MECA 4107 Big Data – Problem Set 1
María Paula Nieto (201612379), Juan Diego López (201425104), Daniela Jaime (201531520)

Problem Set 1: Predicting Income MECA 4107

El código utilizado para desarrollar los modelos predictivos, el *webscrapping* inicial y la generación de los gráficos y tablas se encuentra en el repositorio de GitHub llamado "Problem Set 1". El repositorio mencionado se encuentra en el siguiente link: https://github.com/marianieto198/Problem-Set-1.git

1. Data acquisition

El proceso de adquisición de datos, se deriva de un ejercicio de *webscrapping* a 10 chunks de código correspondientes a los datos de la GEIH para la ciudad de Bogotá en el año 2018. El pseudocode se encuentra en el repositorio de Github con detalle, sin embargo, a continuación se plasman los pasos principales para realizar el procedimiento:
1. En un primer momento se explora el link para verificar la manera en la que está presentada la información y verificar si hay algún tipo de restricción para realizar *webscrapping* (comando Robot.txt).
2. Ahora, con el uso del click derecho se inspecciona la página y se obtiene el url de acceso para el código html desde la ventana de network.
3. Teniendo en cuenta que son 10 chunks separados de información, se realiza un *loop* para realizar webscrapping sobre cada chunk de valores.
3.1. Se define entonces la base vacía que será llenada por la información scrapeada en los url, y para cada chunk de información se extrae un temporal en donde se irá almacenando la información de cada chunk. Para este paso se utiliza el comando "for" junto con el comando "paste0" para concatenar los diferentes url en un único elemento.
3.2. Se utiliza el comando rbind para pegar las filas de información almacenadas en cada archivo temporal y unificarlas en un único elemento.

Así, se obtiene una base de datos con 178 variables y 32.177 observaciones. Se evidencia que no hay restricciones para acceder a los datos y realizar el *scrapping*. En este caso, la muestra de GEIH se presenta en una tabla fija, por lo que es más fácil realizar el scrape y se facilita la lectura y extracción de los datos.

2. Data Cleaning

Para el proceso de limpieza y análisis de la composición de la base de datos se realizarán los siguientes pasos:

1. Se verifica si el tipo de variables fue leído correctamente (numéricos, string, categóricas) y se delimita la muestra de acuerdo con las indicaciones del ejercicio.
2. Analizar los missing values y verificar si se imputan o eliminan.
3. Si se considera pertinente, se normalizan las variables.
4. Verificar por asimetrías en la información y verificar tratamiento de datos atípicos.
5. Se construyen tablas, gráficos y estadísticas descriptivas con las variables de interés.

A continuación, se mostrará el desarrollo de los pasos mencionados con anterioridad.

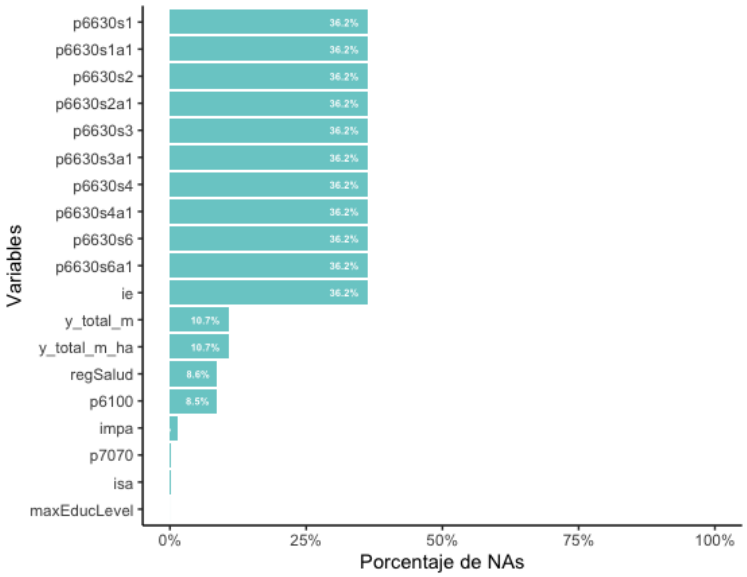
1. **Se verifica si el tipo de variables fue leído correctamente y se delimita la muestra.**

Se observa que la mayoría de las variables fueron leídas como enteros, por lo que se realiza la definición como categóricas de 85 variables. Asimismo, de acuerdo con las instrucciones del ejercicio se mantiene en la muestra las observaciones que correspondan a personas ocupadas de 18 años o mayores. Hay otras variables que pueden definir la ocupación, como por ejemplo si el individuo está desempleado o empleado (dsi) y si pertenece a la población económicamente activa (pea), sin embargo, se elegirá la variable de ocupación (ocu) para delimitar la muestra por cuanto dicha variable contempla a las personas que trabajaron por lo menos una hora remunerada, las personas que cuentan con un trabajo con o sin remuneración. Por su parte, PEA y dsi no se eligieron porque contemplan a menores de edad desde 12 años en zonas urbanas y 10 años en zonas rurales que tienen o están buscando empleo por lo que no incluyen los individuos realmente ocupados. Al realizar la delimitación de las observaciones a ocupados mayores de 18 años en Bogotá, la base queda con 16397 observaciones.

2. Analizar los missing values y verificar si se imputan o eliminan.

Inicialmente, se observa que el 45.2% de las entradas están vacías lo que podría deberse a que algunas variables no aplican para todos los individuos. No obstante, se identifica que variables de interés como el sexo, si está ocupado o no, las horas laborales y la escolaridad no tienen NAs. Después de organizar las variables según el porcentaje de sus observaciones con NAs, se grafican las variables con un menor número de NAs para identificar si se pueden imputar o no (se imputarán aquellas variables con NAs en el 5% de sus observaciones o menos, el resto se eliminarán de la muestra).

