Santiago Becerra Ortiz
Karol Gutierrez
Daniel Raquira Gonzalez

Problem set 2

Predicción del ingreso para la clasificación de pobreza

El Banco Mundial tiene como objetivo acabar con la pobreza extrema para 2030. Para lograr este objetivo, son cruciales las técnicas para determinar qué estrategias de reducción de la pobreza funcionan y cuáles no. Pero medir la reducción de la pobreza requiere medir la pobreza en primer lugar, y resulta que medir la pobreza es bastante difícil. En Colombia particularmente "Para el año 2021, el 42,5% de la población estuvo en condición de pobreza. En total son más de 21,02 millones de personas las que subsisten por debajo de la línea de pobreza en Colombia que es de $331.688 mensuales." (Dane, 2021). Esta cifra es preocupante y nos hace pensar como casi la mitad del país es pobre, esta radiografía nacional es una alerta para ver qué factores están frenando el desarrollo y crecimiento del país. El panorama de la pobreza extrema tampoco es alentador, pues para el mismo año la línea de pobreza extrema fue de $161.099 pesos mensuales, es decir que "el 12,2 % de la población colombiana se encuentra en situación de pobreza extrema." (Dane, 2021).

A lo largo de la historia de Colombia, la desigualdad y la pobreza han sido problemas estructurales enormes que afectan a la mayoría de la población. No es por nada que, según datos del DANE, en 2021 más de 21 millones de personas vivían en condición de pobreza y 7,4 millones más vivían en pobreza extrema (DANE, 2021). Por tanto, en una población de casi 50 millones de personas, que cerca de un 40% de la población se encuentre en pobreza (DANE,2022) indica que es necesario priorizar esta temática en las políticas públicas. Dado esto, se hace de vital importancia poder predecir los niveles de pobreza para poder atacar la problemática. Es por esto que el trabajo a continuación busca lograr estas predicciones para entender como es la tendencia de la pobreza y lograr comprender en amplio espectro dicho tema.

El objetivo principal de este problem set es construir un modelo predictivo de la pobreza de los hogares en Colombia. Hay diversas formas de predecir la pobreza, particularmente esta vez nos enfocaremos en dos: el problema de clasificación y el problema de predicción de ingresos. Ahora bien, para poder realizar las predicciones correspondientes se tienen tres bases de datos pertenecientes de la GEIH con unidad de análisis tanto de individuos como de hogares. Una vez se realiza el merge de la base de datos, la base final continúa agregada tanto como a nivel de personas como a nivel de hogares.

Datos

Para los datos tenemos dos fuentes de la GEIH, una a nivel hogares y la otra a nivel individuos. La variable dependiente en la base desagregada a nivel de hogares es Ingtotugarr, pues esta hace referencia al ingreso total de la persona encuestada con imputación de arriendo a propietarios y usufructuarios; mientras que en la base desagregada a nivel individuos es Ingtot, que es el ingreso total de la persona. Estas variables se escogieron como la medida apropiada del ingreso ya que recoge las diferentes formas y tipos de ingreso que una persona recibe, dando así una visión general de este rubro. Además, se entiende que, en el caso de hogares, Ingtotugarr es la variable que se utiliza para comparar con la línea de pobreza, por lo cual resulta muy útil al momento de predecir que hogares son pobres y que hogares no lo son. Su distribución es la siguiente:

Min. 1st Qu. Median Mean 3rd Qu. Max.

0 900000 1581242 2309190 2786762 88833333

Quiere decir que en promedio un hogar en Colombia tiene un ingreso de 2’309.190

Ambos modelos buscan predecir la pobreza de los hogares con la menor cantidad de variables, para tener una mayor precisión y por tanto también de tener la menor tasa de falsos negativos, es decir que las familias pobres sean clasificadas como no pobres. Primero limpiamos las bases, omitimos las variables redundantes y borramos los NA.

Continuando, para seleccionar las variables independientes nos basamos en que “para medir la pobreza, la mayoría de estas encuestas recopilan datos detallados sobre el consumo de los hogares, desde hábitos alimentarios y de transporte hasta acceso a la atención médica y eventos deportivos” (Banco Mundial, 2018). Por lo tanto, seleccionamos variables como la edad (P6040); educación (p6210) que se refiere al nivel educativo más alto alcanzado, pues “una población con mayor nivel educativo generará productos y/o servicios con mayor valor agregado, los cuales tendrán mayor valor en el mercado” (The economist, 2014). Otra variable relevante es si la vivienda ocupada por el hogar es propia, en arriendo, subarriendo, en usufructo o en posesión sin título, para esto utilizamos la variable “P5090”, junto a esta también consideramos prudente utilizar la variable “P5130” que indica que, si se pagara arriendo, cuanto estima que sería el valor de este; Nper, que es el número de personas en el hogar; la posición ocupacional primaria de la persona (P6430) también la incluimos y las horas trabajadas semanalmente de la persona (P800).

