Trabajo Practico N°4

Lucila Brisighelli

Fecha de entrega: 21 de junio de 2024

1 Parte 1

1.1

La pobreza en Argentina se estudia con caracteristicas de los hogares y de los individuos. La Encuesta Permanente de Hogares (EPH) aborda ambas perspectivas. Algunas variables que pueden ser muy predictivas de pobreza y que seria muy útil incluir para perfeccionar el ejercicio del TP3, Podrian ser variables como el tamaño del hogar (IX_Tot), la edad de los miembros (IX_Men10, IX_Mayeq10), y la relación entre ellos (CH03). Además, incluir variables como las características que describen el tipo de vivienda (IV1), caracteristicas de la propiedad de la vivienda, por ejemplo si es alquilada u otra forma de tenencia. Esta puede indicar estabilidad financiera (V8), asi como la calidad del techo (V4, IV5). Por último, la inclusión de los ingresos totales del hogar (ITF) es una medida directa de la capacidad económica del hogar. Hogares con ingresos más bajos son más propensos a encontrarse en situación de pobreza. Este dato es esencial para calcular líneas de pobreza y evaluar la distribución del ingreso.

1.2

Resuelto en el código.

1.3

Resuelto en el código.

1.4

Resuelto en el código.

1.5

Comenzando con la variable IX_TOT (Cantidad de miembros en el hogar), observamos que el promedio de personas por hogar es 3.78, con un rango que va desde un mínimo de 1 persona hasta un máximo de 12 personas. Este dato sugiere una variabilidad significativa en el tamaño familiar promedio en Argentina, factor que puede influir, tanto

en la capacidad de generación de ingresos como en la distribución de recursos dentro del hogar. Los hogares con un mayor número de miembros podrían enfrentar mayores desafíos económicos debido a la necesidad de sostener a más personas.

En cuanto a ITF_x (Ingreso total familiar), encontramos un ingreso promedio de $298,751.70. Sin embargo, es importante destacar que el rango de ingresos es amplio, desde un mínimo de $0.00 (hogares sin ingresos reportados) hasta un máximo significativo de $8,625,000.00. Esta disparidad llama la atencion indicando la amplitud entre las diferencias económicas entre los hogares encuestados. La variable V4 (Tipo de techo del hogar)es binaria y muestra en promedio un tipo de techo de 2, clasificación ligadas a la calidad de vida y a la vulnerabilidad frente a condiciones climáticas adversas, aspectos que pueden afectar el bienestar económico de los hogares. En cuanto a IV1 (Tipo de vivienda), que también es binaria, vemos que en promedio las viviendas son del tipo departamento. Esto sugiere una predominancia de la vivienda en estructuras urbanas más densas, lo cual puede estar relacionado con accesibilidad a servicios básicos y oportunidades de empleo, factores determinantes en la experiencia de pobreza. Por ultimo, la variable NIVEL_ED (Nivel educativo promedio) revela que, en promedio, el nivel educativo de los miembros del hogar es incompleto para nivel medio. La educación juega un papel crucial en la en la capacidad de generar ingresos suficientes para salir de la pobreza.

1.6

En el eje x (horizontal) representamos el tamaño del hogar (IX_TOT), que indica la cantidad de personas que viven en el hogar. En el eje y (vertical) representamos el ingreso total familiar (ITF_x), que refleja la cantidad total de ingresos percibidos por el hogar en un período determinado.

Podemos observar si hay una tendencia general entre estas dos variables. Por ejemplo, vemos que el ingreso total familiar tiende a disminuir conforme aumenta el tamaño del hogar. Esto contradice la teoria de que podría indicar que los hogares con más miembros tienen una capacidad de ingresos mayor debido a más fuentes de ingreso o a una mayor participación laboral. Por otro lado, no hay una clara asociación si el ingreso total familiar es bajo independientemente del tamaño del hogar, señalando desafíos económicos más profundos que enfrentan estos hogares.

Es importante tener en cuenta, que si los puntos en el gráfico están dispersos y no muestran una tendencia clara, podría sugerir una falta de correlación directa entre estas variables en la muestra de hogares encuestados. Esto podría ser indicativo de que otros factores, como el tipo de ocupación, nivel educativo o acceso a oportunidades económicas, podrían estar influyendo más significativamente en los niveles de ingreso y pobreza observados.

