Universidad de
SanAndrés

Big Data

Profesora: Romero, María Noelia

Tutora: Oubiña, Victoria

Trabajo Práctico 4

Cucher Maximiliano, Saucedo Federico & Soares Gache Manuel

Fecha de entrega: 21/6/2024

Parte I: Análisis de la base de hogares y cálculo de pobreza

1. Las variables que consideramos que pueden ser muy útiles a la hora de predecir la pobreza, sin tener en cuenta las variables de ingreso, ya que estas son eliminadas a la hora de predecir, son: IV1, IV2, IV6, IV9, IV11, IV12_1, II1, II5_1, II7 y IX_Men10 . IV1, que describe el tipo de vivienda, es esencial porque diferentes tipos de vivienda pueden reflejar niveles variados de infraestructura y seguridad habitacional, siendo esto crucial para determinar la calidad de vida y el bienestar económico de los hogares. IV2, que indica el número de ambientes en la vivienda, ofrece una medida del tamaño y la capacidad de alojamiento del hogar, lo cual está directamente relacionado con su composición y su capacidad para proporcionar un entorno adecuado para sus miembros.

IV6, que detalla la fuente de agua, es crucial ya que el acceso a agua potable dentro de la vivienda es fundamental para la salud y el bienestar de los residentes. Por otro lado, IV9 informa sobre la ubicación del baño, un factor determinante para la comodidad y la dignidad de los hogares. IV11, que describe el sistema de desagüe del baño, proporciona *insights* importantes sobre las condiciones sanitarias y ambientales del hogar, elementos clave para evaluar el nivel de calidad de vida.

IV12_1 destaca la proximidad del hogar a basurales, lo cual puede indicar exposición a riesgos ambientales y de salud pública. II1, que captura el número de ambientes exclusivos para el hogar, y II5_1, que mide los ambientes utilizados para dormir, son indicadores directos de la capacidad del hogar para proporcionar un entorno adecuado para sus miembros. II7, que describe el régimen de tenencia de la vivienda, proporciona información sobre la estabilidad residencial y la seguridad habitacional del hogar. Finalmente, IX_Men10, que registra el número de menores de 10 años en el hogar, es crucial porque estos miembros son particularmente vulnerables y su presencia puede influir significativamente en las decisiones financieras y de consumo del hogar.

3. Para limpiar la base, lo primero que hicimos fue ver qué variables tomaban valores negativos. Variables: ('CH06', 'PP06C', 'PP06D', 'PP08D1', 'PP08F1', 'PP08F2', 'PP08J1', 'PP08J2', 'PP08J3', 'P21', 'TOT_P12', 'P47T', 'V2_M', 'V3_M', 'V4_M', 'V5_M', 'V8_M', 'V9_M', 'V10_M', 'V11_M', 'V12_M', 'V18_M', 'V21_M', 'T_VI'). De esta lista nos dimos cuenta que la única variable que no tenía sentido que tomase valores negativos era 'CH06' ya que esta especifica cuántos años tienen las personas. Por otro lado, uno podría pensar que el resto de variables que se tratan del ingreso tampoco tiene sentido que sean negativas, pero

analizando el diccionario de variables encontramos que “Una excepción la constituyen los montos de ingreso, en cuyo caso la no respuesta se identifica con el código -9. Además los montos captados en pp06c y pp06d presentan también los códigos -7 “No tenía esa ocupación en el mes de referencia” y -8 “No tuvo ingresos por el mes de referencia”.” Por lo tanto comprobamos que estas variables si pueden tomar valores negativos.

Una vez hecho esto, eliminamos los *outliers* utilizando el método IQR. Tomando Q1=0.001 y Q3=0.999. Utilizamos estos valores para no eliminar tantas observaciones sino las que verdaderamente están muy alejadas de la muestra. Si hubieras utilizado un enfoque más conservador hubiéramos perdido gran parte de la muestra, eliminando posiblemente a las personas mas pobres de nuestra muestra, un riesgo que no estábamos dispuestos a tomar.

En cuanto a los *missing values*, nuestra decisión fue dejarlos como missing ya que si los hubiésemos reemplazado por la media, en la mayoría de variables (al ser discretas) no tendría sentido. Por ejemplo, no tendría sentido reemplazar un *missing value* de la cantidad de hijos por 2,5 por ejemplo. Por otro lado, las variables que son continuas son las que hacen referencia al ingreso. Asimismo, nuestra decisión fue no reemplazarlas ya que sino esto no tendría sentido con los siguientes ejercicios. Por ejemplo si reemplazamos ITF por su media, tendríamos que todas las personas respondieron a esta pregunta, por lo que la base de norepsondieron estaria vacia.

