PROBLEM SET 1- PREDICTING INCOME

Camila Alejandra Velasco Contreras

Jairo Alexander Torres Preciado

Enlace repositorio: https://github.com/jairolax/ProblemSet_1

Punto 1.

El sitio web que contiene la base de datos de la GEIH para Bogotá funciona muy bien y, en general, no presenta ninguna restricción o dificultad para acceder a los mismos. Sin embargo, dado que la información es de tamaño considerable, los datos toman un tiempo un poco más largo de lo normal para cargarse, pero nada fuera de los límites que impone la paciencia. De igual manera, el tamaño de la base implica que los datos se hayan alojado en 10 enlaces diferentes dentro del sitio web, cada uno con una fracción de la base. Esto obliga a que el proceso de web scrapping se tenga que hacer de forma independiente para cada enlace. A continuación, se describe el proceso:

```
INICIO
Proceso: Web scrapping datos GEIH Bogotá
   Acceder a la URL -> "https://ignaciomsarmiento.github.io/
   GEIH2018 sample/"
   Clic enlace -> Data chunk 1
   Inspeccionar código página web
   Copiar enlace -> "https://ignaciomsarmiento.github.io/
   GEIH2018_sample/pages/geih_page_1.html"
   Leer y guardar datos desde sitio web -> read_html y html_table()
   Guardar como data frame -> as.data.frame
   Repetir para los 10 Data chunks.
   Unir 10 data frames -> rbind
 FIN
```

Punto 2.

Reducción y recodificación de variables. Para empezar, reducimos nuestra base de datos a aquellos individuos mayores a 18 años y que viven en Bogotá, luego escogimos las variables que consideramos más relevantes. Una vez con la base reducida, inspeccionamos el formato de cada variable y lo corregimos. Por ejemplo, las variables *estrato1 maxEducLevel* se recodificaron como factores

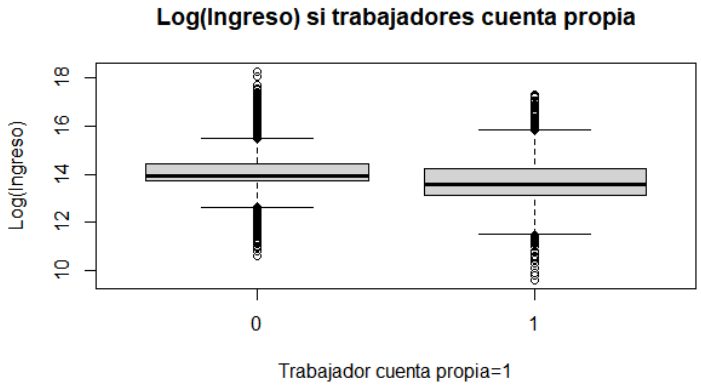
Missings values. Examinamos cuantos missings values tiene cada una de las variables de interés, así como la independiente. Observamos que la cantidad de missings es significativa para las variables: cotiza pensión (*cotPension*), total horas trabajadas (*totalHoursWorked)*y tipo de ocupación (*relab*). En especial, nos parece esencial contar con la variable de horas trabajadas y al ser una proporción considerable de la muestra (33%) consideramos no adecuado imputar los valores y eliminamos las observaciones con missings. Nuestra muestra pasó de tener 24.568 observaciones a 16.541.

