Regresión Logística Robusta y Regularizada

Juliana Restrepo Tobar
jrestrepot@eafit.edu.co

Maria Alejandra Moncada Agudelo
mamoncada@eafit.edu.co

Abelino Sepúlveda Estrada
asepulvede@eafit.edu.co

Andrea Carvajal Maldonado
acarvajal@eafit.edu.co

31 de agosto de 2022

Resumen. La regresión logística es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica en función de las variables independientes o predictoras. Esto nos permite estimar la relación existente entre una variable dependiente y un conjunto de variables independientes. En este trabajo se busca proponer una versión robusta de la regresión logística para que las predicciones no sean sensibles a valores atípicos (outliers). Además, veremos la regresión regularizada para corregir el sobreajuste de las estimaciones.

Índice

1. Introducción

1.1. Regresión logística

La regresión logística es un método que puede mostrar cuál de las diversas covariables que se evalúan tiene la mayor asociación con un resultado y proporciona una medida de la magnitud de la influencia potencial [6].

Las odds asociados a cierto suceso se definen como la razón entre la probabilidad de que dicho suceso ocurra y la probabilidad de que no ocurra; es decir, un número que expresa cuánto más probable es que se produzca frente a que no se produzca el hecho en cuestión [1].

$$Odd = \frac{p}{1-p}$$

$$Logit = Ln(Odd) = Ln\left(\frac{p}{1-p}\right)$$

Se puede suponer un modelo lineal donde n es el número de covariables en el modelo, α es el valor de $Ln(Odd)$ cuando las covariables valen 0, x_n es cada covariable y β_n es su respectivo coeficiente.

$$Ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

$$\frac{p}{1-p} = e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}$$

$$p = (1-p)e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}$$

$$p = e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n} - pe^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}$$

$$p + pe^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n} = e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}$$

$$p(1 + e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}) = e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}$$

Así, podemos despejar la probabilidad de que el suceso ocurra y obtenemos la siguiente ecuación:

$$p = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}}{1 + e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}}$$

Ahora, solo queda encontrar el modelo lineal $\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$, el cual encontramos mediante la regresión lineal múltiple.

1.2. Regresión lineal múltiple

La regresión lineal múltiple trata de ajustar modelos lineales o linealizables entre una variable dependiente y más de una variables independientes.[4]
Se obtiene un modelo lineal de la forma:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

donde y es la variables dependiente, explicada o variable respuesta, x son las variables independientes o explicativas, β_n son los coeficientes estimados que representan el cambio previsto en y por cada unidad de cambio en cada x, y β_0 es el intercepto o término constante.
Para obtener este modelo se utiliza el método de los mínimos cuadrados ordinarios y la eliminación gaussiana la cual, tras su primera iteración, reduce el porblema a

$$\Sigma\hat{\beta} = (Cov(Xi, y))'$$

Al despejar $\hat{\beta}$ y asumiendo que Σ es no singular, se obtiene:

$$\hat{\beta} = \Sigma^{-1}(Cov(Xi, y))'$$

El término constante o β_0 será equivalente a la diferencia entre la media de y y la media de los valores predichos de las estimaciones $X\hat{\beta}$.

1.2.1. Regresión robusta

Como vimos anteriormente, la fórmula de los betas de la regresión lineal múltiple usa la matriz de covarianza de las covariables Σ.

$$\Sigma = \begin{bmatrix} Var(X1) & Cov(X1, X2) & ... & Cov(X1, Xn) \\ Cov(X2, X1) & Var(X2) & ... & Cov(X2, Xn) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ Cov(Xn, X1) & Cov(Xn, X2) & ... & Var(Xn) \end{bmatrix}$$

La covarianza de dos covariables X y Y, o Cov(X,Y) está definida como

$$Cov(X,Y) = \rho_{X,Y}\sigma(X)\sigma(Y)$$

Donde $\rho_{X,Y}$ es el coeficiente de correlación de Pearson de las covariables, y $\sigma(X)$ y $\sigma(Y)$ son las desviaciones estándar de X y Y respectivamente.
La correlación de Pearson es uno de los estimadores estadísticos más usados. Sin embargo, su valor puede verse seriamente afectado por la presencia de solo un outlier [3]. Para hacer la matriz de covarianzas más robusta, reemplazamos esta correlación por los coeficientes de correlación de Spearman y de Kendall, estimadores de correlación no paramétricos y mucho más robustos [3].

