github: https://github.com/camilacely/ProblemSet1_Cely_Ospina

1. Data Acquisition

Se realizó web-scraping del enlace suministrado en el Problem Set. Dicha página web, por ser propiedad del profesor del curso, no contaba con restricciones para el acceso a la información. Teniendo en cuenta que la base de datos suministrada se encontraba dividida en 10 secciones, fue necesario realizar el proceso de fusionar las respectivas secciones para crear la base final.

A grandes rasgos, el proceso de scraping se realizó de la siguiente forma:

Algoritmo 1 Web Scraping

```
for Every data chunk do
    Set up an empty dataframe (R)
    Inspect website
    State url (R)
    Read html using rvest library (R)
    Compile results in element "temp" (R)
    Merge "temp" contents into the empty dataframe
end for
Standardize column names for all 10 dataframes
Merge 10 dataframes
Save final dataframe
State a shortcut for loading the final dataframe
```

2. Data Cleaning

Estadísticas descriptivas

A partir de la base de datos suministrada, de originalmente 32.177 observaciones, se realizó una submuestra que solamente contemplara a los individuos mayores de 18 años que se encontraran empleados, lo cual resultó en 16.542 observaciones.

Se definió como variable dependiente de ingreso total la variable que en la base se denomina "ingtot". Se escoge la variable ingtot pues está teniendo en cuenta tanto valores observados para el ingreso como valores imputados. Se tiene en cuenta el ingreso laboral, ingresos de otras fuentes (como arriendos) e ingresos que deberá tener de acuerdo con las características observadas. Se asume que el DANE hace un ejercicio confiable en la imputación de valores, por lo cual se mantienen los valores imputados propuestos.

En primer lugar, se evalúa como es el comportamiento en cuanto a simetría y distribución de la variable Y escogida. En la Figura 1 se puede observar cómo la escala en la que esta se presenta no permite un análisis claro de su distribución.

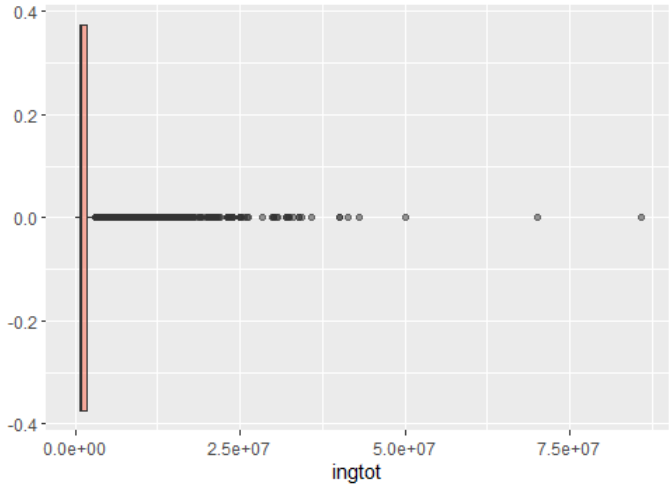


Figura 1. Distribución del ingreso total

Por esta razón, se toma el logaritmo del ingreso, de esta forma se logra normalizar la distribución de las observaciones alrededor de la media y posteriormente esto nos permitirá hacer un mayor análisis de las variables x sobre ella. En la Figura 2 se puede observar el resultado de esta transformación, a pesar de esto, se puede identificar que la variable presenta outliers. Esto se puede observar claramente al presentarse valores de la variable *longintot* iguales a 0, lo cual no concuerda con la información reportada. Esto se sabe ya que la base fue filtrada para contar solo con personas que se declararon como ocupadas y que adicionalmente reportaron horas trabajadas. Adicionalmente, el número de observaciones en esta condición es únicamente de 265, lo cual no representa una perdida grande para una muestra de 16.542 observaciones. Por esta razón, se descartan estas observaciones.

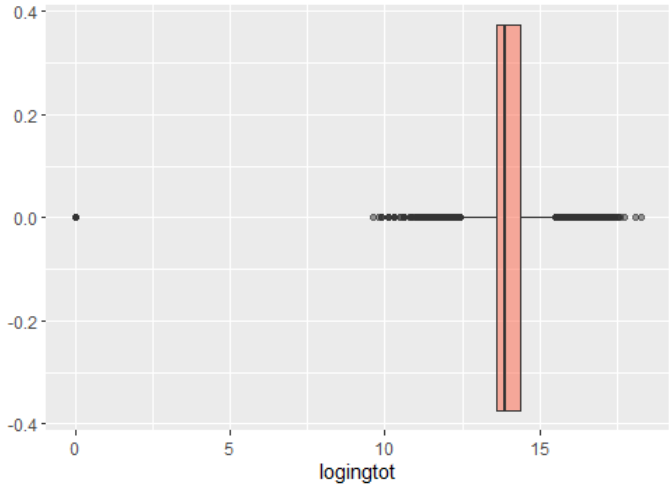


Figura 2. Distribución del logaritmo del ingreso total

En la Figura 3 se puede observar cómo se encuentra la variable una vez se descartaron los valores atípicos, se puede observar un comportamiento normal de los datos, con algunos por fuera del rango intercuartil.

