Modelación estadística: La regresión logística (Parte 2)

Gabriel Cavada Ch.[1,2]
[1]*División de Bioestadística, Escuela de Salud Pública, Universidad de Chile.*
[2]*Facultad de Medicina, Universidad de los Andes.*

Statistical modeling: Logistic regression (Part II)

La regresión logística como discriminadora (diagnóstico y pronóstico)

Desde el punto de vista estadístico, tanto el análisis diagnóstico como el pronóstico, se inscriben en el llamado análisis discriminante, esto es cuando dadas dos muestras pertenecientes a poblaciones distintas y conocida esta pertenencia, determinar el conjunto de variables "descriptoras" (perfil del sujeto) que tiene capacidad de identificar cada una de las poblaciones a las que se hace referencia. Si la pertenencia a cada una de las poblaciones en cuestión la denotamos por los códigos "0 y 1" contenidos en la variable "Y", para el perfil Xβ, podemos escribir:

$$P(Y = 1 \mid X\beta) = e^{X\beta}/(1 + e^{X\beta})$$

Así, dado un perfil X, la idea es determinar si la estimación de probabilidad hecha por la distribución logística está cerca del 0 o cerca del 1. Esta decisión es posible tomarla si se escoge un punto de corte, "p" para la probabilidad estimada, de modo que si $P(Y = 1 \mid X\beta) > p$, el sujeto será clasificado en la población "1", de lo contrario el sujeto será clasificado en la población "0". Con esta conceptualización, si la población de interés (enfermos, muertes, mejorías…) la llamamos "A" y la codificamos con "1", definimos:

S=P(Y=1 | X ϵ A) : Sensibilidad de la discriminación.
E=P(Y=0 | X ϵ A'): Especificidad de la discriminación.

Obviamente que las probabilidades complementarias definen los sucesos "Falso Negativo" y "Falso Positivo", respectivamente, es decir:

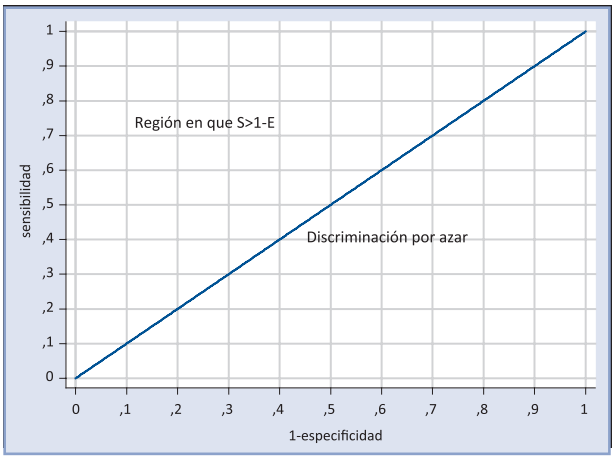
P(Y=0 | X ϵ A) = 1-S: Falso Negativo. (1-S)
P(Y=1 | X ϵ A') =1-E: Falso Positivo. (1-E)

Como nuestro interés es tener buena capacidad de clasificación, es decir, alta sensibilidad y alta especificidad. Lo que se traduce en:

$$P(Y=1 \mid X \in A) > P(Y=1 \mid X \in A')$$

$$S > 1-E$$

Si la discriminación fuera por azar se tendría S=1-E. Representando gráficamente estas relaciones, donde el eje de las abscisas es 1-E y el de las ordenadas S, se tiene:



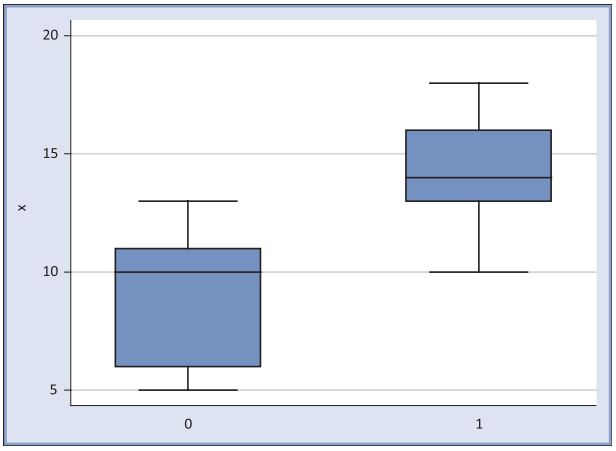
Notar que si la discriminación fuese perfecta, es decir, 100% de sensibilidad y 100% de especificidad, el punto de corte para la discriminación estaría en la intersección de las líneas azules y el área bajo la curva azul sería 1. Obviamente en cualquier aplicación real en que haya buena discriminación, esta área sería menor que 1 pero mayor a 0,5.