1.3. Regresión Regularizada

Como ya sabemos, la regresión logística funciona calculando un coeficiente de regresión para cada covariable independiente, si los coeficientes son muy grandes, la regresión logística sobreestima la probabilidad de que ocurra cierto suceso.Para arreglar este sobreajuste, realizamos una regularización que penaliza los coeficientes elevados

1.3.1. Regularización Ridge

Esta técnica fue propuesta originalmente por Hoerl y Kennard [5] como un método para eludir los efectos adversos del problema de colinealidad en un modelo lineal estimado por mínimos cuadrados, en el contexto p ¡n. Regresión Ridge es muy similar a los mínimos cuadrados, a excepción de que los coefcientes se estiman minimizando una cantidad diferente. Los coefcientes estimados por Ridge, $\hat{\beta}^{ridge}$, son los valores que minimizan [2]

$$\hat{\beta}^{ridge} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda - \sum_{j=1}^{p}\beta_j^2$$

donde $\lambda \geq 0$ es el parámetro de contracción que se determinará por separado.

1.3.2. Regularización Lasso

Lasso(least absolute shrinkage and selection operator) es una técnica de regresión lineal regularizada, como Ridge, con una leve diferencia en la penalización que trae consecuencias importantes.En especial, a partir de cierto valor del parámetro de complejidad el estimador de Lasso produce estimaciones nulas para algunos coeficientes y no nulas para otros, con lo cual Lasso realiza una especie de selección de variables en forma continua, debido a la norma L1. Lasso reduce la variabilidad de las estimaciones por la reducción de los coeficientes y al mismo tiempo produce modelos interpretables por la reducción de algunos coeficientes a cero.[2] Lasso utiliza la siguiente formula para encontrar

sus coeficientes:

$$\hat{\beta}^{lasso} = min_{\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij})^2) \right\} s.a \sum_{j=1}^{p} \beta_j \leq s$$

siendo $s \geq 0$ el respectivo parámetro de penalización por complejidad.

1.3.3. Regularización ElasticNet

Zou y Hastie [10], proponen una nueva técnica de regularización y selección de variables conocido como Elastic Net, la cual retiene las ventajas de Lasso, hace automáticamente selección de variables y contracción continua, y al mismo tiempo supera algunas de sus limitaciones. Con este nuevo método se puede seleccionar grupos de variables correlacionadas. Este método es particularmente útil cuando el número de predictores (P) es mucho más grande que el número de observaciones (n). En primer lugar, los autores definen "Naive Elastic Net"(red elástica simple), que es un método de mínimos cuadrados penalizado utilizando una penalización nueva de Elastic Net. [2] Para cualesquiera λ_1, λ_2 constantes fijas no negativas, se calculan los coeficientes a partir de la siguiente ecuación:

$$\hat{\beta}^{ene} = argmin_{\beta} L(\lambda_1, \lambda_2, \beta)$$

2. Metodología

2.0.1. Preparación de los datos

Se seleccionó un dataset utilizado para predecir si es probable o no que una persona sufra un derrame cerebral. Este dataset obtenido de Kaeggle tiene observaciones de 5110 personas con información sobre 12 características tales como la edad, el género, si es fumador o no y otras enfermedades. Para poder realizar las operaciones se convirtieron los datos de tipo String a Floats y se reemplazaron los valores NaN usando el método fillnan. Como los datos se encontraban desbalanceados en una proproción muy grande, se balancearon con el método undersample. Además, se añadieron valores atípicos o outliers para observar el efecto de estos en los resultados con la regresión robusta.