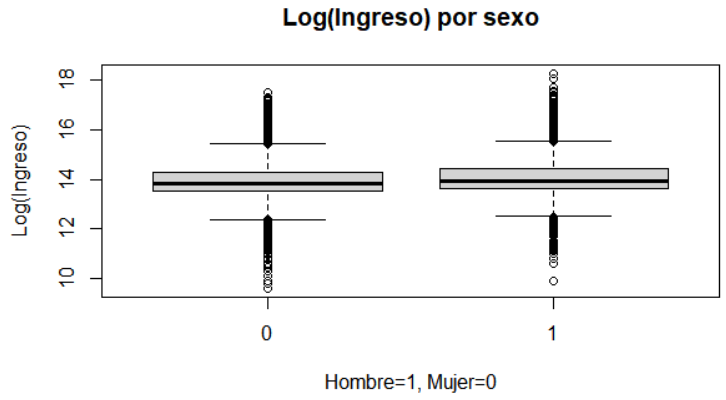
Tabla 2.1 - Missings values

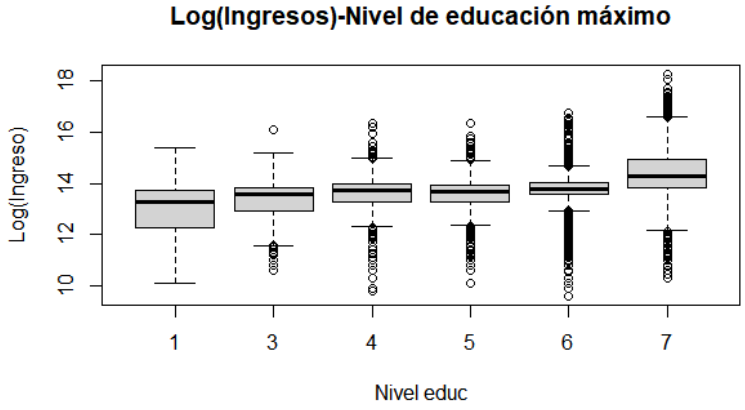
Var	No. Missings	Prop% Muestra
ingtot	0	0,00
age	0	0,00
sex	0	0,00
college	0	0,00
cotPension	8026	32,67
cuentaPropia	0	0,00
dsi	0	0,00
ocu	0	0,00
inac	0	0,00
estrato1	0	0,00
maxEducLevel	2	0,01
relab	8026	32,67
totalHoursWorked	8026	32,67

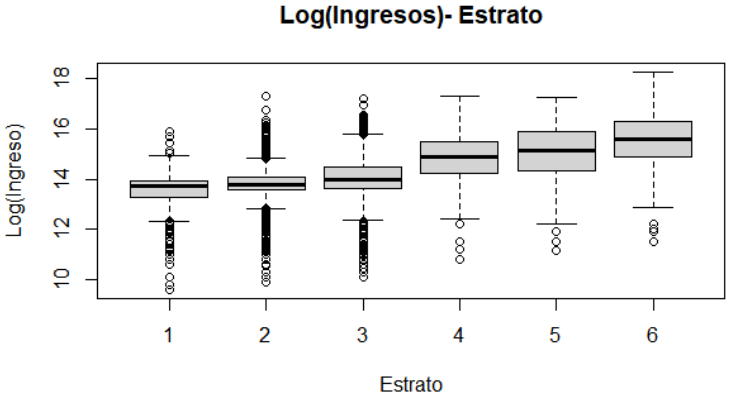
Relación variable dependiente y explicativas

Escogimos la variable *ingtot*= Ingresos totales ya que es la suma de todos los demás ingresos y no tiene missings, ya que el DANE realiza imputación por clase de ingresos. Previo a cualquier análisis nos interesa saber cómo se relacionan las variables independientes con e ingreso. Se evidencia una correlación positiva entre el ingreso y el estrato del indiviuo y el nivel educativo; vemos que hay mayor dispersión en los ingresos de los trabajadores cuenta propia (proxy de informalidad) y una diferencia sutil en el logaritmo del ingreso promedio entre mujeres y hombres.









Punto 3.

La variable que mejor explica el ingreso total de los trabajadores es “ingtot”, ya que esta variable es la suma de todas las fuentes de ingreso registradas en la GEIH. Así, “ingtot” es

la suma de “ingtotob” (suma de ingreso monetario primera actividad, ingreso segunda actividad, ingreso en especie, ingreso monetario desocupados e inactivos, e ingresos provenientes de otras fuentes no laborales) y de “ingtotes”, la cual es ingreso total por persona luego de sumar cada una de las fuentes de ingresos imputadas en la base de datos original.