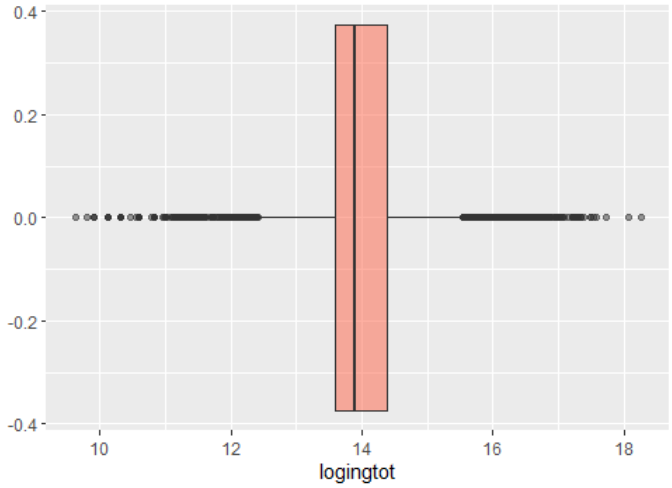


Figura 3. Distribución del logaritmo del ingreso total entre cuartiles

De esta forma, se escoge la variable de logaritmo del ingreso total *(logingtot)* como variable dependiente (y) del modelo. A continuación, en la Figura 4 se observa un análisis de la densidad de esta variable una vez realizados los procesos de limpieza y ajuste descritos. Se evidencia que los datos de la muestra en cuestión tienen una dispersión muy baja y tienden a localizarse cerca a la mediana.

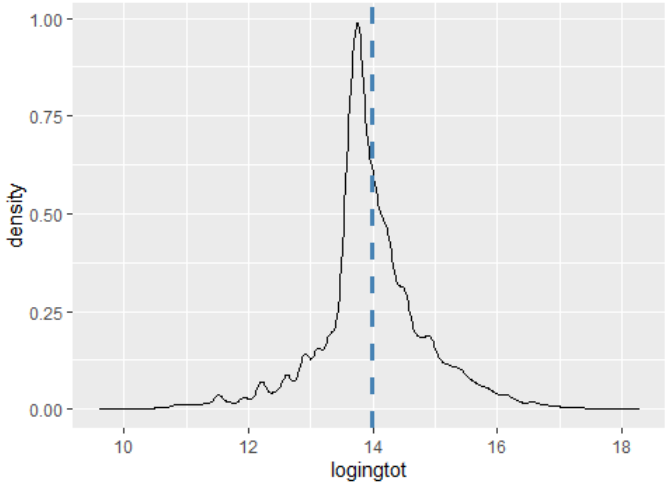


Figura 4. Distribución del logaritmo del ingreso total con su media

En cuanto a las variables independientes (x), se parte por el análisis de la variable de edad (Figura 5) donde se puede ver una distribución muy densa hacia la mediana de la variable de logaritmo del ingreso y hacia arriba y abajo se presentan cada vez menos valores. Igualmente, se observa que entre las edades de 60 y 80 años comienzan a desaparecer observaciones, esto puede corresponder a la edad en la que las personas de pensionan.

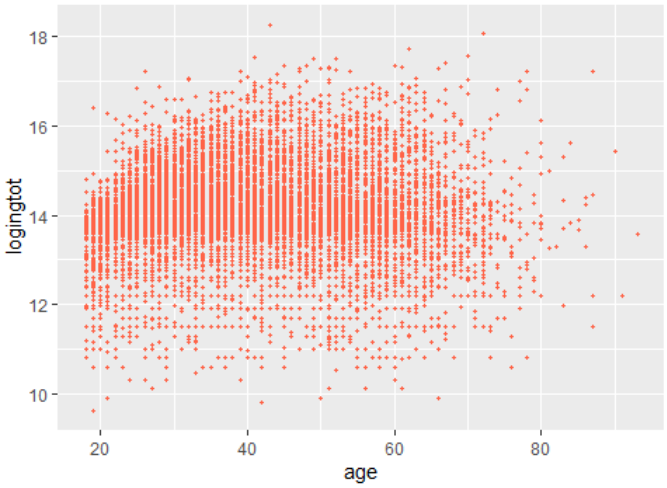


Figura 5. Distribución de la edad contra el logaritmo del ingreso total

En la Figura 6 se presenta la frecuencia del logaritmo del ingreso total diferenciado por género, a partir de la cual se puede observar que los individuos hombres se localizan más en la mediana de la distribución de la variable de ingreso. Al observar las colas de la distribución, se evidencia que la cola inferior de ingreso presenta más mujeres y la cola superior presenta más hombres. Todo lo anterior desde ya da una idea de la disparidad de ingresos por género, un análisis más detallado de esta brecha se desarrollará más adelante.