Creamos la variable Arri que es la combinación entre dos variables relacionadas con el arriendo mensual (P5130 y P5140), encontramos que la media de esta variable es 484.281, es decir que en Colombia en promedio la cantidad destinada al arriendo mensual es de menos de 500.000 pesos.

En el error del modelo podemos tener fuentes de ingresos imputadas, y también variables demasiado específicas que nos arrojaban en la mayor parte de los casos missing values.

En cuanto a las estadísticas descriptivas, en la gráfica 1 podemos ver la relación inversa entre el número de personas en el hogar y el ingreso del mismo, esto quiere decir que entre más personas haya en el hogar, mayor va a ser su probabilidad de ser un hogar pobre. En la gráfica 2 podemos ver el histograma del ingreso, podemos ver que este se concentra en la zona izquierda del plano, lo cual muestra que muchas personas ingresan montos mucho menores en relación con quienes expanden el histograma a la derecha, esto junto a los valores mínimo y máximo soportan la existencia de una fuerte inequidad en el ingreso.

En el gráfico 3 podemos ver la relación entre las horas trabajadas y la edad de las personas, observamos que cercano a los 30 años es la edad en la que más trabajan en promedio las personas.

Modelos de clasificación

En cuanto al enfoque de modelos de clasificación, se realizaron 7 métodos diferentes con el fin de llegar al método que hiciera las predicciones de mejor forma, es decir, aquel que disminuyera lo máximo posible los falsos negativos al momento de realizar la predicción. Para

tal efecto, se realizaron los siguientes modelos: lasso-logit, ridge-reg, logit-reg, upsampling para lasso, upsampling para ridge, downsampling para lasso y downsampling para ridge.

Antes de comparar los modelos, es importante enunciar las variables que se utilizaron para esta sección del trabajo, la especificación que se decidió trabajar tiene como unidad de análisis los hogares, por lo que se decide trabajar con variables con este nivel de agregación, las cuales son: vivienda propia, arriendo, número de cuartos en el hogar y número de personas en el hogar. La decisión de trabajar con estas variables parte del sustento teórico que relaciona la pobreza con las capacidades del hogar y así como por el costo de vida que se tiene, por lo que las variables escogidas presentan características que permiten entender la pobreza desde una perspectiva de vivienda.

Teniendo lo anterior claro, al momento de probar los diferentes modelos de clasificación y escoger el mejor se compararon los resultados presentados en cuanto a precisión del modelo, así como en la tasa de falsos negativos que tiene cada uno y el AUC, el cual es el principal indicador utilizado para comparar los modelos.

Ya al momento de correr los modelos, se llegó a que el mejor método para realizar la predicción para el problem set debido a su desempeño es el modelo upsampling para lasso, con este se llegó a un AUC de 0.7196 (gráfica 4 muestra la curva de ROC) el cual fue el mayor de todos los modelos realizados, ofreciendo también la menor tasa de falsos negativos la cual es de 7.9% y con una precisión de 0.715. En cuanto a la posible razón por la que este haya sido el mejor método de clasificación, es importante mencionar que la base de datos de training hogares mostraba un claro desbalance en cuanto a la cantidad de observaciones pobres y no pobres, lo cual afectaba los modelos sin métodos de remuestreo. Ergo, tiene sentido que el mejor modelo haya sido uno que primero resuelve el desbalance de clases para después realizar la clasificación. Bajo la misma idea, al hacer todos los modelos mencionados anteriormente todos los que tienen métodos de remuestreo antes de hacer la clasificación poseen un mayor AUC y una menor tasa de falsos negativos que los modelos que no utilizan estos métodos, de forma que está claro el beneficio que se obtiene en este trabajo el realizar esto.

Continuando, al momento de comparar el modelo escogido (upsample lasso) con los otros modelos que se entrenaron, se busca enfocar la comparación al poder predictivo de cada uno de estos. Como se mencionó anteriormente, los modelos que mejores resultados mostraron fueron aquellos que resuelven el desbalance de clases antes de realizar la predicción, mostrando en orden descendente los resultados en cuanto al AUC de los modelos, el orden es: upsample lasso (0.7196), downsample lasso (0.7193), upsample ridge (0.6906), downsample ridge (0.6898), lasso-logit (0.621792), logit-reg (0.621788) y, por último, ridge-reg (0.5664).