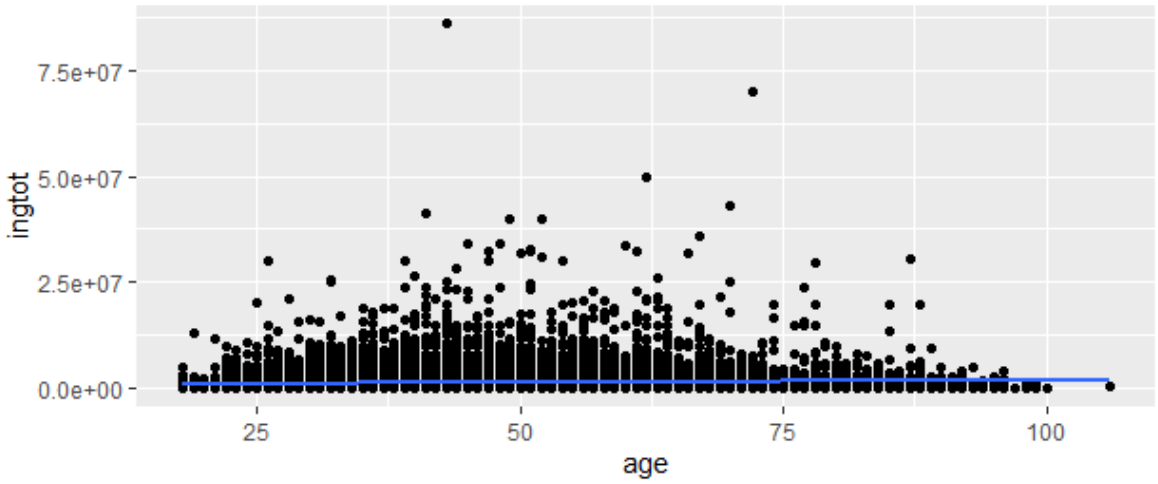
Al estimar el modelo de MCO sugerido, se obtienen los siguientes resultados:

Tabla 3.1 - Resultados OLS

Dependent variable:	
	ingtot
age	98,375.180***
	(4,459.496)
age2	-922.837***
	(46.301)
Constant	-865,500.600***
	(96,907.040)
Observations	24,568
R2	0.022
Adjusted R2	0.022
Residual Std. Error	2,360,588.000 (df = 24565)
F Statistic	281.244*** (df = 2; 24565)
Note:	*p<0.1; **p<0.05; ***p<0.01

El coeficiente de la variable edad es estadísticamente significativo y muestra que, en promedio, una persona recibe un ingreso de $98.375 por cada año de edad adicional. Tanto el coeficiente de determinación R2, como el R2 ajustado, señalan una bondad de ajuste del modelo de 2.2%, es decir, la variable edad apenas explica ese porcentaje del ingreso total de las personas. El valor del error estándar residual (desviación estándar de los residuos) indica que el modelo explica el ingreso total con un error de 2.360.588. El ajuste del modelo no es tan bueno dado que solo se cuenta con la edad como variable explicativa para predecir el ingreso total. El modelo es estadísticamente significativo en conjunto, según el resultado del estadístico F.

Figura 3.1. Edad vs Ingreso total



La Figura 2 que relaciona los valores predichos de edad e ingreso total en el modelo, también muestra que, en promedio, las personas de Bogotá tienen el mayor nivel de ingresos totales en los años previos y posteriores a la edad de 50 años. Esto significa que el ingreso de una persona, en promedio, va aumentando a lo largo de su vida hasta que alcanza este punto máximo y luego se mantiene o se reduce ligeramente.

