# La regresión logística: Una alternativa al análisis discriminante ante la ausencia de normalidad multivariante

*Gabriel Cavada Chacón**
*Luis Valverde Fallas***

**Resumen**

*Tradicionalmente, la técnica para discriminar entre dos o más poblaciones a través de una colección de descriptores continuos es el Análisis Discriminante de Fisher, sin embargo cuando se relajan los supuestos en que descansa dicho análisis, persiste el uso de aquel. El objetivo de este trabajo es mostrar a través de algunos ejemplos, cómo la técnica de discriminación basada en la Regresión Logística discrimina mejor que el Análisis Discriminante de Fisher, mejorando ostensiblemente las proporciones de buena clasificación.*

*Palabras claves: Análisis Discriminante, Regresión Logística, Normalidad Multivariante.*



* Ph.D. División de Bioestadística, Escuela de Salud Pública, Universidad de Chile. Académico, Escuela de Ingeniería, UCINF.
** Mg. Sc. Escuela de Matemáticas, Universidad de Costa Rica.

## Introducción

Como es sabido, la técnica para discriminar entre dos o más poblaciones, a través de una colección de descriptores continuos, es el Análisis Discriminante de Fisher. El método propuesto exige el supuesto de Normalidad Multivariante conjunta para el juego de variables predictoras y, si además se desea una función lineal discriminante, se debe agregar el supuesto de igualdad de matrices de covarianza para cada una de las poblaciones en estudio. Sin embargo, cuando se relaja el supuesto de Normalidad Multivariante, la tendencia es inclinarse nuevamente al Análisis Discriminante de Fisher, considerando la robustez del mismo.

El objetivo de este trabajo es mostrar, por medio de algunos ejemplos, que la Regresión Logística no sólo es una alternativa a tomar en cuenta, sino que, incluso, podría mejorar la capacidad de discriminación cuando no se cumplen los supuestos de Fisher. Cuando el juego de variables descriptoras sigue una distribución normal multivariante, los coeficientes de la Regresión Logística pueden ser estimados a través de la función lineal discriminante, pues se tiene la relación:

$$\beta_0 = \ln\left(\frac{p}{1-p}\right) - \frac{1}{2}\beta(\mu_1 + \mu_0)$$

donde p es la probabilidad de pertenecer a la población 1, 1-p la probabilidad de pertenecer a la población 0 y $\beta = (\mu_1 - \mu_0)'\Sigma^{-1}$ que es el vector de coeficientes de la función lineal discriminante (Hosmer y Lemeshow, 1989). Es decir, cuando los supuestos de Fisher se cumplen, discriminar a través de la Regresión Logística y de la Función Lineal Discriminante es equivalente.

Si no se cumplen los supuestos de Fisher, no es posible encontrar relaciones entre ambos procedimientos y, como se planteó al inicio, la tendencia es usar la Función Lineal Discriminante, debido a la robustez de dicho método. Sin embargo, en las condiciones expuestas, la experiencia con algunas bases de datos es que la Regresión Logística entrega mayores porcentajes de buena clasificación.

## Resultados

Como primer ejemplo ilustrativo se presenta la propuesta metodológica aplicada a los clásicos datos de Fisher, referentes a las flores de Iris (1936) (SAS, 1997):

- Estadísticas descriptivas de las variables descriptoras desagregadas por especie:

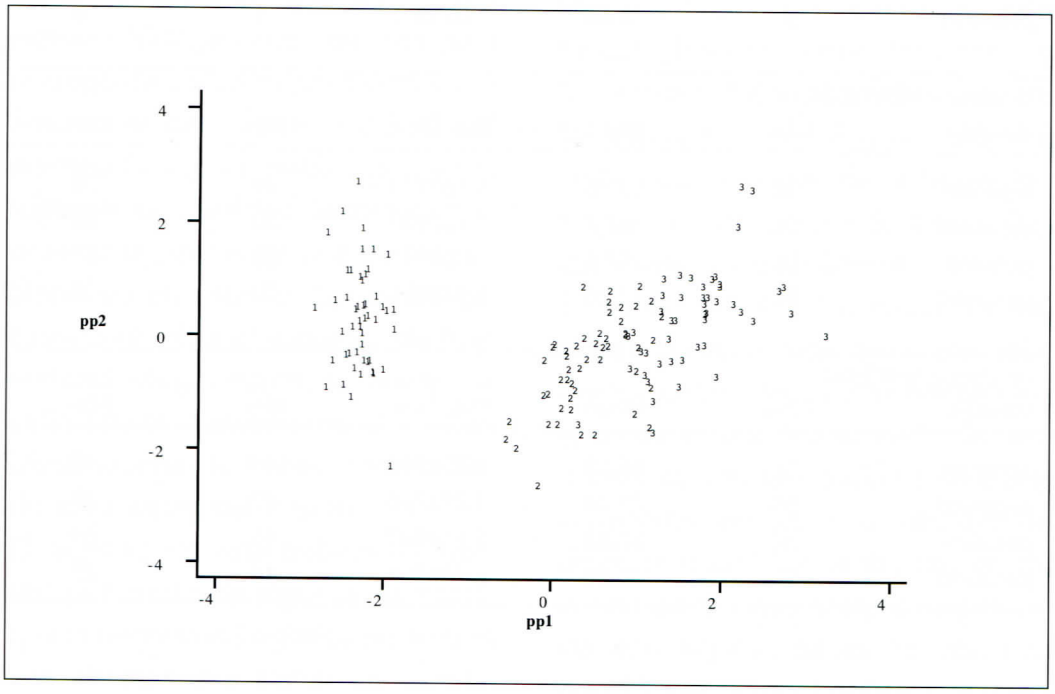
| -> clase= SETOSA | | | | | |
|---|---|---|---|---|---|
| Variable | Obs | Mean | Std. Dev. | Min | Max |
| sepallen | 50 | 50.06 | 3.524897 | 43 | 58 |
| sepalwid | 50 | 34.28 | 3.790644 | 23 | 44 |
| petallen | 50 | 14.62 | 1.73664 | 10 | 19 |
| petalwid | 50 | 2.46 | 1.053856 | 1 | 6 |
| -> clase= VERSICOL | | | | | |
| Variable | Obs | Mean | Std. Dev. | Min | Max |
| sepallen | 50 | 59.36 | 5.161711 | 49 | 70 |
| sepalwid | 50 | 27.7 | 3.137983 | 20 | 34 |
| petallen | 50 | 42.6 | 4.69911 | 30 | 51 |
| petalwid | 50 | 13.26 | 1.977527 | 10 | 18 |
| -> clase= VIRGINIC | | | | | |
| Variable | Obs | Mean | Std. Dev. | Min | Max |
| sepallen | 50 | 65.88 | 6.358796 | 49 | 79 |
| sepalwid | 50 | 29.74 | 3.224966 | 22 | 38 |
| petallen | 50 | 55.52 | 5.518947 | 45 | 69 |
| petalwid | 50 | 20.26 | 2.746501 | 14 | 25 |



- Dócima de Shapiro-Wilk para probar la normalidad de datos:

| Shapiro-Wilk W test for normal data | | | | | |
|---|---|---|---|---|---|
| Variable | Obs | W | V | z | Pr > z |
| sepallen | 150 | 0.97940 | 2.397 | 1.981 | **0.02377** |
| sepalwid | 150 | 0.99098 | 1.049 | 0.108 | 0.45682 |
| petallen | 150 | 0.88019 | 13.940 | 5.973 | **0.00000** |
| petalwid | 150 | 0.92197 | 9.079 | 5.001 | **0.00000** |

los p-value en negrita muestran el no cumplimiento del supuesto de normalidad multivariante de los datos. Usando análisis de componentes principales, graficando los dos primeros puntajes principales, se expone el traslape de las especies Versicolor y Virgínica, las que posteriormente serán discriminadas por el método de la Función Lineal Discriminante y de Regresión Logística para comparar su capacidad de discriminación:




Donde 1: Setosa, 2: Versicolor y 3: Virgínica

Para proceder a comparar los métodos de discriminación, se recodificaron las variables para las especies Versicolor y Virgínica con 0 y 1, respectivamente, dejando fuera del análisis la especie Setosa.

En estas condiciones, el Análisis Discriminante de Fisher entrega la siguiente tabla de clasificación:

| Clasif. | clase Versicolor | Virgínica | Total |
|---|---|---|---|
| Mal | 4<br>8.00% | 3<br>6.00% | 7<br>7.00% |
| Bien | 46<br>92.00% | 47<br>94.00% | 93<br>93.00% |
| Total | 50<br>100.00% | 50<br>100.00% | 100<br>100.00% |

Es decir, para la clase Versicolor se clasifica correctamente el 92% de las flores, para la clase Virgínica se clasifica correctamente el 94% de las flores, generándose un total de buena clasificación de 93%.