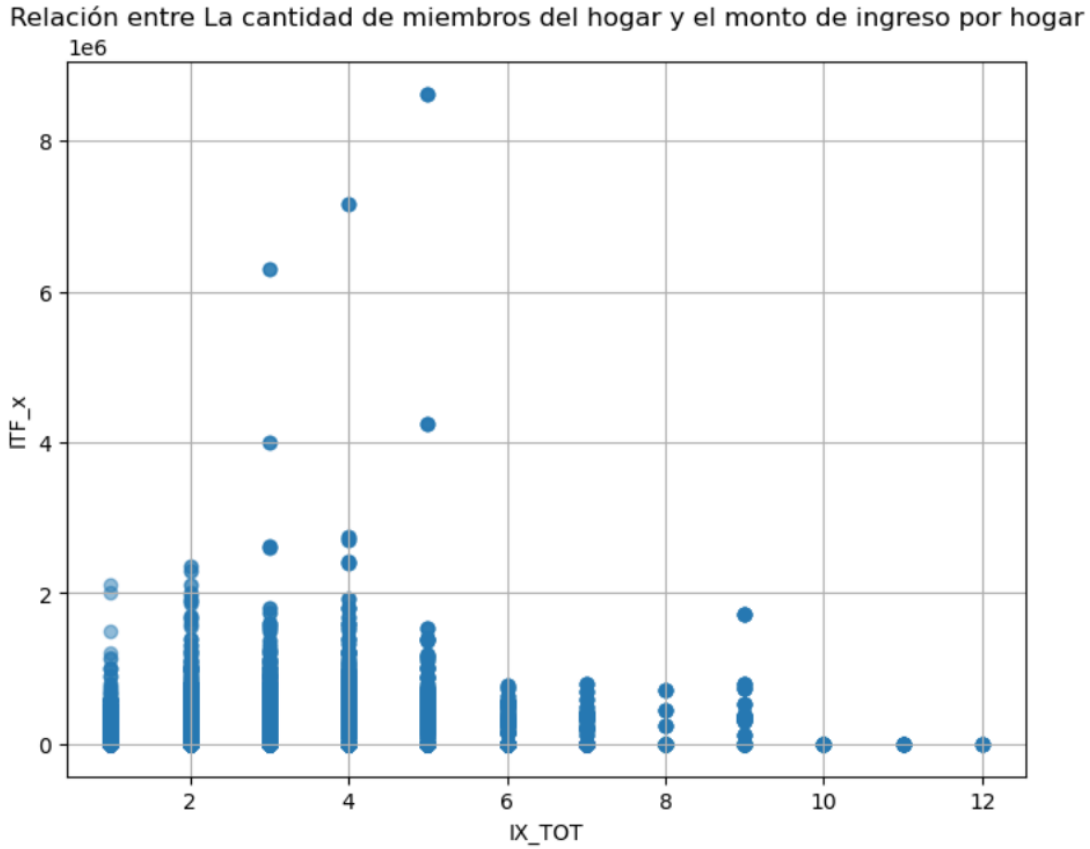

Figure 1: Gráfico de Dispersión

1.7

Resuelto en el código.

1.8

Resuelto en el código.

1.9

La tasa de hogares viviendo bajo la línea de pobreza es del 33,33%. Al comparar este resultado con el informado por el INDEC, que es del 31,8%, hay una diferencia del 4,8% más alto que el dato oficial.

2 Parte 2

Resuelto en el código.

3 Parte 3

3.1

Resuelto en el código.

3.2

Resuelto en el código.

3.3

La elección del parámetro de regularización en modelos como LASSO, sirve para encontrar el entre el ajuste del modelo y su capacidad para generalizar a nuevos datos. En la validación cruzada, el conjunto de datos se divide en k subconjuntos de igual tamaño. El modelo se entrena k veces, cada vez usando $k-1$ de estos subconjuntos para el entrenamiento y el subconjunto restante para la validación. Este proceso se repite para diferentes valores de landa, y se evalúa el rendimiento del modelo en cada iteración. Por ultimo, se toma el valor de landa que obtiene el mejor rendimiento promedio a lo largo de todas las iteraciones. No usaríamos el conjunto de prueba (test) para seleccionar landa es que el conjunto de prueba está destinado a ser un "proxy" de datos no vistos todavía. Utilizarlo para la selección de hiperparámetros podría sobreajustaria el modelo, entonces generalizaria a nuevos datos. La validación cruzadaes más objetiva, ya que el proceso de entrenamiento y evaluación se realiza en diferentes particiones del conjunto de datos de entrenamiento, manteniendo el conjunto de prueba reservado para la evaluación final.

3.4

La elección del número de particiones k en la validación cruzada tiene implicancias en el rendimiento del modelo y en la estimación del error. Si k es muy pequeño, cada subconjunto de validación será grande, y el modelo se entrenará en un conjunto de datos relativamente pequeño en cada iteración. Esto puede llevar a estimaciones con alta varianza que no aproveche toda la información disponible en los datos.