Tabla 3.2 Bootstrap e intervalos de confianza

```
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = geih_fil, statistic = bootGEIH.fn, R = 1000)


Bootstrap Statistics :
original       bias    std. error           ci
t1* -865500.6184 1720.5980704 77140.13571 -1055444.29 -675556.9508
t2*   98375.1804  -84.6726376  4127.48859    89634.30  107116.0627
t3*    -922.8369    0.4607371    45.30622    -1013.59    -832.0835
```

La Tabla 3.2 muestra los errores estándar para el intercepto, y las variables de edad y edad al cuadrado, los cuales señalan qué tan precisos son los parámetros estimados. Así mismo, se muestran los intervalos de confianza (al 95%) para cada variable. En este caso, por ejemplo, se puede afirmar que, en promedio, cada año de edad adicional implica un mayor ingreso total de al menos $86.634.30 y máximo $107.116.06 con un 95% de confianza.

Punto 4.

Tabla 4.1 Modelo earnings gap

```
Resultados earnings gap
================================================
```

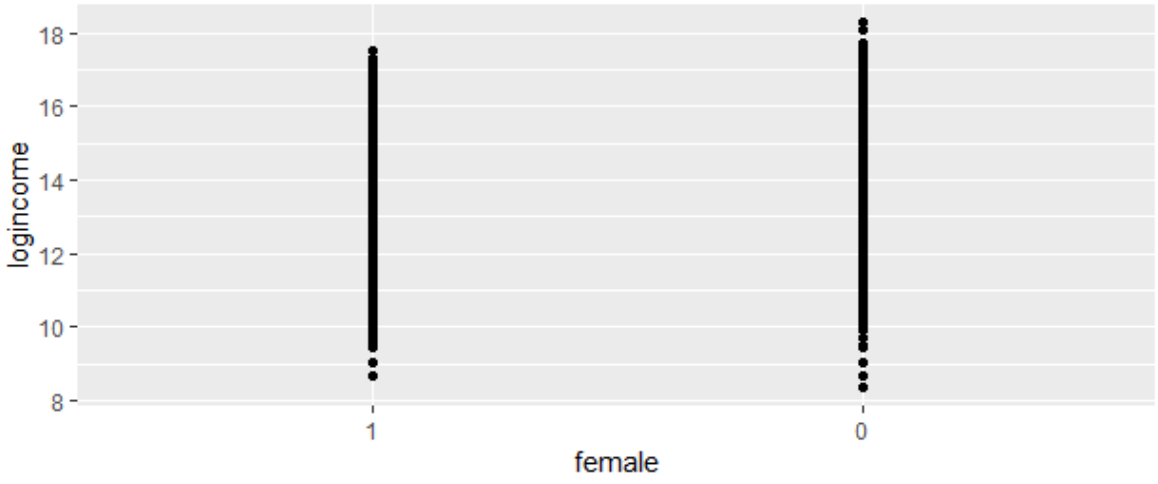
```
Dependent variable:
---------------------------
logincome
-----------------------------------------------
female0                       0.242***
(0.014)

Constant                      13.747***
(0.010)

-----------------------------------------------
Observations                   19,801
R2                              0.015
Adjusted R2                     0.015
Residual Std. Error      0.969 (df = 19799)
F Statistic          309.299*** (df = 1; 19799)
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

La Tabla 4.1 muestra los resultados del modelo MCO de earnings gap. El coeficiente de "female" significa que las mujeres tienen en promedio, un ingreso total que es 24.2% superior al de los hombres en la muestra de la GEIH de Bogotá. Si bien, la variable "female" es estadísticamente significativa al igual que el modelo en su conjunto, según el estadístico F, el ajuste no es tan alto dado que el R2 y el R2 ajustado son 0.015. Esto puede explicarse debido a que el modelo solo cuenta con una variable explicativa para el ingreso total.

Figura 4.1 Earnings por género



La Figura 4.1 muestra que los ingresos totales en hombres tienen un intercepto menor que el de las mujeres, es decir, tienen ingresos que arrancan desde un nivel inferior.

Tabla 4.2. Bootstrap e intervalos de confianza

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = geih_fil, statistic = bootGEIH.fn, R = 1000)


Bootstrap Statistics :
      original        bias    std. error          ci
t1* 13.7474765  1.911869e-06  0.01064020  13.7282497 13.7667032
t2*  0.2421841 -2.930540e-04  0.01422167   0.2151924  0.2691759
```