Al discriminar con la Regresión Logística se obtiene la siguiente tabla de clasificación:

**LOGISTIC MODEL FOR ESPECIE**



| | True | | |
|---|---|---|---|
| Classified | D | ~D | Total |
| + | 49 | 1 | 50 |
| - | 1 | 49 | 50 |
| Total | 50 | 50 | 100 |
| Classified + if predicted Pr(D) >= .5<br>True D defined as espe ~= 0 | | | |
| Sensitivity | | Pr ( + D) | 98.00% |
| Specificity | | Pr ( - ~D) | 98.00% |
| Positive predictive value | | Pr ( D +) | 98.00% |
| Negative predictive value | | Pr (~D -) | 98.00% |
| False + rate for true ~D | | Pr ( + ~D) | 2.00% |
| False - rate for true D | | Pr ( - D) | 2.00% |
| False + rate for classified + | | Pr (~D +) | 2.00% |
| False - rate for classified - | | Pr ( D -) | 2.00% |
| Correctly classified | | | 98.00% |

Se observa que para la clase Versicolor se clasifica correctamente el 98% de las flores, para la clase Virgínica también se clasifica correctamente el 98% de las flores, resultanto un total de buena clasificación de 98%. En el ejemplo analizado se advierte que la Regresión Logística aumenta la buena clasificación de 93% a un 98%.

Un segundo ejemplo discrimina, basándose en seis variables descriptoras, entre dos grupos: veinte aprendices de ingeniero y veinte pilotos (Rencher, 1995). La variable grupo está codificada como 0: aprendiz de ingeniero y 1: piloto. Las seis variables descriptoras son: $y_1$: intelligent, $y_2$: form relation, $y_3$: dinamomether, $y_4$: dotting, $y_5$: censory motor cordination y $y_6$: perseveration. En este contexto se tiene que:

- Estadísticas descriptivas de las variables descriptoras desagregadas por grupo:

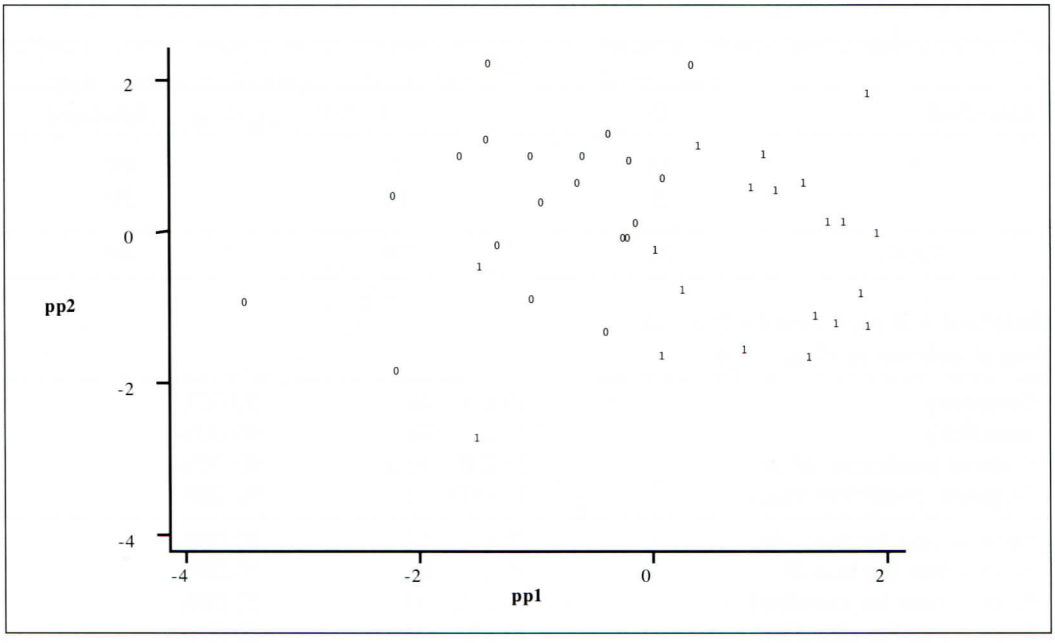
| -> clase= 0 | | | | | |
|---|---|---|---|---|---|
| Variable | Obs | Mean | Std. Dev. | Min | Max |
| y1 | 20 | 124.5 | 19.60263 | 77 | 154 |
| y2 | 20 | 38.1 | 8.258329 | 21 | 55 |
| y3 | 20 | 76.2 | 11.00526 | 52 | 91 |
| y4 | 20 | 192.75 | 31.6259 | 152 | 266 |
| y5 | 20 | 53.65 | 17.95689 | 29 | 88 |
| y6 | 20 | 250.3 | 21.68458 | 209 | 300 |
| -> clase= 1 | | | | | |
| Variable | Obs | Mean | Std. Dev. | Min | Max |
| y1 | 20 | 129.3 | 26.22594 | 47 | 164 |
| y2 | 20 | 31.7 | 7.189905 | 17 | 47 |
| y3 | 20 | 87.4 | 10.53016 | 67 | 105 |
| y4 | 20 | 236.6 | 28.15259 | 183 | 291 |
| y5 | 20 | 44.25 | 13.17843 | 27 | 66 |
| y6 | 20 | 280.2 | 35.4083 | 217 | 324 |