Figura 1. Porcentaje de NAs en variables de interés



En la Figura 1 se observa que las variables con NAs en menos del 5% del total de sus observaciones son el ingreso monetario de la primera y segunda actividad, las ganancias del mes pasado y el máximo nivel de educación alcanzado. A partir de lo anterior, se procede a eliminar las variables con más del 5% de sus observaciones como *missing* y a imputar las variables mencionadas con anterioridad. La razón por la cuál se eliminan las variables con más del 5% es que aquellas observaciones sin información pueden producir sesgos y reducir el poder explicativo de los modelos. Asimismo, varias de las variables que se eliminarán tienen una alta probabilidad de no contar con información bajo la causal de "Informarte inadecuado", es decir, cuando la persona entrevistada no está en posibilidad de proporcionar la información que se le demanda (Cochran, 1977). Como por ejemplo, en el caso de no contar con información sobre si recibió ingresos por prima de servicios, de navidad, viáticos, o preguntas

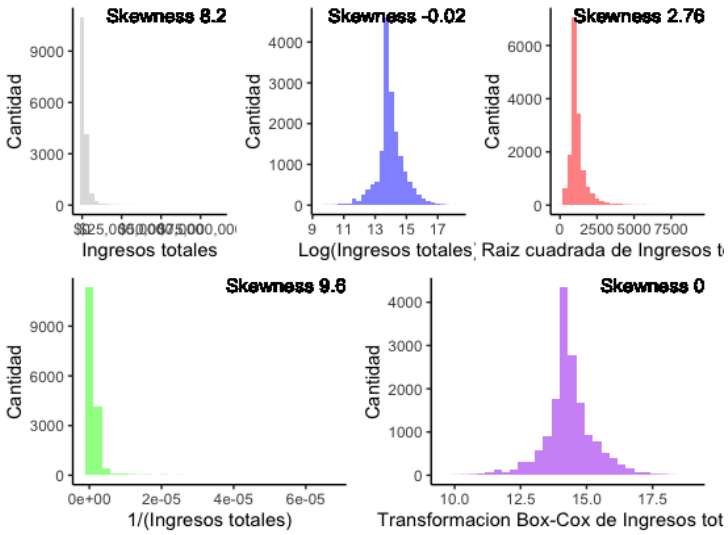
para desempleados cuando el individuo está empleado, entre otros. Después de imputar la muestra queda con 62 variables, dado que todas las variables a imputar son numéricas a excepción del nivel educativo, para las numéricas los NAs se reemplazarán con la mediana y la educación con la moda.

3. **Si se considera pertinente, se normalizan las variables:** El código para realizar la normalización se agregó a manera descriptiva, sin embargo, para facilitar la interpretación del coeficiente en los próximos ejercicios no se normalizarán los datos.

4. **Verificar por asimetrías en la información y verificar tratamiento de datos atípicos.**

Ahora, se verifica el *skewness* para analizar qué variables cuentan con una mayor asimetría, en este caso son: p7500s2a1, p7510s5a1 y iof1; los cuales corresponden a valores recibidos por intereses de préstamos, pensiones, entre otros. Un *skewness* alto se puede deber a que cuentan con varias observaciones de valor cero. Estas variables no serán relevantes para el análisis posterior por lo que no se tratarán estos valores atípicos.
En cuanto a la variable de ingreso total el análisis da cuenta que tiene un *skewness* de 8.2, por lo que se verificarán posibles transformaciones que reduzcan esta asimetría para una futura regresión. Las demás variables numéricas de interés como la edad y las horas trabajadas cuentan con skewness menores a 1.
Se utiliza entonces la transformación box-cox para hallar un lambda que haga a la variable de ingreso total más simétrica. Las transformaciones realizadas se muestran en la siguiente figura:

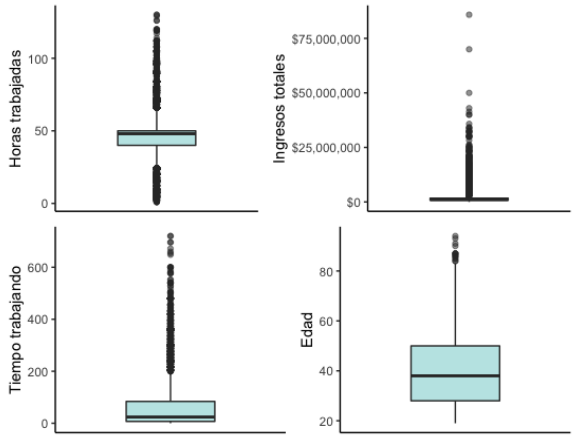
Figura 2. Transformaciones sobre la variable Ingresos Totales



Se observa que las transformaciones que generan una mayor simetría son el logaritmo de los ingresos totales y la transformación box-cox, cuyo lambda es igual a 0.003 (Figura 2).

Para analizar posibles datos atípicos se realizarán gráficos de tablas y bigotes sobre las variables de interés. Asimismo, se verificarán las correlaciones existentes entre las variables. En primer lugar, la teoría económica nos dice que el ingreso de un trabajador depende principalmente de dos factores. Por un lado, su productividad y, en segundo lugar, de las condiciones del mercado laboral y su contexto (Mincer, 1962). A continuación se presenta la distribución de cuatro variables relacionadas con estos postulados:

Figura 3. Gráficas de tablas y bigotes de variables de interés



La variable de horas trabajadas cuenta con una distribución de los datos cuya mediana se encuentra aproximadamente en 50 horas trabajadas, la edad mediana es de aproximadamente 40 años. Si bien todas las variables de interés graficadas cuentan con datos atípicos, es necesario analizar si estos son incorrectos (como edades o ingresos negativas) o si realmente tiene sentido lo extremas que son. Para el caso de los ingresos, existen dos valores que representan ingresos mayores a $50 millones, lo cual para el contexto colombiano sí tiene coherencia. Sobre la edad, el valor máximo es de 94 años por lo que no representan un valor imposible o incorrecto. Para el caso del tiempo que el individuo lleva trabajando en el mismo oficio o trabajo, la unidad de medición es la cantidad de meses. El número máximo es 720, correspondiente a trabajar por más de 60 años seguidos. Lo anterior comparándolo con la distribución de la variable edad, guarda coherencia con lo que podría alcanzar un individuo.