2.0.2. Implementación

Ya teniendo conocimiento de los conceptos definidos anteriormente, podemos empezar con la aplicación de éstos.
Se programaron en el lenguaje Python dos métodos para calcular los coeficientes β_n y el intercepto β_0: la regresión lineal múltiple y la regresión regularizada. Para la regresión regularizada se usaron las librerias de Python correspondientes a cada técnica (Ridge, lasso y ElasticNet) retornado los coeficinetes en cada

caso. Para la regresión lineal se programó esta desde cero obteniendo así los valores por el método de la matriz de covarianza. Además, se reemplazó la matriz de covarianza por sus formas robustas usando los coeficientes de Kendall y de Spearman.
Tras obtener los coeficientes, se aplicó la función sigmoide o link function a los modelos obtenidos para convertirlos en una regresión logística. De esta forma, se obtuvieron 6 resultados: la regresión logistíca no robusta (coeficiente de Pearson), la regresión logística robusta con coeficiente de Kendall, la regresión logística robusta con coeficiente de Spearman, la regresión regularizada con técnica Ridge, la regresión regularizada con técnica Lasso y la regresión regularizada con técnica ElasticNet.

3. Resultados

A continuación se presentan las gráficas del accuracy score y del AUC (Area under the ROC Curve) para cien iteraciones en cada caso.

3.1. Librería:

La regresión logística de la librería de Python arroja resultados excelentes, tiene una precisión acertada y categoriza de manera exitosa si la persona tendrá o no un derrame en la gran mayoría de los casos. Se puede apreciar que alcanza un accuracy score y un AUC perfecto en varias iteraciones consecutivas.

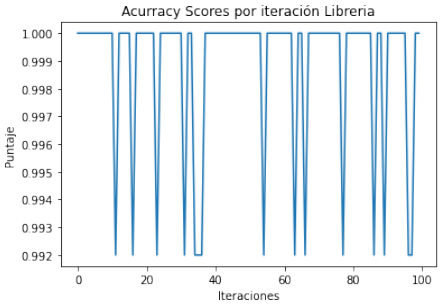

Figura 1: Accuracy Score Librería

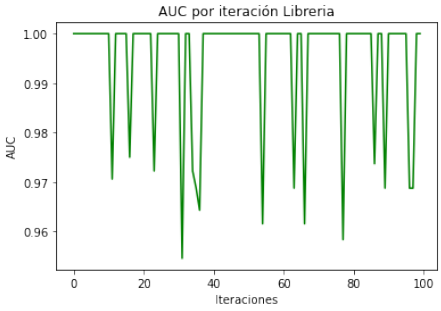

Figura 2: Área bajo la curva ROC Librería

3.2. Coeficientes de correlación

Es importante aclarar que en estas iteraciones se desecharon los casos donde la matriz de covarianzas Σ es singular.
El código que reproduce estos resultados se encuentra en:

https://github.com/jrestrepot/ModelacionSimulacion4/blob/main/RegresionLogisticaRobusta/regresi%C3%B3nlog%C3%ADstica.py

Los resultados de los accuracy score usando los diferentes tipos de coeficientes de correlación son similares. Por un lado la gráfica que tiene un mayor rango es la de Kendall, alcanzando los valores más altos y bajos. Sin embargo, es la gráfica que tiene menos variación. La gráfica de Pearson es parecida a la de Spearman, pero no alcanza resultados tan altos.
Los AUC de los coeficientes robustos dieron mejores resultados que los de Pearson, alcanzando valores de hasta de 0.90. Se puede evidenciar que el mejor AUC es el de Spearman, pues nunca llega a valores más bajos que 0.50 y alcanza valores altos.