La Tabla 4.2. muestra el error estándar de la variable “female” y de la constante, así como los intervalos de confianza al 95%. El error estándar señala la precisión de los parámetros estimados, mientras que los intervalos de confianza muestran en dónde está el verdadero valor de cada parámetro. En el caso del coeficiente de “female”, se puede asegurar que este se encuentra entre 0.21 y 0.26 con un 95% de confianza. Los intervalos de confianza no se superponen dado que solo se cuenta con una variable y el intercepto.

Tabla 4.3 Modelo earning gap con controles

Resultados earnings gap controles

	Dependent variable:
	logincome
female0	0.308***
	(0.011)
maxEducLevel	0.172***
	(0.006)
cuentaPropia	-0.350***
	(0.020)
formal	0.578***
	(0.014)
oficio	-0.006***
	(0.0002)
age	0.012***
	(0.0005)
relab	0.065***
	(0.006)
Constant	12.229***
	(0.047)

Observations	16,276
R2	0.366
Adjusted R2	0.366
Residual Std. Error	0.699 (df = 16268)
F Statistic	1,342.231*** (df = 7; 16268)
Note:	*p<0.1; **p<0.05; ***p<0.01

La Tabla 4.3. incluye múltiples covariables con características del trabajo y de la persona que contribuyen a explicar el ingreso total. En particular, se incluyó el máximo nivel educativo, si la persona es trabajadora a cuenta propia (trabajador independiente), si tiene un empleo formal, el tipo de trabajo u oficio que desempeña, la edad, y el tipo de ocupación (empleador, jornalero, empleado de gobierno, empleado doméstico, trabajador sin remuneración, etc.). Todas las variables son estadísticamente significativas al 99%, por lo que individualmente contribuyen a explicar el ingreso total. El modelo en su conjunto también es estadísticamente significativo, tal como lo indica el estadístico F. En esta nueva estimación, aumenta el coeficiente de "female", pasando de 0.24 a 0.30, lo que significa que, en promedio, una mujer tiene un ingreso total 30% superior al del hombre. Tanto el R2 como el R2 ajustado aumentaron y el error residual estándar se redujo, lo que muestra que este modelo con controles tiene un mejor ajuste que el anterior.