5. **Se construyen tablas, gráficos y estadísticas descriptivas con las variables de interés.**

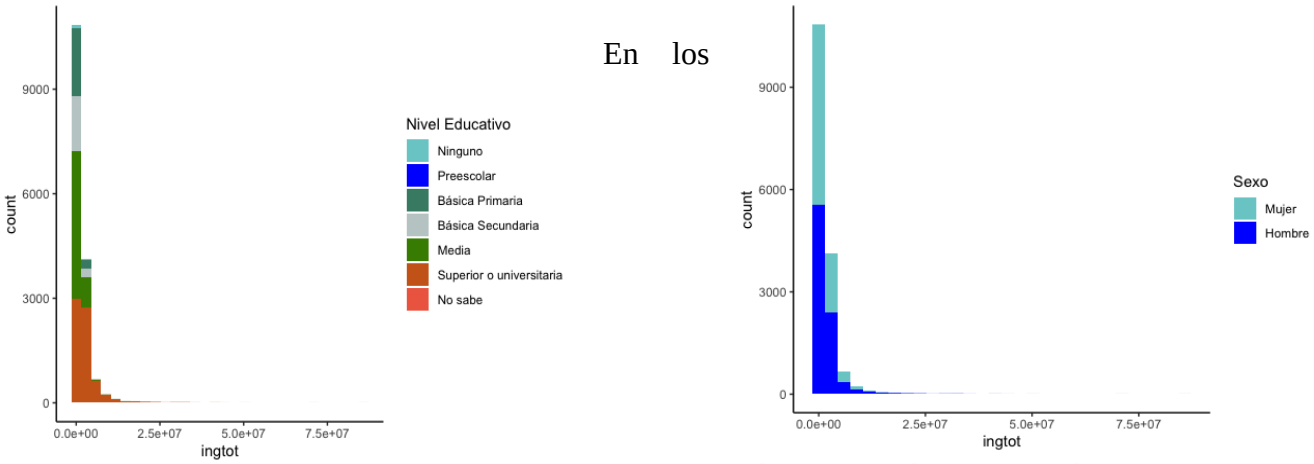
Las variables que se considera pueden influir en el ingreso total es si se encuentran en una zona urbana-rural, el sexo, el número de horas trabajadas, el nivel máximo de educación alcanzado, el oficio y la experiencia en el oficio actual. Lo anterior, a partir de la teoría económica y los modelos de Becker (1964) y Mincer (1962), sin embargo, adicionalmente se observarán los datos para verificar si existe de manera preliminar alguna correlación entre los ingresos totales y las variables destacadas. No obstante, dado que todas las observaciones se encuentran en la zona urbana, no se podrá analizar cómo influye la variable de clase en el modelo.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Ingresos Totales	16397	179040 4.65	267959 4.778	0	8e+05	173654 4.333	8583333 3.333
Horas trabajadas	16397	47.441	15.616	1	40	50	130
Experiencia	16397	64.244	89.724	0	7	84	720
Edad	16397	39.626	13.39	19	28	50	94

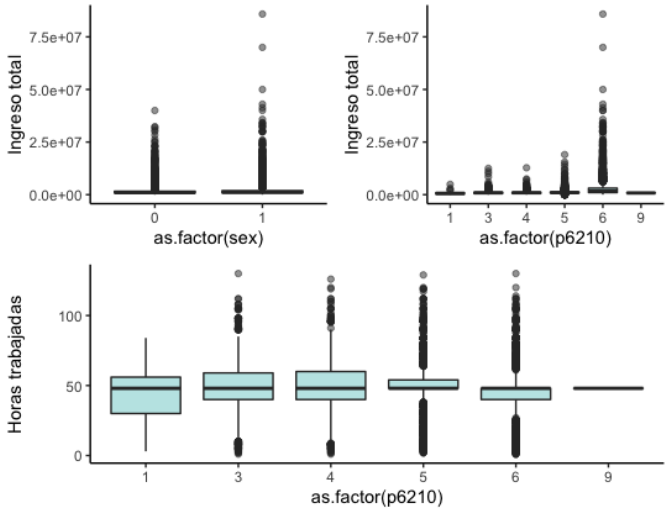
En la tabla que se observa de manera previa, se muestran las estadísticas descriptivas para las variables de interés de los próximos modelos. En primer lugar, en promedio los individuos de la muestra tienen 40 años y cuentan con un ingreso total de $ 1.790.404 COP. El tiempo promedio en el que los individuos de la muestra han trabajado en el mismo oficio es de 5 años aproximadamente y trabajan de manera semanal 47 horas. Las desviaciones estándar nos indican que hay una alta variación de los datos de los meses trabajados alrededor de la media,

al igual que para los ingresos totales, es decir los datos de estas variables cuentan con una dispersión alta.

Los siguientes histogramas muestran la distribución de la variable de ingreso total por las categorías de sexo y nivel educativo. De manera preliminar se observa que las personas con un mayor nivel educativo son las que tienen mayores ingresos. Mediante la distribución del sexo se observa que una mayor cantidad de mujeres tienen ingresos menores que los hombres. Esto se explica, pues históricamente ha implicado una discriminación negativa hacia las mujeres en el mercado laboral, al mismo tiempo que se le dan mayores responsabilidades en el hogar, lo que puede afectar sus ingresos totales (Mincer, 1962; Cruces & Galiani, 2007).

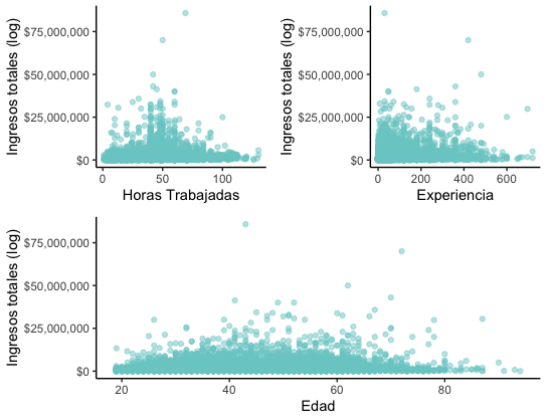


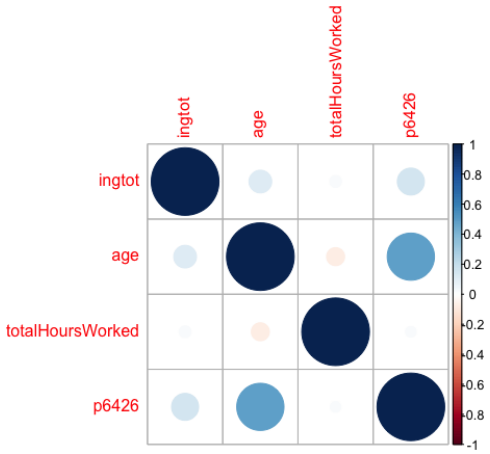
En los siguientes diagramas de cajas y bigotes se observa la distribución del ingreso por niveles de educación y sexo. De primera mano se observa que los hombres cuentan con una mayor dispersión en el ingreso total, contando con los valores más grandes y con una mayor cantidad de valores atípicos. Por otro lado, las personas que cuentan con una mayor dispersión en los ingresos totales y con ingresos más altos, son las personas que cuentan con una educación superior o universitaria. Finalmente, la mediana de las horas trabajadas en la muestra es muy parecida para todos los niveles de educación, sin embargo, la gran mayoría de valores atípicos tanto superiores a la mediana como inferiores, se encuentran en los niveles de educación media y superior o universitaria.



