Big Data: Trabajo Práctico 4

Autores: Sofía Ellenberg, Vicente Zervino, Sophie Schulzen

Profesores: Maria Noelia Romero

Tutor: Victoria Oubiña

Parte 1:

Ejercicio 1:

La pobreza en Argentina además de ser medida a nivel individual, también se puede analizar desde una perspectiva a nivel hogar. Si bien este método no logra abarcar cuestiones intrahogar, como por ejemplo cómo se distribuye el ingreso familiar dentro de los individuos de cada vivienda, hay ciertos fenómenos que son fundamentales para analizar la relevancia de la pobreza en Argentina. La encuesta EPH enfatiza en ambas dinámicas, tanto individual y hogar, por eso, es útil determinar algunas variables del hogar para poder determinar los niveles de pobreza. En ese sentido, es sabido que las personas más pobres tienen una mayor probabilidad de, no solo tener más hijos, sino también cohabitar. De este modo, algunas variables a tener en cuenta para perfeccionar el trabajo anterior podrían ser aquellas que hablan sobre las características habitacionales de la vivienda, como por ejemplo, IX_Tot que menciona la cantidad de miembros del hogar, IX_Men10 y IX_Mayeq10 que hace referencia a cuantas personas dentro del hogar son mayores o menores de 10 años. Además, dentro de estas características también es fundamental analizar qué tipo de relación tienen las personas que viven juntas, ya sea un vínculo familiar o de otro tipo (CH03). Luego, es importante incluir algunas características propias del hogar, como por ejemplo, cuántos ambientes tiene la vivienda (IV2), que tipo de vivienda es (IV1), ya sea departamento, casa y si es un hogar propio o alquilado (V8). Cabe destacar que, muchos estudios previos han determinado que el tipo de techo de una vivienda tiene una alta correlación con los niveles de pobreza, por eso, es importante incluir las variables V4 y IV5 que hablan sobre las características del techo del hogar. Finalmente, las variables de ingreso a nivel hogar (ITF) también son importantes para incluir en el trabajo y el tipo de educación que tienen las personas dentro de ese hogar, aunque estas variables ya estén incluidas en las características individuales de las personas. En conjunto, estas características pueden dar indicios sobre los niveles de pobreza y cómo viven las personas en Argentina además de las ya analizadas en el trabajo anterior.

Ejercicio 5:

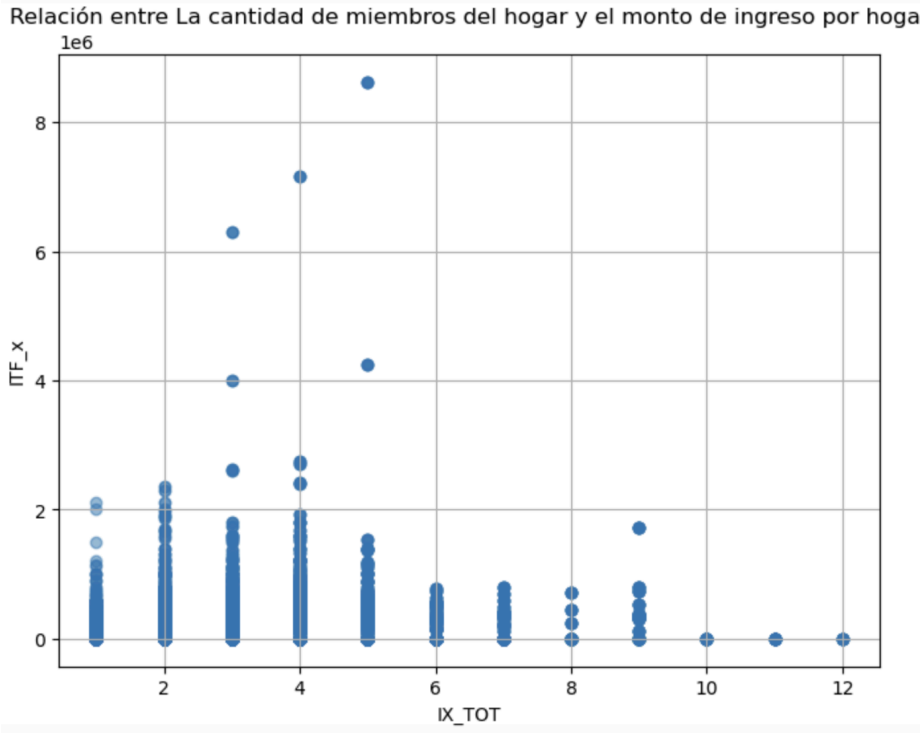
Elegimos realizar estadística descriptiva sobre estas 5 variables: IX_TOT, ITF_x, V4, IV1 y NIVEL_ED. Estas hacen referencia a la cantidad de miembros en el hogar, el ingreso total por hogar, el tipo de techo del hogar y el nivel educativo de las personas. Vemos que el promedio de personas en un hogar es de 3.78 personas con un máximo nivel de 12 personas por hogar y un mínimo nivel de 1 persona por hogar. Además el ingreso total familiar es de

