Evaluación de la clasificación mediante de la técnica estadística Regresión Logística en datos simulados bajo distintos escenarios, para distintos tamaños de muestra.

Celina Beltrán; Ivana Barbona; Ciminari, Jesica

Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina

cbeltran2510@gmail.com

Abstract

This research proposes the study of the multivariate statistical classification technique, Logistic Regression, to evaluate its performance when it is used in simulated data under different scenarios and different sample sizes.

Simulation generated 500 data files for each of the following sample sizes: 30, 75, 200, 400, 600, 1000. Each set contains 6 columns (variables) under different conditions or scenarios. In each sample, 20% of the observations were "marked" to be used as a test group and the remaining 80% for the estimation of the models evaluated in each case. A total of 12,000 simulated data sets were defined, with 6 different sample sizes and 4 scenarios with the following characteristics defined by the structure of the correlation matrix. Scenario 1 corresponds to data from a population in which the predictors are strongly correlated with the response but not with each other. Scenario 2 proposes a simulation based on a population with little correlation of the response with the predictor variables but these correlated with each other. In scenario 3, the correlation present in the source population of the simulation is important both between the predictors and between them and the response. Finally, scenario 4 corresponds to an original population in which there is no type of correlation of significant magnitude between the variables, neither of the predictors with the response nor between them.

From this analysis it is concluded that, in conditions where the predictor variables are highly correlated with the response (scenarios 1 and 3), regardless of the correlation between the predictors, the Logistic Regression technique works satisfactorily. However, , when the predictors are poorly correlated with the response (scenarios 2 and 4), the percentage of correct classification is much lower. This difference between the two

groups of scenarios in terms of the correlation of the response with the predictors is accentuated as the sample size increases.

As a final conclusion, it can be said that, regardless of the sample size, when the response variable is poorly correlated with the predictor variables, the Logistic Regression technique does not have a good classification of the observations.

Keywords: Logistic regression; simulation

Resumen

En esta investigación se propone el estudio de la técnica estadística multivariada de clasificación, Regresión Logística, donde se quiere evaluar el desempeño de la misma cuando es utilizada en datos simulados bajo distintos escenarios y bajo distintos tamaños de muestra.

Se generaron mediante simulación 500 archivos de datos para cada uno de los siguientes tamaños de muestra: 30, 75, 200, 400, 600, 1000. Cada conjunto contiene 6 columnas (variables) bajo distintas condiciones o escenarios. En cada muestra se "marcó" el 20% de las observaciones para ser utilizadas como grupo de test y el restante 80% para la estimación de los modelos evaluados en cada caso. Quedaron definidos un total de 12000 conjuntos de datos simulados, con 6 tamaños de muestra diferentes y 4 escenarios con las siguientes características definidos por la estructura de la matriz de correlaciones. El escenario 1 corresponde a datos provenientes de una población en la que los predictores están fuertemente correlacionados con la respuesta pero no entre ellos. El escenario 2 plantea una simulación a partir de una población con poca correlación de la respuesta con las variables predictoras pero éstas correlacionadas entre sí. En el escenario 3, la correlación presente en la población origen de la simulación es importante tanto entre las predictoras como entre éstas y la respuesta. Por último, el escenario 4 corresponde a una población original en la que no existe ningún tipo de correlación de magnitud importante entre las variables, ni de los predictores con la respuesta ni entre ellos.

De este análisis se concluye que, en condiciones donde las variables predictoras están altamente correlacionadas con la respuesta (escenarios 1 y 3), sin importar la correlación entre las predictoras, la técnica de Regresión Logística funciona satisfactoriamente. Sin

embargo, como se puede observar en el gráfico 1, cuando las predictoras están poco correlacionadas con la respuesta (escenarios 2 y 4) el porcentaje de clasificación correcta es bastante más bajo. Esta diferencia entre los dos grupos de escenarios en cuanto a la correlación de la respuesta con las predictoras se va acentuando a medida que el tamaño de muestra se hace más grande.

Como conclusión final se puede decir que, sin importar el tamaño de muestra, cuando la variable respuesta está poco correlacionada con las variables predictoras la técnica de Regresión Logística no tiene una buena clasificación de las observaciones.