Correlaciones entre las variables de interés

En un primer momento, de manera gráfica se observa que no existe una correlación clara entre los ingresos totales y las variables de interés, lo cual es confirmado por los valores de la tabla de correlación. En este sentido, si bien la edad está correlacionada con la experiencia, tienen una importancia particular debido a que las personas, en general, presentan cambios en sus intereses y en sus metas a medida que crecen. En teoría, una persona joven puede tener menos experiencia, pero está dispuesta a ofrecer más horas de trabajo con el objetivo de cumplir sus ambiciones y aumentar su ingreso total. Por otro lado, una persona de mayor edad tiene más experiencia y conocimientos, pero puede no estar dispuesta a ofrecer la misma cantidad de horas en el mercado laboral debido a que su salud no lo permite o por querer más tiempo con su familia (Kaufman, 2008).
A continuación se evidencia gráficamente la magnitud de las correlaciones siendo el color azul y rojo oscuro, una correlación alta entre las variables ya sea positiva o negativa. Las variables de número de meses en un mismo oficio y la edad representan una correlación de 0.49, el cual es explicado porque entre más años tiene la persona, más años de experiencia puede tener. Asimismo, hay una correlación de 0.11 entre la edad y los ingresos totales, mientras que la correlación entre los meses en un mismo oficio y el ingreso total es de 0.16. Estas correlaciones son de una pequeña magnitud y se sustentan por lo observado gráficamente.

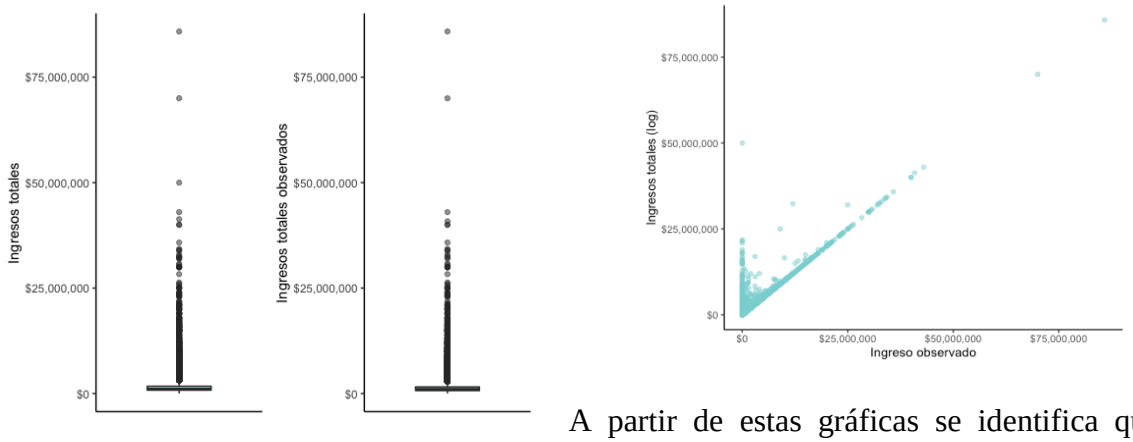




Ingresos totales	Edad	Horas de trabajo	Experiencia
1			
0.11	1		
0.03	-0.07	1	
0.16	0.49	0.02	1

3. *Age-earnings profile*

Para elegir la variable de ingreso que mejor representa las ganancias totales de los trabajadores se tienen que tener en cuenta tres variables: el ingreso total observado que es la suma de los ingresos percibidos en diferentes fuentes, el ingreso total imputado que es la suma de cada una de las fuentes de ingresos imputadas a los registros faltantes y el ingreso total es la suma de cada una de las fuentes de ingresos tanto observadas como imputadas. Teniendo en cuenta que el ingreso total imputado se eliminó por cuenta de la cantidad de NAs se discutirá la elección entre las variables de ingreso total e ingreso total observado como la mejor representación de las ganancias de los individuos. El ingreso total entonces, comprende el ingreso total observado y los ingresos contemplados en la variable ingtotes. A partir de esto, se extrajo la correlación entre estas dos variables siendo de 0.92, por lo que cuentan con valores similares. Las siguientes gráficas es para observar de qué manera difieren las variables:

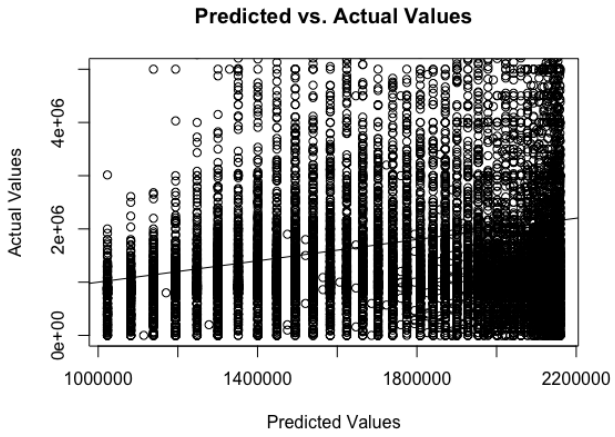


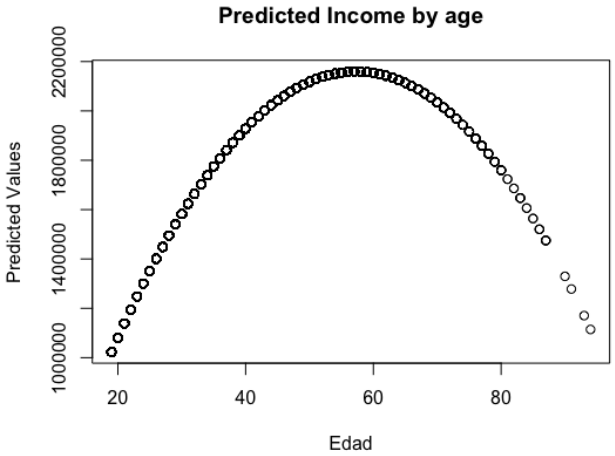
A partir de estas gráficas se identifica que el comportamiento y distribución de las variables es similar, sin embargo, hay varias observaciones dentro de la variable de ingreso observado que son iguales a cero a diferencia de sus valores en la variable de ingresos totales. Lo anterior, se asocia con los individuos que cuentan solamente con ingresos imputados, por lo que con el fin de contemplar todos los ingresos tanto imputados como observados y no perder las ganancias obtenidas por los individuos con solo ingresos imputados se utilizará la variable de ingresos totales.

