Universidad de los Andes

Problem Set 2 – GRUPO 6

Víctor Dulio Chique

Víctor Iván Sánchez

Natalia Castro

PREDICCIÓN DE PROBREZA EN COLOMBIA

1. Introducción:

La medición de la pobreza busca identificar los grupos de la población más vulnerables para realizar intervenciones de política pública que permitan mejorar sus carencias. Si bien en la actualidad el concepto de pobreza no aborda solamente el ingreso y se busca entender también las carencias sociales de los hogares, la medida de pobreza monetaria continúa siendo muy utilizada ya que es un indicador de un mínimo de subsistencia alimentaria y no alimentaria.

El ingreso usualmente se mide a través de encuestas las cuales requieren de un gran esfuerzo logístico, de diseño y de implementación. De igual manera exige un trabajo de medición de canastas alimentarias y no alimentarias que no son fáciles de generalizar pues cada territorio y cada familia tienen costumbres propias. Adicionalmente, las familias pueden esconder algunos de sus ingresos si saben que sus respuestas favorecerán subsidios o ser incluidos en programas sociales. Las variables que deben ser incluidas en los estudios de medición de pobreza monetaria son entonces numerosas y utilizarlas todas en un modelo puede resultar ser una tarea muy compleja.

El objetivo de este trabajo es precisamente buscar modelos y variables diferentes al ingreso que logren clasificar correctamente a los hogares colombianos entre aquellos que se encuentran en situación de pobreza o no, con el fin de tener herramientas de clasificación más precisas, menos complejas y más efectivas. Se utilizaron entonces dos enfoques: 1. Clasificación utilizando variables diferentes al ingreso y 2. Clasificación utilizando la predicción de ingresos de los hogares.

En el enfoque de clasificación sin ingreso, seleccionamos un modelo que a partir de las características del hogar permite predecir pobreza con un *Accuracy* del 81,9% en los datos de entrenamiento y de aproximadamente el 81,5% en los datos de prueba.

En el modelo de ingreso y posterior clasificación contrastándola con la línea de pobreza se encuentra que la predicción del ingreso no es un medio adecuado para identificar hogares pobres porque termina subclasificándolos. En efecto, este resultado es el esperado, pues los determinantes de la pobreza van más allá de lo monetario, tales como las condiciones

habitacionales y de vivienda, el acceso a servicios públicos y características socio demográficas de los que integran el hogar.

2. Base de datos:

La base de datos que se utilizó en este trabajo fue la de la Medición de Pobreza Monetaria "Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE del Departamento Administrativo Nacional de Estadística – DANE. Los datos se encuentran divididos en dos grandes muestras: la de entrenamiento y la de testeo. A su vez cada una se divide en dos grupos: hogares y personas. Esta base de datos es conveniente porque provee información no sólo sobre los diferentes tipos de ingreso de los hogares sino sobre las características de contexto propias de cada hogar. Lo anterior permite proponer modelos con variables diferentes al ingreso total de cada hogar y seleccionar aquellos que estiman las predicciones con las mejores métricas de clasificación.

La base de datos de entrenamiento tenía 164.960 observaciones y la de testeo 68.168. Como primer paso se unieron las bases de hogares y personas para entrenamiento y posteriormente se hizo lo mismo para las bases de testeo. De la base a nivel de personas se eligieron variables que caracterizan al jefe de hogar para luego ser empalmada con la base a nivel de hogar, tanto en train como en test. Las variables que se utilizaron en los modelos fueron:

Cabecera_resto	Hombre	Horas de trabajo
Ingreso	Edad	Más tiempo de trabajo
Cuartos_hogar	Entidad_salud	Tamaño de la empresa
Vivienda_propia	Nivel Educativo	
Personas_hogar	Tiempo de trabajo	

Cuadro 1: Variables seleccionadas 1

Debido a que existían datos faltantes se imputaron las variables: *entidad_salud, mas_trabajo, tipo_de_trabajo, tiempo_trabajando, horas_trabajo*, tamaño de la empresa con la media y moda de cada una de estas.

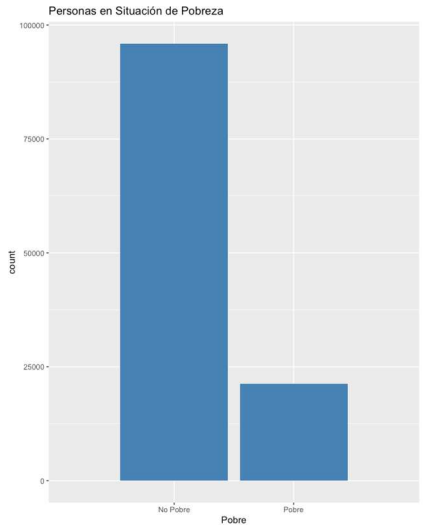
3. Estadísticas Descriptivas

A continuación se realiza una breve descripción de las variables que se incluyeron en los modelos de clasificación y predicción. Comenzamos con la variable objetivo del estudio: *Pobre*. El cuadro 2 muestra que la muestra tiene un desbalance importante que se puede clasificar como moderado (1%-20%). El problema con la prevalencia del resultado *No pobre* es que los algoritmos tenderán a categorizar las predicciones dentro de esta clase mayoritaria. De esta manera la medida de *Accuracy* (proporción de predicciones correctas dentro del total de predicciones) tenderá a ser muy alta pero el modelo no necesariamente predecirá con exactitud a aquellos que sí lo son. Esto es un problema porque si solamente se toman en cuenta estas predicciones para realizar

intervenciones de política pública, seguramente se estaría beneficiando a personas para las cuales no están destinados esos recursos.

No Pobre	Pobre
81.88%	18.11%

Cuadro 2: Porcentaje personas pobreza 1



Gráfica 1: Personas en situación de pobreza

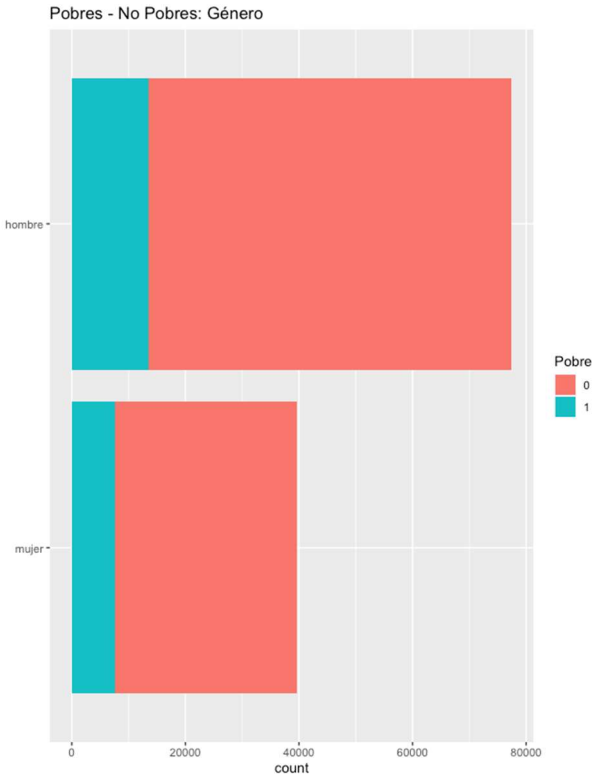
Los gráficos del 1 al 9 presentan la distribución de las variables categóricas que se utilizaron en los modelos. Se observa una población en la que más del 60% son hombres, más del 85% tiene acceso a salud, además de las horas de trabajo 92% quiere trabajar más, la mayoría son trabajadores por cuenta propia o son empleados u obreros de empresas particulares, el 52% sólo cuenta con educación primaria y/o secundaria, cerca del 50% vive en arriendo y en su mayoría son hogares de cabeceras de Colombia.

Así mismo el cuadro 2 muestra la media, mediana, cuartiles, mínimo y máximo de las variables.

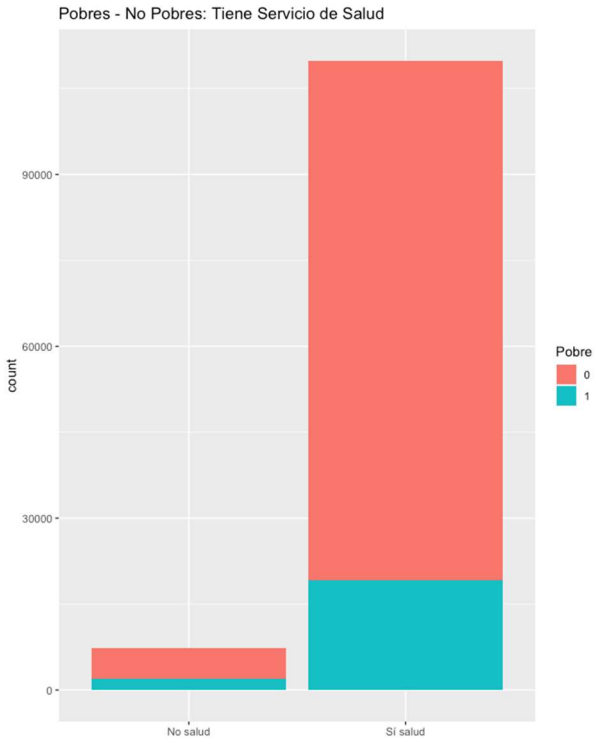
Estadísticas

Statistic	Min	Pctl(25)	Mean	Pctl(75)	Max	Median	St. Dev.
cabecera_resto	0	1	0.892	1	1	1	0.310
income	0.000	875,000.000	2,197,958.000	2,583,190.000	85,833,333.000	1,500,000.000	2,632,234.000
cuartos_hogar	1	3	3.294	4	43	3	1.182
vivienda_propia	1	1	2.574	3	6	3	1.216
personas_hogar	1	2	3.337	4	22	3	1.723
Pobre	0	0	0.181	0	1	0	0.385
hombre	0	0	0.661	1	1	1	0.473
edad	14	35	45.444	55	99	45	13.464
entidad_salud	0	1	0.938	1	1	1	0.241
nivel_educativo	1	3	4.513	6	9	5	1.347
tiempo_trabajando	0	13	110.372	168	948	60	128.721
tipo_de_trabajo	1	1	2.976	4	9	4	1.620
horas_trabajo	1	40	46.912	52	130	48	15.320
mas_trabajo	0	0	0.081	0	1	0	0.273
tam_emp	1	1	3.825	8	9	2	3.311