Por último, en cuanto a las variables categóricas y strings nuestra decisión fue dejarlas de este modo ya que eran únicamente 10, las cuales analizamos y vimos que no eran de gran importancia a la hora de predecir la pobreza, pero si las transformamos a dummies nos generará problemas para los siguientes ejercicios.

4. Las variables creadas para predecir pobreza que no están en la EPH son:

Proporción de niños menores a 10 años en el hogar: la proporción resultante es una variable relevante para predecir la pobreza porque cuanto más hijos menores a 10 años tengan las familias en promedio (personas que en la gran mayoría de los casos no trabajan por su corta edad) más dinero deben destinar las familias en gastos de necesidades (tales como alimentación, ropa, entre otras variables) y como consecuencia para el mismo nivel de ingresos de padres (suponiendo que ambos padres perciben ingresos, si solo tienen un ingreso sería peor la situación) la cantidad de dinero por hijo sería mucho menor (trade-off entre cantidad y calidad) por lo que podríamos usar como *proxy* de que más joven estos hijos

podrían salir al mercado laboral, dejando sus estudio y teniendo menos capital humano que en un caso donde los padres tengan pocos hijos y no necesitan que trabajen), percibirán menos ingresos a lo largo de su vida en comparación con su contrafactual con más capital humano por lo cual la movilidad intergeneracional podría ser más baja y la persistencia de la pobreza entre generaciones puede ser mayor.

Proporción de personas en el hogar que trabajan en oficina: esta variable puede ser relevante ya que suele estar vinculada a una mayor seguridad laboral, ingresos más altos y beneficios adicionales como seguro de salud y pensiones. Estos trabajos tienden a ser más estables, lo que se traduce en ingresos constantes y previsibles, proporcionando a los hogares una mayor capacidad para planificar a largo plazo y enfrentar emergencias económicas sin caer en la pobreza. Además, las personas que trabajan en oficinas generalmente tienen acceso a recursos financieros, como crédito, que pueden utilizarse para inversiones significativas o para manejar imprevistos, así como mayores oportunidades de capacitación y educación que incrementan su empleabilidad y potencial de ingresos futuros.

Proporción de personas en el hogar que trabajan en la calle: esta variable refleja una mayor vulnerabilidad económica y social. Los trabajos informales o callejeros, como la venta ambulante o el trabajo independiente, suelen ser inestables y están sujetos a variaciones significativas en los ingresos debido a factores externos como el clima, la demanda del mercado y regulaciones locales. Estos trabajos a menudo no ofrecen beneficios como seguro de salud o jubilaciones, dejando a los trabajadores y sus familias más expuestos a riesgos económicos. La falta de ingresos regulares y la ausencia de un historial financiero sólido limitan el acceso a servicios financieros formales, dificultando la posibilidad de obtener préstamos o créditos para mejorar la calidad de vida o manejar crisis económicas. Además, los hogares que dependen de trabajos en la calle tienen menos acceso a oportunidades educativas y de capacitación, perpetuando el ciclo de pobreza y disminuyendo las oportunidades de movilidad económica a largo plazo.

Proporción del ingreso que viene dado por ayuda social: La proporción del ingreso que proviene de la ayuda social es relevante para predecir la pobreza porque un mayor porcentaje de ingresos provenientes de estas ayudas indica una mayor dependencia de los hogares vulnerables en el apoyo estatal para satisfacer sus necesidades básicas. En Argentina, donde el 11% del PBI se destina a la ayuda social (según Zarazaga et al., 2021) [Ver **Anexo 1**] esta dependencia es alta, pero la persistencia de altos niveles de pobreza sugiere que la eficacia de

estos programas puede no ser óptima. Comparado con países como Chile y Uruguay, que gastan menos en ayuda social y tienen menores tasas de pobreza, se puede inferir que no solo el monto de la ayuda es importante, sino también la eficiencia con la que se implementan estas ayudas y las políticas complementarias que se aplican. Por lo tanto, la proporción del ingreso derivado de la ayuda social es un indicador crucial para entender y predecir los niveles de pobreza.

5.

	IV1	IV2	IV8	V1	IX_TOT
count	7262.000000	7262.000000	7262.000000	7262.000000	7262.000000
mean	1.321261	3.061002	1.002066	1.110989	3.667447
std	0.547873	1.102262	0.045404	0.314140	1.863689
min	1.000000	1.000000	1.000000	1.000000	1.000000
25%	1.000000	2.000000	1.000000	1.000000	2.000000
50%	1.000000	3.000000	1.000000	1.000000	3.000000
75%	2.000000	4.000000	1.000000	1.000000	4.000000
max	6.000000	8.000000	2.000000	2.000000	12.000000

Figura 1: Estadística descriptiva de ciertas variables.