Observemos el siguiente ejemplo: Supongamos que la variable X discrimina dos poblaciones, disponemos de 10 valores para la población "1" y 10 valores para la población "0", los datos son los siguientes:

id	y	x
1	0	6
2	0	11
3	0	10
4	0	10
5	0	11
6	0	6
7	0	10

Rincón de la Bioestadística

8	0	5
9	0	10
10	0	13
11	1	13
12	1	11
13	1	14
14	1	16
15	1	18
16	1	14
17	1	14
18	1	18
19	1	10
20	1	16



Al ajustar el modelo de regresión logística se obtiene la siguiente salida STATA:

y	Coef.	Std. Err.	z	P >|z|	[95% Conf. Interval]	
x	,9371287	,4217506	2,22	0,026	,1105127	1,763745
_cons	-10,99978	4,905968	-2,24 0,025	-20.6153	-1,384262	

O en términos de la función de probabilidades logística:

$$P(y = 1|X) = \frac{e^{-10,99978 + 0,9371287 \cdot X}}{1 + e^{-10,99978 + 0,9371287 \cdot X}}$$

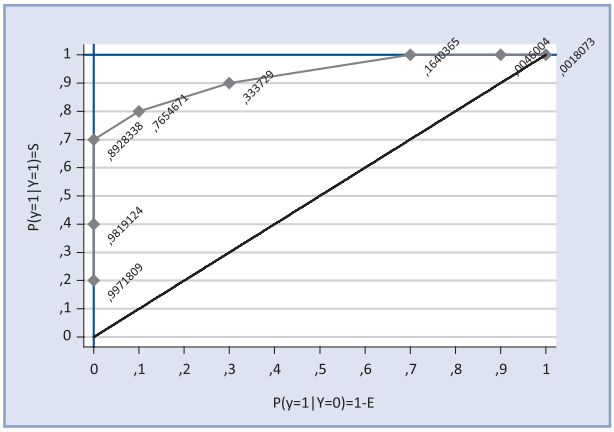
fórmula que al ser evaluada en cada observación entrega las siguientes probabilidades:

id	y	x	P(y=1|x)
1	0	6	0,004600
2	0	11	0,333729
3	0	10	0,164037
4	0	10	0,164037
5	0	11	0,333729
6	0	6	0,004600
7	0	10	0,164037
8	0	5	0,001807
9	0	10	0,164037
10	0	13	0,765467
11	1	13	0,765467
12	1	11	0,333729
13	1	14	0,892834
14	1	16	0,981912
15	1	18	0,997181
16	1	14	0,892834
17	1	14	0,892834
18	1	18	0,997181
19	1	10	0,164037
20	1	16	0,981912

La información anterior la podemos resumir en la siguiente tabla:

Pto. corte	n de sujetos (y=1|Y=0)	n de sujetos (y=1|Y=1)	P(y=1|Y=0)=1-E	P(y=1|Y=1)=S
0,0018073	10	10	1,000	1,000
0,0046004	9	10	0,900	1,000
0,1640365	7	10	0,700	1,000
0,333729	3	9	0,300	0,900
0,7654671	1	8	0,100	0,800
0,8928338	0	7	0,000	0,700
0,9819124	0	4	0,000	0,400
0,9971809	0	2	0,000	0,200

Al graficar la sensibilidad *versus* 1-especificidad para los distintos puntos de corte se obtiene la curva ROC (*Receiver Operating Characteristic)*.