Cuadro 2: Descripción de variables



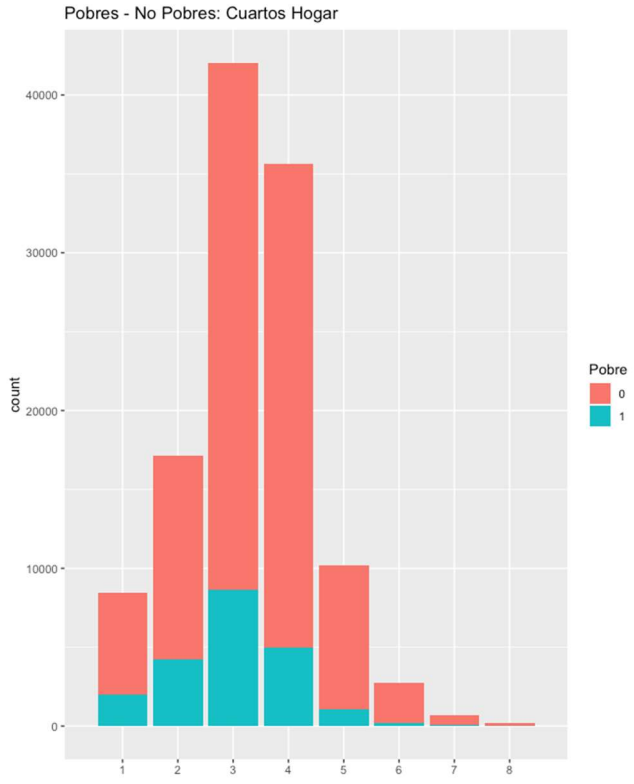
Gráfica 2: Género



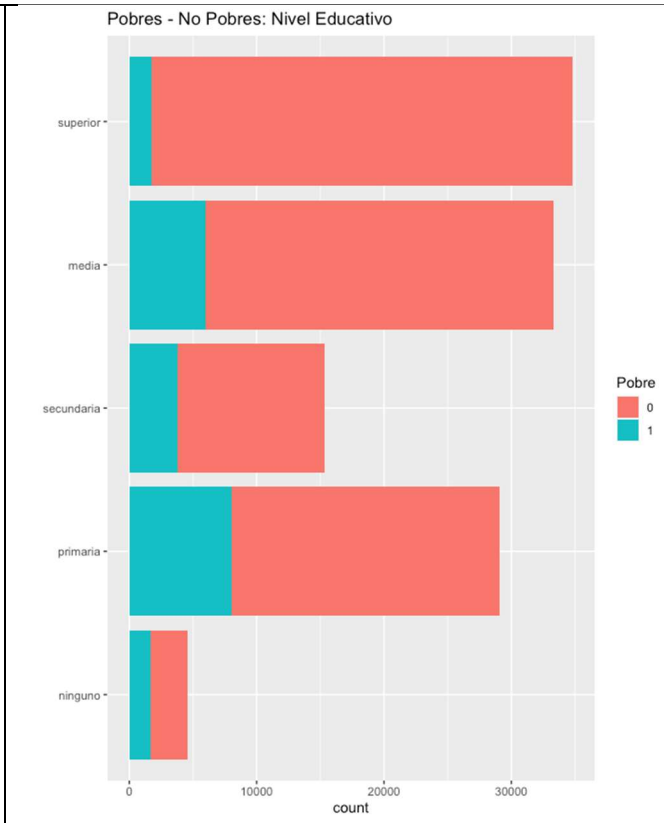
Gráfica 3: Personas con servicio salud

Las gráficas 2 muestra una mayor cantidad de hombres en situación de pobreza respecto a las mujeres (barras verdes) pero se debe tomar en cuenta que la base de datos que se construyó se tomó como referente al jefe de hogar el cual usualmente es la persona que más ingresos gana en el hogar. Si los hombres ganan más que las mujeres, el referente va a ser en su mayoría un hombre. Por lo anterior no se puede afirmar que la cantidad de hombres en situación de pobreza sea mayor que la cantidad de mujeres. La gráfica muestra una distribución de la base de datos desbalanceada por género.

La mayoría de los hogares cuentan con servicios de salud indiferentemente de si el hogar es clasificado como pobre o no. No se esperaría entonces que esta variable fuera un predictor importante para clasificar a las personas en situación de pobreza.



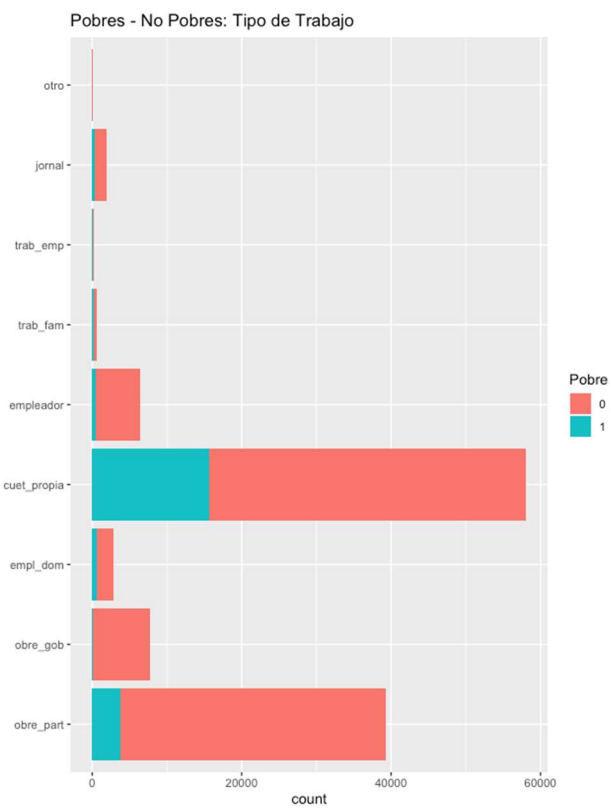
Gráfica 4: Cuartos en el hogar



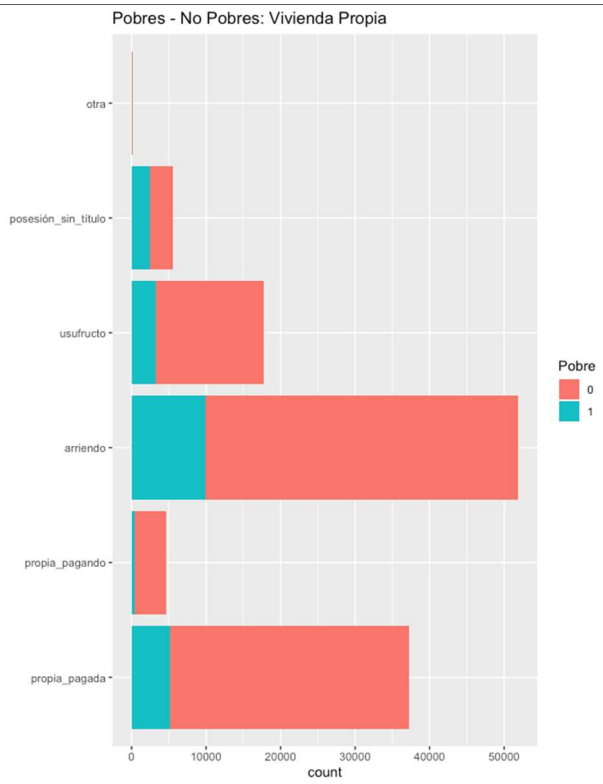
Gráfica 5: Nivel Educativo

La gráfica 4 muestra el número de cuartos con los que cuentan los hogares. En promedio la mayoría de las personas clasificadas pobre o no pobres cuenta con 3 cuartos. La distribución es similar para ambos grupos, por lo cual no pareciera ser un predictor importante para segmentar. Sin embargo, no estamos teniendo en cuenta la cantidad de personas que viven en cada hogar y que deben compartir un mismo cuarto. Tampoco para qué se utiliza cada cuarto. Una variable con interacción que tomara en cuenta el número de cuartos, el número de personas y la función de cada espacio podría ser un potencial buen predictor. En este trabajo no se incluyó dicha interacción pero se sugiere para otro estudio.

El nivel educativo refleja que quienes cuentan con educación superior en su gran mayoría no están en situación de pobreza y quienes son pobres en su mayoría sólo cuentan con educación primaria. Se espera que esta variable sea importante en los modelos de clasificación.