Ahora, se estimará el siguiente modelo: Income = β1 + β2Age + β3Age2 + u, mediante OLS. Los resultados son:

	Dependent variable:
	Ingreso Total
Edad	87,464.690***
	(9,122.961)
$Edad^2$	−757.561***
	(105.089)
Constant	−350,107.700*
	(184,187.800)
Observations	16,397
R^2	0.016
Adjusted R^2	0.016
Residual Std. Error	2,658,319.000 (df = 16394)
F Statistic	132.749*** (df = 2; 16394)
Note:	*p<0.1; **p<0.05; ***p<0.01

El ajuste del modelo se analizará mediante el R cuadrado, el cual es de 0.016 y representa que el modelo explica un 1.6% de la variabilidad de las observaciones. El poco ajuste del modelo a los datos también se evidencia en el siguiente gráfico, en el que la mayoría de los datos observados de los ingresos totales no están cerca de los valores predichos por el modelo de regresión lineal. Asimismo, las conclusiones preliminares de este modelo es que contemplando un 5% de significancia, la edad es significativa para los ingresos totales de los individuos. Es decir, un aumento de un año en la edad del individuo aumentará en promedio sus ingresos en $87.464 COP. Asimismo, gracias al signo del coeficiente de la edad al cuadrado se respalda que los ingresos tienen rendimientos decrecientes a medida que aumenta la edad. A continuación se grafican los valores predichos de ingresos totales por la edad de los individuos que se obtuvieron de la estimación del modelo destacado con anterioridad:





Preliminarmente, se puede observar que la "edad
pico" en la que se obtienen los ingresos más altos se encuentra cercano a los 60 años. Por lo tanto, de acuerdo con la evidencia empírica de la economía laboral, desde esa edad la tendencia de los ingresos totales es decreciente. Más específicamente, la "edad pico" obtenida después de derivar el modelo econométrico e igualar a cero es de 57.7 años. Ahora bien, se utilizará un "bootstrap" para estimar los errores estándar y los intervalos de confianza, los resultados son los siguientes:

- Media: 87587.48
- Error Estándar: 12981.24
- Intervalos de confianza: 2.5% 97.5%

60829.97 111820.64

La variabilidad cambia a 12.981, que es distinta al 9.122 del modelo de regresión lineal inicial. Asimismo, indica que con un 95% de confianza el verdadero valor está entre 60.829 y 111.820. Gracias al bootstrap, no se está suponiendo que el modelo definido anteriormente está correctamente especificado, por lo que la diferencia entre errores estándar puede estar sugiriendo que el modelo lineal no es el más adecuado para estimar los coeficientes y para explicar completamente la variabilidad de los datos.

4. The earnings gap

En primer lugar, para indagar acerca de la brecha de ingresos entre hombres y mujeres, se estima el siguiente modelo: (Income) = β1 + β2Female + u, mediante OLS. Los resultados son:

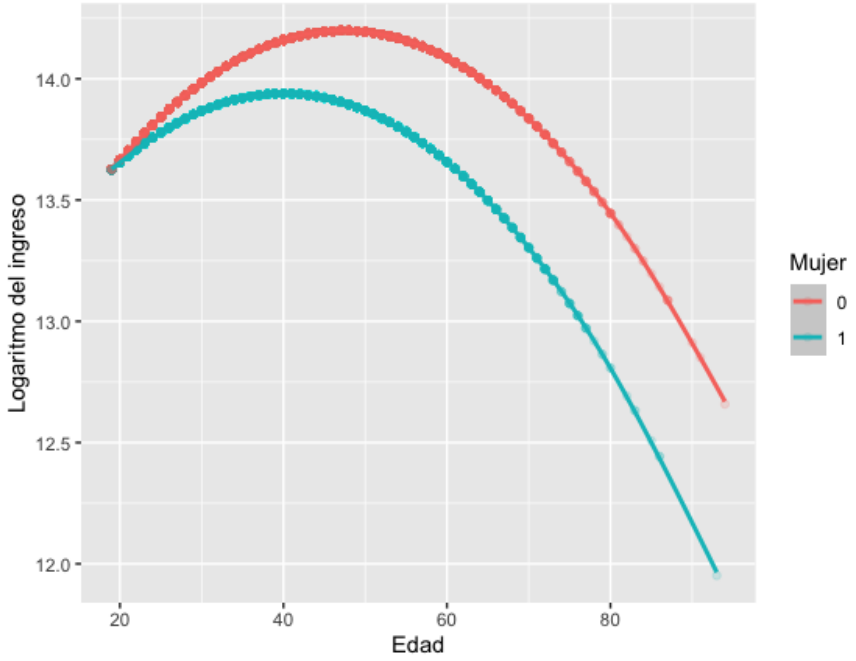
	Dependent variable:
	logIngreso
female	−0.203***
	(0.019)
Constant	14.023***
	(0.013)
Observations	16,397
R^2	0.007
Adjusted R^2	0.007
Residual Std. Error	1.193 (df = 16395)
F Statistic	118.095*** (df = 1; 16395)
Note:	*p<0.1; **p<0.05; ***p<0.01

En este primer modelo, es posible identificar que el efecto el sexo sobre el ingreso es significativo y tiene un impacto negativo. Específicamente, el coeficiente β2 puede ser interpretado como la penalidad que reciben las mujeres en razón de su sexo sobre su ingreso total. Específicamente, este modelo sugiere que el hecho de ser mujer está asociado con una reducción en promedio de 20.3% en el ingreso a comparación de los hombres. En cuanto ajuste del modelo, se observa un R^2 de 0.007 lo que implica que solo el 0.7% de la varianza en el ingreso es explicado por la varianza en la variable *Female*. En ese sentido, esta medida sugiere que el modelo planteado tiene un ajuste limitado y es necesario mejorarlo para tener una mejor aproximación a la brecha salarial entre hombres y mujeres.