un promedio de 2.987517e+05 con un máximo de 8.625000e+06 y un mínimo de 0 ingresos. Por otro lado, las variables V4, IV1 y Nivel_ED son variables binarias, por lo que evaluar su estadística descriptiva no es de suma relevancia, aunque vemos que en promedio las personas tienen un techo de tipo 2, que es un techo de V4 de losa sin cubierta, su tipo de vivienda suele ser un departamento y el nivel educativo de las personas, en promedio, es "Secundario incompleto".

Ejercicio 6:

Para este inciso, elegimos realizar un gráfico de regresión, que representa la relación entre la cantidad de miembros de un hogar y el nivel de ingresos por hogar. Antes de realizar el gráfico, intuimos que sería muy relevante correr esta regresión ya que si bien podría ser lógico pensar que mientras más miembros convivan en un hogar, mayor deberían ser los ingresos para poder mantener a la vivienda en su conjunto, diversos estudios han corroborado que las personas más pobres suelen vivir con otras personas fuera de su relacion familiar y ademas, tener mas hijos. Esto se relaciona fuertemente con los niveles de educación de las personas y para cubrir gastos.

Sin embargo, al realizar el gráfico de regresión, la relación no es muy evidente:



En principio, intuimos que el gráfico se ve así debido a las medidas de medición y a las cantidades registradas para la variable ingresos y para la cantidad de miembros de un hogar, que es como máximo de 12 personas. Sin embargo, teniendo en cuenta que los valores de los

ingresos son más altos, podemos observar que los hogares con menor cantidad de miembros tienen ingresos más altos.

Ejercicio 9:

Según nuestro análisis, la tasa de hogares bajo la línea de pobreza es de 33,33%. Si comparamos este resultado con el reportado por INDEC en sus informes, vemos que el resultado de ellos es de 31,8%. Esto quiere decir que a nosotros nos dio un 4,8 % más alto que al INDEC. Es decir, si bien los resultados con respecto al INDEC se asemejan casi en su totalidad, según nuestro análisis la tasa de hogares bajo la línea de pobreza es mayor.

Parte 3:

Ejercicio 3

El parámetro λ, en validación cruzada es útil para determinar la mejor elección del modelo. En casos como el método de regularización LASSO, λ Es el parámetro de penalización, donde se penaliza por la elección de variables irrelevantes. A diferencia de otros métodos de regularización, LASSO es capaz de imponer un coeficiente igual a cero, a las variables que son menos relevantes para predecir el modelo. De este modo, en la validación cruzada, el proceso se realiza primero mediante una división de k subconjuntos de datos de igual tamaño, y se entrena la cantidad de veces necesaria utilizando k-1 de los subconjuntos de datos para el entrenamiento y el restante para evaluar la validación. Luego se considera el valor de lambda para el cual el modelo tiene un mejor rendimiento promedio a lo largo de todas las iteraciones. Sin embargo, el conjunto de prueba (test) no se utiliza porque podría sobre ajustar al modelo al seleccionar el parámetro λ.