Aquello que observamos del tipo de vivienda (IV1)[1] es que en promedio, las personas son más propensas a tener una casa (el valor de la media es de 1,3) y posee un desvió estándar de 0,5 por lo que en la mayoría de los casos (95% de significancia) las personas dicen vivir en una casa(1) o departamento[2], para el cuartil más bajo y la mediana observamos que el 50% de las observaciones reportan tener como vivienda una casa y para el cuartil más alto el 75% de las observaciones están por debajo del valor (2), es decir, casas y departamentos. Resulta relevante comprender que el tipo de vivienda en el que viven las personas es relevante para predecir la pobreza. Sin embargo, bajo estas estadísticas descriptivas simples y al no relacionarlos con ingresos puede llegar a ser poco informativo.

La variable referente a IV2 (¿Cuántos ambientes/habitaciones tiene la vivienda en total? (sin contar baño/s, cocina, pasillo/s, lavadero, garaje) se observa que en promedio las viviendas tienen 3 habitaciones con un desvió estándar de 1 habitación; el valor mínimo es 1 habitación mientras que el máximo son 8 habitaciones en una vivienda. El cuarto más bajo tiene 2 habitaciones, la mediana 3 y el cuarto más alto 4. Esta variable es relevante para analizar la pobreza dado que si la comparamos con la cantidad de miembros del hogar (IX_TOT) observamos casos en los que por habitación viven más de una persona y como correlaciona con otros *outcomes* relevantes para analizar la pobreza.

La variable IV8 hace referencia a si en el hogar hay baño/letrina, toma el valor de 1 cuando la respuesta es sí y 2 en el caso de negativo. La media de esta variable es 1.002 y el desvío estándar es 0.045. En esta línea el valor de los percentiles no es tan informativo dado que lo

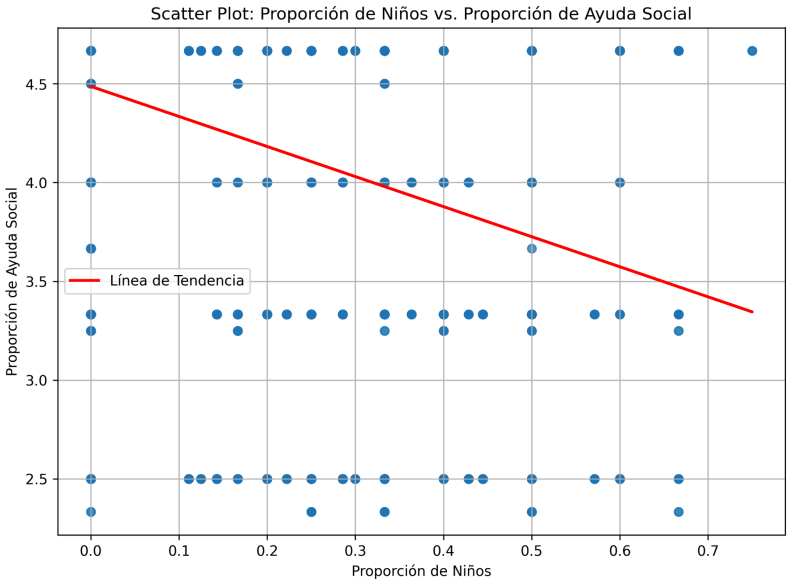
[1] Para la variable (IV1) los valores son los siguientes: 1=casa, 2=departamento, 3=pieza de inquilinato, 4=pieza en hotel/pensión, 5=local no construido para habitación y 6=otros, especificar.
[2] Hacemos la media +- 2 desvíos estándar.

que ocurre por encima de 75% de las observaciones no lo comprendemos bien (es decir, que porcentaje respondió que no) aunque observando la media y su desviación estándar vemos que la gran mayoría de las observaciones respondieron que poseen baño. Esta es una variable relevante para analizar la pobreza cerca de la "cola inferior" dado que si los hogares reportan no tener baños habla de que puede estar también relacionado con un ambiente insalubre en el cual se desarrolla la vida de las personas correspondientes a ese hogar.