Ahora, para continuar indagando acerca de las diferencias entre hombres y mujeres en ingresos, se muestra el perfil edad-ingresos por género predicho. En este caso, se decidió estimar este modelo teniendo en cuenta la interacción entre la variable de mujer con la edad para tener en cuenta si la edad tiene un efecto distinto dependiendo del género de la persona. Así, los resultados del modelo sugieren que para las mujeres la edad tiene un efecto negativo sobre sus ingresos.

	Dependent variable:
	logIng
female1	0.196***
	(0.058)
age	0.067***
	(0.004)
age2	−0.001***
	(0.00005)
agefem	−0.010***
	(0.001)
Constant	12.605***
	(0.087)
Observations	16,397
R^2	0.024
Adjusted R^2	0.024
Residual Std. Error	1.183 (df = 16392)
F Statistic	99.830*** (df = 4; 16392)
Note:	*p<0.1; **p<0.05; ***p<0.01

En términos gráficos, el perfil de edad-ingresos por género demuestra que si bien el comportamiento de ingreso con la edad es similar tanto para los hombres como para las mujeres, existe una brecha en los ingresos que inicia alrededor de los 20 años y crece a medida que aumenta la edad. En ese sentido, en términos de las pendientes, se observa que la pendiente para los hombres tiene un crecimiento más pronunciado y que, luego de alcanzar el *Peak Age*, tiene un decrecimiento menos pronunciado en los ingresos. Por el otro lado, para la pendiente de la curva de las mujeres tiene un crecimiento menos pronunciado antes del *Peak Age* *y* un decrecimiento acelerado después. De esta manera, este análisis contribuye a evidenciar que sí existe una brecha de género y que esto se mantiene, o incluso, se profundiza teniendo en cuenta los efectos de la edad. En cuanto al intercepto, es posible observar que las mujeres inician con intercepto mayor al de los hombres, pero a medida que aumenta la edad el intercepto disminuye hasta que alrededor de los 18 años los interceptos son iguales y después los hombre tienen un intercepto mayor cuando ingresan al mercado laboral.



Ahora bien, para complementar el análisis anterior, se estimará la "edad pico" para hombres y mujeres y se complementará con una estimación de errores estándar e intervalos de confianza con Bootstrap. Utilizando el modelo estimado, se encuentra que el Peak Age para los hombres es 47.40 años y el Peak Age para las mujeres es 40.06. Al realizar el Bootstrap, es posible identificar que los errores estándar disminuyen. Así mismo, en la siguiente gráfiac es posible identificar que los intervalos de confianza no se cruzan.

Intervalos de confianza	2.5%	97.5%
IC_FMod2	0.082	-0.001
IC_AgeMod2	0.313	-0.001
IC_Age2Mod2	0.057	-0.013
IC_AgeFemMod2	0.080	-0.013

Al considerar la brecha salarial por género, se ha señalado que ésta podría responder a un problema de selección y no a la discriminación. Específicamente, se ha sugerido que estas diferencias en los ingresos entre hombres y mujeres pueden ser explicadas por diferencias en los oficios elegidos por los agentes. Con esto en mente, a continuación se incluirá como un control la variable categórica de *Oficio* con el propósito de poner a prueba esta hipótesis. Así,

se estima el siguiente modelo: log(Income) = β1 + β2Female + θOficio + u.

Como se puede evidenciar, al incluir este control el coeficiente de *Female* es -0.180 y es significativo al 99% de confianza. Así las cosas, resulta claro que incluyendo este control se mantiene la relación observada en el primer modelo: ser mujer está asociado con un efecto negativo sobre los ingresos. Así mismo, en cuanto al ajuste del modelo con este control, el R^2 es de 0.185.

	Dependent variable:		
	logIng		ingreso_tilde
	(1)	(2)	(3)
female1	−0.180***		
	(0.020)		
oficio2	−0.229		
	(0.256)		
female_tilde		−0.180***	−0.180***
		(0.022)	(0.020)
Constant	15.143***	13.928***	0.000
	(0.249)	(0.009)	(0.008)
Observations	16,397	16,397	16,397
R^2	0.185	0.004	0.005
Adjusted R^2	0.181	0.004	0.005
Residual Std. Error	1.084 (df = 16316)	1.195 (df = 16395)	1.081 (df = 16395)
F Statistic	46.327*** (df = 80; 16316)	65.058*** (df = 1; 16395)	79.521*** (df = 1; 16395)
Note:			*p<0.1; **p<0.05; ***p<0.01

No obstante, dado que en este caso la relación de interés es entre el género y el ingreso, se demostrará el teorema FWL. De esta manera, lo que se busca es demostrar que es posible estudiar la brecha salarial de una manera más intuitiva y sencilla, al convertir el modelo de multivariado a univariado. Los resultados de la demostración del teorema, se encuentran en la tabla anterior. Específicamente, se observa que el coeficiente estimado en el primer modelo es equivalente a los coeficientes estimados con los residuales. En este caso, en los tres modelos el coeficiente es el mismo.

Finalmente, en cuanto a la interpretación del coeficiente β2 reiteramos que podría ser interpretado como la "penalidad" asociada al hecho de ser mujer sobre el salario que recibe una persona. Como se mostró a lo largo de este punto, los modelos dan cuenta que el hecho de ser mujer tiene un efecto negativo y significativo sobre los ingresos. Al controlar por las características de trabajo, encontramos que el efecto significativo y negativo asociado a la variable *Female* se mantiene. No obstante, resulta pertinente mencionar que, con este control, el efecto sobre el salario disminuye de 20.3% a 18% y el ajuste del modelo mejora. Lo anterior, da cuenta de que las mujeres colombianas sí se enfrentan a un cultura de discriminación en el mercado laboral que lleva a que su esfuerzo no lleve a que reciban el mismo pago por el mismo trabajo. Además, esta discriminación también fue identificada en el perfil edad-ingresos diferenciado por género: en cualquier momento de su vida las mujeres reciben un menor salario que los hombres.

5. Predicting earnings

Para esta sección se realizarán varios modelos a partir de la muestra de entrenamiento y se probarán en la muestra de testeo. Dicha partición se realizó dividiendo el 70% de entrenamiento

y el 30% de testeo. Pero antes de esto, se corren los modelos de los puntos anteriores y se comparan.