Ejercicio 4

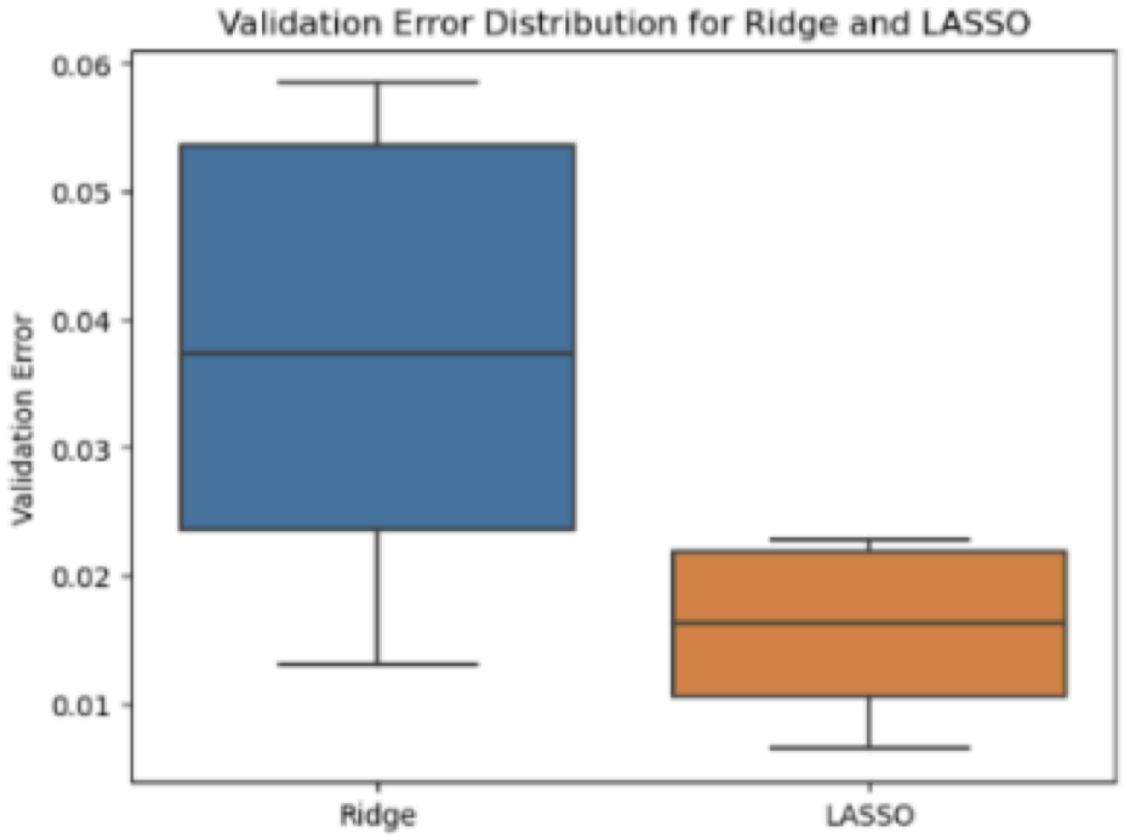
La elección de las k particiones tiene implicancias tanto en el rendimiento del modelo como en la estimación del error. En otras palabras, si k es pequeño, los subconjuntos de validación serán muy grandes, por lo que los conjuntos de datos son relativamente pequeños en cada iteración y tiene efectos en el entrenamiento del modelo. Dichos efectos tienen que ver con una alta variabilidad en las estimaciones y un mayor sesgo. En el otro extremo, cuando k es muy grande, los subconjuntos de validación tendrán una única muestra, lo cual maximizará la cantidad de datos usados para entrenar al modelo y por lo tanto podría llevar a estimaciones de error con baja variabilidad y menor sesgo. Sin embargo, un k muy grande es computacionalmente muy costoso. Además, un k muy grande puede llevar a que el modelo se

sobreajuste a los datos de entrenamiento. Finalmente, cuando k=n, el modelo se estima n veces, con n igual al número de muestras y en cada iteraciones se usa una sola muestra como conjunto de prueba y las demás son el conjunto de entrenamiento.