La variable V1 reporta si en los últimos tres meses las personas de este hogar han vivido/ o no de lo que ganan en el trabajo. El valor 1 corresponde al valor afirmativo mientras que el 2 al negativo. Observamos una media de 1.11 y un desviación estándar de 0.31 por lo que existe un porcentaje no menor de la muestra que no vive de lo que gana en el trabajo por lo cual podemos descomponerlo en personas jubiladas o personas que un gran porcentaje de su ingreso se debe a transferencias por parte del gobierno. Comprender esta dinámica es relevante para analizar la pobreza porque si bien esta ayuda puede ser necesaria para estas personas las cuales pueden necesitar un empujón para despegar y salir de una potencial "trampa de pobreza" puede que también esté llevando al desincentivo al trabajo.[3] El análisis de cuartiles resulta ser poco informativo dado que es trivial que si existe algún valor =2 el cuantil más alto dará un valor de 75%.

La variable que hace referencia a la cantidad de miembros del hogar (IX_TOT) tiene un valor promedio que resulta relevante para analizar la pobreza el cual es cercano a 4 miembros de la encuesta a nivel hogares de la EPH (3,67) y con un desvió estándar de casi miembros. El valor mínimo de la muestra es de 1 miembro y máximo de 12 miembros. El 25% de las observaciones poseen al menos dos miembros por hogar, para la mediana se observa 3 miembros y para el percentil más alto se observan 4 miembros por hogar.

6.



[3] Para comprender los mecanismos detrás de la respuesta de V1=2 están las variables siguientes tales en la EPH (desde V2 hasta V19_B).

Figura 2: Scatter Plot de Proporción de Niños vs. Proporción de Ayuda Social.

Podemos observar una tendencia negativa entre la proporción de niños en el hogar y la proporción de ingresos por ayuda social. Esto indica que a medida que aumenta la proporción de niños menores de diez años en el hogar, la proporción de ingresos de las familias provenientes de ayuda social es menor.

Este hallazgo podría resultar preocupante, ya que está demostrado que la primera infancia es el período en el que las intervenciones tienen mayor impacto[4]. Por lo tanto, deberíamos esperar una mayor ayuda del Estado para los hogares con más niños pequeños, ya que intervenir desde una edad temprana es crucial para maximizar el efecto de nuestras acciones.

9. Tasa de hogares bajo la línea de pobreza: 34.19%

Según el INDEC, en el segundo semestre de 2023, el 31,8% de los hogares estaban en situación de pobreza. Sin embargo, nuestros cálculos indican que la tasa de hogares bajo la línea de pobreza es del 34,19%, un porcentaje mayor que el reportado por el INDEC. Esto puede deberse a varios factores.

En primer lugar, es posible que hayamos eliminado los outliers, y al hacerlo, podríamos haber excluido a los hogares más pobres de nuestra muestra, reduciendo así el porcentaje de hogares en pobreza.En segundo, el peso utilizado en nuestra muestra, PONDIH, podría no ser tan preciso, lo que también podría afectar nuestros resultados.

Parte III: Clasificación y regularización

2.

	modelo	parametros	matrices_de_confusion	curvas_roc	promedio_auc	accuracies	promedio_accuracy
0	Regresion Logistica	{'C': 0.01}	[[[308, 42], [92, 164]], [[316, 43], [63, 184]...	[([0.0, 0.0, 0.0, 0.002857142857142857, 0.0028...	0.874829	[0.7788778877887789, 0.8250825082508251, 0.797...	0.795380
1	LDA	{}	[[[305, 45], [103, 153]], [[320, 39], [82, 165...	[([0.0, 0.0, 0.0, 0.002857142857142857, 0.0028...	0.857584	[0.7557755775577558, 0.8003300330033003, 0.795...	0.782178
2	k-NN	{'n_neighbors': 3}	[[[307, 43], [97, 159]], [[318, 41], [86, 161]...	[([0.0, 0.005714285714285714, 0.04571428571428...	0.843257	[0.768976897689769, 0.7904290429042904, 0.7739...	0.781188

Figura 3: Matriz de confusión, auc y accuracy de LDA, Regresión logística y K-NN

En la Figura 3 podemos comparar la precisión en la predicción de tres modelos: la regresión logística con un parámetro de 0.01, el análisis discriminante lineal (LDA) y K-Nearest Neighbors (KNN) con K=3. Para realizar esta comparación, se toman como medidas el

[4] Heckman, J. J., & Mosso, S. (2014). The economics of human development and social mobility. Annu. Rev. Econ., 6(1), 689-733.

promedio de AUC (Área Bajo la Curva ROC) y el promedio de *accuracy* (precisión) de distintas simulaciones realizadas para cada modelo.