Tabla 4.4 Modelo FWL

Resultados earnings gap FWL

	Dependent variable:
	logincome
res2(female)	0.312***
	(0.011)
res3	-0.324***
	(0.019)
res4	0.582***
	(0.013)
res5	-0.006***
	(0.0002)
res6	0.119***
	(0.0004)
res7	0.059***
	(0.006)
Constant	0.105***
	(0.006)

```
------------------------------------------------
      Observations                    16,276
       R2                              0.258
       Adjusted R2                     0.258
   Residual Std. Error       0.699 (df = 16269)
 F Statistic           945.12*** (df = 6; 16269)
================================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

La Tabla 4.4 muestra el resultado de estimar el mismo modelo con covariables, pero esta vez haciendo uso del teorema de Frisch-Waugh-Lowell, es decir, incluyendo las estimaciones de los residuos de cada variable que contribuyen a explicar el ingreso total. Los resultados de este modelo no coinciden exactamente con los del anterior modelo, pero son semejantes en magnitud. El ajuste se redujo un poco, ya que el R2 y el R2 ajustado ahora son 0.258. Por su parte, el error estándar residual conserva el mismo valor de 0.699 y, en conjunto, todo el modelo es estadísticamente significativo, de acuerdo con estadístico F.

En este modelo se considera que r1 es el residuo de la regresión de "logincome" y "maxEducLevel", r2 el residuo de "female" y "maxEducLevel", r3 el residuo de "cuentaPropia" y "maxEducLevel", r4 el residuo de "formal" y "maxEducLevel", r5 el residuo de "oficio" y "maxEducLevel", r6 el residuo de "age" y "maxEducLevel", y r7 el residuo de "relab" y "maxEducLevel". El coeficiente de res2(female) es el β_2 de interés, el cual es bastante similar al del modelo anterior. En este caso, este parámetro significa que las mujeres tienen en promedio un ingreso total que es 31.2% superior al de los hombres. Dado que la reducción en la brecha es prácticamente no representativa, se puede inferir que efectivamente hay una brecha de ingresos de género en Bogotá. Si bien estos resultados parecen ir en contra de la lógica que sugiere que las hombres tienen mayores ingresos que las mujeres, en el caso de este ejercicio, la brecha a favor de las mujeres se puede deber al hecho de que la variable de ingreso total que tuvimos en cuenta para las estimaciones tiene en cuenta a todo tipo de ingresos, no solo los salarios, y es posible que algunas mujeres tengan ingresos adicionales por cuenta de programas de asistencia y similares.