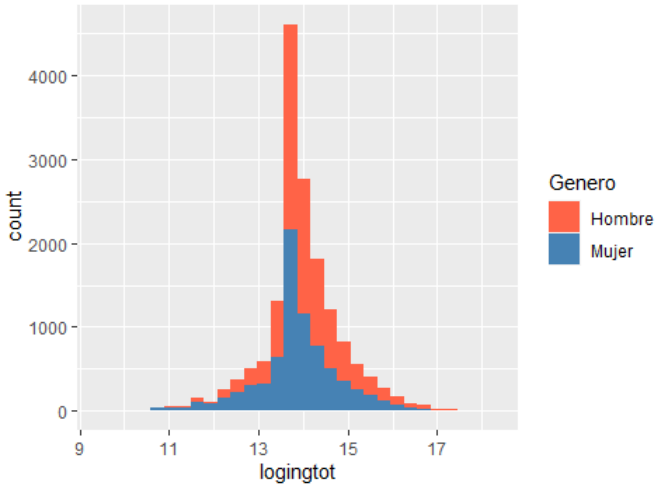


Figura 6. Frecuencia del logaritmo del ingreso total diferenciado por género

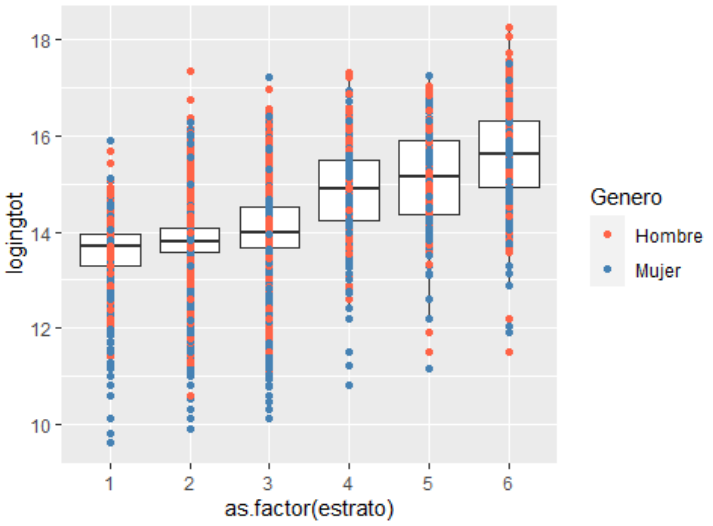


Figura 7

En la Figura 7 se puede observar la distribución del logaritmo del ingreso contra el estrato de las observaciones de la muestra diferenciado por el género de los encuestados. De forma general se observa que conforme aumenta el estrato se evidencian mayores ingresos totales, y, de forma congruente con lo analizado en la Figura 6, los valores más altos de la distribución corresponden a individuos hombres y los valores más bajos corresponden a individuos mujeres. Además, estos valores más bajos se hacen especialmente presentes en los estratos más bajos, lo cual puede estar dando cuenta de dificultades de acceso al mercado laboral en mujeres de estos estratos, lo cual se desarrollará a fondo más adelante.

En este documento se compilan los comentarios más relevantes respecto al análisis de las variables escogidas, sin embargo, en el script adjunto en el repositorio Github se encuentra el código comentado, el cual suministra información adicional sobre las mismas.

Missing values

Las variables elegidas como explicativas no presentaban missing values, exceptuando un único valor de missing value en la variable maxEducLevel. En la variable dependiente no había missing values pero sí se presentaban varios valores de cero, lo cual no era coherente pues se trataba de personas que reportaban estar empleadas y trabajar determinado número de horas a la semana, sin embargo no reportaban ningún ingreso. Teniendo en cuenta que la variabled ingtot captura muchos tipos de ingreso, no solamente monetario, este valor de cero no tenía realmente sentido, y como se trataba de 265 observaciones (1,6% de la muestra), las anteriores observaciones según nuestro criterio no representaban problemas al ser eliminadas.

3. Age-earnings profile

Como se describió en el numeral anterior, la variable escogida como variable dependiente es la que corresponde a ingreso total en la base suministrada, incluyendo los valores imputados por el DANE y realizando una submuestra que no incluyera los datos por debajo del primer cuartil, con lo cual se descartaron 265 observaciones (1,6% del total original). Así mismo, las

estadísticas descriptivas de esta variable y su relación con las variables independientes escogidas se presentan a detalle en el numeral anterior.

Habiendo definido las variables enunciadas, se procedió a estimar el siguiente modelo mediante Mínimos Cuadrados Ordinarios (OLS):

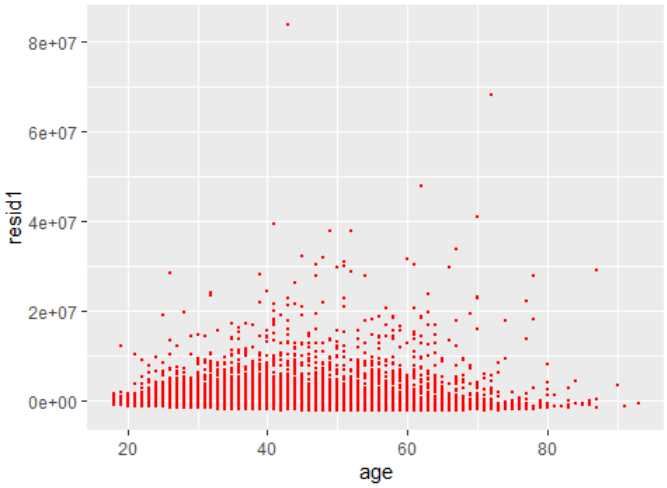
$$Income = \beta_1 + \beta_2 Age + \beta_3 Age^2 + u$$