Palabras clave: Regresión logística; simulación

1. Introducción

Una de las herramientas estadísticas más usadas en el campo de la epidemiología y de la salud pública para analizar el grado de asociación entre una exposición de interés y una enfermedad es la regresión logística, cuya medida de asociación es el odds ratio. Este tipo de regresión presenta grandes ventajas, pero también tiene algunos inconvenientes (Almendros Morón, 2018). La multicolinealidad y la falta de traslape o separación extrema en los grupos son problemas que afectan la inferencia basada en el modelo de regresión logística. Hay diversos estudios que se ocuparon de evaluar cómo los coeficientes estimados del modelo se ven afectados. En esta investigación se propone el estudio de la técnica estadística multivariada de clasificación, Regresión Logística, donde se quiere evaluar el desempeño de la misma cuando es utilizada en datos simulados bajo distintos escenarios y bajo distintos tamaños de muestra.

2. Metodología

2.1. Simulación de los datos

Se generaron mediante simulación 500 archivos de datos para cada uno de los siguientes tamaños de muestra: 30, 75, 200, 400, 600, 1000. Cada conjunto contiene 6 columnas (variables) bajo distintas condiciones o escenarios. En cada muestra se “marcó” el 20%

de las observaciones para ser utilizadas como grupo de test y el restante 80% para la estimación de los modelos evaluados en cada caso. La simulación se realizó a partir de distribuciones normales multivariadas con matriz de correlaciones según cuatro estructuras diferentes. Se consideró la primer columna (X1) como la variable respuesta y las restantes variables (X2 a X6) como las variables predictoras o explicativas. Luego de la generación de los ficheros se transformó la variable respuesta a dicotómica utilizando la mediana de la distribución.

Quedaron definidos un total de 12000 conjuntos de datos simulados, con 6 tamaños de muestra diferentes y 4 escenarios con las siguientes características definidos por la estructura de la matriz de correlaciones. El escenario 1 corresponde a datos provenientes de una población en la que los predictores están fuertemente correlacionados con la respuesta pero no entre ellos. El escenario 2 plantea una simulación a partir de una población con poca correlación de la respuesta con las variables predictoras pero éstas correlacionadas entre sí. En el escenario 3, la correlación presente en la población origen de la simulación es importante tanto entre las predictoras como entre éstas y la respuesta. Por último, el escenario 4 corresponde a una población original en la que no existe ningún tipo de correlación de magnitud importante entre las variables, ni de los predictores con la respuesta ni entre ellos. El proceso de simulación de ficheros de datos como las aplicaciones estadísticas subsiguientes se llevó a cabo en el software R versión 3.4.0.

2.2. Regresión Logística

Uno de los problemas de los que se ocupan las técnicas estadísticas multivariadas es la clasificación de objetos o unidades en grupos o poblaciones. Dos enfoquen agrupan a las técnicas de clasificación. Uno de ellos es cuando se conocen los grupos o categorías y se pretende ubicar los individuos dentro de estas categorías a partir de los valores de ciertas variables. El segundo enfoque, que no es el utilizado en este trabajo, ocurre cuando no se conocen los grupos de antemano y se pretende establecerlos a partir de los datos observados.

La técnica evaluada en este trabajo tienen por objetivo construir un sistema que permita clasificar unidades en una de las categorías definidas y conocidas previamente en función de las variables relevadas, como así también otras variables que demuestren un aporte significativo en la predicción del grupo de pertenencia.

Esta técnica de Regresión Logística es un caso particular de los modelos lineales generalizados, modela la probabilidad de que una unidad experimental pertenezca a un grupo en particular considerando información medida o registrada en dicha unidad.

La regresión logística es utilizada en situaciones en las cuales el objetivo es describir la relación entre una variable respuesta categórica, en este caso dicotómica, y un conjunto de variables explicativas que pueden ser tanto categóricas como cuantitativas.

Sea **x** un vector de p variables independientes, esto es, **x**′= (x1, x2,…, xp). La probabilidad condicional de que la variable y tome el valor 1 (presencia de la característica estudiada), dado valores de las covariables **x** es:

$$P(y=1/X)=\pi(x)=\frac{e^{g(x)}}{1+e^{g(x)}}$$

donde

$$g(x)=\beta_0+\beta_1x_1+...+\beta_px_p$$

β_0 es la constante del modelo o término independiente

p el número de covariables

β_i los coeficientes de las covariables

x_i las covariables que forman parte del modelo.