Punto 5

R: Se dividió la base entre dos muestras, entrenamiento (70%) y prueba (30%), y se estimaron los modelos especificados en la Tabla 1. Para los modelos a los cuales se les aplicó logaritmo a la variable dependiente se eliminaron las observaciones con ingresos totales=0.

Tabla 5.1. Modelos - Variable dependiente e independientes

	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
y=Ingresos totales	x	x								
y=log(Ingresos totales)			x	x	x	x	x	x	x	x
Constante	x	x	x	x	x	x	x	x	x	x
Edad		x		x	x	x	x	x	x	x
Edad^2		x			x	x	x	x	x	x
Sex			x	x	x	x	x	x	x	x
Total horas trabajadas				x	x	x	x	x	x	x
Estrato (categórica)				x	x	x	x	x	x	x
Cuenta prompia (si=1, no=0)				x	x	x	x	x	x	x
Max. Nivel Educ (categórica)				x	x	x	x	x	x	x
CuentaPropia*Sex						x	x	x	x	x
Max. Nivel Educ*Sex							x	x	x	x
Estrato*Sex								x	x	x
Total horas trabajadas^2									x	x
CuentaPropia*Estrat										x

Luego de estimar los distinntos modelos se calculó el ingreso total estimado usando los datos de la muestra de prueba. Para los modelos en que la variable dependiente está en logaritmo se le aplicó la función exponente. Una vez estimado el ingreso total para cada observación de la base de prueba fue posible hallar el error cuadrático medio, sacando el promedio del cuadrado la resta del ingreso estimado menos el valor observado de los ingresos, A continuación se reportan los errores cuadráticos medios estimados para cada modelo (Tabla 2)

Tabla 5.2 Error cuadrático medio (MSE) por modelo - Fuera y

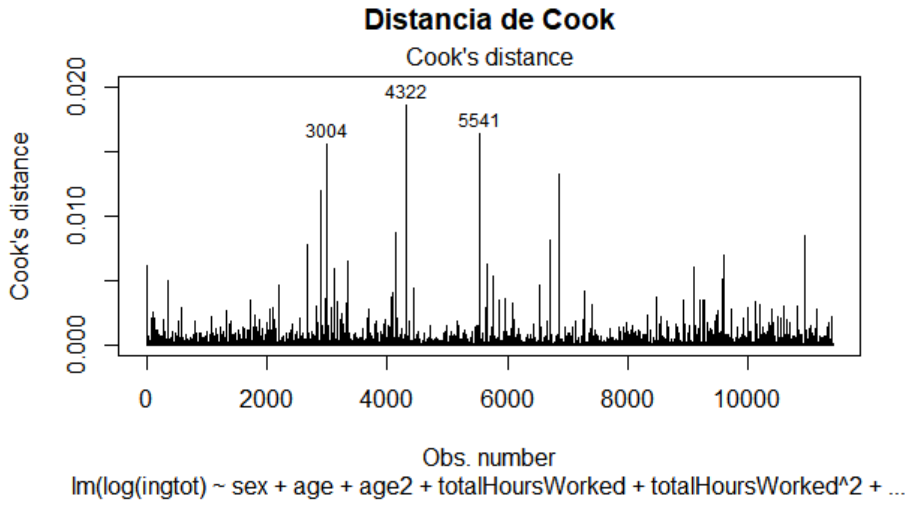
	MSE Fuera de muestra	MSE_dentro de muestra	MSE_FM/MSE_DM
Modelo 1	7,86088E+14	6,85825E+14	1,15
Modelo 2	7,73567E+14	6,73699E+14	1,15
Modelo 3	8,23679E+14	7,18288E+14	1,15
Modelo 4	5,43596E+14	4,54176E+14	1,20
Modelo 5	5,42314E+14	4,54706E+14	1,19
Modelo 6	5,41754E+14	4,55424E+14	1,19
Modelo 7	5,42691E+14	4,56893E+14	1,19
Modelo 8	5,33935E+14	4,46319E+14	1,20
Modelo 9	5,33935E+14	4,46319E+14	1,20
Modelo 10	5,33658E+14	5,33658E+14	1,00