Con los resultados anteriores se entiende que el modelo de clasificación con mejor poder predictivo es el upsample lasso tal como se mencionó anteriormente, también se comprueba el hecho de que el desbalance dentro de la base de datos penaliza mucho a los modelos que no corrigen este desbalance antes de realizar la predicción ya que, en efecto, son los 3 modelos con menor AUC. Finalmente, debido a que las estrategias de submuestreo resultaron ser tan importantes al momento de la predicción, se considera necesario aclarar que el método de upsampling lo que hace es simular datos adicionales a la clase que es minoría (en este caso a los hogares que son pobres) para mejorar el balance entre las clases, lo cual con los resultados se observa claramente que es un método efectivo para mejorar el poder predictivo del modelo.

Por último, se presentan los resultados de la predicción del mejor modelo de clasificación en la gráfica 5.

Modelos de regresión de ingresos

Para los modelos de regresión de ingresos se optó por realizar 7 modelos diferentes en el cual se viera o como se iba prediciendo mejor los ingresos. Para esto, se utilizaron ambas bases de datos por aparte, de tal forma que se hicieron dos modelos tradicionales, dos modelos con Ridge y otros dos modelos por medio de la metodología Lasso; uno por cada base (a nivel de personas y a nivel de hogares). En suma, todos los modelos de la misma base tenían las mismas variables explicativas que aparentemente deberían afectar la predicción. Estas variables fueron *edad* (edad que tiene la persona), *posicion* (posición que ocupa en la empresa), *educ* (educación de la persona), *horastr* (horas que trabaja a la semana)*;* respecto a la base de personas; y *vivienda* (Si la vivienda es en arriendo, propio, entre otras), *Nper* (el número de personas que viven juntos), *cuartos*, *Arri* (el arriendo que pagan o el arriendo que pagarían)*;* respecto a los modelos que utilizaron la base de hogares. De los resultados que se expondrán más adelante se extrajeron las variables explicativas que mejor predicen el ingreso *(Ingtot e Ingtotugarr)* y, con ese análisis, se construyó el modelo elegido. De este modelo vale recalcar que se utilizó metodología Lasso y se unificaron las bases de personas y hogares para poder hacer una predicción "robusta" reflejada en los resultados con base en el objetivo planteado inicialmente.

En primer lugar, se corrieron los modelos tradicionales para cada base y sus respectivas variables que aparentemente debían afectar el modelo. Los resultados de los coeficientes son los expuestos por la gráfica 6 y 7 respectivamente; en donde se puede apreciar que variables están influyendo en el ingreso. Ahora bien, como no queremos el enfoque econométrico y caer en consecuencias de overfit, es preferible metodologías de regularización que permitan minimizar la complejidad de los modelos. De esta forma, se pueden tener modelos simples que permitan una mejor generalización. Asimismo, no se recae en encontrar soluciones que solo funcionen para el entrenamiento, sino que permite un buen rendimiento con datos nuevos que vayan surgiendo. Entonces, se procede a utilizar el ridge en ambas especificaciones; encontrando los lambas óptimos y los respectivos coeficientes. Estos resultados se pueden observar en las gráficas 8 y 9 en el caso del ridge para la base personas y las gráficas 10 y 11 para el ridge para la especificación de los hogares. De igual forma, se regulariza por medio de la metodología Lasso en ambas bases y se reportan los resultados en las gráficas 12 y 13 para personas y 14 y 15 para hogares.

Ahora bien, con estos resultados se creó el modelo 7, es decir, el modelo elegido para poder hacer nuestras predicciones sobre pobreza con base en las regresiones del ingreso. De los resultados obtenidos en la metodología y expuestos en las gráficas anteriormente propuesta se tomó la decisión de incluir en el modelo final las siguientes variables: *cuartos, Nper, vivienda, educ, posicion.* Esto debido a que en los modelos anteriores mostraron ser las variables que mejor predecirían el ingreso. La anterior elección va acorde a nuestra intuición como investigadores, como, por ejemplo, el hecho de pensar que el número de personas de un hogar debe afectar los ingresos netos y por ende darnos predicciones de pobreza. Por último, nuestra

variable de interés es la variable Ingtotugarr que tiene en cuenta los ingresos del hogar a la vez de los descuentos por pago en arriendos y otros gastos.

En concordancia, se tomó la decisión de estimar este modelo por medio de la metodología Lasso por dos razones teóricas, la primera, como ya se explicó, es necesario regularizar el modelo para evitar que haya sobreajuste y permitir una mayor validez por fuera de la muestra de nuestra estimación. En segundo lugar, debido a que igualmente se utilizaron las explicativas con una variable de interés distinta en cada base, se tiene la sospecha que haya posibles atributos que sean irrelevantes, por lo que Lasso ayuda en este sentido. De esta forma y con esta metodología se espera que el modelo mejore la generalización y se aseguren los atributos correctos para la predicción. Por último, en los modelos de clasificación encontramos un AUC de 0.7196 para esta metodología; por tanto, nuestra evidencia empírica también sugiere la utilización de este método. Los resultados de este Lasso se pueden ver en las gráficas 16 y 17, en donde se puede observar como las variables escogidas, en totalidad sí están afectando el ingreso.