Si alguna de las variables independientes es una variable discreta con k niveles se debe incluir en el modelo como un conjunto de k-1 “variables de diseño” o “variables dummy”. El cociente de las probabilidades correspondientes a los dos niveles de la variable respuesta se denomina odds y su expresión es:

$$\frac{P(y=1|X)}{1-P(y=1|X)}=e^{\beta_0+\beta_1x_1+...+\beta_px_p}$$

Si se aplica el logaritmo natural, se obtiene el modelo de regresión logística:

$$\log\left(\frac{P(y=1|X)}{1-P(y=1|X)}\right)=\log(e^{\beta_0+\beta_1x_1+...+\beta_px_p})=\beta_0+\beta_1x_1+...+\beta_px_p$$

En la expresión anterior el primer término se denomina logit, esto es, el logaritmo de la razón de proporciones de los niveles de la variable respuesta.

Los coeficientes del modelo se estiman por el método de Máxima Verosimilitud y la significación estadística de cada uno de los coeficientes de regresión en el modelo se puede verificar utilizando, entre otros, el test de Wald y el test de razón de verosimilitudes.

Una cuestión importante en este tipo de análisis es determinar si todas las variables consideradas en la función de discriminante contienen información útil y si solamente algunas de ellas son suficientes para diferenciar los grupos. Dado que las variables utilizadas para explicar la respuesta es probable que estén correlacionadas, es posible también que compartan información. Por lo tanto, se puede buscar un subgrupo de variables mediante algún criterio de modo tal que las variables excluidas no contengan ninguna información adicional. Existen varios algoritmos de selección de variables, entre ellos podemos citar:

Método forward: comienza por seleccionar la variable más importante y continúa seleccionando las variables más importantes una por vez, usando un criterio bien definido. Uno de estos criterios involucra el logro de un nivel de significación deseado pre-establecido. El proceso termina cuando ninguna de las variables restantes encuentra el criterio pre-especificado.

Método backward: comienza con el modelo más grande posible. En cada paso descarta la variable menos importante, una por vez, usando un criterio similar a la selección forward. Continúa hasta que ninguna de las variables pueda ser descartada.

Selección paso a paso: combina los dos procedimientos anteriores. En un paso, una variable puede entrar o salir desde la lista de variables importantes, de acuerdo a algún criterio pre-establecido.

En este trabajo se utilizó como evaluación del ajuste del modelo el test de Hosmer-Lemeshow y, dado que el modelo es utilizado para clasificar unidades, se utilizó también la tasa de mala clasificación calculada sobre la muestra independiente excluida en la etapa de estimación.

3. Resultados

Como primera instancia se evaluó si se observaban diferencias significativas entre los escenarios, dentro del mismo tamaño de muestra, utilizando el test de Kruskal-Wallis. Se concluyó que, para todos los tamaños de muestra al menos un escenario difiere de los demás (p-valor < 0,01 en todos los test de Kruskal-Wallis). Mediante una técnica de comparaciones múltiples no paramétrica se evaluó cuáles de los escenarios diferían dentro de cada tamaño de muestra y se observó que los únicos escenarios donde no se encontraron diferencias significativas son los escenarios 2 y 4.

De este análisis se concluye que, en condiciones donde las variables predictoras están altamente correlacionadas con la respuesta (escenarios 1 y 3), sin importar la correlación entre las predictoras, la técnica de Regresión Logística funciona satisfactoriamente. Sin embargo, como se puede observar en el gráfico 1, cuando las predictoras están poco correlacionadas con la respuesta (escenarios 2 y 4) el porcentaje de clasificación correcta es bastante más bajo. Esta diferencia entre los dos grupos de escenarios en cuanto a la correlación de la respuesta con las predictoras se va acentuando a medida que el tamaño de muestra se hace más grande.

Respecto a la convergencia de los procedimientos iterativos de estimación de los modelos, en los escenarios 1 y 3, debido a la estructura de correlaciones que los generan, existe cierta separación de grupos evidenciada en las simulaciones de tamaños de muestras bajos. Esto ocasiona que no exista convergencia en el proceso de estimación del modelo haciendo imposible la interpretación de los coeficientes pero no afectaría en su tarea de predicción (Tabla 1).