	Dependent variable:			
	ingtot		logIng	
	(1)	(2)	(3)	(4)
age		81,379.530***		0.074***
		(10,819.810)		(0.005)
age2		−695.818***		−0.001***
		(124.477)		(0.0001)
agefem				−0.010***
				(0.002)
female			−0.213***	0.170**
			(0.023)	(0.071)
Constant	1,787,381.000***	−222,354.900	14.023***	12.479***
	(24,738.100)	(218,816.200)	(0.016)	(0.108)
Observations	11,478	11,478	11,478	11,478
R^2	0.000	0.015	0.008	0.026
Adjusted R^2	0.000	0.015	0.007	0.026
Residual Std. Error	2,650,327.000 (df = 11477)	2,630,915.000 (df = 11475)	1.219 (df = 11476)	1.208 (df = 11473)
F Statistic		85.996*** (df = 2; 11475)	87.655*** (df = 1; 11476)	76.123*** (df = 4; 11473)
Note:				*p<0.1; **p<0.05; ***p<0.01

Se puede observar que el modelo con solo la constante muestra el promedio de ingresos de la muestra de entrenamiento, pero hay un R^2 de 0. Más adelante se puede ver los otros modelos de los puntos anteriores que ya se explicaron.

Con el propósito de explorar las transformaciones a las variables independientes para aumentar su complejidad. Con esto en mente, se plantean los siguientes 5 modelos.

1. log (Income)=β1+β2Female+ β3Estrato+ β4Experiencia + β5Educación + u.
2. log (Income)=β1+β2Female+ β3Educación+ β4Experiencia + β5Educación xFemale + u.
3. log (Income)=β1+β2Female+ β3Edad+ β4Edad^2 + β5Experiencia xEdad + u.
4. log (Income)=β1+β2Female+ β3Experiencia+ β4PreferenciaMásHorasTrabajo + β5ActividadOcupóMásTiempo + β6 TotalHoursWorked^2 + β7TotalHoursWorked^3 + β8Informal + u
5. log (Income)=β1+β2Female+ β3 TotalHoursWorked + β4Edad + β5 TotalHoursWorked xFemale + β6 PreferenciaMásHorasTrabajoxFemale + u

A. Validation Set Approach

A continuación, se reportan los errores de predicción promedio de cada uno de los modelos. Así, es posible identificar que el modelo que tiene el menor RMSE es el modelo 4. En particular, el RSME dentro de muestra es de 1.04 y por fuera de muestra 1.12. Como se mostró anteriormente, este se construye incluyendo la experiencia (p6426), el género, la actividad en la que ocupó más tiempo la semana pasada (p6240), si la persona desea trabajar más horas en la semana (p7090), si el trabajo de la persona es informal y se incluyen variables polinómicas de grado 2 y 3 del número de horas trabajadas.

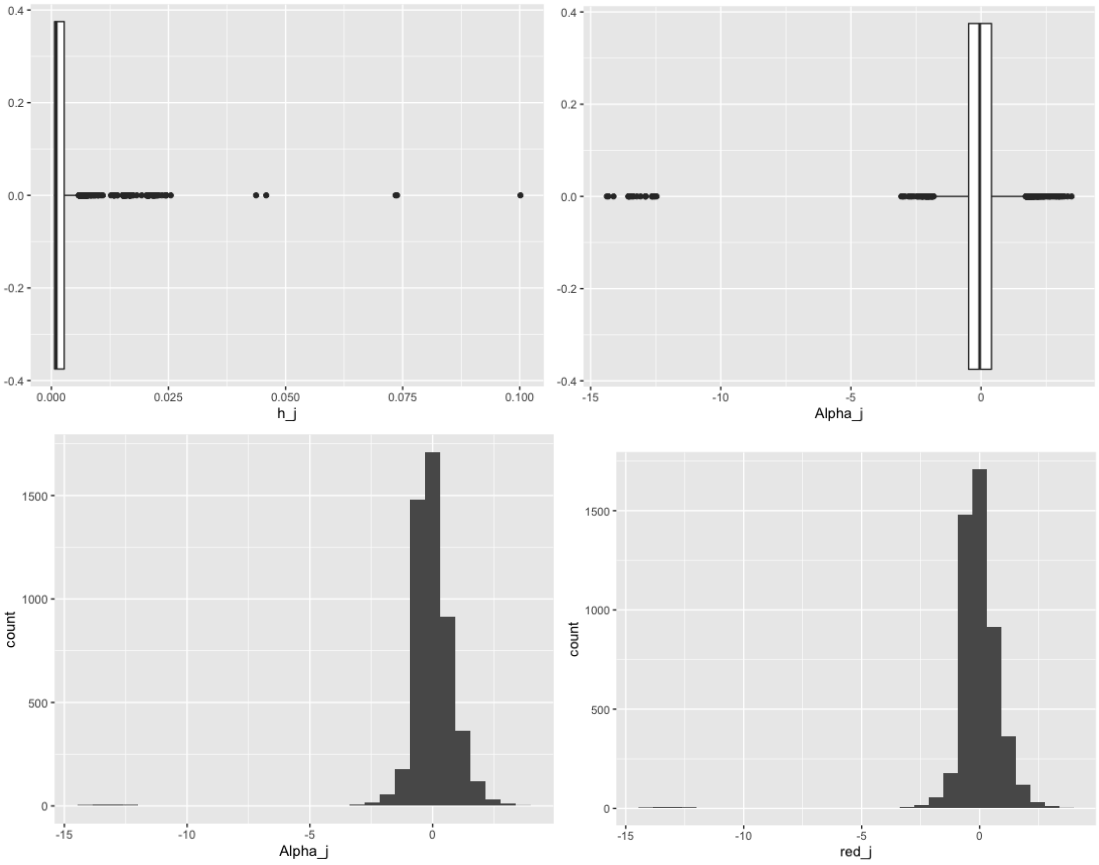
```
RMSEDentroMuestra,RMSE_FueraMuestra
Modelos_C_1,     1.04708926898685,         1.15576492131202
```

Modelos_C_2,	1.08264459421991,	1.19122628884471
Modelos_C_3,	1.14638715571178,	1.23966835706381
Modelos_C_4,	1.04295834591973	1.12834796665255
Modelos_C_5,	1.13236682726234	1.22069357992452

Para identificar los valores atípicos en el Modelo 4, se realiza el análisis del leverage statistic. En primer lugar, para calcular los h_j de la muestra test, fue necesario obtener la Matriz H:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top$$