Ejercicio 5:

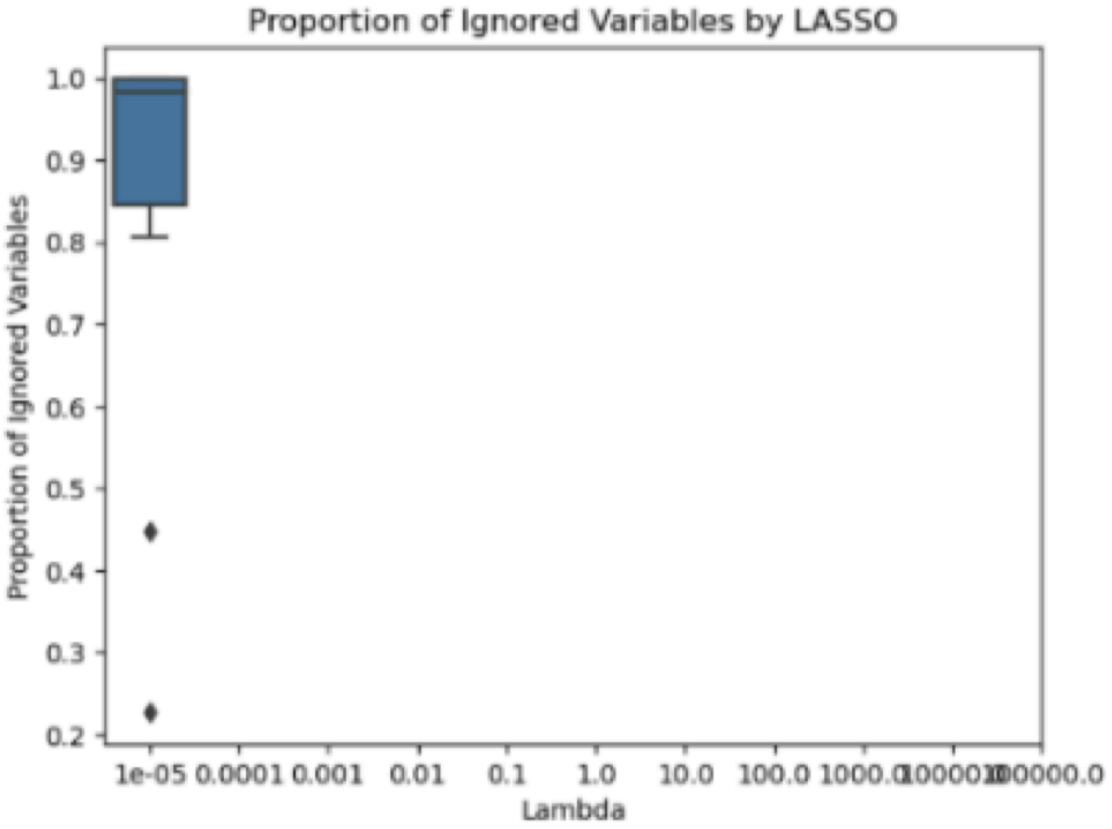
Comparación de la distribución del error de validación para LASSO y Ridge:

- LASSO muestra un mejor rendimiento promedio porque tiene una Mediana de error más Baja en comparación con ridge
- Ridge tiene una mayor dispersión de los errores, lo que hace que sea más inestable en las particiones de los datos.



Proporción de variables ignoradas por LASSO en función de diferentes valores de landa

- A medida que λ aumenta, LASSO ignora más variables, porque penaliza más. A menores valores de λ, LASSO es menos restrictivo con respecto a la inclusión de variables.



Ejercicio 6:

Para el valor óptimo de λ, las variables que LASSO descarto fueron:

LASSO: [25 26 28 ... 5720 5721 5722]

Ejercicio 7:

- ECM promedio para Ridge: 0.037987012987012986
- ECM promedio para LASSO: 0.014935064935064935

Como el objetivo para medir qué método de regularización es mejor es a través del método que menor error cuadrático medio tiene, podemos concluir que LASSO funcionó mejor como método de regularización en términos de ECM

Ejercicio 8:

El modelo LASSO otorga un error cuadrático medio menor que el modelo RIDGE, por lo que concluimos que LASSO predice mejor.

Ejercicio 9:

La proporción de hogares que son pobres en la submuestra de los que no respondieron es 0.40059077482390365, según nuestra predicción del modelo LASSO.