Si k muy grande: Si k es muy grande (por ejemplo, el número total de muestras), cada subconjunto de validación contendrá una única muestra. Esto maximiza la cantidad de datos usados para entrenar el modelo en cada iteración, y computacionalmente seria muy costoso llevando a estimaciones de error con baja varianza y alta carga computacional. La validación cruzada tiende a proporcionar una estimación menos sesgada del error de generalización, el alto costo computacional y la posibilidad de mayor varianza pueden ser grandes implicancias.

3.5

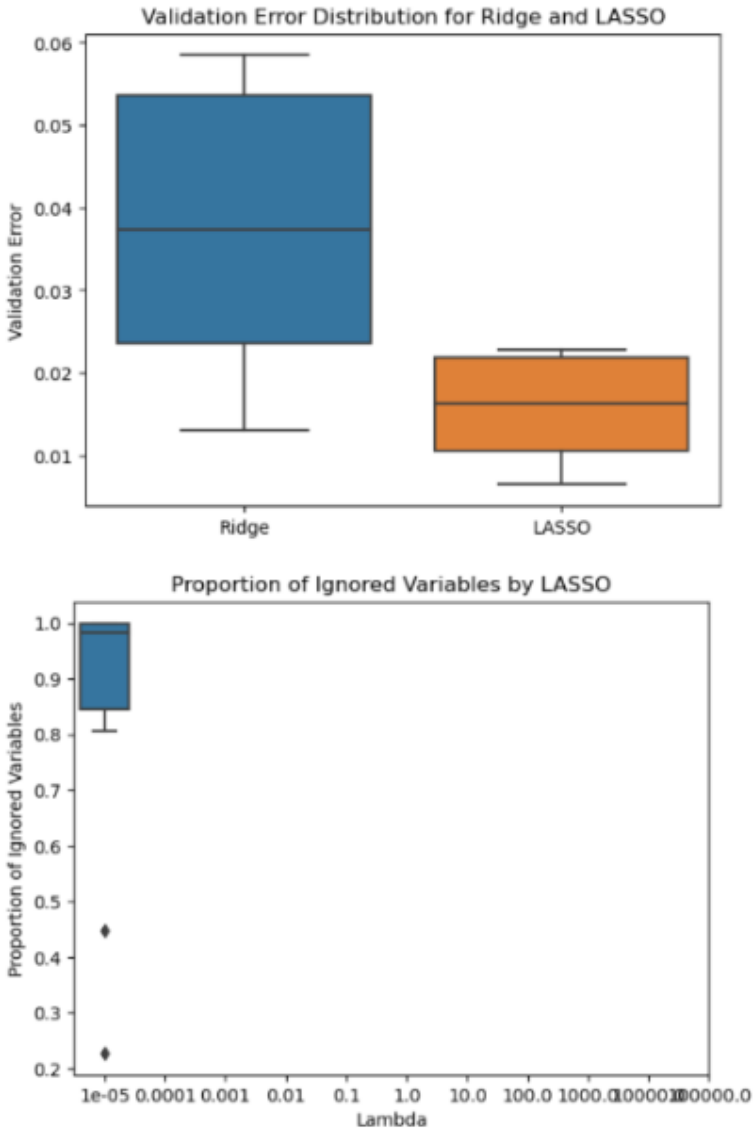

Figure 2: Box-plots

El boxplot de arriba compara la distribución del error de validación para los modelos Ridge y LASSO. Se observa que LASSO tiene una mediana de error de validación más baja y menos variabilidad en comparación con Ridge. Esto sugiere que LASSO ofrece un mejor rendimiento promedio y también es más consistente. Ridge, en cambio, presenta mayor dispersión en los errores, indicando inestabilidad en diferentes particiones de los datos. El segundo boxplot muestra la proporción de variables ignoradas por LASSO en función de diferentes valores de landa. A medida que landa aumenta, LASSO ignora más variables (coeficientes cero). Para valores pequeños de landa, pocas variables son descartadas, mientras que para valores grandes, la mayoría de las variables son ignoradas. Esto destaca la capacidad de LASSO para realizar una selección de variables efectiva.

3.6

Índices de las variables descartadas por LASSO: [25 26 28 ... 5720 5721 5722]

3.7

ECM promedio para Ridge: 0.037987012987012986 ECM promedio para LASSO: 0.014935064935064935
LASSO funcionó mejor que Ridge en términos de ECM.

3.8

El el modelo LASSO el cuadrático medio es menor que en el modelo RIDGE, siendo LASSO un mejor predictor.

3.9

Según el modelo LASSO, la proporción de hogares que son pobres en la submuestra de los que no respondieron es de 0.4005907748.