El AUC mide la capacidad del modelo para distinguir entre clases positivas y negativas. Cuanto mayor es el AUC, mejor es el modelo en identificar correctamente los verdaderos positivos en relación con los falsos positivos. Por otro lado, la accuracy indica la proporción de predicciones correctas sobre el total de predicciones. Un valor alto de accuracy significa que el modelo hace más predicciones correctas.

En cuanto a los resultados, observamos que la regresión logística tiene un promedio de AUC de 0.875 y un promedio de accuracy de 0.795. El análisis discriminante lineal (LDA) presenta un promedio de AUC de 0.858 y un promedio de accuracy de 0.782. Por su parte, K-Nearest Neighbors (KNN) con K=3 muestra un promedio de AUC de 0.843 y un promedio de accuracy de 0.781.

Podemos interpretar que la regresión logística es el modelo que mejor predice, ya que tiene el mayor AUC y el mayor promedio de accuracy entre los tres modelos. Esto significa que la regresión logística clasifica correctamente los verdaderos positivos el 87.5% del tiempo y predice correctamente el 79.5% de la muestra. En comparación, LDA y KNN tienen un rendimiento ligeramente inferior, con la LDA mostrando un buen rendimiento pero no tan alto como la regresión logística, y KNN teniendo el rendimiento más bajo.

En conclusión, la regresión logística es el modelo más efectivo en esta comparación, ya que tiene la mayor capacidad para distinguir entre clases y hacer predicciones precisas, reflejando su alto AUC y promedio de accuracy.

3. La elección del valor de λ se realizará bajo *cross validation* y los modelos que utilizan son regresiones logísticas con regulaciones (Lasso y Ridge) los cuales se utilizan para problemas de linealidades, métodos de análisis discriminantes y vecinos cercanos.

La manera en la que elegimos λ es mediante *k-fold cross validation algorithm* en la cual realizamos los siguientes pasos: 1) partimos los datos al azar en K partes, 2) ajustamos el modelo dejando afuera una de las particiones, 3) repetimos este proceso para cada posible valor de λ 4) computamos los errores de predicción de los datos que no utilizamos 4) repetimos k veces y luego promediamos los errores

Las observaciones son elegidas en dos roles (*test* y *training*). Y no utilizamos el conjunto de prueba para la elección de λ porque puede llevar a un problema de sobreajuste al conjunto de datos particulares, si usamos múltiples subconjuntos de datos para la validación tienes una mejor estimación de generalización del modelo, no estamos atados a los datos observados buscando mejor desempeño en aquellos datos no vistos.

Entonces, a la hora de elegir la cantidad de particiones de la muestra tenemos el trade-off que, cuanto menos particiones hacemos, maximizamos los datos de entrenamientos pero somos sensibles a valores particulares. Mientras que cuando la partición es grande maximizamos los

datos de testeo pero el modelo está estimado de manera menos precisa. En la regla se utilizan 5 o 10 particiones.

4. Usar un k muy pequeño, como k=2, puede llevar a una mayor varianza en la estimación porque cada conjunto de validación es grande y cada conjunto de entrenamiento es pequeño, lo que significa que el modelo puede estar subentrenado en cada iteración y los resultados pueden variar significativamente entre diferentes particiones de los datos. Esto puede resultar en un modelo menos generalizado y con mayor sesgo. Por otro lado, usar un k muy grande, como k=10, tiende a reducir tanto el sesgo como la varianza, proporcionando una estimación más estable y representativa del rendimiento del modelo. Sin embargo, esto incrementa el costo computacional porque el modelo debe ser entrenado y validado más veces, aumentando el tiempo y los recursos necesarios para completar el proceso.

Cuando k=n, es decir, cuando se realiza *Leave-One-Out*, el modelo se entrena n veces, utilizando n-1 muestras para entrenar y una muestra para validar en cada iteración. Esto minimiza el sesgo porque el modelo se entrena con casi todos los datos disponibles en cada iteración, pero puede aumentar la varianza debido a que cada estimación del error se basa en una sola muestra de validación, lo que puede hacer que las estimaciones del rendimiento sean muy variables. Esto puede llevar a una interpretación ambigua del rendimiento del modelo y dificultar la detección de *overfitting*. Además, *Leave-One-Out* es computacionalmente muy costoso, especialmente para grandes conjuntos de datos, ya que requiere entrenar el modelo tantas veces como el número de muestras en el conjunto de datos.

5. El valor de lambda óptimo elegido para Lasso fue de 10 mientras que para Ridge fue de 100. Es decir, en el modelo de Ridge, es más óptimo penalizar más a los coeficientes. Esto puede ser debido a que Ridge no puede llevar los coeficientes a 0.