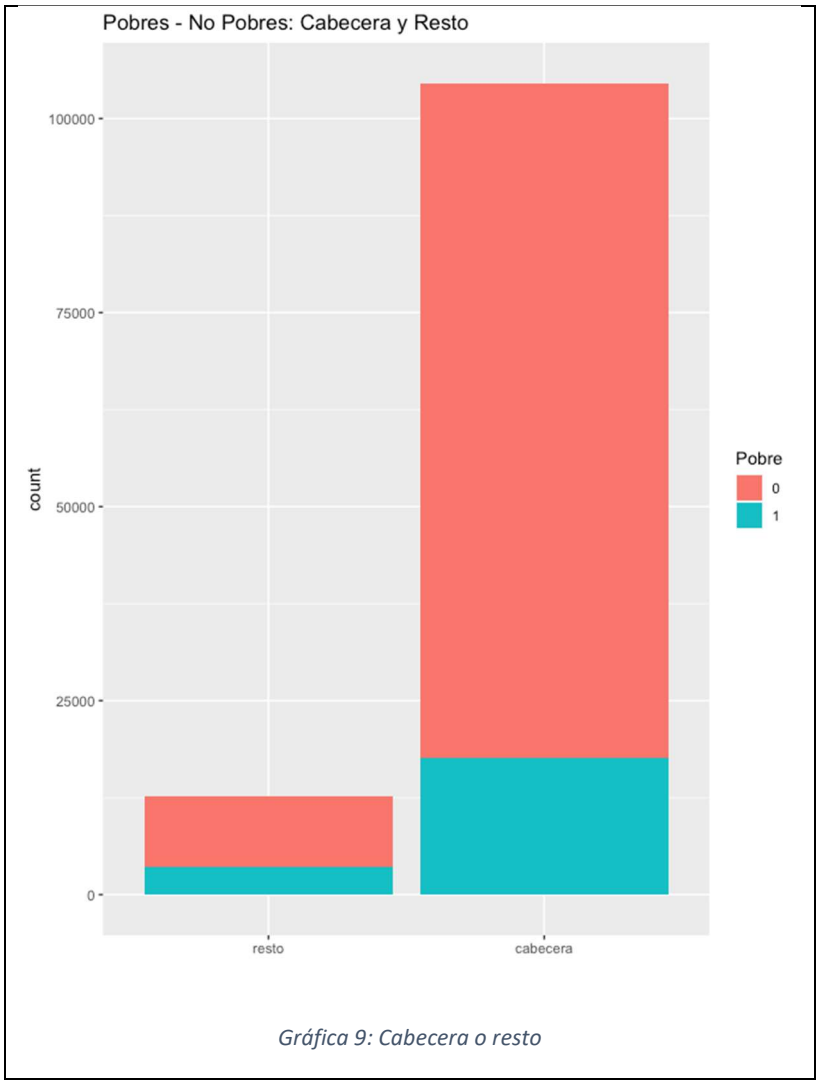
Gráfica 6: Tipo de trabajo



Gráfica 7: Tipo de vivienda

Como se puede observar en las gráficas 6 y 7 el grupo de personas en situación de pobreza trabaja en su mayoría por cuenta propia o como obrero. De igual manera, viven en arriendo. Estás también podrían ser variables significativas a la hora de clasificar a los más vulnerables.

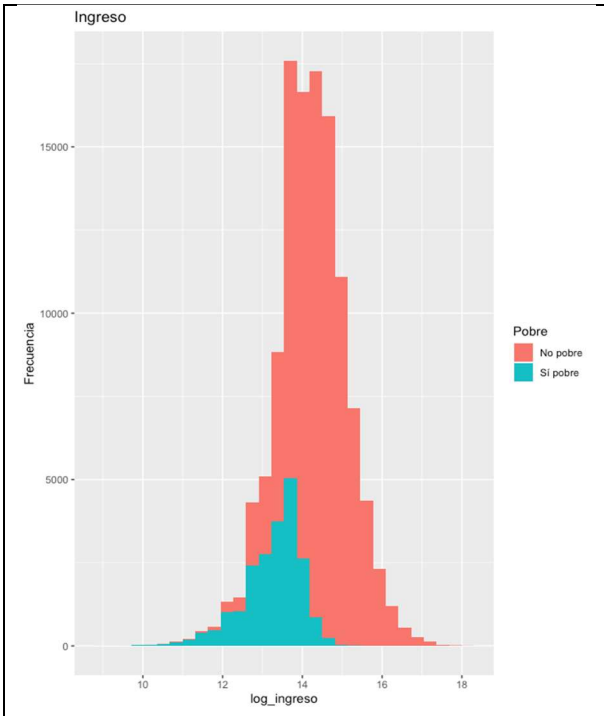
La gráfica 9 refleja una mayor cantidad de personas en pobreza en las cabeceras respecto al resto de territorios. Sin embargo, esta gráfica se debe analizar con cautela ya que la mayoría de las observaciones se realizaron en cabeceras.



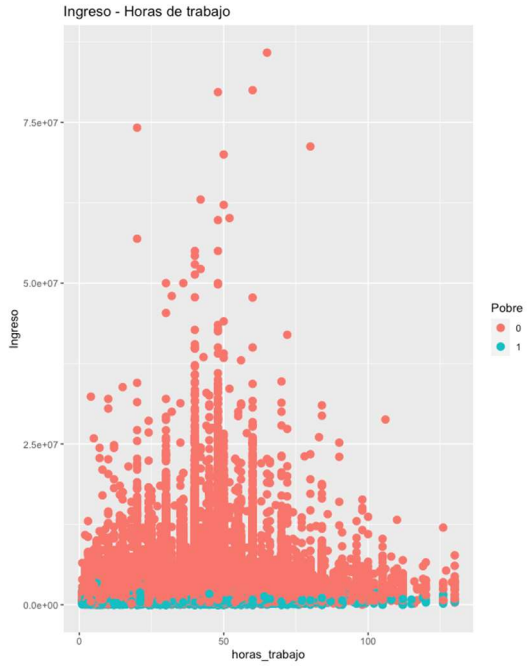
Gráfica 9: Cabecera o resto

Las gráfica 10 muestra la distribución logaritmo del ingreso para los dos grupos. En la gráfica 11 se observa que el ingreso no tiene mayor variabilidad respecto a las horas de trabajo para quienes están en situación de pobreza. Es decir que trabajar más horas no va a representar un ingreso mayor para ese grupo.

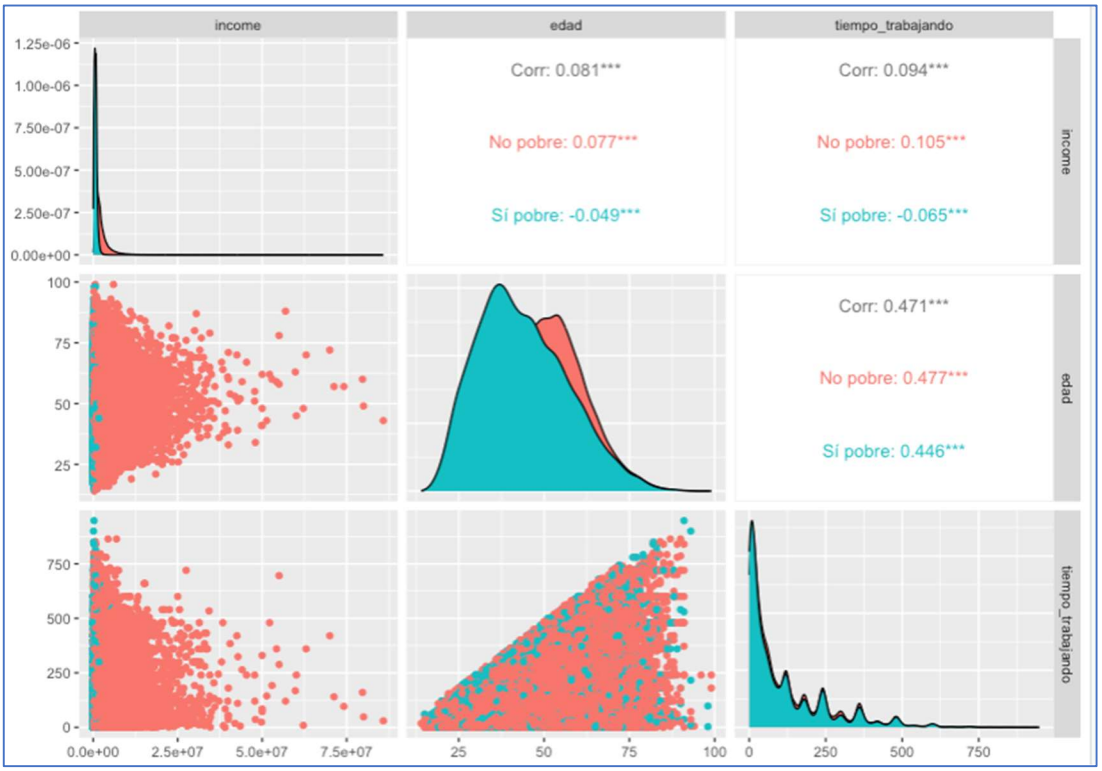
En cuanto a las variables continuas se observa una muy baja correlación (aunque significativa) entre ingreso, edad y tiempo trabajado. Inclusive la correlación es negativa entre los años que lleva trabajando una persona en su empresa y el ingreso para quienes se encuentran en situación de pobreza. Lo mismo sucede entre la variable edad y el ingreso.



Gráfica 10: Logaritmo del Ingreso



Gráfica 11: Ingreso - Horas de Trabajo

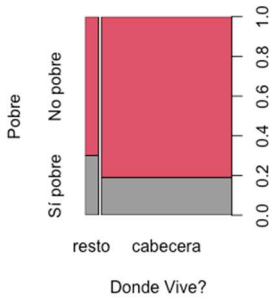


Gráfica 10: Distribución y correlación 1

4. Modelos de Predicción y Clasificación:

4.1. Modelo de Clasificación sin Ingreso

La estrategia para construir el modelo de clasificación consistió en evaluar los predictores adecuados para realizar la estimación, por tanto basándonos en el análisis de los datos observamos algunas de las variables relevantes mediante los siguientes gráficos.



Gráfica 11: Dónde vive?