Para finalizar, con estos resultados predichos de ingresos se realiza el nexo de tal forma que, si se cumplen cierta condición, se determine la clasificación de 1 si es pobre y 0 de lo contrario. La condición que se establece para la clasificación es que si el ingreso total relativo al número de personas que se benefician de ese ingreso (de la misma unidad de gasto) es menor a la línea de pobreza Lp propuesta en los datos suministrados, se clasificara como "pobre". De esta forma se conforma una nueva "base de datos" que recopila las predicciones hechas y clasifica a los hogares, según la condición mencionada, entre pobres y no pobres. Lo anterior es útil en la focalización de políticas públicas y seguimiento a población tratada que, al parecer, van a seguir en pobreza según sus condiciones influyentes de su ingreso. Los resultados se encuentran en la recopilación de nuestra información en la carpeta de views.

Conclusión y anotaciones finales

En conclusión, a pesar de que medir la pobreza es difícil, requiere mucho tiempo y es costoso. Al construir mejores modelos, podemos realizar encuestas con menos preguntas y más específicas que miden de manera rápida y económica la efectividad de las nuevas políticas e intervenciones. Cuanto más precisos sean nuestros modelos, con mayor precisión podremos orientar las intervenciones e iterar las políticas, maximizando el impacto y la rentabilidad de estas estrategias.

En cuanto a la predicción a través de métodos de clasificación, se concluye que el mejor método para predecir la pobreza bajo este contexto es el método de upsample lasso, el cual tiene un AUC de 0.7196 indicando así mejor poder predictivo.

Ahora bien, este resultado se toma en cuenta a la hora de la metodología de predicción para el ingreso. Puesto que el modelo final utiliza Lasso para encontrar los coeficientes respectivos y los valores predichos del ingreso. Para la formación del modelo se utilizaron diferentes modelos con diferentes metodologías para ver la influencia de las variables explicativas sobre la de interés. Luego de analizar los diferentes resultados, se decidió utilizar en el modelo final las siguientes variables: *cuartos, Nper, vivienda, educ, posicion*; entendiendo a esta como las mejores predictoras del ingreso según nuestros resultados.

Con los resultados de este modelo final, denominado por nosotros como modelo 7, se permitió lograr una clasificación de pobreza por medio de los valores predichos de ingreso. Esta

construcción se basó en la comparación entre la linea de pobreza de los datos originales y el ingreso total de los hogares relativo al número de personas que se benefician de ese ingreso. De tal forma, un ingreso final, luego de esta consideración, menor a la línea de pobreza deja como conclusión la clasificación de “Pobre” dentro de nuestro sistema.

Para finalizar, es importante volver al por qué es importante dicha clasificación y predicción. Colombia es de los países más desiguales del mundo y, como ya se mencionó, gran parte de nuestra población se encuentra en pobreza y pobreza extrema. Debido a que nuestros resultados también están pensados para poder extrapolar fuera de la muestra al regularizar y así “simplificar el modelaje, se aporta a una nueva forma de clasificación de pobreza. Este aporte puede verse como una nueva forma de enfocar políticas públicas y programas estatales que busquen reducir la brecha de la desigualdad en Colombia. El poder de la predicción no es quedarse en un computador, es poder plasmarlo en la realidad para la mejoría de la sociedad.