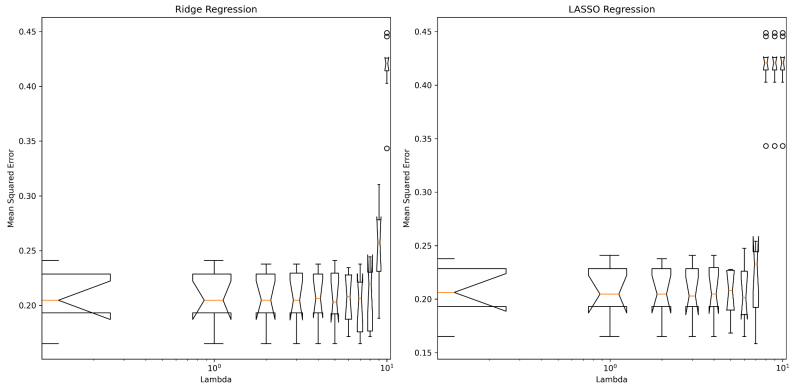


Figura 4: Box plot para Ridge y Lasso.

En los gráficos de los boxplots, observamos la variabilidad del error cuadrático medio (MSE) en función de diferentes valores de lambda para las regresiones Ridge y LASSO. En el caso de Ridge Regression, notamos que para valores muy bajos de lambda, el MSE es relativamente alto con mayor variabilidad, lo que indica menor precisión y consistencia en las predicciones. A medida que la lambda aumenta, el MSE disminuye, alcanzando un valor mínimo alrededor de $\lambda = 1$. Sin embargo, más allá de $\lambda = 1$, el MSE vuelve a aumentar,

especialmente en lamda = 10, donde se observa un incremento significativo del MSE y su variabilidad, sugiriendo un ajuste excesivo del modelo que no generaliza bien a nuevos datos. Esto implica que la penalización adecuada de los coeficientes mejora el rendimiento del modelo hasta cierto punto, pero una penalización demasiado alta puede ser contraproducente.

Por otro lado, en la regresión LASSO, se observa un patrón similar. Para valores muy bajos de lambda, el MSE es alto y variable. Conforme lambda aumenta, el MSE disminuye hasta alcanzar su valor mínimo alrededor de $\lambda = 1$, indicando que la regularización ayuda a mejorar la precisión del modelo. No obstante, al igual que en Ridge, cuando lamda es demasiado alto ($\lambda = 10$), el MSE aumenta significativamente, indicando que el modelo está siendo penalizado en exceso, lo que resulta en una pérdida de información relevante al establecer muchos coeficientes en cero. Estos resultados demuestran la importancia de seleccionar un valor óptimo de lambda que equilibre adecuadamente la regularización y el ajuste del modelo para obtener predicciones precisas y generalizables.

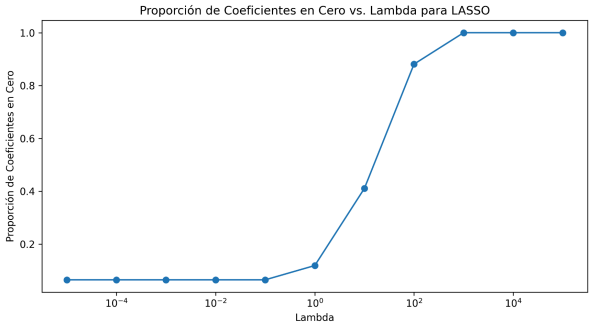


Figura 5: Proporción de coeficientes en cero vs Lambda para Lasso

En este gráfico podemos ver cómo a medida de que aumenta la penalidad que le otorgamos al modelo LASSO vamos eliminando más variables.

6. Las variables omitidas por Lasso son las siguientes:

ANO4, TRIMESTRE, H15, REGIÓN, PONDERA, CH03, CH15_COD, CH16, ESTADO, IMPUTA, PP02C2, PP02C3, PP02C4, PP02C6, PP02C7, PP02C8, PP02H, PP02I, PP03C, PP03D, PP03H, PP03I, PP04A, PP04B2, PP04B3_MES, PP04B3_ANO, PP04B3_DIA, PP04C, PP04C99, PP04G, PP05B2_MES, PP05B2_ANO, PP05B2_DIA, PP05C_2, PP05E, PP05F, PP06A, PP06H, PP07D, PP07F1, PP07F2, PP07F3, PP07F4, PP07G1, PP07G2, PP07G3, PP07G4, PP07G_59, PP07H, PP07I, PP09B, PP09C, PP10C, PP10D, PP11A, PP11B_COD, PP11B2_MES, PP11B2_DIA, PP11C, PP11G_DIA, PP11L, PP11L1, PP11M, PP11O, PP11P, PP11R, REALIZADA, IV4, II3, II4_1, II6, II9, V21, V5, V12, V19_A, V19_B, IX_MEN10, IX_MAYEQ10, VII2_4, trabaja.