Gráfica 12: Tipo de Vivienda

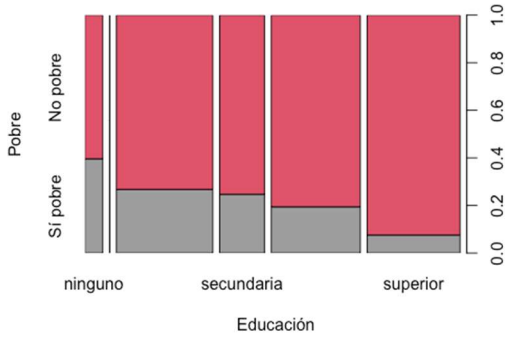


Gráfico 13: Educación

Encontramos que en cabeceras se encuentra menor proporción de pobreza, de igual forma, hogares con vivienda propia, ya sea pagada o pagándola son menos pobres que aquellos que viven en arriendo o en otras formas, así mismo la educación es relevante pues a niveles de educación mas altos se encuentran menor proporción de pobreza. Para complementar el análisis realizamos una regresión logística para verificar la significancia de los coeficientes:

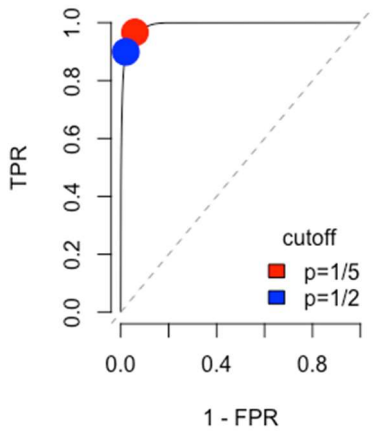
Resultados de la Regresion

	Dependent variable:
	Pobre
ingreso	-0.0000135***
	(0.0000001)
cabecera	3.0649050***
	(0.0504595)
cuartos	-0.2887299***
	(0.0180297)
v. propia pagadando	2.8788520***
	(0.1110614)
v. arriendo	3.5085430***
	(0.0480096)
v.usufructo	0.1266176***
	(0.0435014)
posesion sin titulo	3.2603130***
	(0.0641315)
otro tipo vivienda	3.1097760***
	(0.3844306)
No personas hogar	6.3796790***
	(0.0542698)
hombre	0.0545318*
	(0.0302638)
edad	0.0087683
	(0.0160031)
educación superior	-0.5413813***
	(0.0473782)
horas de trabajo	-0.1669183***
	(0.0413525)
Reg_segsocSubsidiado	0.4316623***
	(0.0346088)
horas_trabajo	-0.0731255***
	(0.0145741)
desempleadoSi	-0.1866937**
	(0.0782935)
tam_emp	0.1310738***
	(0.0206026)
Constant	5.4895030***
	(0.0768016)
Observations	164,960
Log Likelihood	-16,012.2900000
Akaike Inf. Crit.	32,060.5800000
Note:	*p<0.1; **p<0.05; ***p<0.01

Cuadro 3

Encontramos que, salvo la edad, todos los predictores escogidos luego de depurar nuestra base de datos son estadísticamente significativos, por lo tanto, definimos el siguiente modelo:

La curva ROC en datos de entrenamiento del modelo planteado en la regresión lo podemos observar en la grafica 12.



Gráfica 14: Curva ROC dentro de muestra

Si bien, la curva ROC muestra una forma de boomerang bien pronunciado, recordemos que esto se está evaluando dentro de muestra, lo cual nos puede indicar que el modelo esta sobre ajustado, en ese sentido, y teniendo en cuenta que la variable ingreso no está en la base test, procedemos a retirar dicha variable y nos quedamos con el siguiente modelo.

Predictor	Descripción
Cabecera	Indica si el hogar vive en cabecera o fuera de ella
Cuartos	No de cuartos de la vivienda
v. propia pagando	Vivienda propia pero la esta pagando
v. arriendo	Viven en arriendo
v.usufructo	En usufructo
posesión sin titulo	Tienen posesión pero no titulo
otro tipo de vivienda	Vivienda Otro tipo
No personas hogar	Personas que viven en la vivienda
hombre	1 si es hombre 0 si es mujer
edad	Edad del jefe de hogar
educación superior	Educación del jefe del hogar
horas de trabajo	Horas trabajadas por semana
Regimen Seguridad	En que régimen de seguridad social esta
Desempleado	1 si esta desempleado 0 de lo contrario
Tamaño de la empresa	Tamaño de la empresa empleadora de jefe

Cuadro 4: Predictores y descripciones.

Estimamos entonces el modelo logit utilizando técnica de Cross Validation para evaluar la capacidad de generalización del modelo, usamos 5 folds, estimamos tres modelos diferentes donde combinamos diferentes predictores que nos permitieran tener un trade-off adecuado entre métricas dentro de la muestra, en particular haciendo énfasis en recall o sensibilidad pues nuestro objetivo es predecir la pobreza. Evaluando dicho modelo en Caret nos encontramos con un Private-Score = 81,9%, se genero un modelo que inclui la construcción de una variable predicha de ingreso en la base de test, no obstante como anticipamos, dicho modelo estaba sobre ajustado y genero un Private – Score = 80,16%.

5. Modelo final:

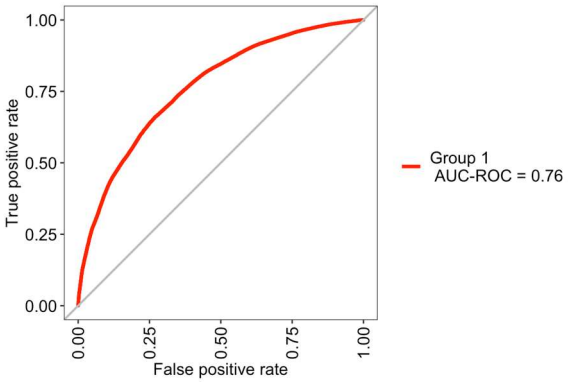
Out of Sample Prediction

Para evaluar el desempeño de los modelos fuera de muestra se procedio a partir la base de datos train hogares, en Entrenamiento y Prueba (70% y 30% respectivamente) y se procedo a estimar el modelo descrito en el cuadro 4. Ahora haciendo énfasis en el desempeño out of sample. En esta etapa evaluamos el modelo Logit y hacemos una variación con LDA, no obstante este ultimo solo incluimos variables continuas dado que incluirlas implica hacer transformaciones a variables numéricas, pues en LDA asumimos que todas tienen una distribución normal y una matriz de covarianza igual entre grupos. Dado que las variables categóricas no tienen distribución normal, no las incluimos en el modelo LDA por tanto usamos el siguiente modelo:

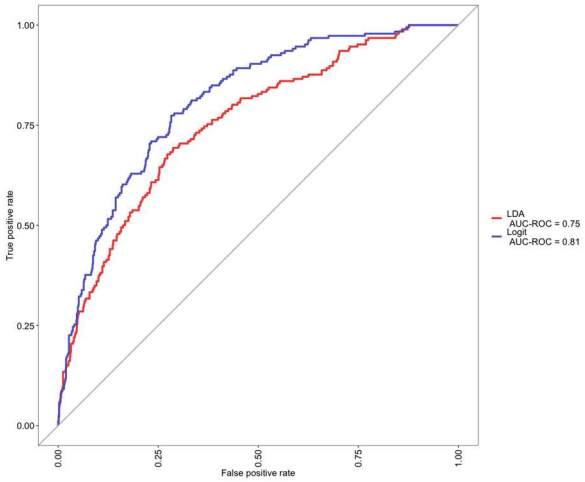
Predictor	Descripción
Cabecera	Indica si el hogar vive en cabecera o fuera de ella
Cuartos	No de cuartos de la vivienda
No personas hogar	Personas que viven en la vivienda
Tiempo trabajando	Años trabajando en su empleo actual
horas de trabajo	Horas trabajadas por semana
tam_emp	Tamaño de la empresa empleadora de jefe

Cuadro 5. Modelo LDA

Evaluando su performance in sample encontramos que si bien la curva ROC de LDA es de 0.76 (Grafica 13), contrastándola con la capacidad predictiva dentro de muestra del modelo Logit, el performance de este



Grafica 15. Curva ROC para LDA



Grafica 16. Curva ROC LDAvsLogit

Estimamos y comparamos con nuestro modelo Logit, dado que una de nuestras métricas de mayor interés es la sensibilidad pues es la que nos muestra la True Positive Rate (TPR) pues nuestra intención es clasificar bien los casos en los cuales el hogar es pobre , vemos que el modelo logit se comporta mejor que LDA en la prueba del modelo (Out of Sample)

	Out Sample						
Modelo	AUC	Precision	Recall	F	ROC	Sens	Spec
Logit	0.9470346	0.8257389	0.9720061	0.8929222	0.8251503	0.9720061	0.1804784
LDA	0.9206166	0.8271456	0.9698585	0.8928351	0.7647142	0.9692774	0.1926920

Desbalance de clases

Uno de los principales retos para la predicción es el desbalance de clases, nuestro conjunto de datos de entrenamiento es del 20,01%, por tanto tiene un grado de desbalance leve pero muy cerca de ser moderado, para verificar que nuestro modelo pueda predecir bien fuera de muestra dividimos nuestro conjunto de datos de entrenamiento en; entrenamiento, evaluación y prueba, asi:

Conjunto	No Obs
Training	115.473
Testing	32.990
Evaluation	16.497

De esta forma se procedio a entrenar el modelo en los datos de training, usamos target de los hiperparametros relevantes; Sensitivity, Specificity, ROC y Accuracy, para que nos permitieran validar a su vez que existía forma de optimizar la capacidad predictiva del modelo.

In Sample - Hiperparametros

Modelo	Target	Lambda	ROC	Sens	Spec	Accuracy	Kappa
Logit	N/A		0.8244825	0.952315	0.3054892	0.8228244	0.3178178
Logit	Spec		0.8244825	0.952315	0.3054892	0.8228244	0.3178178
Logit_lasso	Sepc	0.0060565070	0.8241050	0.9620599	0.261063168	0.8217246	0.286102116
Logit_lasso	Sens	1,023293	0.8111025	1,0000000	0.000000	0.7998060	0.000000
Logit_lasso	Accuracy	0.0060565070	0.8241050	0.9620599	0.261063168	0.8217246	0.286102116
Logit_lasso	Roc	0.0060565070	0.8241050	0.9620599	0.261063168	0.8217246	0.286102116

Alternative Cut Off

Evaluamos el punto de corte optimo, que mas se acerca a nuestro ideal y nos arroja lo siguiente:

threshold	specificity	sensitivity
0.7817639	0.7344838	0.7482189

	Hat Con regla de Bayes	
pobre	No	Si
No	525	12669
Si	890	2413

	Hat Con Threshold > 0,78	
pobre	No	Si
No	3322	9872
Si	2426	877

El threshold me cambia de forma importante los verdaderos positivos(disminuyen) pero los verdaderos negativos aumentan.

Remuestreo

Se aplicaron las siguientes técnicas de remuestreo; upsampling y downsampling, cuyos indicadores son inferiores a los modelos estimados anteriormente y a nuestro modelo benchmark, esto puede deberse a que el desbalance de clases es leve por tanto el remuesreo no mejora las métricas de forma significativa.