Obteniendo como resultado de la regresión:

	Dependent variable:
	ingtot
age	89,516.850***
	(9,078.271)
age2	−771.785***
	(105.215)
Constant	−391,598.400**
	(181,969.700)
Observations	16,277
R^2	0.018
Note:	*p<0.1; **p<0.05; ***p<0.01

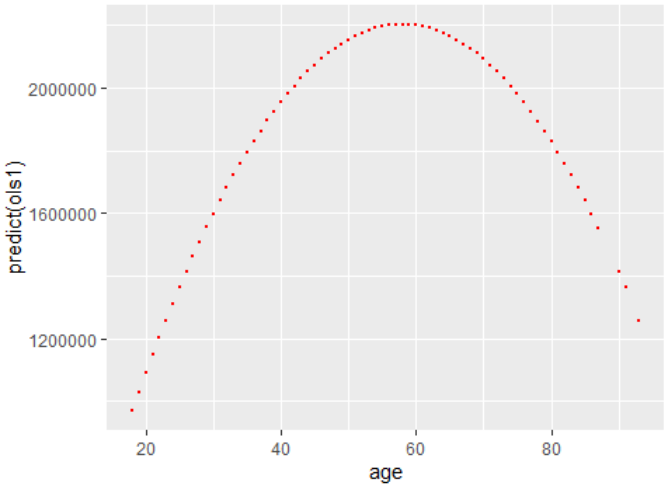
Según este modelo, por cada año adicional de vida, las personas ganan en promedio 89.500 pesos adicionales, ceteris paribus. Edad^2 tiene coeficiente negativo, por lo cual sabemos que esta función no es lineal sino decreciente. Entre paréntesis podemos observar los errores estándares. El R^2 es de 0.018, este corresponde a la fracción de la variabilidad total en la respuesta que es explicada por el modelo, es decir, las variables dependientes propuestas explican alrededor de 1,8% de la variable independiente.

Al analizar el ajuste de este modelo, se realiza un análisis gráfico de los residuales. Se obtiene lo siguiente:



Para interpretar este gráfico se debe tener en cuenta que, si el modelo tuviera un muy buen ajuste, los residuales presentarían un comportamiento “aleatorio”. Sin embargo, en este caso los valores no parecen comportarse de manera aleatoria pues se acumulan casi todos cerca del valor cero. Lo anterior sumado al R^2 de la regresión permite concluir que este modelo no tiene muy buen ajuste con esta muestra. La intuición es que hay otros factores que pesan mucho más en la distribución de ingresos que la edad, los cuales exploraremos más adelante.

Dando continuidad al análisis, se presenta la gráfica de la predicción del perfil de edad ingreso, en la que se evidencia que la variable de edad tiene un comportamiento cuadrático, lo cual se comprueba al observar en la regresión que la variable age^2 es significativa y tiene coeficiente negativo. Se observa que parece haber un pico de edad en la que se maximizan los ingresos, poco antes de los 60 años.



Posterior a la gráfica, se calcula ese pico de edad en el que se maximizan los ingresos, a partir de derivar la función planteada (derivar ingreso con respecto a edad), igualar a cero y despejar el valor de edad que maximiza el ingreso. Esta operación se realiza extrayendo los coeficientes de la regresión que se acaba de correr. Como resultado se obtiene que dicha edad pico es a los 57.99338 años en esta muestra.

Para determinar los intervalos de confianza, se utiliza la metodología de Bootstrap para remuestrear a partir de la muestra que tenemos. Con las estadísticas obtenidas de esta metodología se obtiene el error estándar del remuestreo. Para el caso de todas las variables, el error estándar aumenta por fuera de muestra. Como el error estándar aumenta al hacer Bootstrap concluimos que puede que este modelo tampoco ajuste bien fuera de muestra, pues los errores estándares altos muestran que el promedio de la muestra está muy disperso con respecto al promedio de la población. Los intervalos de confianza construidos, debido a que los errores estándares son tan altos, resultan en que la edad pico variaría entre 29 años y 129 años, lo cual es un indicador adicional de que este modelo no tiene buen ajuste.

