**ESTIMACIÓN DE LA REGRESIÓN LASSO**

**Miguel Coto-García1 & Natalia Díaz-Ramírez1**

1Estudiante de Maestría Profesional en Estadística, Universidad de Costa Rica, San Pedro, Costa Rica

1. **INTRODUCCIÓN**

La selección de variables es un procedimiento utilizado para incluir las variables con mayor significancia, entre los métodos más utilizados se encuentran los métodos de eliminación por stepwise que sólo mejora la exactitud de las predicciones en ciertos casos, como cuando hay variables con un efecto importante en la variable independiente.

Tibshirani propuso un nuevo método llamado LASSO (Least Absolute Shrinkage and Selection Operator), el cual es una técnica de regresión lineal regularizada que combina un modelo de regresión con un procedimiento de contracción de algunos parámetros hacia cero y selección de variables, imponiendo una penalización sobre los coeficientes de regresión para mejorar la exactitud e interpretabilidad del modelo estadístico producido por este.

Tibshirani (1996) en su artículo propone el modelo minimizando la suma residual de cuadrados sujeto a que la suma del valor absoluto de los coeficientes sea menor que una constante, dicha restricción, produce coeficientes que tienden a 0 y, por lo tanto, da modelos interpretables.

En este método el parámetro de penalización lambda se elige mediante validación cruzada, donde se selecciona un conjunto de valores para lambda, se calcula el error de validación para cada valor y se elige el valor de lambda con menor error, finalmente se reajusta el modelo con el valor de lambda escogido.

A continuación, se muestra la sección del modelo en la cual se describe la función de verosimilitud y propiedades del modelo como la potencia de la prueba, estimación de coeficientes, los residuos de deviancia y el umbral del parámetro lambda. Además, se describen los resultados de la aplicación del modelo a un ejemplo empírico donde se evaluó la selección de variables y medidas de clasificación comparando los modelos con la regresión lasso y regresión logística. Finalmente se presenta una sección de conclusiones.

1. **MODELO**
   1. **Función de verosimilitud**

Función de densidad de es , donde g es una función de enlace conocida.

Se denota la función condicional de log-verosimilitud de por .

La función de log verosimilitud penalizada es:

Para obtener un estimador de máxima verosimilitud penalizado de β se minimiza la función anterior respecto a β para algún umbral del parámetro λ.

* 1. **Propiedades del modelo**

Para evaluar las propiedades del modelo se utilizó una función de enlace logito en la función de log verosimilitud.

Se evaluó la potencia de la prueba de los coeficientes con simulaciones, se consideraron tres tamaños de muestra 50, 100 y 1000, con mil simulaciones se obtuvo la proporción de veces en las que cada uno de los once coeficientes son significativos. En el gráfico 1 se puede observar que a mayor tamaño de muestra la proporción de coeficientes significativos aumenta y la cantidad de coeficientes significativos es menor en tamaños de muestra pequeños que en tamaños grandes.

|  |
| --- |
| **Gráfico 1. Potencia de la prueba para los coeficientes variando el tamaño de la muestra y con mil simulaciones** |
|  |

En la estimación de los coeficientes se consideraron tres tamaños de muestra 50, 100 y 1000, con mil simulaciones se obtuvo la estimación de los coeficientes para un modelo lasso y un modelo logístico, comparando las estimaciones obtenidas con un histograma. Se consideraron dos casos, uno en el que el coeficiente es significativo, b1, y otro en el que no lo es, b4.

En el gráfico 2, se muestran los resultados para el caso en el que el coeficiente es significativo. Se puede observar que, a mayor tamaño de muestra, las estimaciones del coeficiente son más pequeñas en ambos modelos. También se muestra que las estimaciones en la regresión logística tienden a ser mayores que en la regresión lasso.

|  |  |  |
| --- | --- | --- |
| **Gráfico 2. Histograma de la estimación del coeficiente significativo (b1) variando el tamaño de muestra con mil simulaciones** | | |
|  | **Modelo** | |
| **Tamaño muestra** | **Lasso** | **Logística** |
| **n=50** |  |  |
| **n=100** |  |  |
| **n=1000** |  |  |

En el gráfico 3, se muestran los resultados para el caso del coeficiente no significativo. Se puede observar que, para la regresión lasso a mayor tamaño de muestra las estimaciones del coeficiente tienden a valores muy pequeños cercanos a cero. Mientras que para la regresión logística se observa que a mayor tamaño de muestra las estimaciones del coeficiente son más pequeñas, pero no tienden a valores cercanos a cero.

|  |  |  |
| --- | --- | --- |
| **Gráfico 3. Histograma de la estimación del coeficiente no significativo (b4) variando el tamaño de muestra con mil simulaciones** | | |
|  | **Modelo** | |
| **Tamaño muestra** | **Lasso** | **Logística** |
| **n=50** |  |  |
| **n=100** |  |  |
| **n=1000** |  |  |

En el gráfico 4 se obtienen los gráficos de residuos de deviancia contra los valores predichos para el modelo lasso, en tres escenarios variando el valor del parámetro lambda y el tamaño de muestra. En los gráficos para los posibles escenarios y tamaños de muestra no se observa un patrón aleatorio, se muestra una ligera tendencia negativa, conforme los valores predichos aumentan los residuos disminuyen. También se observa que a mayor valor de lambda los predichos muestran valores bajos.

|  |  |  |  |
| --- | --- | --- | --- |
| **Gráfico 4. Residuos de deviancia vs Predicho variando el tamaño de muestra y el valor de lambda** | | | |
|  | **Parámetro Lambda** | | |
| **Tamaño muestra** | **Lambda =0.001** | **Lambda = 0.02** | **Lambda = 0.07** |
| **n=50** |  |  |  |
| **n=100** |  |  |  |
| **n=1000** |  |  |  |