[Ver **Anexo 2** para que signifiquen estas variables]. Algunas de estas variables puede ser que de por sí sean muy importantes, pero tal vez al estar muy correlacionadas con otras variables pierden su efecto de por sí y por eso son eliminadas.

Resulta importante destacar que de las 10 variables que creíamos importante para predecir la pobreza en el inciso 1 de la Parte I: IV1, IV2, IV6, IV9, IV11, IV12_1, II1, II5_1, II7 y IX_Men10, la única variable que eliminó LASSO fue IX_Men10. Por lo tanto, nuestras primeras estimaciones de las variables importantes a la hora de predecir la pobreza no fueron desacertadas.

7. Notamos que el error cuadrático medio para el óptimo de Ridge es de 0.2016 mientras que para el Lasso óptimo el error cuadrático medio es de 0.2029. Por lo tanto, Ridge predice mejor que Lasso en este modelo. Esto puede suceder debido a la alta multicolinealidad de muchas variables que hace que algunas de ellas se borren arbitrariamente.

8. Cuando comparamos los errores cuadráticos medios de los modelos utilizando k=10 validación cruzada encontramos los siguientes resultados:

Regresión Logística: 0.20759075907590757, Análisis de Discriminante Lineal: 0.21155115511551154, Ridge Regression (λ=100): 0.20165016501650163, LASSO Regression (λ=10): 0.201980198019802

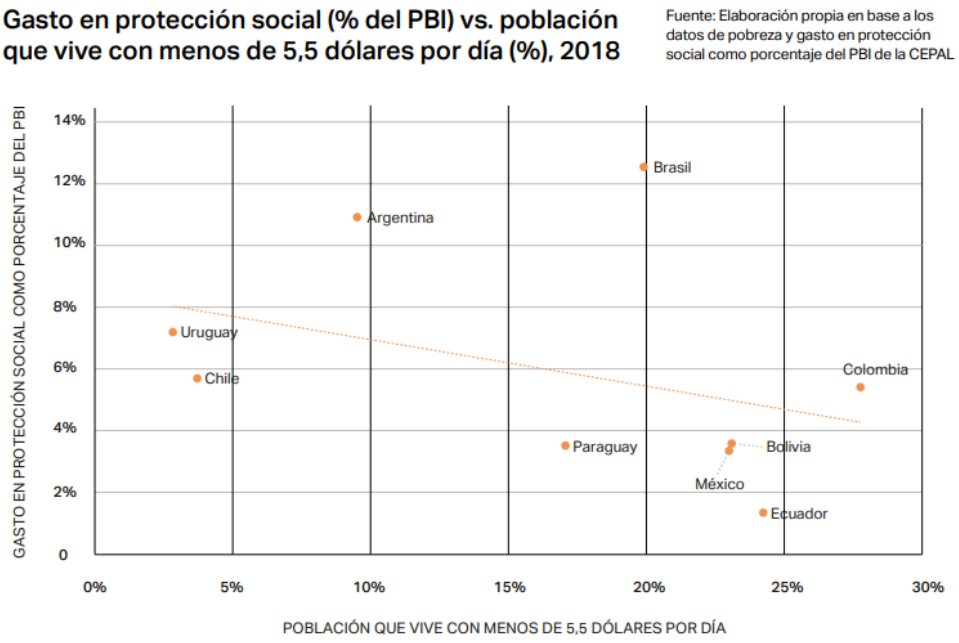
Por lo tanto, podemos afirmar que el mejor modelo para predecir es el modelo de Ridge con λ=100 ya que es el que menor error cuadrático tiene. Notar que en este caso los ECM de Ridge y Lasso nos dieron distintas que en el inciso anterior ya que aquí las evaluamos con un cross validation con k=10, algo tal vez más acertado. De todos modos, Ridge siguió siendo mejor que Lasso.