4. The earnings gap

Para aproximarse a la brecha de ingresos dependiendo del género, y a realizar el respectivo análisis en la muestra escogida, en primer lugar, se estima el siguiente modelo:

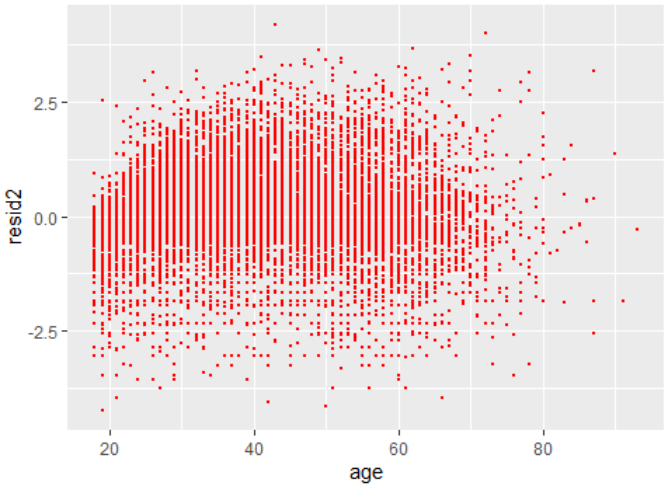
$$log(Income) = \beta_1 + \beta_2 Female + u$$

Obteniendo como resultado:

	Dependent variable:
	logingtot
fem	−0.193***
	(0.014)
Constant	14.064***
	(0.009)
Observations	16,277
R^2	0.012
Note:	*p<0.1; **p<0.05; ***p<0.01

Se puede observar que el coeficiente asociado a ser mujer es significativo y negativo, lo que quiere decir que las mujeres ganan en promedio 19% menos que los hombres. Entre paréntesis podemos observar los errores estándares. El R^2 de este modelo es de 0.012, es decir que el hecho de ser mujer solamente explica el ingreso total en 1,2%.

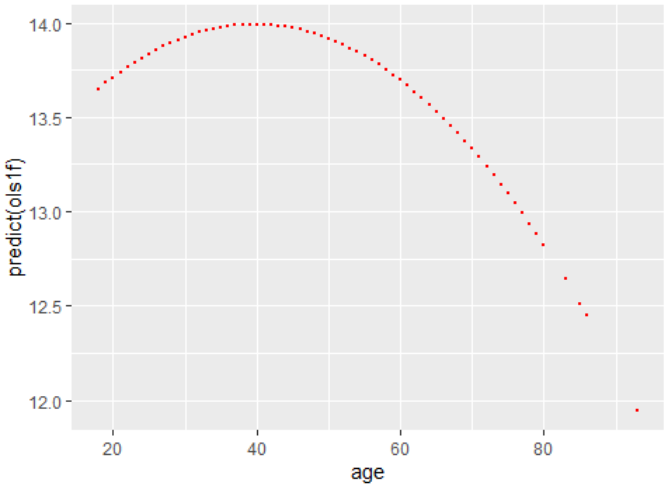
Al analizar el comportamiento de los residuales del modelo, obtenemos lo siguiente:



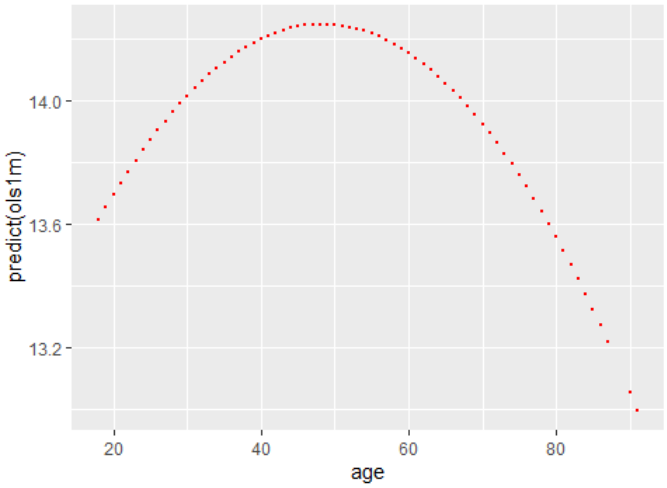
Aquí también observamos que los datos no se distribuyen aleatoriamente, sin embargo y pese al R^2 menor que en el modelo anterior, pareciera que se distribuyen un poco mejor que los residuales de ese primer modelo. En todo caso, el ajuste de este modelo a la muestra tampoco es muy bueno y da cuenta de la importancia de incluir variables de control, lo cual se hará más adelante.

Teniendo en cuenta que se evidencia que para las mujeres de la muestra el coeficiente asociado a su género es negativo y significativo, indicando que obtienen en promedio menos ingresos que los hombres, ceteris paribus, se procede a analizar la predicción edad ingreso de manera separada para hombres y para mujeres, es decir, se generan dos submuestras dependiendo del género de los encuestados. Esta división en dos muestras distintas se realiza para poder analizar de manera diferenciada las edades pico y también para analizar los intervalos de confianza con la misma metodología con la que se realizó para la totalidad de la muestra, ya que de esta manera se nos facilita más la comparación.

En ese sentido, se obtienen las siguientes gráficas:



En esta primera gráfica se observa la edad pico que maximiza ingresos para las mujeres. Aquí vemos que la edad que maximiza el ingreso se observa hacia los 40 años.



En esta segunda gráfica se observa la edad pico que maximiza ingresos para los hombres. Aquí vemos que la edad que maximiza el ingreso se observa hacia los 50 años. Además, entre los 18 años y la edad que maximiza el ingreso, observamos que para los hombres el crecimiento tiene una pendiente mayor, indicando que los hombres estarían aumentando su ingreso total de manera más rápida que las mujeres, cuya pendiente es más horizontal. Con respecto al intercepto, se observa que el de las mujeres es mayor al inicio de la vida laboral, pero precisamente porque la pendiente es menor ellas no alcanzan un nivel tan alto como el de los hombres al llegar a la edad que maximiza su ingreso.