	In Sample - Remuestreo					
Modelo	Lambda	ROC	Sens	Spec	Accuracy	Kappa
Logit_lasso_upsamp	0.0168957618	0.8246148	0.7302287	0.7585430	0.7443858	0.4887717
Logit_lasso_downsamp	0.0168957618	0.8225518	0.7267810	0.7579269	0.7423540	0.4847079

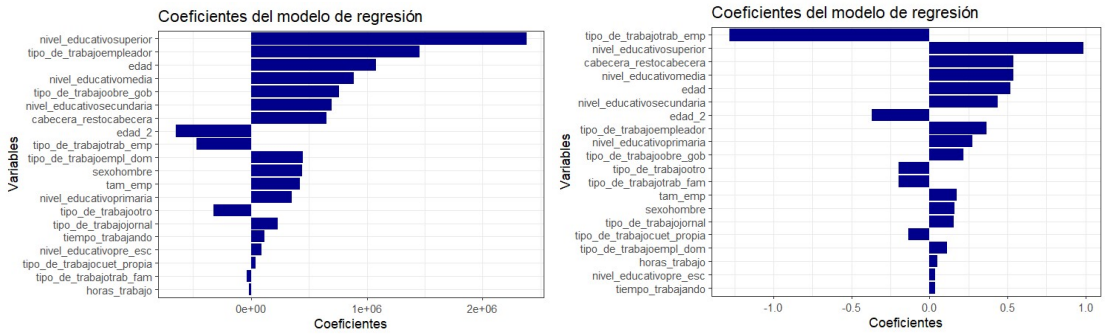
4.2 Modelo de Regresión de Ingreso

La medición del ingreso es una variable muy importante a la hora de estudiar las condiciones de vida de las personas y los hogares, pues mediante ella se puede conocer si las personas o los hogares logran satisfacer la mayor parte de sus necesidades, el no hacerlo revela condiciones de los hogares difíciles para al menos cubrir lo esencial de sus necesidades.

En esta parte del trabajo, el objetivo es estimar un modelo de ingreso de los hogares para identificar las características que lo determinan o explican su comportamiento. Se realiza estimaciones del ingreso en función de sus predictores, luego ver si se puede reducir el espacio de sus determinantes a un sub grupo de variables que verdaderamente sean relevantes para después predecir el ingreso y clasificar indirectamente si los hogares son pobres o no.

Variable dependiente: es el ingreso a nivel de hogar de las familias, agrega todos los ingresos de los individuos que conforman una familia.
Variables independientes: se toman variables de las características del del jefe del hogar como edad, sexo, nivel educativo, el tiempo que lleva trabajando, las horas que trabaja a la semana, el tipo de trabajo que tiene y el tamaño de la empresa donde trabaja, y otra variable de la ubicación del hogar, es decir si se ubica en cabecera o no.

Las primeras aproximaciones son modelos lineales del ingreso, el primero (lado izquierdo) tiene el ingreso expresado en niveles y el segundo (lado derecho) el ingreso en logaritmos para reducir la amplitud de la variable dependiente.



Gráfica 17: Coeficientes

En ambos modelos las características del jefe de hogar y la ubicación del hogar son determinantes del ingreso al ser estadísticamente significativas. El ajuste de los modelos es bueno, medido a partir del R-cuadrado, en el primero es de 54,9% y en el modelo semilogarítmico es de 99%, lo que implica que las variables incluidas en el modelo explican el comportamiento del ingreso, aunque el último observa sobreajuste.

La variable que más destaca en el modelo lineal, según la magnitud de los coeficientes, es el nivel de educación alcanzando por el jefe del hogar, en particular el tener una educación superior determina un mayor nivel de ingresos, le sigue el tipo de trabajo, en este caso el ser dueño del negocio o patrón hace que el ingreso aumente; también está la ubicación, es decir si el hogar se encuentra en cabecera municipal sus ingresos aumentan, la edad es importante pero su efecto es no lineal, en otras palabras llegado a una edad determinada los ingresos comienzan a disminuir, también existen diferencias en el ingreso del hogar dependiendo si el jefe del hogar es hombre o mujer, siendo los ingresos más altos en el caso de los hombres. Las otras variables como la educación media y el tipo de trabajo que retribuye un pago, aun aportan en el nivel de ingresos, y las variables que le aportan menos al ingreso son el tamaño de la empresa donde trabaja, el tiempo que viene trabajando y las horas que trabaja a la semana.

El modelo semilogarítmico confirma los resultados del modelo lineal en cuanto a que el nivel de educación determina el nivel de ingresos del hogar -cuanto mayor educación tenga el jefe del hogar los ingresos del hogar aumentan-, también si la familia se encuentra en cabecera municipal, sus ingresos aumentan. No obstante, el tipo de trabajo en empresa que no es remunerado le afecta negativamente al ingreso del hogar y de igual modo la ocupación en la familia sin remuneración.

Los modelos de regresión múltiple tienen inconvenientes cuando se incorporan predictores correlacionados (multicolinealidad) y no seleccionan predictores relevantes. Para ello se puede usar modelos de regularización y ajustar los modelos lineales, que consiste en ajustar el modelo con todos los predictores y penalizar de tal modo que las estimaciones de los coeficientes de la regresión tiendan a cero, así evitar el sobre ajuste, reducir la varianza y reducir el efecto de los predictores menos relevantes.

Como el objetivo del modelo es predecir el ingreso e indirectamente clasificar si el hogar es pobre o no, necesitamos obtener un modelo que tenga el mejor poder predictivo, lo cual se hará mediante métodos de regularización como Ridge, Lasso y Elastic Net.

Modelo Ridge:

Este método consiste en estimar el modelo de regresión dependiendo del hiperparametro lambda que determina el grado de penalización. El valor que se utiliza abarca el rango 10^{10} a 10^{-2}, lo que significa que va desde un modelo muy restrictivo (no tiene ningún predictor) hasta un modelo equivalente al estimado por mínimos cuadrados. Se parte del modelo de regresión lineal anterior.

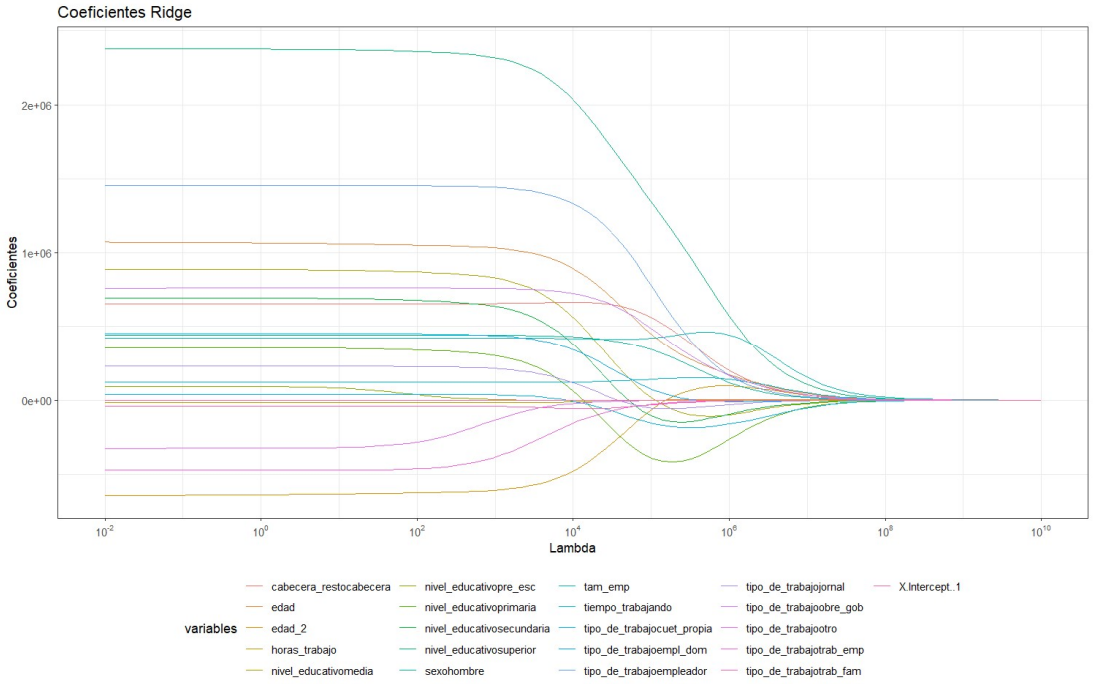


Gráfico 18: Ridge

El graficó muestra que a medida que aumenta el valor de lambda, el valor de los coeficientes tiende a cero porque la regularización es mayor. Como se esperaba con la estimación por mínimos cuadrados ordinarios, con este modelo Ridge vemos que ninguno de los coeficientes es cero.

Modelo Lasso:

La regularización mediante Lasso a diferencia de Ridge fuerza a que los coeficientes de los predictores lleguen a cero, al igual que Ridge el grado de penalización está controlado por el hiperparametro lambda (rango 10^{10} a 10^{-2}).