3.2.1. Coeficiente de correlación de Pearson

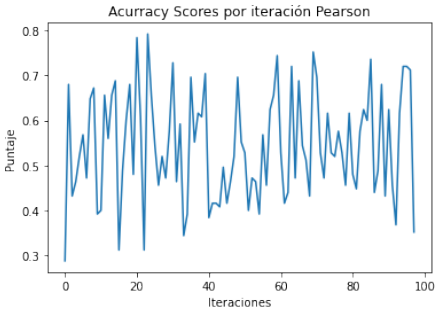

Figura 3: Accuracy Score Pearson

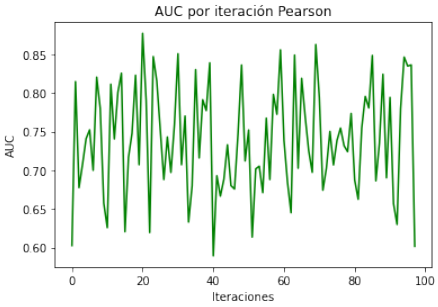

Figura 4: Área bajo la curva ROC Pearson

3.2.2. Coeficiente de correlación de Kendall

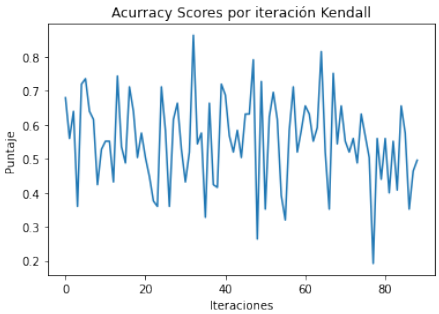

Figura 5: Accuracy Score Kendall

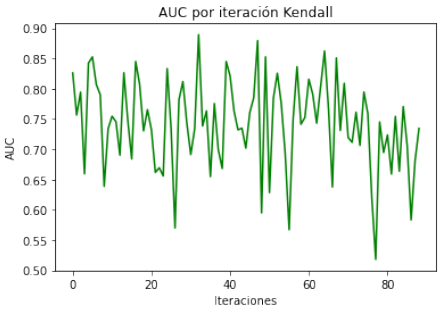

Figura 6: Área bajo la curva ROC Kendall

3.2.3. Coeficiente de correlación de Spearman

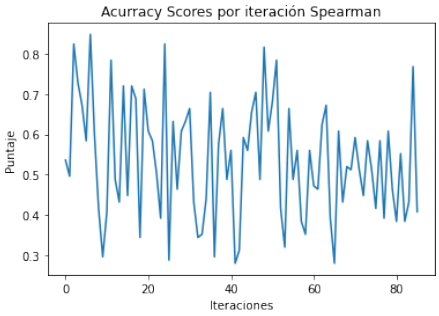

Figura 7: Accuracy Score Spearman

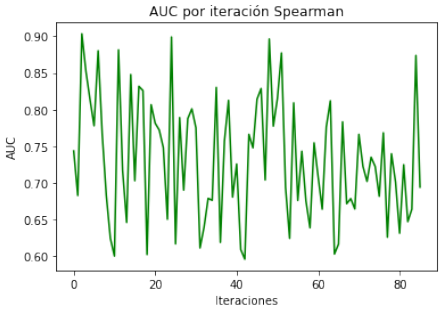

Figura 8: Área bajo la curva ROC Spearman

3.3. Regularización

El código que reproduce los siguientes resultados se encuentra en `https://github.com/jrestrepot/ModelacionSimulacion4/blob/main/RegresionLogisticaRobusta/regresionregularizada.py`

En las siguientes gráficas podemos ver los resultados de la regresión logística con los coeficientes de cada una de las técnica de regresión regularizada. Como se puede apreciar en las figuras ninguna técnica se destaca por su buen resultado, en todas el accuary score siempre esta por debajo de 0.7 y el AUC en general oscila entre valores de 0.5 y 0.6.