9. Utilizando el método de Ridge con un λ igual a 100, nuestro modelo estimó que el 39,48% de las personas que no respondieron son pobres, y el 35,7% de los hogares. Lo que resulta relevante de este resultado es que, si bien bien la cantidad de personas pobres predichas por el modelo está por debajo de del valor que informa el INDEC (41,7%), lo cual resulta sorprendente, ya que uno esperaría que si la persona no respondió la pregunta de ITF, es porque no quiere revelar cuánto gana, muchas veces debido a que es pobre y puede sentir vergüenza de decirlo. Por otra parte, la proporción de hogares pobres, para la base no respondieron está por encima de las estimaciones del INDEC, que como mencionamos anteriormente es del 31,8%. Estas diferencias en las estimaciones se puede deber a dos cuestiones. En primer lugar, pueden deberse a la falta de accuracy del modelo, que si bien en promedio fue del 80%, no es perfecto. En segundo lugar, puede suceder que verdaderamente existan diferencias entre la base no respondieron y respondieron.

Anexo

Anexo 1: Gráfico Zarazaga et al (2011) proporción del gasto y niveles de pobreza % PBI.

Gráfico 23



Anexo 2: Variables eliminadas por LASSO.

ANO4: Año en cuatro dígitos, TRIMESTRE: Trimestre del año, H15: Tipo de hogar (puede ser la categoría del hogar en función de características específicas), REGIÓN: Región geográfica, PONDERA: Ponderador para expandir la muestra, CH03: Sexo del individuo, CH15_COD: Código de la relación con el jefe del hogar, CH16: Código de la situación conyugal, ESTADO: Estado civil, IMPUTA: Indicador de imputación de datos, PP02C2: Nivel de educación alcanzado, PP02C3: Tipo de institución educativa, PP02C4: Situación de asistencia a establecimiento educativo, PP02C6: Turno en que asiste a clases, PP02C7: Jornada escolar, PP02C8: Situación de asistencia a actividades extracurriculares, PP02H: Nivel educativo alcanzado por los padres, PP02I: Situación laboral de los padres, PP03C: Categoría ocupacional del individuo, PP03D: Situación laboral, PP03H: Tipo de jornada laboral, PP03I: Número de horas trabajadas, PP04A: Ingreso mensual del trabajo principal, PP04B2: Ingreso por otro trabajo, PP04B3_MES: Mes de ingreso por otro trabajo, PP04B3_ANO: Año de ingreso por otro trabajo, PP04B3_DIA: Día de ingreso por otro trabajo, PP04C: Ingreso por trabajos ocasionales, PP04C99: Ingreso por trabajos ocasionales en el año 99, PP04G: Ingreso no laboral, PP05B2_MES: Mes de ingreso no laboral, PP05B2_ANO: Año de ingreso no laboral, PP05B2_DIA: Día de ingreso no laboral, PP05C_2: Ingreso por renta, PP05E: Ingreso por pensión, PP05F: Ingreso por subsidio, PP06A: Ingreso total del hogar, PP06H: Ingreso per cápita del hogar, PP07D: Indicador de pobreza, PP07F1: Condición habitacional, PP07F2: Acceso a servicios básicos, PP07F3: Materiales de construcción de la vivienda, PP07F4: Condiciones del entorno de la vivienda, PP07G1: Equipamiento del hogar, PP07G2: Servicios de comunicación en el hogar, PP07G3: Servicios de transporte en el hogar, PP07G4: Acceso a servicios educativos, PP07G_59:

Acceso a servicios de salud, PP07H: Acceso a servicios de seguridad social, PP07I: Acceso a servicios de recreación, PP09B: Gasto total del hogar, PP09C: Gasto en alimentos, PP10C: Gasto en transporte, PP10D: Gasto en salud, PP11A: Gasto en educación, PP11B_COD: Código de gastos no especificados, PP11B2_MES: Mes de gasto no especificado, PP11B2_DIA: Día de gasto no especificado, PP11C: Gasto en vivienda, PP11G_DIA: Día de gasto en vivienda, PP11L: Gasto en recreación, PP11L1: Gasto en comunicaciones, PP11M: Gasto en cultura, PP11O: Gasto en transporte, PP11P: Gasto en servicios, PP11R: Gasto en otros bienes y servicios, REALIZADA: Indicador de encuesta realizada, IV4: Indicador de infraestructura vial, II3: Indicador de infraestructura de servicios, II4_1: Indicador de calidad de vida, II6: Indicador de seguridad, II9: Indicador de acceso a tecnologías, V21: Indicador de vulnerabilidad, V5: Indicador de condición económica, V12: Indicador de acceso a educación, V19_A: Indicador de salud, V19_B: Indicador de bienestar, IX_MEN10: Número de personas menores de 10 años en el hogar, IX_MAYEQ10: Número de personas mayores de 10 años en el hogar, VII2_4: Indicador de acceso a vivienda, trabaja: Indicador de trabajo.