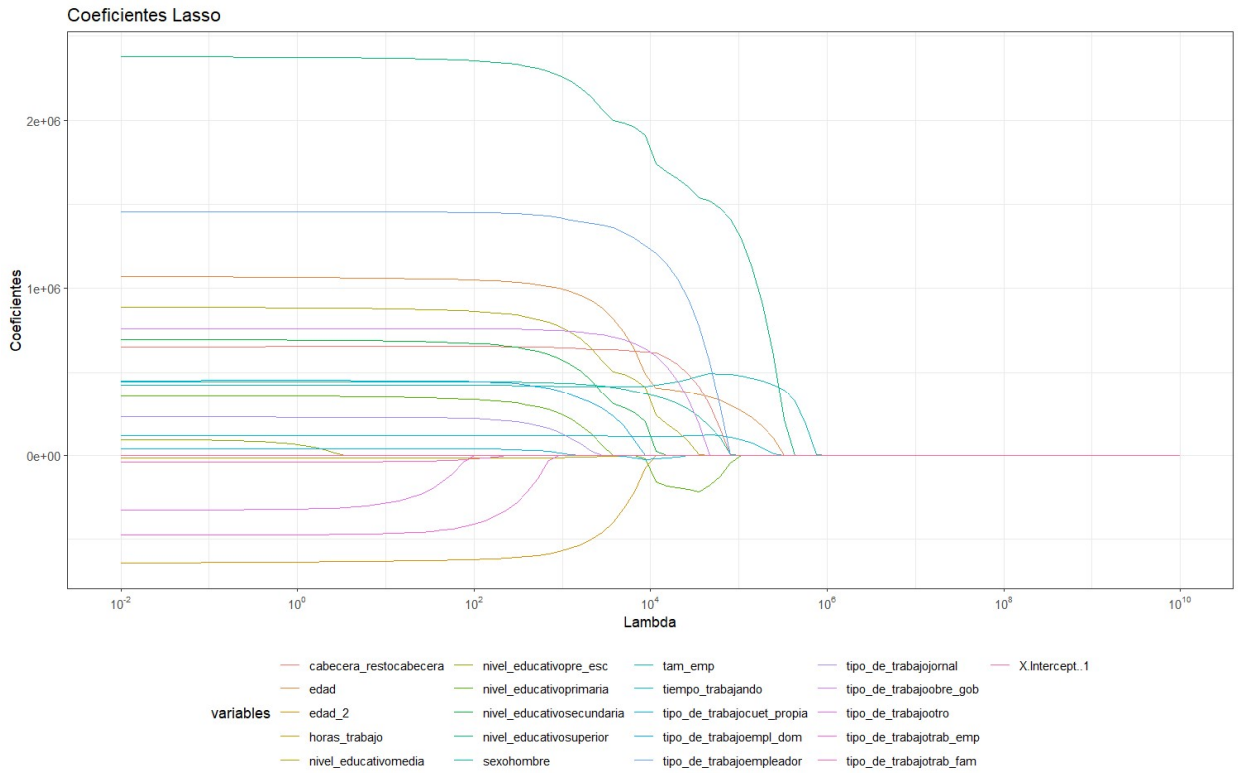


Gráfico 19: Lasso 1

Cuando lambda es igual a cero, el modelo resultante es equivalente al modelo lineal por mínimos cuadrados ordinarios y conforme aumenta lambda, mayor es la penalización y mas predictores quedan excluido. En efecto, Lasso excluye del modelo de ingreso característica que no son levantes como el tipo de trabajo gratuito en empresa o familia, las horas de trabajo a la semana, entre otros.

Elección de parámetros de penalización:

En esta sección se plantean diversos modelos de ingreso del hogar complejos añadiendo formas polinómicas y de interacción entre las características del hogar. Los resultados que se muestran en los siguientes gráficos y métricas corresponden al modelo de ingreso con variables independientes expresadas en el caso de la edad con un polinomio de grado 3 y su interacción con las otras variables, y la interacción entre estas (anexo: mejor modelo). Para identificar el valor de lambda que arroja el mejor modelo se recurre a validación cruzada con ocho folds.

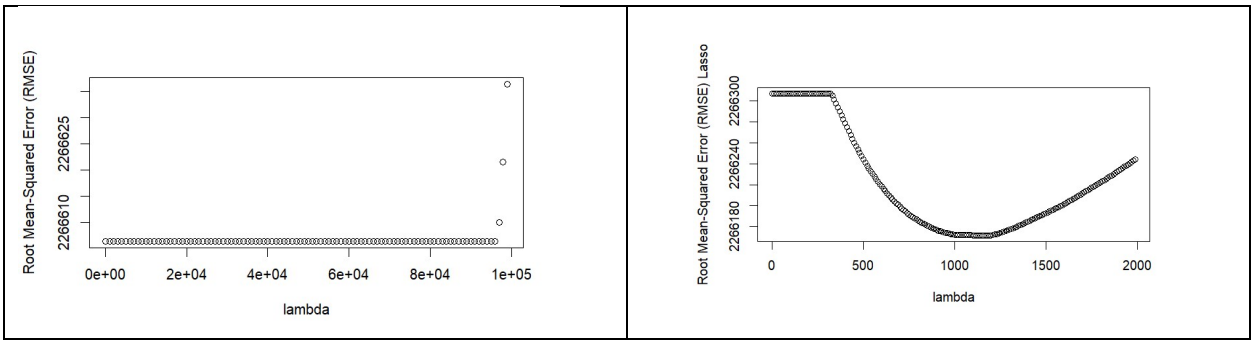


Gráfico 17: Lambda

Los gráficos muestran el Root Mean-Squared Error y los lambdas del modelo plantado. En el modelo Ridge el lambda óptimo es 96000.1 y en el modelo Lasso es 1220.01, y los hiperparametros en el Elastic Net son alpha 1 y lambda 1000.1. El menor RMSE le corresponde al modelo Lasso, que tiene el menor RMSE, superando a los modelos OLS, Ridge y Elastic Net.

Root Mean-Squared error			
OLS	*RIDGE*	*LASSO*	*EN*
2,267,717	2,266,606	2,266,307	2,268,232

Tabla 1

Por lo tanto, el mejor modelo para predecir el ingreso es el regularizado por Lasso por tener un mejor ajuste. Además, Lasso regulariza el modelo de ingreso (en niveles) plateado, excluyendo variables como: nivel educativo escolar, el trabajado como jornalero y algunas interacciones entre edad y nivel educativo, lo que significa que los menores niveles de educación alcanzados por el jefe de hogar y estar en un trabajo con un pago por jornal no le aportan a generar mas ingresos, tampoco el contar con experiencia en el mismo trabajo y su edad implicará incrementos en el ingreso.

Clasificación de pobreza con ingreso predicho:

Los resultados con los que se predice el ingreso corresponden al modelo de regularización Lasso, tal como se muestra en el anexo (mejor modelo). La clasificación de si un hogar es pobre o no, se obtiene de los ingresos estimados a nivel de hogar y luego se contrasta con la línea de pobreza de la base de datos train_hogares. Si la predicción del ingreso es mayor a la línea de pobreza, entonces se clasifica como no pobre y si el ingreso está por debajo de esta línea se clasifica como pobre. Se encuentra que el modelo logra predecir a no más del 1% de la muestra del test_hogares como pobres, cuando en la muestra train_hogares la proporción de pobres llega a más del 18%. Con ello se evidencia que clasificar si un hogar es pobre o no a partir del ingreso no es lo más conveniente, se corre el riesgo de subestimar la clasificación de pobres. Esto se explica en la mucha concentración de hogares con ingresos alrededor de la línea de pobreza y ante un leve shock cambian rápidamente de condición, además el reporte de ingresos de los miembros del hogar está sujeto a problemas como la subdeclaración de ingresos, la no respuesta a la encuesta (no es aleatoria), entre otros, que agrava el problema de clasificación

También hicimos estimaciones de modelos con otras especificaciones más complejas, como el ingreso expresado en logaritmos, cuyas métricas no son superiores a las del modelo de regularización Lasso, en particular la RMSE es más alto en los otros modelos. Asimismo, con los modelos semilogarítmicos se llega a clasificar como pobres a menos de 100 hogares de la muestra train.

Estos resultados nos permite recomendar a no usar modelos de predicción de ingresos de hogares para clasificar si un hogar se encuentra en condición de pobreza o no, ya que existen otros factores que determinan la pobreza, tales como condiciones habitacionales y de vivienda, el acceso a servicios públicos y características socio demográficas de los que integran el hogar.

5. Conclusiones:

Anexos

Resultados (mejor modelo)