- Dócima de Shapiro-Wilk para probar la normalidad de datos:

| Shapiro-Wilk W test for normal data | | | | | |
|---|---|---|---|---|---|
| Variable | Obs | W | V | z | Pr > z |
| y1 | 40 | 0.92914 | 2.801 | 2.168 | **0.01510** |
| y2 | 40 | 0.98409 | 0.629 | -0.977 | 0.83560 |
| y3 | 40 | 0.98188 | 0.716 | -0.702 | 0.75861 |
| y4 | 40 | 0.96838 | 1.250 | 0.469 | 0.31950 |
| y5 | 40 | 0.93884 | 2.418 | 1.858 | **0.03161** |
| y6 | 40 | 0.93684 | 2.497 | 1.926 | **0.02708** |

Nuevamente, los p-value en negrita descartan la Normalidad Multivariante de los datos. A partir de un análisis de componentes principales, se genera un gráfico con los dos primeros puntajes principales, donde se muestra el traslape de los grupos en estudio. Procedemos entonces a discriminar por el método de la Función Lineal Discriminante y el de Regresión Logística:






Donde 0: aprendiz de ingeniero y 1: piloto

Al realizar Análisis Discriminante, se encuentra la siguiente tabla de clasificación:

| Clasif. | clase 0 | 1 | Total |
|---|---|---|---|
| Mal | 3<br>15.00 | 3<br>15.00 | 6<br>15.00 |
| Bien | 17<br>85.00 | 17<br>85.00 | 34<br>85.00 |
| Total | 20<br>100.00 | 20<br>100.00 | 40<br>100.00 |

El Análisis Discriminante de Fisher proporciona en este caso un 85% de buena clasificación en ambos grupos y también un 85% de buena clasificación en el total de los casos.

Por otra parte, la Regresión Logística entrega la siguiente tabla de clasificación:

**LOGISTIC MODEL FOR CLASE**

| | True | | |
|---|---|---|---|
| Classified | D | ~D | Total |
| +<br>- | 18<br>2 | 2<br>18 | 20<br>20 |
| Total | 20 | 20 | 40 |
| Classified + if predicted Pr(D) >= .5<br>True D defined as clase ~= 0 | | | |
| Sensitivity<br>Specificity<br>Positive predictive value<br>Negative predictive value | | Pr ( + D)<br>Pr ( - ~D)<br>Pr ( D +)<br>Pr (~D -) | 90.00%<br>90.00%<br>90.00%<br>90.00% |
| False + rate for true ~D<br>False - rate for true D<br>False + rate for classified +<br>False - rate for classified - | | Pr ( + ~D)<br>Pr ( - D)<br>Pr (~D +)<br>Pr ( D -) | 10.00%<br>10.00%<br>10.00%<br>10.00% |
| Correctly classified | | | 90.00% |

El resultado anterior muestra que la buena clasificación mejora nuevamente con la Regresión Logística, en un 5% para cada grupo y en igual porcentaje para el total de los datos.

## Conclusión

A través de los ejemplos analizados se ha dado cuenta de que la técnica de discriminación a través de la Regresión Logística, cuando el supuesto de Normalidad Multivariante no se satisface, puede mejorar significativamente la capacidad de discriminación de los datos; aunque, si bien lo expuesto no es generalizable, nuestro planteamiento es que la Regresión Logística debería considerarse como una alternativa para mejorar la discriminación ante la carencia de algunos de los supuestos de Fisher.

## Bibliografía


Fisher, R.A. "The Use of Multiple Measurement in Taxonomic Problems". *Annals of Eugenics* 7 (1936): 179-84.

Hosmer, David and Stanley Lemeshow. *Applied Logistic Regression*. New York: John Wiley & Sons, 1989.

*SAS User's Guide*. Chapell Hill, North Carolina: SAS Institute Inc., 1997.

Rencher, Alvin. *Method of Multivariate Analysis*. New York: John Wiley & Sons,1995.