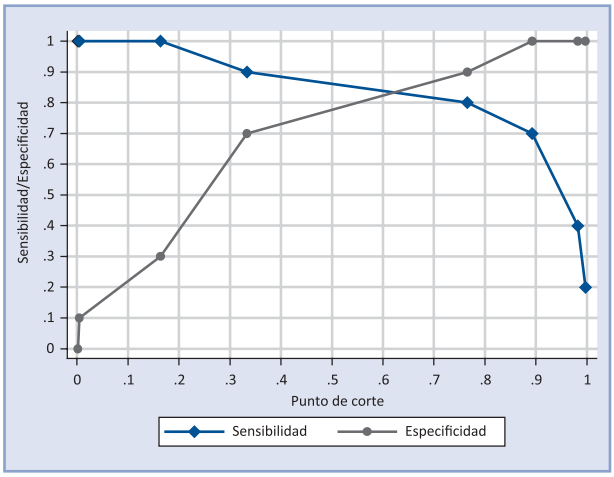


Como se sugirió, la capacidad de discriminación está dada por el área bajo la curva ROC, que en nuestro caso es 0,9250. El mejor punto de corte es aquel que está más cerca de la discriminación perfecta.

Según Hosmer y Lemeshow ("Applied Logistic Regression" Second Edition, p. 162), la capacidad de discriminación puede clasificarse según:

Área bajo la curva ROC	Discriminación
0,5	por azar
0,7 a 0,8	aceptable
0,8 a 0,9	muy buena
0,9 a 1	excelente

Sin embargo, para encontrar el mejor punto de corte, es preferible usar el siguiente gráfico:



Recordando la definición de los Likelihood Ratios:

$$LR+ = \frac{P(y=1|Y=1)}{P(y=1|Y=0)} = \frac{sensibilidad}{1-especificidad}$$

$$LR- = \frac{P(y=0|Y=1)}{P(y=0|Y=0)} = \frac{1-sensibilidad}{especificidad}$$

Podemos leer a cabalidad la siguiente salida de STATA:

Detailed report of Sensitivity and Specificity					
			Correctly		
Cutpoint	Sensitivity	Specificity	Classified	LR+	LR-
(> = ,0018073)	100,00%	0,00%	50,00%	1,0000	
(> = ,0046004)	100,00%	10,00%	55,00%	1,1111	0,0000
(> = ,1640365)	100,00%	30,00%	65,00%	1,4286	0,0000
(> = ,333729)	90,00%	70,00%	80,00%	3,0000	0,1429
(> = ,7654671)	80,00%	90,00%	85,00%	8,0000	0,2222
(> = ,8928338)	70,00%	100,00%	85,00%	0,3000	
(> = ,9819124)	40,00%	100,00%	70,00%	0,6000	
(> = ,9971809)	20,00%	100,00%	60,00%	0,8000	
(> ,9971809)	0,00%	100,00%	50,00%	1,0000	

	ROC		-Asymptotic Normal-	
Obs	Area	Std. Err.	[95% Conf.	Interval]
20	0,9250	0,0575	0,81231	1,00000

Si se ha escogido como punto de corte 0.7654671 el punto de corte para la variable original, X, se obtiene despejando su valor de la ecuación:

$$\ln = \left(\frac{p}{(1-p)} \right) = a + b \bullet X$$

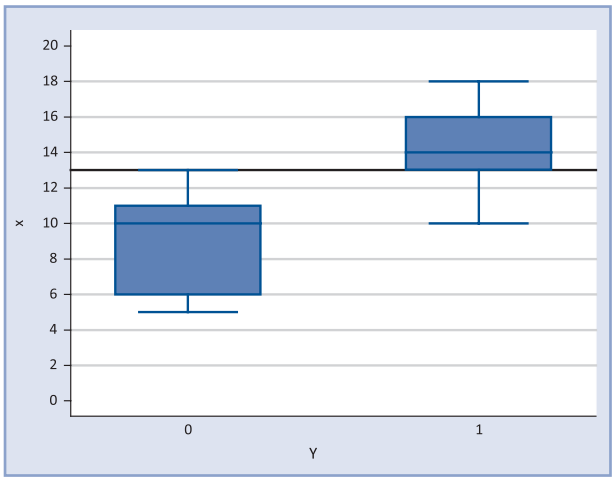
Rincón de la Bioestadística

De donde:

$$X = \frac{\ln = \left(\frac{p}{(1- p)} \right) - a}{b}$$

En nuestro caso:

$$X = \frac{\ln = \left(\frac{,7654671}{1- ,7654671} \right) - 10,99978}{,9371287} 12,999997 = 13$$



Obviamente, que conocido un desenlace, se puede utilizar una metodología idéntica, para evaluar un pronóstico.