	Variable Dependiente: Ingreso del hogar			
	OLS	RIDGE	LASSO	ELASTIC_NET
Constante	2,197,958	2,197,958	2,197,958	2,197,958
cabecera_restocabecera	202,562	206,483	203,663	203,493
nivel_educativopre_esc	12,948	2,240	0	0
nivel_educativoprimaria	38,348	-37,247	27,201	28,984
nivel_educativosecundaria	64,127	3,861	54,205	55,804
nivel_educativomedia	97,185	19,212	83,618	85,883
nivel_educativosuperior	647,612	544,105	634,451	636,689
tiempo_trabajando	117,154	112,739	115,647	115,949
tipo_de_trabajoobre_gob	148,866	140,163	143,698	144,569
tipo_de_trabajoempl_dom	32,381	14,662	27,032	27,902
tipo_de_trabajocuet_propia	46,144	-6,055	24,635	28,030
tipo_de_trabajoempleador	339,155	272,060	320,776	323,790
tipo_de_trabajotrab_fam	-589	-8,099	-3,097	-2,741
tipo_de_trabajotrab_emp	-5,878	-10,329	-8,382	-7,967
tipo_de_trabajojornal	-6,438	-2,098	0	0
tipo_de_trabajootro	-3,853	-5,169	-3,748	-3,797
horas_trabajo	26,884	15,022	15,056	16,614
tam_emp	448,672	413,172	443,134	444,057
`poly(edad, 3)1:sexomujer`	13,976	7,812	-4,678	-5,013
`poly(edad, 3)2:sexomujer`	22,794	3,780	37,127	37,665
`poly(edad, 3)3:sexomujer`	-4,773	-13,198	-12,683	-12,843
`poly(edad, 3)1:sexohombre`	185,295	160,998	150,406	151,245
`poly(edad, 3)2:sexohombre`	-31,469	-52,878	-7,373	-7,251
`poly(edad, 3)3:sexohombre`	12,741	627	0	0
`sexohombre:nivel_educativopre_esc`	-28,927	-4,313	0	0
`sexohombre:nivel_educativoprimaria`	41,929	5,042	13,427	17,168
`sexohombre:nivel_educativosecundaria`	55,334	26,251	33,853	36,703
`sexohombre:nivel_educativomedia`	157,389	110,972	127,406	131,334
`sexohombre:nivel_educativosuperior`	430,137	395,118	401,984	405,659
`poly(edad, 3)1:nivel_educativopre_esc`	-3,511	4,237	0	0
`poly(edad, 3)2:nivel_educativopre_esc`	22,608	4,901	1,700	2,082
`poly(edad, 3)3:nivel_educativopre_esc`	NA	7,106	1,185	1,485
`poly(edad, 3)1:nivel_educativoprimaria`	23,960	22,562	40,777	41,054
`poly(edad, 3)2:nivel_educativoprimaria`	20,206	21,529	0	0
`poly(edad, 3)3:nivel_educativoprimaria`	-15,061	-11,614	-3,408	-3,932
`poly(edad, 3)1:nivel_educativosecundar`	79,358	75,973	89,934	90,172
`poly(edad, 3)2:nivel_educativosecundar`	36,828	35,904	22,749	23,207
`poly(edad, 3)3:nivel_educativosecundar`	-7,263	-5,719	-355	-673
`poly(edad, 3)1:nivel_educativomedia`	156,272	147,083	170,616	171,048
`poly(edad, 3)2:nivel_educativomedia`	55,206	50,957	35,975	36,615
`poly(edad, 3)3:nivel_educativomedia`	-10,967	-10,460	-3,069	-3,258
`poly(edad, 3)1:nivel_educativosuperior`	496,040	461,299	510,753	511,234
`poly(edad, 3)2:nivel_educativosuperior`	59,665	44,174	40,030	40,781
`poly(edad, 3)3:nivel_educativosuperior`	53,843	47,538	58,063	58,511
`sexohombre:tipo_de_trabajoobre_gob`	-3,961	16,365	1,573	781
`sexohombre:tipo_de_trabajoempl_dom`	6,888	9,666	7,019	7,023
`sexohombre:tipo_de_trabajocuet_propi`	-52,298	-21,952	-33,089	-35,995
`sexohombre:tipo_de_trabajoempleador`	-43,455	12,551	-25,179	-28,168
`sexohombre:tipo_de_trabajotrab_fam`	-12,385	-7,698	-9,473	-9,920
`sexohombre:tipo_de_trabajotrab_emp`	-16,573	-13,695	-15,175	-15,345
`sexohombre:tipo_de_trabajojornal`	21,101	8,966	12,043	12,546
`sexohombre:tipo_de_trabajootro`	-1,008	-480	-7	-160
`poly(edad, 3)1:tipo_de_trabajoobre_got`	29,276	40,312	31,108	31,144
`poly(edad, 3)2:tipo_de_trabajoobre_got`	-2,321	-525	0	0
`poly(edad, 3)3:tipo_de_trabajoobre_got`	-14,619	-17,439	-13,145	-13,112
`poly(edad, 3)1:tipo_de_trabajoempl_do`	-1,812	-160	0	0
`poly(edad, 3)2:tipo_de_trabajoempl_do`	-21,485	-17,878	-18,372	-18,810
`poly(edad, 3)3:tipo_de_trabajoempl_do`	-4,506	-3,334	-2,844	-3,025
`poly(edad, 3)1:tipo_de_trabajocuet_pro`	-127,240	-103,998	-122,000	-122,523
`poly(edad, 3)2:tipo_de_trabajocuet_pro`	-40,885	-22,692	-34,639	-35,430
`poly(edad, 3)3:tipo_de_trabajocuet_pro`	8,676	15,353	4,516	5,412
`poly(edad, 3)1:tipo_de_trabajoemplead`	1,533	17,095	2,342	2,336
`poly(edad, 3)2:tipo_de_trabajoemplead`	-5,432	-501	-1,847	-2,365
`poly(edad, 3)3:tipo_de_trabajoemplead`	-13,626	-8,580	-13,378	-13,367
`poly(edad, 3)1:tipo_de_trabajotrab_fam`	-5,152	-3,007	-2,898	-3,017
`poly(edad, 3)2:tipo_de_trabajotrab_fam`	2,117	4,761	0	0
`poly(edad, 3)3:tipo_de_trabajotrab_fam`	-3,436	-2,682	-1,439	-1,563
`poly(edad, 3)1:tipo_de_trabajotrab_emp`	-11,491	-9,547	-9,148	-9,523
`poly(edad, 3)2:tipo_de_trabajotrab_emp`	-2,119	-447	-94	-453
`poly(edad, 3)3:tipo_de_trabajotrab_emp`	3,404	3,944	674	1,179
`poly(edad, 3)1:tipo_de_trabajojornal`	-33,485	-31,211	-31,490	-31,716
`poly(edad, 3)2:tipo_de_trabajojornal`	-1,305	2,051	0	0
`poly(edad, 3)3:tipo_de_trabajojornal`	6,065	7,664	4,014	4,426
`poly(edad, 3)1:tipo_de_trabajootro`	6,100	5,715	5,059	5,257
`poly(edad, 3)2:tipo_de_trabajootro`	2,184	2,600	1,640	1,741
`poly(edad, 3)3:tipo_de_trabajootro`	-1,175	-1,346	-1,148	-1,155
`sexohombre:horas_trabajo`	-47,071	-19,985	-19,902	-23,295

Root Mean-Squared error

OLS	*RIDGE*	*LASSO*	*EN*
2,267,717	2,266,606	2,266,307	2,268,232

Resultados

Variable Dependiente: Ingreso del hogar				
	OLS	*RIDGE*	*LASSO*	*ELASTIC_NET*
(Intercept)	2,197,958	2,197,958	2,197,958	2,197,958
cabecera_restocabecera	186,639	199,876	188,825	188,616
edad	763,583	359,371	774,034	730,750
edad_2	-426,905	-4,863	-418,984	-381,167
hombre	-173,491	-80,135	-102,568	-157,839
nivel_educativopre_esc	-4,806	-4,860	-1,945	-3,331
nivel_educativoprimaria	127,700	-3,817	147,758	127,799
nivel_educativosecundaria	195,861	85,273	212,951	196,589
nivel_educativomedia	298,220	149,219	322,566	299,185
nivel_educativosuperior	749,798	588,801	775,562	750,681
tiempo_trabajando	-72,181	223,439	231,287	213,493
tipo_de_trabajoobre_gob	219,751	207,672	217,252	218,925
tipo_de_trabajoempl_dom	48,540	32,760	47,917	47,979
tipo_de_trabajocuet_propia	54,723	4,497	50,648	51,187
tipo_de_trabajoempleador	361,031	293,606	354,826	356,670
tipo_de_trabajotrab_fam	4,203	-5,223	2,488	2,750
tipo_de_trabajotrab_emp	-7,130	-12,845	-7,222	-7,465
tipo_de_trabajojornal	24,674	13,999	27,152	25,165
tipo_de_trabajootro	-3,952	-4,872	-3,868	-3,899
horas_trabajo	25,672	29,170	27,220	26,406
tam_emp	428,322	402,040	426,655	427,199
`edad:hombre`	267,834	136,581	229,557	260,738
`edad_2:hombre`	-3,243	42,689	8,812	-223
`hombre:nivel_educativopre_esc`	6,160	3,982	3,255	4,627
`hombre:nivel_educativoprimaria`	36,457	18,259	5,377	28,665
`hombre:nivel_educativosecundaria`	47,495	33,227	22,495	40,924
`hombre:nivel_educativomedia`	128,670	102,444	91,572	118,854
`hombre:nivel_educativosuperior`	430,786	406,010	394,043	421,577
`edad:tiempo_trabajando`	662,553	40,135	14,291	73,403
`edad_2:tiempo_trabajando`	-493,628	-169,400	-140,497	-184,256
`hombre:tipo_de_trabajoobre_gob`	-37,686	-15,474	-36,429	-38,292
`hombre:tipo_de_trabajoempl_dom`	4,441	6,753	4,241	4,358
`hombre:tipo_de_trabajocuet_propia`	-56,521	-33,151	-56,053	-56,439
`hombre:tipo_de_trabajoempleador`	-42,482	15,528	-37,501	-39,394
`hombre:tipo_de_trabajotrab_fam`	-12,296	-9,148	-11,362	-11,748
`hombre:tipo_de_trabajotrab_emp`	-18,348	-16,857	-18,352	-18,410
`hombre:tipo_de_trabajojornal`	10,138	9,375	6,718	9,145
`hombre:tipo_de_trabajootro`	-2,425	-1,962	-2,182	-2,266
`hombre:horas_trabajo`	-73,781	-78,345	-79,713	-76,058