Gráfico 1: Mediana del porcentaje de clasificación correcta según escenario y tamaño de muestra

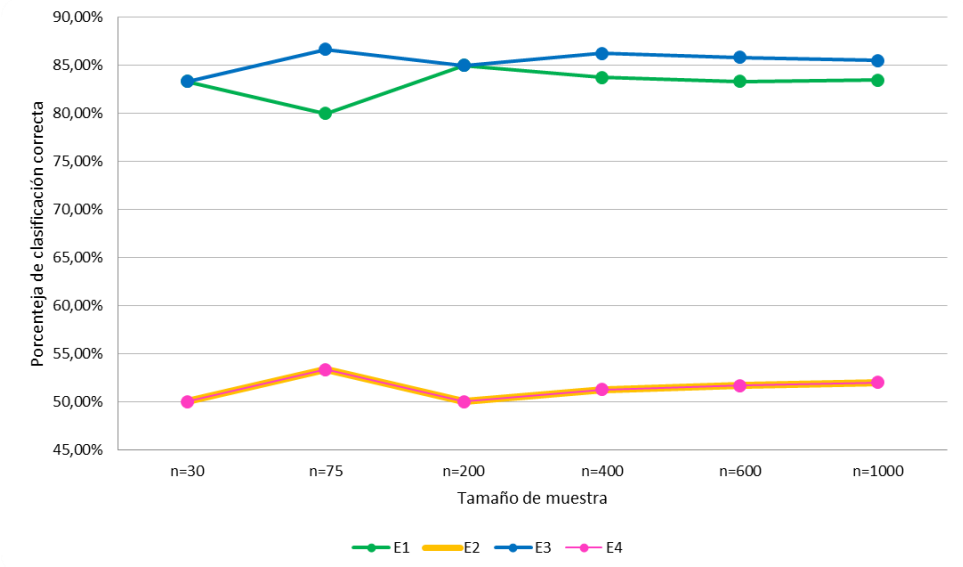

Tabla 1: Porcentaje de convergencia según escenario y tamaño de muestra

n	E1	E2	E3	E4
30	56%	99%	47%	99%
75	100%	100%	98%	100%
20	100%	100%	100%	100%
400	100%	100%	100%	100%
600	100%	100%	100%	100%
1000	100%	100%	100%	100%

4. Discusión

En este trabajo se ha evaluado el desempeño de la técnica de Regresión Logística en datos simulados bajo distintas condiciones que diferían en la estructura de correlaciones entre las variables y el tamaño de muestra.

En las situaciones en las que las variables predictoras están altamente correlacionadas con la respuesta, la técnica de Regresión Logística funciona satisfactoriamente, independientemente de la existencia o no de multicolinealidad. Sin embargo, cuando las predictoras están poco correlacionadas con la respuesta el desempeño de la técnica es inferior. Esta diferencia entre los escenarios en cuanto a la correlación de la respuesta con las predictoras se va acentuando a medida que el tamaño de muestra se hace más grande.

Respecto a la convergencia de los procedimientos iterativos de estimación de los modelos, cuando existe cierta separación de grupos y tamaños de muestras bajos, se imposibilita la convergencia en el proceso de estimación del modelo haciendo absurda la interpretación de los coeficientes.

Como conclusión final se puede decir que, sin importar el tamaño de muestra, cuando la variable respuesta está poco correlacionada con las variables predictoras la técnica de Regresión Logística no tiene una buena clasificación de las observaciones.

5. Bibliografía

Almendros Morón, O. (2018). El problema de la separación en modelos de regresión logística. Estadística e Investigación Operativa. Universitat Politècnica de Catalunya y Universitat de Barcelona (UPC-UB)

Hair, J.F., Anderson, R.L.,Tatham, R.L., Black, W.C. 1999. Análisis Multivariante. Prentice Hall Iberia, Madrid, España.

Hastie, T., Tibshirani, R., Friedman, J. 2009.The Elements of Statistical Learning.Data Mining, Inference, and Predicion.Springer Series in Statistics.

Pérez López, C. 2004. Técnicas de Análisis Multivariante de Datos. PEARSON EDUCACIÓN, S.A., Madrid, España.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.