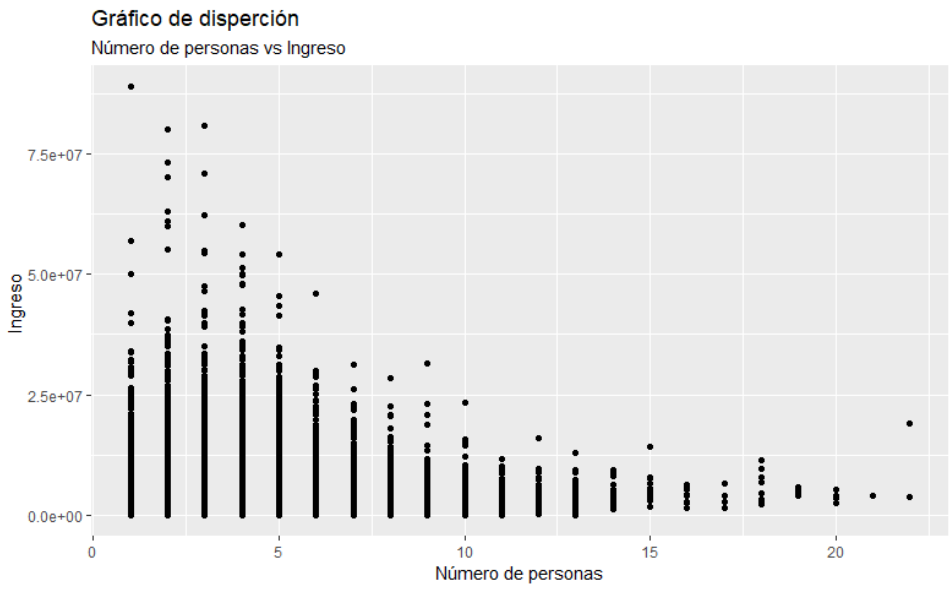
Referencias

La República. (2021). “Más de 21 millones de personas viven en la pobreza y 7,4 millones en pobreza extrema”. Recuperado en: Más de 21 millones de personas viven en la pobreza y 7,4 millones en pobreza extrema (larepublica.co)

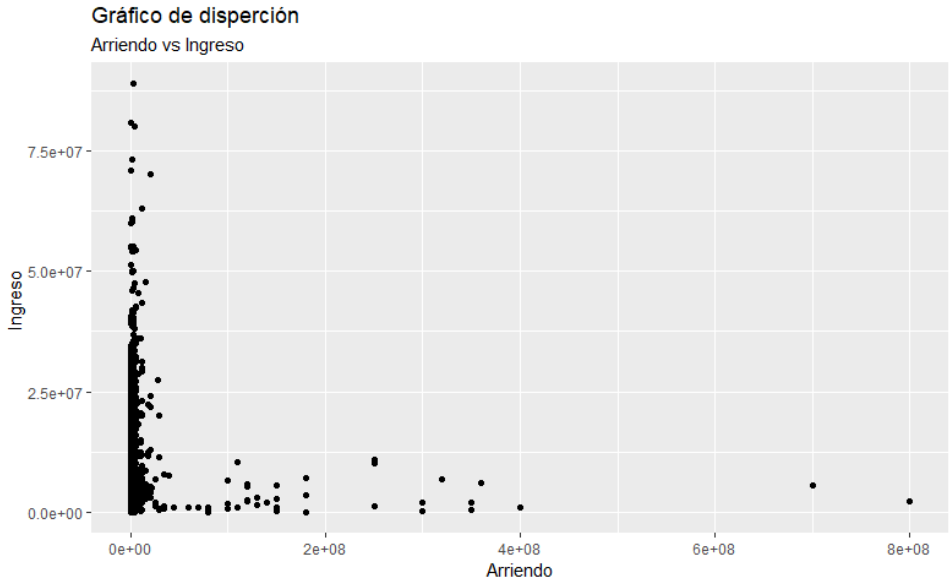
DANE (2022). Información pobreza monetaria nacional 2022. Recuperado de: https://www.dane.gov.co/index.php/estadisticas-por-tema/pobreza-y-condiciones-de-vida/pobreza-monetaria

Anexos

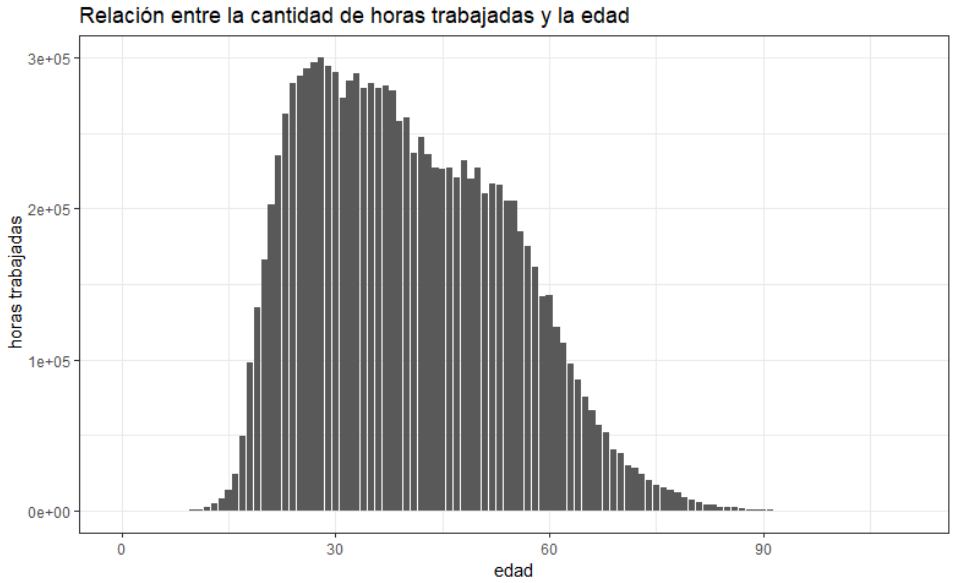
Grafica 1.