3.3.1. Ridge

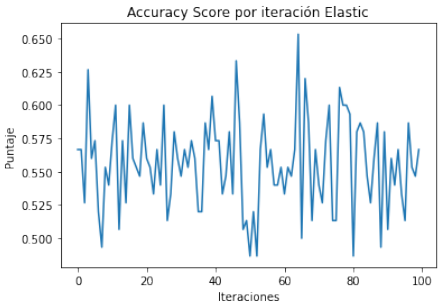

Figura 9: Accuracy Score Ridge

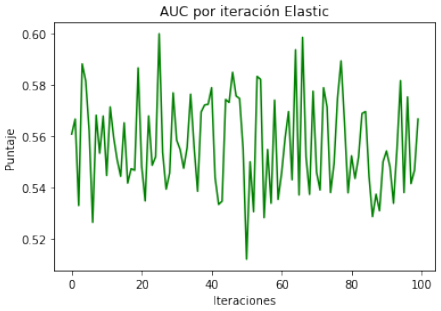

Figura 10: Área bajo la curva ROC Ridge

3.4. Lasso

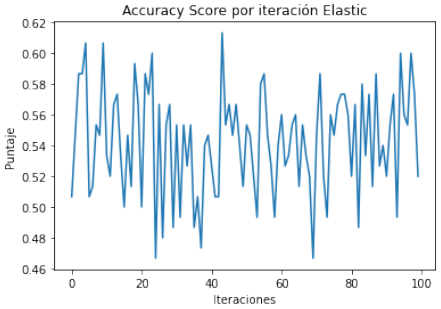

Figura 11: Accuracy Score Lasso

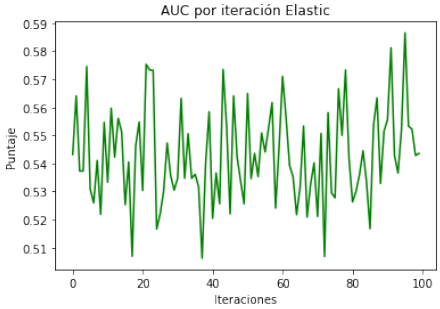

Figura 12: Área bajo la curva ROC Lasso

3.5. ElasticNet

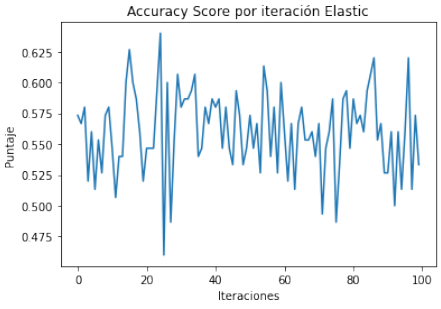

Figura 13: Accuracy Score ElasticNet

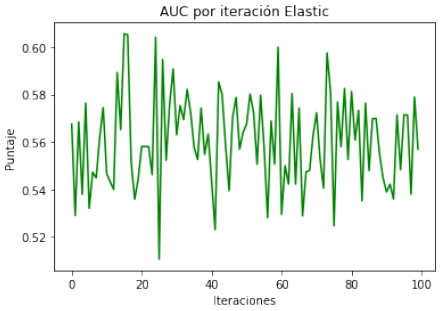

Figura 14: Área bajo la curva ROC ElasticNet

4. Conclusiones

Se puede concluir que los resultados no son los esperados pues no alcanzan un nivel de precisión muy alto. Esto puede ser consecuencia de las múltiples ocasiones en las que la matraiz Σ es singular o tiene un determinante muy cercano a cero dificultando el proceso de invertir la matriz. También se considera que la baja dependencia lineal de los datos usados puede estar afectando considerablemente los resultados. Se espera continuar por esta línea de investigación avanzando en el desarrollo del proyecto y obteniendo una predicción mucho más acertada.

5. Referencias.

Referencias

[1] Luis Carlos Silva Ayçaguer. *Excursión a la regresión logıstica en ciencias de la salud*. Ediciones Dıaz de Santos, 1994.

[2] Marıa Carrasco Carrasco. "Técnicas de regularización en regresión: implementación y aplicaciones". En: (2016).

[3] Catherine Dehon Christophe Croux. "Influence functions of the Spearman and Kendall correlation measures". En: *Stat Methods Appl* 19.1 (2010), págs. 497-515.

[4] Roberto Montero Granados. "Modelos de regresión lineal múltiple". En: *Granada, España: Departamento de Economıa Aplicada, Universidad de Granada* (2016).

[5] Arthur E Hoerl y Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". En: *Technometrics* 12.1 (1970), págs. 55-67.

[6] Juliana Tolles y William J Meurer. "Logistic regression: relating patient characteristics to outcomes". En: *Jama* 316.5 (2016), págs. 533-534.