Al igual que en el numeral anterior, se utiliza la misma metodología para calcular exactamente el valor de la edad que maximiza el ingreso para ambos géneros. Para mujeres el valor es de 39.76167. Para hombres el valor es de 48.35461. Se comprueba la intuición obtenida de las gráficas, ya que la edad que maximiza el ingreso para los hombres es mayor y además se corresponde con un ingreso más alto.

Utilizando Bootstrap, tal como en el numeral anterior, se calculan los intervalos de confianza. Según los resultados, se obtiene que la edad pico para las mujeres estaría entre los 22 y los 86 años, y la de los hombres estaría entre los 35 y los 66 años.

Resultan importantes las siguientes observaciones:

- Los intervalos de hombres y de mujeres si se solapan
- En todo caso la edad pico de mujeres continúa siendo menor que la de hombres
- Los intervalos de confianza son demasiado grandes
- Para el caso de los hombres el intervalo de confianza es más acotado, esto daría cuenta de que los datos para los hombres tienen mejor ajuste
- Lo anterior lo comprobamos al ver que el r^2 del modelo de solo mujeres es de 0.23 y el de solo hombres es de 0.45. Esto quiere decir que en el caso de las mujeres se requieren aún más variables explicativas y de control que en el caso de los hombres para explicar el ingreso.

Ya que en los análisis realizados hasta ahora se evidencia la necesidad de incorporar variables de control, se plantea incorporar variables relacionadas con características del trabajo y de los trabajadores para intentar analizar si al incorporarlas en el modelo, la brecha de ingresos entre hombres y mujeres se modifica o desaparece.

Teniendo en cuenta el análisis de variables realizado al comienzo, se plantean modelos de Mínimos Cuadrados Ordinarios que incorporan cada vez más variables de control. Los resultados principales se presentan en la siguiente tabla.

	Dependent variable:			
	logingtot			
	(1)	(2)	(3)	(4)
age	0.060***	0.058***	0.046***	0.043***
	(0.003)	(0.003)	(0.002)	(0.002)
age2	−0.001***	−0.001***	−0.0004***	−0.0004***
	(0.00003)	(0.00003)	(0.00003)	(0.00003)
fem	−0.206***	−0.249***	−0.307***	−0.290***
	(0.014)	(0.012)	(0.011)	(0.011)
maxEducLevel		0.319***	0.165***	0.109***
		(0.005)	(0.006)	(0.005)
oficio			−0.007***	−0.004***
			(0.0002)	(0.0002)
formal			0.616***	0.587***
			(0.012)	(0.011)
estrato2				0.053***
				(0.018)
estrato3				0.174***
				(0.019)
estrato4				0.703***
				(0.027)
estrato5				0.959***
				(0.041)
estrato6				1.380***
				(0.037)
Constant	12.863***	10.823***	11.969***	12.178***
	(0.059)	(0.063)	(0.063)	(0.060)
Observations	16,277	16,276	16,276	16,276
R^2	0.037	0.210	0.362	0.445
Note:				*p<0.1; **p<0.05; ***p<0.01

Se observa que para todos los casos la variable asociada a ser mujer tiene coeficientes negativos y significativos, y pese a la inclusión de variables de control, el valor del coeficiente no presenta grandes saltos. La inclusión de variables de control sí resulta importante en lo relacionado con el ajuste del modelo, ya que se observa que el R^2 aumenta progresivamente conforme se complejiza el modelo.

Habiendo realizado el ejercicio de incorporar variables de control, se realiza un ejercicio para comprobar, mediante FWL Theorem, si el “desaparecer” estas variables de control tiene algún efecto en el coeficiente asociado a ser mujer.

En primer lugar, se realiza el análisis solo incorporando la variable de género y la variable de estrato. Hay que tener cuidado porque estrato podría ser una variable endógena (menor ingreso lleva a elegir menor estrato de residencia). Sin embargo, vivir en barrios de menores estratos puede estar generando dificultad de acceso a trabajos mejor pagos, por lo tanto,

afectando el ingreso. Teniendo en cuenta que, de ser cierto el último supuesto, estrato se convierte en una variable que explica buena parte de la diferencia en ingreso, se realiza el siguiente ejercicio:

	Dependent variable:	
	logingtot	logingtot
	(1)	(2)
fem	−0.193***	
	(0.014)	
fem		−0.226***
		(0.012)
estrato2		0.190***
		(0.021)
estrato3		0.469***
		(0.021)
estrato4		1.266***
		(0.029)
estrato5		1.533***
		(0.046)
estrato6		1.997***
		(0.041)
Constant	14.064***	13.661***
	(0.009)	(0.019)
Observations	16,277	16,277
R^2	0.012	0.246
Note:	*p<0.1; **p<0.05; ***p<0.01	

En la tabla se observan los resultados de dos regresiones, ambas con el logaritmo del ingreso como variable dependiente. El primer modelo solo tiene como variable independiente la variable de género, y el segundo controla por estrato, según la intuición que se explicó anteriormente.

