

# 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.

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.

En el enfoque de clasificación, seleccionamos un modelo que a partir de las características del hogar, sin tener en cuenta el ingreso, 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.

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 *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 no 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

|  |
| --- |
| Chart, bar chart  Description automatically generated |

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.

Table

Description automatically generated

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 |

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.

|  |  |
| --- | --- |
|  |  |

Chart, scatter chart

Description automatically generated

Gráfica 10: Distribución y correlación 1

4. Modelos de Predicción y Clasificación:

4.1 Modelo de clasificación

Podemos inicialmente identificar algunas características de los hogares clasificados como pobres en los datos de entrenamiento:

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

*Gráfica 11 Gráfica 12*

La Gráfica 11 muestra que en promedio, personas que viven por fuera de cabeceras están clasificadas en mayor proporción como pobres que aquellas que viven en las cabeceras, asi mismo, existe una menor proporción de personas pobres que viven en vivienda propia ( sea pagada o pagando). Hacemos una primera estimación logit tomando las siguientes variables:

Resultados de la Regresion

===============================================

Dependent variable:

---------------------------

Pobre

-----------------------------------------------

ingreso -0.0000135\*\*\*

(0.0000001)

cabecera 3.0062840\*\*\*

(0.0498550)

cuartos -0.3090874\*\*\*

(0.0178814)

v. propia pagadando 2.7819130\*\*\*

(0.1104976)

v. arriendo 3.4532500\*\*\*

(0.0475044)

v.usufructo 0.1216218\*\*\*

(0.0433736)

posesion sin titulo 3.2577080\*\*\*

(0.0638888)

otro tipo vivienda 3.0300070\*\*\*

(0.3805796)

No personas hogar 6.3759240\*\*\*

(0.0539241)

hombre 0.0080015

(0.0298929)

edad 0.0317639\*\*

(0.0157935)

educación superior -0.5573191\*\*\*

(0.0457574)

horas de trabajo -0.0664711\*\*\*

(0.0143651)

Constant 5.7969590\*\*\*

(0.0706046)

-----------------------------------------------

Observations 164,960

Log Likelihood -16,104.3100000

Akaike Inf. Crit. 32,236.6300000

===============================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

Se observa que todas son significativas a excepción de la variable, hombre. Esto indica que si el jefe de hogar es hombre o mujer, esto no tiene implicaciones sobre la probabilidad de que el hogar sea clasificado como pobre. Procedemos a retirar la variable ingreso dado que no la vamos a tener en nuestra base de datos de prueba y procedemos a estimar con las demás variables, usamos el modelo logit y la técnica de validación cruzada (5 folds) para garantizar que los datos son independientes de la partición entre entrenamiento y prueba. El modelo nos arroja los siguientes resultados:

Text

Description automatically generated

Procedemos de igual forma a evaluar el ROC que muestra la Grafica 13, el cual nos muestra la capacidad predictiva del modelo.

Chart, line chart

Description automatically generated

*Grafica 13*

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.

Grafico # 14

Chart, funnel chart

Description automatically generatedChart, funnel chart

Description automatically generated

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 a , 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.

Grafico #15

Chart, diagram

Description automatically generated with medium confidence

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 a ).

Grafico #16

Diagram, schematic

Description automatically generated

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.

Grafico #17

Chart

Description automatically generatedHistogram

Description automatically generated with low confidence

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.

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.

Anexos