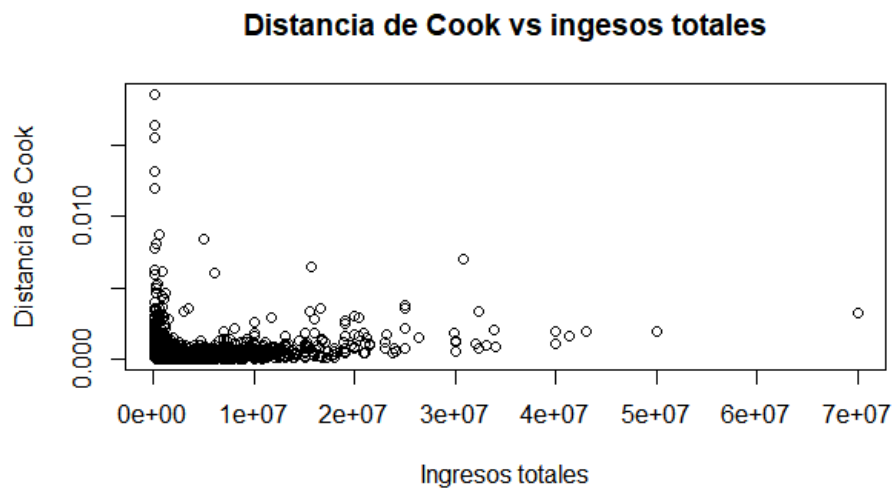
Como se puede ver en la tabla el modelo 10 es el que tiene el error cuadrático medio por fuera de muestra más bajo entre todos los modelos, aunque no es el menor dentro de muestra.

(comentar modelo11)

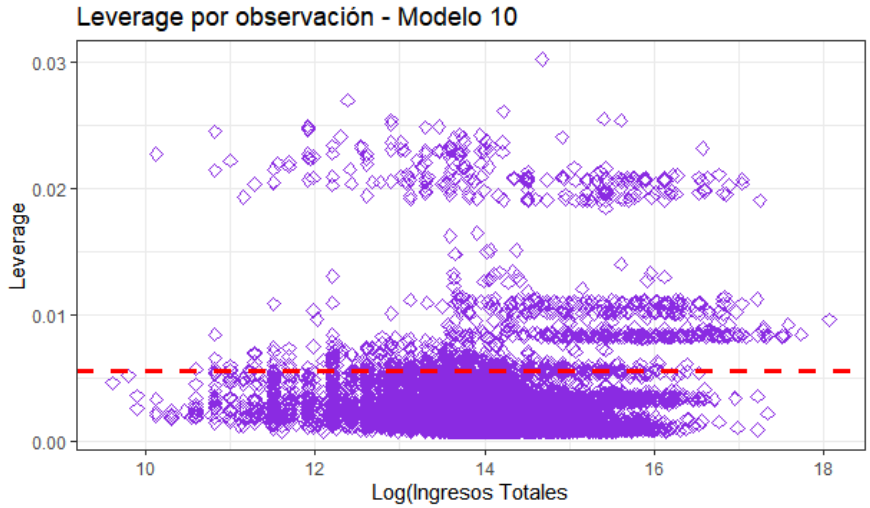
Análisis outliers

En primer lugar, examinamos la distancia de cook de cada observación, está es otra medida alternativa indicadora de la influencia de la observación en la estimación del modelo. Se evidencian en las dos gráficas a continuación que hay incidencias desiguales entre observaciones, además, las observaciones con una distancia de cook más alta son aquellos individuos que reportan ingresos más bajos.





Para complementar el análisis previo se calculó el leverage de cada observación dentro de la muestra de entrenamiento. Si este leverage es mayor a 2*(k+1)/n la observación es considerada un outlier, es decir un valor atípico que está afectando considerablemente la estimación del modelo. El modelo 10 se estima con 31 variables (k=31). En total 973 observaciones tienen un leverage por encima del valor esperado, lo que equivale a 8.5% de la muestra. (Ver gráfica)



Ahora, estás observaciones atípicas podrían ser individuos que subreportan sus ingresos para pagar menos impuestos. Sin embargo, vale la pena decir que el modelo tiene un r cuadrado bajo (0.45).

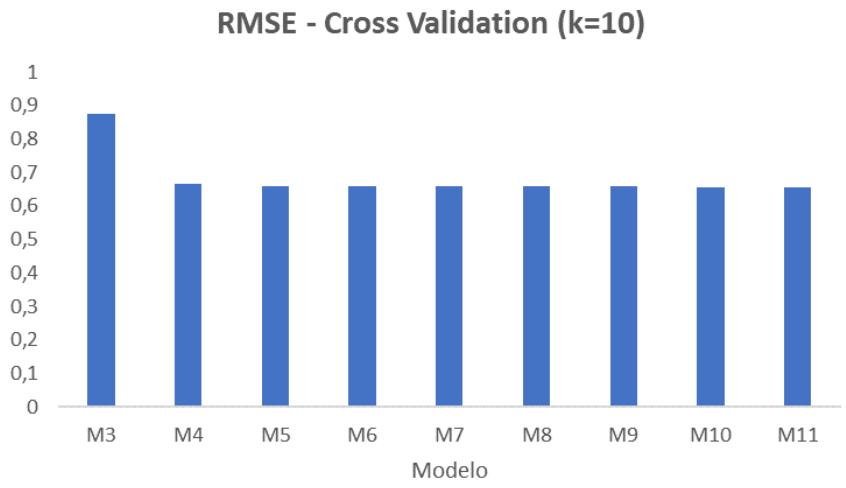
b) K-fold cross-validation

En un proceso similar al anterior realizamos el siguiente proceso: i) generar una participación de los datos en dos, la muestra de entrenamiento (70%) y la muestra de prueba (30%); ii) se hace la partición en los k-folds (K=10 en este caso) y se estima cada modelo las diez veces,

cada una excluyendo el k-esimo fold (fusión *train* en r); iii) sacamos el promedio de la raíz del error cuadrático medio de cada fold para cada modelo y comparamos este valor entre modelos. A continuación se presentan lo RMSE obtenidos para cada modelo:

Tabla 5.3 - RMSE - CV

Modelo	RMSE
M3	0,8735
M4	0,6647
M5	0,6603
M6	0,6592
M7	0,6582
M8	0,6575
M9	0,6582
M10	0,6571



Como se observa en la tabla, de nuevo el modelo con el menor es el RMSE es el 10.

c) LOOCV

(En el script se encentra el código, sin embargo este no nos funcionó)