Mediante el FWL Theorem se realiza la regresión de un modelo de residuales en residuales, obteniendo lo siguiente:

	Dependent variable:		
	logingtot	logingtot	res_y_e
	(1)	(2)	(3)
fem	−0.193***		
	(0.014)		
fem		−0.226***	
		(0.012)	
estrato2		0.190***	
		(0.021)	
estrato3		0.469***	
		(0.021)	
estrato4		1.266***	
		(0.029)	
estrato5		1.533***	
		(0.046)	
estrato6		1.997***	
		(0.041)	
res_x_e			−0.226***
			(0.012)
Constant	14.064***	13.661***	0.000
	(0.009)	(0.019)	(0.006)
Observations	16,277	16,277	16,277
R^2	0.012	0.246	0.021
Note:		*p<0.1; **p<0.05; ***p<0.01	

En el modelo de residuales en residuales vemos que el coeficiente asociado a ser mujer no cambia (en este tercer modelo lo debemos analizar en la variable res_x_e), pero el r^2 aumenta de 0.012 a 0.021, por lo cual se puede decir que el modelo ajusta mejor. Este mejor ajuste estaría indicando que el problema no es de selección y que efectivamente hay una brecha de ingresos para las mujeres.

A partir del ejercicio anterior se decide correr un modelo más complejo que incorpore más variables, en este caso las de edad y edad^2, con el fin de comprobar si el FWL Theorem continúa aplicando al analizar modelos de residuales en residuales con más de una variable. Utilizando la misma metodología, se obtiene como resultado lo siguiente:

	Dependent variable:		
	logingtot	logingtot	res_y_e
	(1)	(2)	(3)
fem	−0.193***		
	(0.014)		
fem		−0.240***	
		(0.012)	
age		0.053***	
		(0.003)	
age2		−0.001***	
		(0.00003)	
estrato2		0.181***	
		(0.020)	
estrato3		0.463***	
		(0.021)	
estrato4		1.255***	
		(0.029)	
estrato5		1.531***	
		(0.046)	
estrato6		2.000***	
		(0.041)	
res_x_e			−0.240***
			(0.012)
Constant	14.064***	12.704***	0.000
	(0.009)	(0.054)	(0.006)
Observations	16,277	16,277	16,277
R^2	0.012	0.267	0.025
Note:		*p<0.1; **p<0.05; ***p<0.01	

Se observa que, al incorporar la variable categórica de estrato, así como las variables de edad y edad^2 utilizadas en análisis anteriores, en el modelo de residuales en residuales nuevamente se obtiene el mismo coeficiente para la variable de género, el cual es negativo y significativo, con la particularidad de que después de aplicar el FWL Theorem aumenta el R^2.

5. Predicting earnings

a.

i. Se estima un modelo únicamente con la constante, en este intercepto se encuentra la media del logaritmo del ingreso.

	Dependent variable:
	logingtot
Constant	13.974***
	(0.008)
Observations	11,394
R^2	0.000
Adjusted R^2	0.000
Residual Std. Error	0.873 (df = 11393)
Note:	*p<0.1; **p<0.05; ***p<0.01

ii. Se corren los modelos anteriores. se estiman los modelos de perfil de ganancias por edad, el de efecto de ser mujer y el que incluye ambos.

	Dependent variable:			
	logingtot			
	(1)	(2)	(3)	(4)
age		0.003***		0.003***
		(0.001)		(0.001)
poly(age, 2)1				
poly(age, 2)2		−13.823***		−14.356***
		(0.862)		(0.857)
fem			−0.195***	−0.206***
			(0.016)	(0.016)
Constant	13.974***	13.837***	14.064***	13.938***
	(0.008)	(0.025)	(0.011)	(0.026)
Observations	11,394	11,394	11,394	11,394
R^2	0.000	0.025	0.012	0.039
Adjusted R^2	0.000	0.025	0.012	0.038
Residual Std. Error	0.873 (df = 11393)	0.862 (df = 11391)	0.867 (df = 11392)	0.856 (df = 11390)
F Statistic		145.629*** (df = 2; 11391)	142.929*** (df = 1; 11392)	153.025*** (df = 3; 11390)
Note:				*p<0.1; **p<0.05; ***p<0.01