Root Mean-Squared error

OLS	*RIDGE*	*LASSO*	*EN*
2,293,922	2,296,142	2,294,158	2,297,345

Resultados

Variable Dependiente: logaritmo del ingreso del hogar				
	OLS	*RIDGE*	*LASSO*	*ELASTIC_NET*
Constante	14.1893	14.1893	14.1893	14.1893
cabecera_restocabecera	0.1640	0.1661	0.1641	0.1640
edad	0.4549	0.1645	0.3382	0.3741
edad_2	-0.3059	-0.0231	-0.1889	-0.2251
sexohombre	0.0185	-0.0351	0	0
nivel_educativopre_esc	-0.0013	-0.0021	0	0
nivel_educativoprimaria	0.1191	0.0163	0.0976	0.1078
nivel_educativosecundaria	0.1446	0.0590	0.1276	0.1356
nivel_educativomedia	0.2198	0.1032	0.1965	0.2074
nivel_educativosuperior	0.3903	0.2635	0.3668	0.3772
tiempo_trabajando	0.1850	0.0999	0.0998	0.1007
tipo_de_trabajoobre_gob	0.0706	0.0732	0.0657	0.0697
tipo_de_trabajoempl_dom	0.0126	0.0099	0.0110	0.0123
tipo_de_trabajocuet_propia	-0.0685	-0.0719	-0.0709	-0.0694
tipo_de_trabajoempleador	0.0993	0.0841	0.0887	0.0970
tipo_de_trabajotrab_fam	-0.0103	-0.0122	-0.0108	-0.0106
tipo_de_trabajotrab_emp	-0.0285	-0.0298	-0.0287	-0.0287
tipo_de_trabajojornal	0.0012	0.0049	0	0.0011
tipo_de_trabajootro	-0.0035	-0.0032	-0.0023	-0.0031
horas_trabajo	0.0531	0.0576	0.0542	0.0540
tam_emp	0.1787	0.1747	0.1792	0.1792
`edad:hombrehombre`	0.0086	0.0464	0.0336	0.0354
`edad_2:hombrehombre`	0.0184	-0.0077	0	0
`hombrehombre:nivel_educativopre_esc`	0.0018	0.0015	0	0.0003
`hombrehombre:nivel_educativoprimaria`	0.0013	0.0348	0.0054	0.0067
`hombrehombre:nivel_educativosecundaria`	0.0042	0.0311	0.0067	0.0081
`hombrehombre:nivel_educativomedia`	0.0299	0.0666	0.0339	0.0357
`hombrehombre:nivel_educativosuperior`	0.0788	0.1197	0.0820	0.0846
`edad:tiempo_trabajando`	-0.2058	0.0051	0	0
`edad_2:tiempo_trabajando`	0.0510	-0.0865	-0.0766	-0.0764
`hombrehombre:tipo_de_trabajoobre_gob`	-0.0239	-0.0203	-0.0171	-0.0222
`hombrehombre:tipo_de_trabajoempl_dom`	0.0050	0.0052	0.0046	0.0049
`hombrehombre:tipo_de_trabajocuet_propia`	-0.0038	-0.0052	0.0000	-0.0020
`hombrehombre:tipo_de_trabajoempleador`	-0.0206	-0.0056	-0.0087	-0.0177
`hombrehombre:tipo_de_trabajotrab_fam`	-0.0064	-0.0063	-0.0052	-0.0061
`hombrehombre:tipo_de_trabajotrab_emp`	-0.0332	-0.0324	-0.0324	-0.0330
`hombrehombre:tipo_de_trabajojornal`	0.0209	0.0123	0.0204	0.0206
`hombrehombre:tipo_de_trabajootro`	0.0005	0.0004	0	0.0002
`hombrehombre:horas_trabajo`	0.0009	-0.0078	0.0001	0

Root Mean-Squared error

OLS	*RIDGE*	*LASSO*	*EN*
0.8599	0.8617	0.8602	0.8618

Resultados

Variable Dependiente: logaritmo del ingreso del hogar

	OLS	RIDGE	LASSO	ELASTIC_NET
(Intercept)	14.1893	14.1893	14.1893	14.1893
cabecera_restocabecera	0.1659	0.1662	0.1668	0.1667
nivel_educativopre_esc	0.0094	0.0017	0	-0.0002
nivel_educativoprimaria	0.0904	-0.0021	0.0455	0.0502
nivel_educativosecundaria	0.1082	0.0324	0	0
nivel_educativomedia	0.1713	0.0691	0	0
nivel_educativosuperior	0.3535	0.2418	0.3066	0.3111
tiempo_trabajando	0.0380	0.0352	0.0367	0.0371
tipo_de_trabajoobre_gob	0.0603	0.0610	0.0564	0.0579
tipo_de_trabajoempl_dom	0.0069	0.0037	0.0057	0.0061
tipo_de_trabajocuet_propia	-0.0721	-0.0738	-0.0708	-0.0709
tipo_de_trabajoempleador	0.0978	0.0810	0.0865	0.0899
tipo_de_trabajotrab_fam	-0.0141	-0.0142	-0.0126	-0.0130
tipo_de_trabajotrab_emp	-0.0418	-0.0396	-0.0386	-0.0395
tipo_de_trabajojornal	-0.0082	0.0005	0	0
tipo_de_trabajootro	-0.0041	-0.0039	-0.0024	-0.0026
horas_trabajo	0.0500	0.0529	0.0519	0.0519
tam_emp	0.1829	0.1749	0	0.1824
`poly(edad, 3)1:sexomujer`	0.0540	0.0048	0	0.0003
`poly(edad, 3)2:sexomujer`	-0.0395	-0.0169	0	-0.0001
`poly(edad, 3)3:sexomujer`	0.0064	-0.0065	-0.0062	-0.0066
`poly(edad, 3)1:sexohombre`	0.1054	0.0355	0.0254	0.0270
`poly(edad, 3)2:sexohombre`	-0.0672	-0.0330	0	0
`poly(edad, 3)3:sexohombre`	0.0295	0.0103	0	0.0092
`sexohombre:nivel_educativopre_esc`	-0.0205	-0.0037	0	0
`sexohombre:nivel_educativoprimaria`	0.0176	0.0262	0.0223	0.0221
`sexohombre:nivel_educativosecundaria`	0.0182	0.0267	0.0218	0.0218
`sexohombre:nivel_educativomedia`	0.0511	0.0617	0.0555	0.0557
`sexohombre:nivel_educativosuperior`	0.0948	0.1092	0	0
`poly(edad, 3)1:nivel_educativopre_esc`	-0.0014	0.0042	0.0001	0.0004
`poly(edad, 3)2:nivel_educativopre_esc`	0.0151	0.0032	0.0005	0.0008
`poly(edad, 3)3:nivel_educativopre_esc`	NA	0.0046	0	0
`poly(edad, 3)1:nivel_educativoprimaria`	-0.0105	0.0299	0.0334	0.0339
`poly(edad, 3)2:nivel_educativoprimaria`	0.0232	-0.0008	-0.0139	-0.0139
`poly(edad, 3)3:nivel_educativoprimaria`	-0.0116	-0.0006	0	0.0007
`poly(edad, 3)1:nivel_educativosecundaria`	0.0205	0.0476	0.0509	0.0511
`poly(edad, 3)2:nivel_educativosecundaria`	0.0300	0.0133	0.0030	0.0041
`poly(edad, 3)3:nivel_educativosecundaria`	-0.0060	0.0011	0	0.0016
`poly(edad, 3)1:nivel_educativomedia`	0.0370	0.0726	0.0798	0
`poly(edad, 3)2:nivel_educativomedia`	0.0420	0.0175	0.0056	0.0069
`poly(edad, 3)3:nivel_educativomedia`	-0.0163	-0.0070	-0.0046	-0.0045
`poly(edad, 3)1:nivel_educativosuperior`	0.0813	0.1127	0.1252	0.1256
`poly(edad, 3)2:nivel_educativosuperior`	0.0382	0.0119	0.0013	0.0029
`poly(edad, 3)3:nivel_educativosuperior`	0.0060	0.0134	0	0
`sexohombre:tipo_de_trabajoobre_gob`	-0.0173	-0.0141	0	0
`sexohombre:tipo_de_trabajoempl_dom`	0.0059	0.0060	0.0053	0.0054
`sexohombre:tipo_de_trabajocuet_propia`	0.0009	-0.0050	0.0000	0.0000
`sexohombre:tipo_de_trabajoempleador`	-0.0210	-0.0072	-0.0097	-0.0131
`sexohombre:tipo_de_trabajotrab_fam`	-0.0074	-0.0078	-0.0060	-0.0065
`sexohombre:tipo_de_trabajotrab_emp`	-0.0401	-0.0385	-0.0383	-0.0389
`sexohombre:tipo_de_trabajojornal`	0.0257	0.0123	0.0162	0.0166
`sexohombre:tipo_de_trabajootro`	0.0011	0.0009	0.0000	0.0000
`poly(edad, 3)1:tipo_de_trabajoobre_gob`	-0.0016	0.0018	0.0009	0.0004
`poly(edad, 3)2:tipo_de_trabajoobre_gob`	-0.0020	-0.0020	0.0000	-0.0005
`poly(edad, 3)3:tipo_de_trabajoobre_gob`	-0.0024	-0.0032	-0.0006	-0.0011
`poly(edad, 3)1:tipo_de_trabajoempl_dom`	0.0076	0.0075	0.0069	0.0073
`poly(edad, 3)2:tipo_de_trabajoempl_dom`	-0.0098	-0.0094	0	-0.0101
`poly(edad, 3)3:tipo_de_trabajoempl_dom`	0.0024	0.0022	0.0004	0.0009
`poly(edad, 3)1:tipo_de_trabajocuet_propia`	-0.0218	-0.0179	-0.0150	-0.0162
`poly(edad, 3)2:tipo_de_trabajocuet_propia`	-0.0248	-0.0226	-0.0242	-0.0249
`poly(edad, 3)3:tipo_de_trabajocuet_propia`	0.0020	0.0037	0.0007	0.0009
`poly(edad, 3)1:tipo_de_trabajoempleador`	-0.0057	-0.0013	0	0
`poly(edad, 3)2:tipo_de_trabajoempleador`	0.0050	0.0050	0	0.0031
`poly(edad, 3)3:tipo_de_trabajoempleador`	-0.0077	-0.0059	-0.0049	-0.0058
`poly(edad, 3)1:tipo_de_trabajotrab_fam`	0.0014	0.0016	0.0002	0.0007
`poly(edad, 3)2:tipo_de_trabajotrab_fam`	0.0089	0.0084	0.0063	0.0070
`poly(edad, 3)3:tipo_de_trabajotrab_fam`	-0.0041	-0.0033	-0.0011	-0.0020
`poly(edad, 3)1:tipo_de_trabajotrab_emp`	0.0232	0.0208	0	0
`poly(edad, 3)2:tipo_de_trabajotrab_emp`	0.0137	0.0121	0.0109	0.0117
`poly(edad, 3)3:tipo_de_trabajotrab_emp`	-0.0062	-0.0037	-0.0025	-0.0036
`poly(edad, 3)1:tipo_de_trabajojornal`	-0.0132	-0.0126	-0.0106	-0.0110
`poly(edad, 3)2:tipo_de_trabajojornal`	0.0024	0.0026	0.0007	0.0011
`poly(edad, 3)3:tipo_de_trabajojornal`	0.0031	0.0035	0.0018	0.0021
`poly(edad, 3)1:tipo_de_trabajootro`	0.0045	0.0041	0.0027	0.0030
`poly(edad, 3)2:tipo_de_trabajootro`	0.0007	0.0008	0.0003	0.0005
`poly(edad, 3)3:tipo_de_trabajootro`	0.0010	0.0008	0.0000	0.0000
`sexohombre:horas_trabajo`	0.0141	0.0013	0	0.0072

Root Mean-Squared error

OLS	*RIDGE*	*LASSO*	*EN*
0.8575	0.8578	0.8569	0.8574