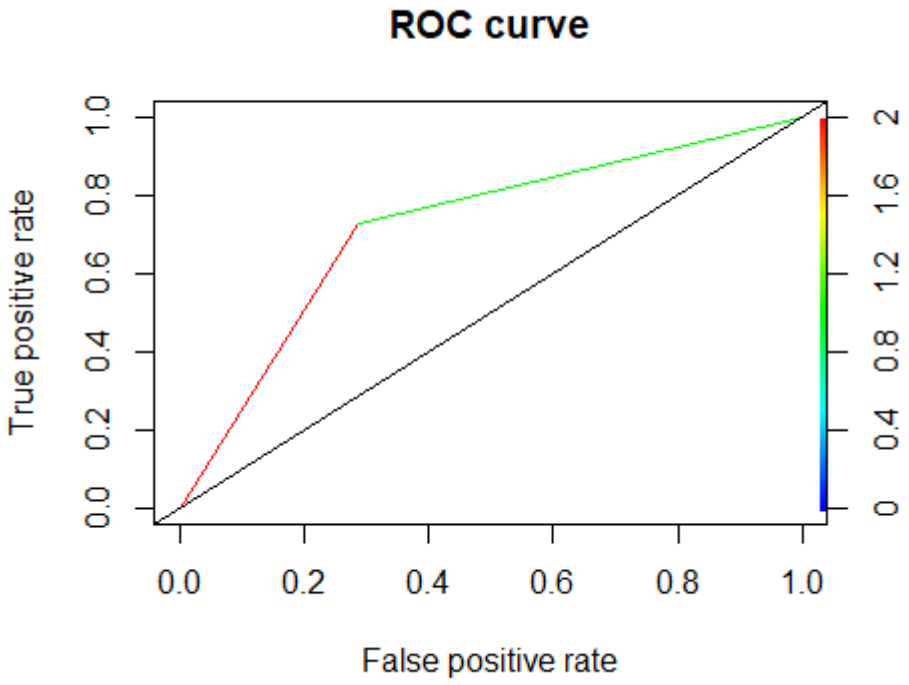
Gráfica 2:



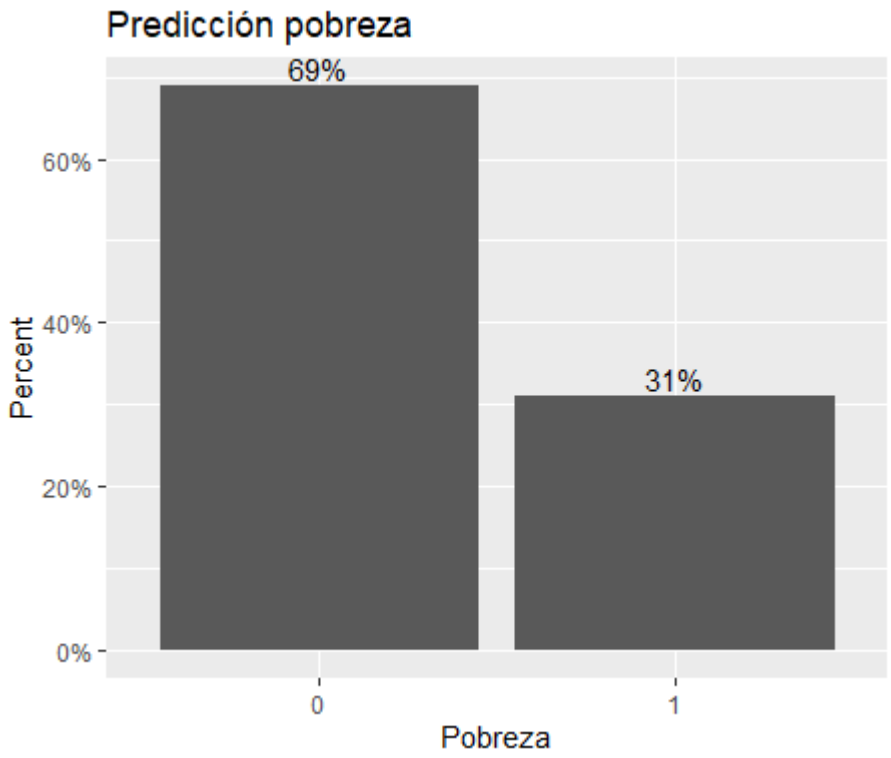
Grafica 3:



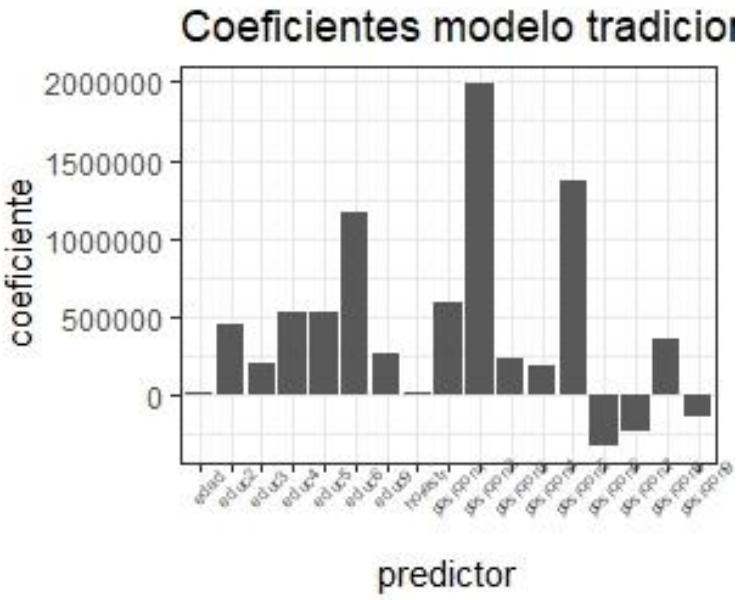
Gráfica 4:



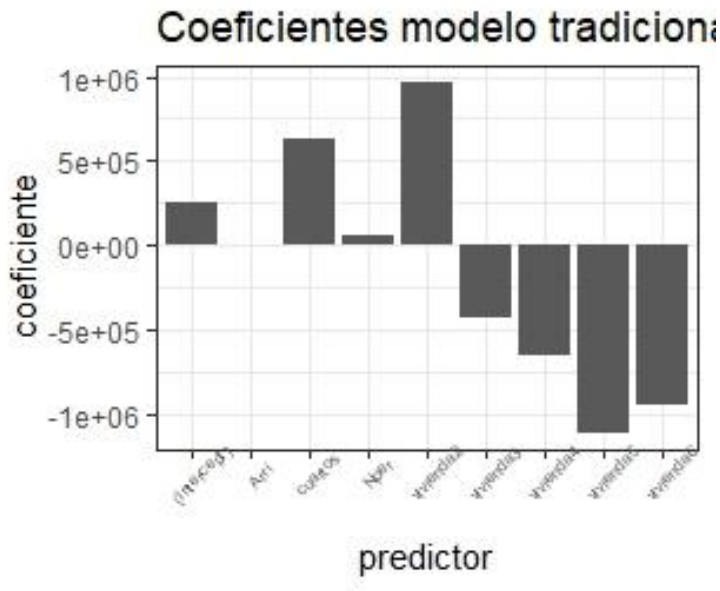
Gráfica 5:



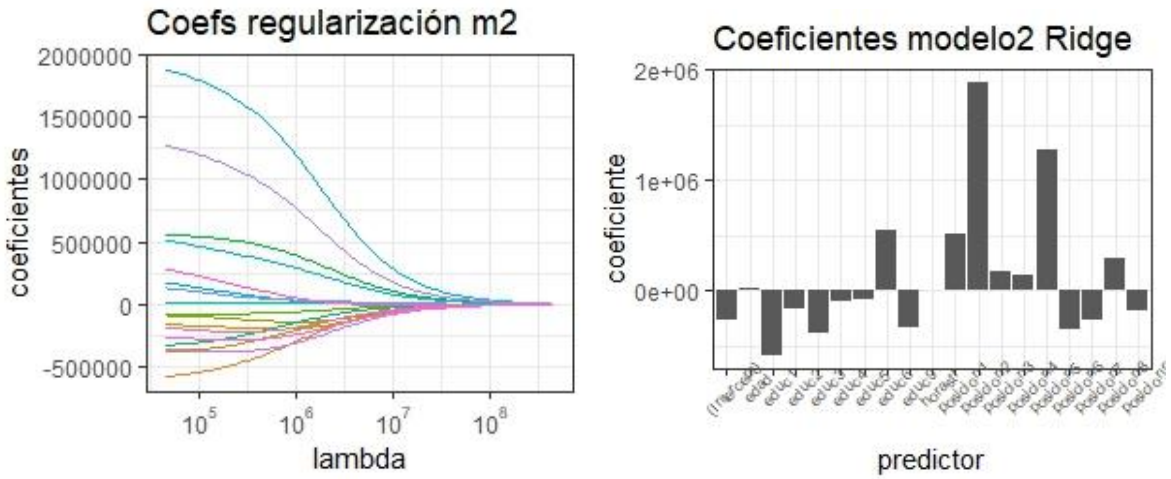
Grafica 6. Tradicional con base de personas