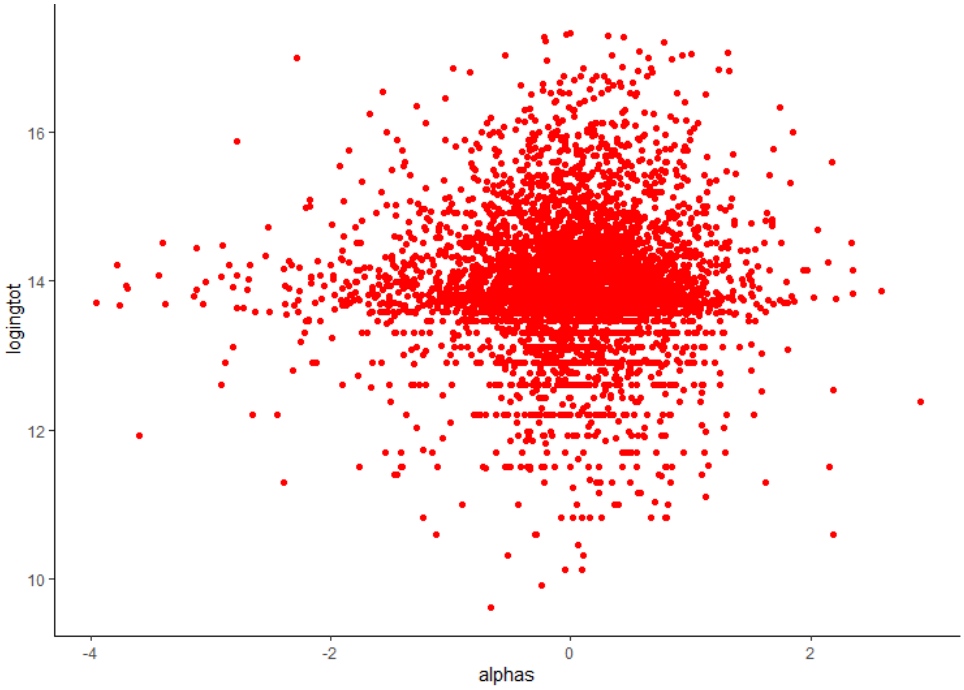
iii. Se comienza a complejizar el modelo por medio de la adición y transformación de variables X:

	Dependent variable:				
	logingtot				
	(1)	(2)	(3)	(4)	(5)
fem	−0.250***	−0.260***	−0.324***	0.044	0.099*
	(0.014)	(0.014)	(0.014)	(0.029)	(0.054)
age	0.007***	0.006***	0.006***	0.007***	0.007***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
poly(age, 2)1					
poly(age, 2)2		−11.951***	−12.390***	−12.027***	−12.028***
		(0.730)	(0.717)	(0.711)	(0.711)
maxEducLevel	0.221***	0.213***	0.165***	0.161***	0.160***
	(0.007)	(0.007)	(0.007)	(0.007)	(0.007)
estrato1	0.287***	0.289***	0.252***	0.247***	0.256***
	(0.008)	(0.007)	(0.008)	(0.008)	(0.010)
oficio			−0.006***	−0.004***	−0.004***
			(0.0003)	(0.0003)	(0.0003)
fem:oficio				−0.008***	−0.008***
				(0.001)	(0.001)
fem:estrato1					−0.017
					(0.014)
Constant	11.782***	11.838***	12.539***	12.453***	12.429***
	(0.048)	(0.047)	(0.058)	(0.057)	(0.061)
Observations	11,393	11,393	11,393	11,393	11,393
R^2	0.289	0.305	0.330	0.343	0.343
Adjusted R^2	0.289	0.305	0.330	0.342	0.342
Residual Std. Error	0.736 (df = 11388)	0.728 (df = 11387)	0.714 (df = 11386)	0.708 (df = 11385)	0.708 (df = 1138
F Statistic	1,155.900*** (df = 4; 11388)	999.917*** (df = 5; 11387)	936.041*** (df = 6; 11386)	847.412*** (df = 7; 11385)	741.694*** (df = 8; 1
Note:					*p<0.1; **p<0.05; ***p

iv. Una vez se corren todos los modelos se encuentra que el modelo 9 es el que presenta un MSE más pequeño, por lo cual se selecciona para computar el peso estadístico de cada una de las observaciones en el siguiente punto.

mse1	3187321032652.28
mse2	0.774354980001545
mse3	0.781800494475404
mse4	0.763863716358705
mse5	0.557425955878508
mse6	0.546570863672189
mse7	0.528099249274231
mse8	0.519938654104733
mse9	0.519807276669949

v. Una vez se selecciona el modelo 9 de acuerdo con los resultados obtenidos en el punto anterior, se encuentra que efectivamente se encuentran outliers dentro de las observaciones lo cual puede llevar a pensar que las personas están reportando menores ingresos que los reales.



b. Al correr el modelo por medio del método k-fold, se obtiene en este caso que es el modelo 8 el modelo con el menor MSE. En este caso, se escogerá este modelo para la aplicación del LOOCV, debido a que este obtiene un Lamda mucho más ajustado al correr múltiples alternativas por medio del programa. De igual forma, se está obteniendo de una muestra más robusta, lo cual genera mucha más confianza.

```
> model8cv
Linear Regression

16277 samples
    5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 13020, 13020, 13021, 13022, 13021
Resampling results:

  RMSE       Rsquared   MAE
  0.7117622  0.3422278  0.5172198

Tuning parameter 'intercept' was held constant at a value of TRUE
> model9cv
Linear Regression

16277 samples
    5 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 13022, 13020, 13021, 13020, 13021
Resampling results:

  RMSE       Rsquared   MAE
  0.7120185  0.3419766  0.5173301
```