Al obtener dicha matriz, se obtiene su diagonal, la cual será el *leverage score* para cada observación. Estos valores h_j se tendrán en cuenta para los siguientes puntos. Al observar los datos relevantes, se encuentra que h_j tiene un valor pequeño lo que sugiere que la distancia entre el valor de x_i en el outlier es corta respecto al valor promedio del predictor x. Adicionalmente, al tener en cuenta el comportamiento de Alpha y de los residuales, se observa que ambos tienen el mismo comportamiento. Lo anterior, sugiere que los valores atípicos identificados para este modelo no tienen influencia sobre el mismo y, por lo tanto, no son *high leverage points.*



vars	mean	sd	median	min	max
female	0.48160195	0.49971219	0	0	1
p6426	63.7686522	89.8392568	24	0	720
informal*	1.40841634	0.49159085	1	1	2
p6240*	1.56922139	1.42539708	1	1	6
p7090*	1.90404554	0.29455872	2	1	2
totalHoursW	2493.51291	1620.83822	2304	1	16900
totalHoursW	142409.614	156100.52	110592	1	2197000
logIng	13.8973294	1.26257865	13.8547313	0	18.0640058
red_j	-0.03573696	1.12789655	-0.0520272	-14.3695473	3.47504166
h_j	-0.00110249	0.00392301	-0.00239345	-0.00963737	0.02258771
Alpha_j	-0.03591948	1.13033294	-0.05188574	-14.302228	3.46744131
Yh_j	13.9330663	0.53393761	14.2100157	12.1839183	15.0898056

Con esto en mente, es posible sostener que la DIAN no debería enfocarse en analizar los casos de los individuos que aparecen como valores atípicos dado que no tienen influencia sobre el modelo. Así, como se mostró anteriormente, es probable que estos valores atípicos resulten de un modelo defectuoso.

B. K-Fold Cross Validation

En este punto, se busca identificar cuál es el mejor modelo usando el método de K-Fold Cross Validation. A continuación, se presentan los principales resultados. En cuanto al modelo con el menor error de predicción promedio, es posible identificar que el Modelo 4 tiene el menor RMSE. Al respecto, resulta pertinente mencionar que el valor del RMSE es igual con este método que con el enfoque de separar los datos.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
RMSE Fuera de Muestra	1.155765	1.191226	1.239668	1.128348	1.220694

vars	mean	sd	median	min	max
logYtest	13.8973294	1.26257865	13.8547313	0	18.0640058
Modelo4_cv_Yhat	13.9330663	0.53393761	14.2100157	12.1839183	15.0898056
h_j	-0.00110249	0.00392301	-0.00239345	-0.00963737	0.02258771
Resid_j_CV	-0.03573696	1.12789655	-0.0520272	-14.3695473	3.47504166
Alpha_j_CV	-0.03591948	1.13033294	-0.05188574	-14.302228	3.46744131

Por el otro lado, frente al análisis del leverage statistic, con este método se llega a la misma conclusión: los valores atípicos no inciden significativamente en modelo predicho o en sus coeficientes. Al igual que en el punto anterior, el h_j tiene un valor -0.001lo que sugiere que la distancia entre el valor observado del predictor y el valor promedio es pequeña. Por el otro lado, el valor del alpha y de los residuales es muy cercano.

Ahora bien, frente a las diferencias y similitudes entre ambos métodos, se debe destacar que el K-Fold Cross Validation parte de dividir aleatoriamente las observaciones en k grupos o folds, del mismo tamaño aproximadamente.

Con un training set es posible que la tasa de error tenga una alta varianza dependiendo de ls observaciones que entre en los sets de training y validación. En ese sentido, k-fold Cross Validation ofrece una ventaja debido a que reduce el impacto de cómo se separan los subsets de la muestra. Lo anterior, porque cada observación puede hacer parte tanto del subset de entrenamiento y de test. Sin embargo, dado que la creación de subsets se lleva a cabo k veces, este método tiene un mayor costo computacional que el primero.

C. Leave- One- Out Cross Validation (LOOCV)

En LOOCV cada observación es considerada una validación del entrenamiento en cada iteración. Este método es útil ya que elimina la aleatoriedad de entrenamiento y testeo, ya que cada observación se utiliza en ambos grupos. Esto elimina la variabilidad que se puede tener en K-fold Cross Validation. Sin embargo, esto tiene un costo computacional significativo. Cuando el dataset es muy grande, como en este caso, puede generar overfitting de los datos porque se tiene en cuenta todos los datos de entrenamiento menos la *i-ésima* observación. Además, como solo una observación se usa como validación puede que la variabilidad aumente. Para este ejercicio se utilizaron la muestra de entrenamiento y se realiza el análisis. Este método se demora entre 30 y 45 corriendo, por lo que si se demuestra un alto costo computacional. Al finalizar, se obtiene un RMSE de 1.04, mostrando que efectivamente si se reduce el error cuadrático medio en comparación al modelo original sin ningún método de validación.

vars	mean	sd	median	min	max
logYtest	13.8973294	1.26257865	13.8547313	0	18.0640058
Modelo4_cv_Yhat	13.9330663	0.53393761	14.2100157	12.1839183	15.0898056
h_j	-0.00110249	0.00392301	-0.00239345	-0.00963737	0.02258771
Resid_j_CV	-0.03573696	1.12789655	-0.0520272	-14.3695473	3.47504166
Alpha_j_CV	-0.03591948	1.13033294	-0.05188574	-14.302228	3.46744131

Por otro lado, se muestra que bajo este método este modelo tampoco *tiene high leverage points*, ya que el Alpha se mantiene en una misma proporción igual que los residuales. No obstante, es necesario analizar el modelo de una manera más profunda para encontrar el poco cambio en apalancamiento bajo los diferentes métodos.

Referencias

Cochran, W. G., (1977), Sampling Techniques, Second Edition, John Wiley and Sons, Inc.

Cruces, G. y S. Galiani (2007), “Fertility and female labor supply in Latin America: new causal evidence”, Labour Economics, vol. 14, N° 3, Amsterdam, Elsevier

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2003, May 26). Fifty Years of Mincer earnings regressions. NBER. Retrieved June 27, 2022, from

Kaufman, B.E. (2008). Jacob Mincer's Contribution to Modern Labor Economics: A Review Essay. Andrew Young School of Policy Studies Research Paper Series.

Mincer, Jacob, (1962), Labor Force Participation of Married Women: A Study of Labor Supply, p. 63-105 in, Aspects of Labor Economics, National Bureau of Economic Research, Inc, https://EconPapers.repec.org/RePEc:nbr:nberch:0603.