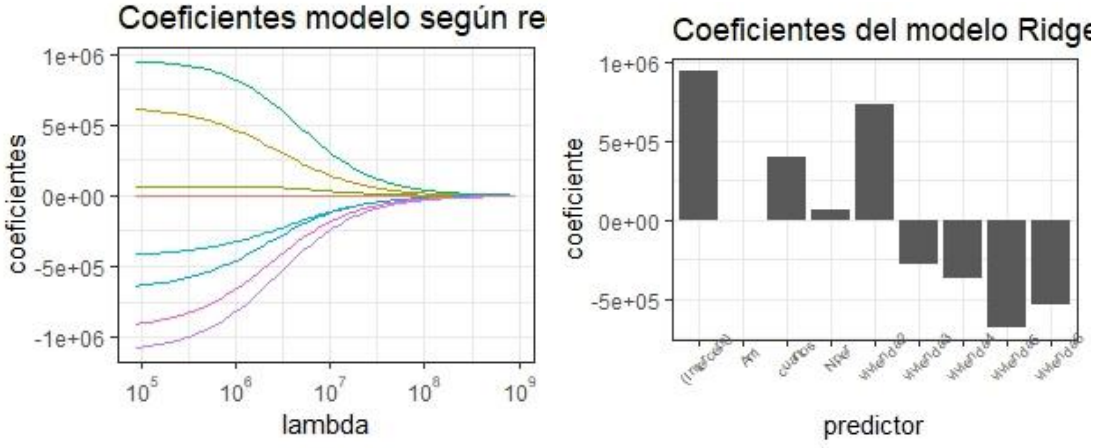
Gráfica 7. Tradicional con base de hogares.



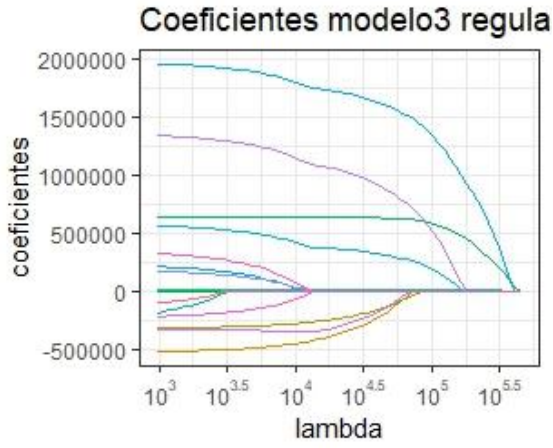
Grafica 8 y 9. Modelo Ridge con personas

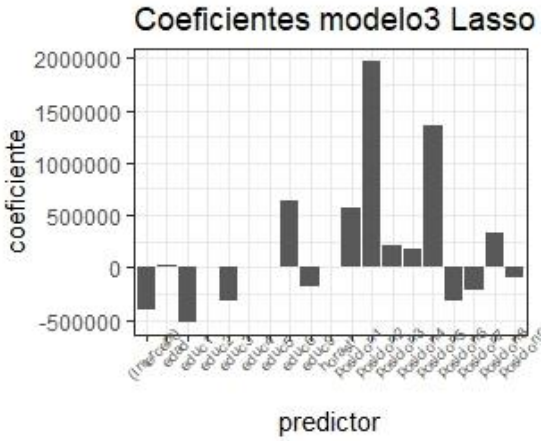


Gráfica 10 y 11. Modelo Ridge con hogares

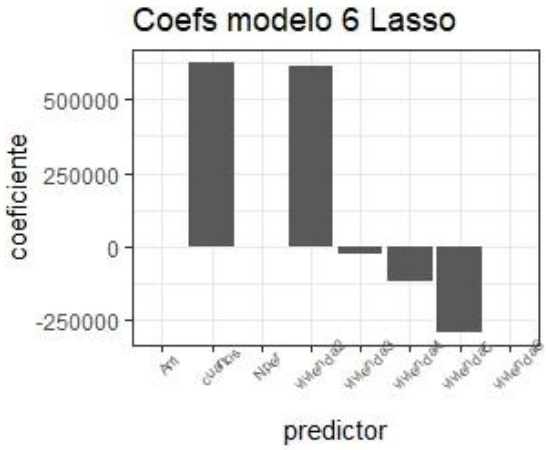


Gráfica 12 y 13. Modelo Lasso con personas





Gráfica 14 y 15. Modelo Lasso con hogares



Gráfica 16 y 17. Modelo elegido con metodología Lasso