En el gráfico 5 se muestran los gráficos de normalidad qq plot donde se puede observar que, para las diferentes combinaciones de escenarios, los residuos de deviancia no se distribuyen como una normal. En tamaños de muestra altos se forma dos agrupaciones, una con residuos de deviancia positivos y altos, y otra con residuos de deviancia negativos.

|  |  |  |  |
| --- | --- | --- | --- |
| **Gráfico 5. Gráfico de normalidad QQ-Plot por cantidad de predictores y variando el tamaño de muestra** | | | |
|  | **Parámetro Lambda** | | |
| **Tamaño muestra** | **Lambda =0.001** | **Lambda =0.02** | **Lambda =0.07** |
| **n=50** |  |  |  |
| **n=100** |  |  |  |
| **n=1000** |  |  |  |

También se consideró el umbral lambda, donde se obtuvo la cantidad de variables significativas para distintos valores de lambda. En el gráfico 6 se observa que a mayor lambda menor es la cantidad de variables significativas. Se suele elegir los valores del umbral lambda.1se o el lambda.min. (detallar más estos con un artículo)

|  |
| --- |
| **Gráfico 6. Cantidad de variables significativas según el lambda** |
|  |

* 1. **Ejemplo empírico**

Se utilizaron datos recolectados por el Banco Interamericano de Desarrollo (BID) que contiene variables sobre 9558 personas de 2973 hogares para Costa Rica. La variable objetivo se refiere a la categorización usando el PMT que clasifica a los hogares en “No vulnerables”, “Vulnerables”, “Pobres” y “En extrema pobreza”, en este caso se categorizó como 1 hogares pobres o en extrema pobreza y como 0 las categorías restantes.

Las variables utilizadas fueron:

* ninnos:
* cocina\_elec:
* banno:
* sin\_educ:
* urbano:
* renta:
* sani\_alcant:
* casadx:
* dependencia:
* casa\_propia:

Además, se realizó una comparación de los resultados con la regresión logística.

Al comparar los coeficientes significativos de los dos modelos se observa que los obtenidos por la regresión lasso también son significativos en el modelo de la regresión logística y las estimaciones de dichos coeficientes son similares, así como los respectivos errores estándar. En este caso la regresión logística presenta una mayor cantidad de variables significativas al compararla con la regresión lasso.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Cuadro 1. Variables significativas con regresión lasso y logística** | | | | |
| **Modelo** | **Variable** | **Coeficiente** | **Error Estándar** | **Valor P** |
| **Lasso** |  |  |  |  |
|  | ninnos | 0.356 | 0.038 | < 2.2e-16 |
|  | cocina\_elec | -0.313 | 0.095 | 0.001 |
|  | renta | -0.010 | 0.001 | 2.333e-16 |
|  | casadx | -0.325 | 0.095 | 0.001 |
|  | dependencia | 0.052 | 0.021 | 0.012 |
|  | casa\_propia | -0.427 | 0.120 | 0.001 |
| **Logística** |  |  |  |  |
|  | ninnos | 0.497 | 0.040 | < 2e-16 |
|  | cocina\_elec | -0.466 | 0.101 | 3.73e-06 |
|  | banno | -1.284 | 0.532 | 0.016 |
|  | sin\_educ | 0.990 | 0.110 | 7.14e-07 |
|  | renta | -0.009 | 0.001 | 3.47e-14 |
|  | casadx | -0.553 | 0.102 | 5.37e-08 |
|  | dependencia | 0.098 | 0.021 | 4.97e-06 |
|  | casa\_propia | -0.465 | 0.125 | 0.0002 |

En cuanto al error cuadrático medio se obtuvo que el modelo de regresión logística es el que tiene un valor menor de 0.30.

|  |  |
| --- | --- |
| **Cuadro 2. Error cuadrático medio para la regresión lasso y la regresión logística** | |
| **Modelo** | **MSE** |
| Logística | 0.302 |
| Lasso | 0.427 |

Al evaluar las medidas de precisión para la clasificación se utilizó como punto de corte la media de la variable respuesta que corresponde a 0.22. Se obtuvo que el modelo de regresión lasso tiene una sensibilidad de 76% mientras que el modelo de la regresión logística tiene un 65%.

|  |  |  |
| --- | --- | --- |
| **Cuadro 3. Medidas de precisión de los modelos de regresión lasso y regresión logística** | | |
| **Modelo** | **Sensibilidad** | **Especificidad** |
| Logística | 0.655 | 0.711 |
| Lasso | 0.761 | 0.518 |

En el gráfico 5, se muestran las curvas ROC para ambos modelos, en los cuales se obtuvo un área bajo la curva similar de 0.73.

|  |  |
| --- | --- |
| **Gráfico 7. Curva ROC para la regresión lasso y la regresión logística** | |
| **Regresión Lasso** | **Regresión Logística** |
|  |  |

1. **CONCLUSIONES**

El modelo de regresión lasso permite realizar selección de variables cuando se tiene una gran cantidad.

En el ejemplo utilizado se obtuvo que el modelo de regresión lasso da una mejor sensibilidad para la categoría de hogares pobres que el modelo de logístico. En cuanto a la selección de variables con la regresión lasso se obtuvo que las variables se redujeron de 10 a 6, donde las que resultaron significativas fueron: niños, cocina\_elec, renta, casadx, dependencia, casa\_propia.

1. **BIBLIOGRAFÍA**

Fan, J., Li, R. (1999). Variable Selection via Penalized Likelihood. California Digital Library: Universidad de California.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of Royal Statistical Society, B, 58, 267-288.

1. **ANEXO**

**Código de R**