Modelado estadístico de datos: Práctica 2

Emilio Letón y Elisa M. Molanes-López

1. (CALC) (2 puntos) Se ha realizado un estudio para ver si la utilización del biomarcador "acid phosphatase" en sangre (x100) (exp) influye a la hora de detectar la presencia de nódulos infectados (rta=1: sí, rta=0: no). Para ello se han tomado 20 individuos con nódulos infectados y 33 sin nódulos infectados. Los datos experimentales se dan en Le (2006) y se han reproducido en la tabla 1. Además dichos datos se pueden encontrar en el fichero le.txt alojado en el curso virtual. Se pide rellenar la tabla 2 y calcular a partir de dicha tabla el *auc*. Adicionalmente, utilizando R, dibujar la curva *roc*.

2. (2 puntos) Se pide analizar los datos de la tabla 1 con regresión logística evaluando su comportamiento a través de la curva *roc*. ¿Cuál es la relación con el ejercicio anterior?

3. (CALC) (2 puntos) Se pide calcular el *or* y su IC para los datos del fichero d_d_3.txt, que se encuentran resumidos en la tabla 3.

   Adicionalmente, utilizando R, análicese dichos datos con regresión logística ¿Cuál es la relación con el cálculo del *or* y de su IC? Compruébese que

   $$exp(\widehat{\beta_0}) = lrn\frac{prev.\ muestral}{1 - prev.\ muestral}$$

   .

4. (CALC) (1 punto) Se supone que los datos experimentales están dados por la tabla 4. En esta situación, se pide demostrar que la función de verosmilitud en el modelo de regresión logística con una única variable explicativa dicotómica es

   $$\ell(\beta_0, \beta_1) = \Big(\frac{1}{1+e^{-(\beta_0+\beta_1)}}\Big)^a \Big(\frac{1}{1+e^{-\beta_0}}\Big)^b \Big(1 - \frac{1}{1+e^{-(\beta_0+\beta_1)}}\Big)^c \Big(1 - \frac{1}{1+e^{-\beta_0}}\Big)^d.$$

5. (CALC) (1 punto) Se pide demostrar que el modelo de regresión logística es lineal en el logit.

6. (2 puntos) Se ha realizado un estudio para ver si el hecho de fumar (fumar=1: Sí, fumar=0: No), tomar café (cafe=1: Sí, cafe=0: No) o tomar un fármaco (trat=1: Sí, trat=0: No) influye en la presencia de una enfermedad (enf=1: Sí, enf=0: No). Han participado en el estudio 90 individuos. Los datos experimentales están en el fichero d_ddd.txt alojado en el curso virtual.

   Se pide:

   - En un análisis de regresión logística bivariante, variable a variable explicativa, ¿qué ocurre?
   - En un análisis de regresión logística multivariante con todas las variables, ¿qué ocurre?
   - Aplicar al análisis de regresión logística multivariante un proceso de selección automático, ¿qué ocurre?
   - ¿Hay "confusión"? ¿Hay "interacción"?
   - ¿Cuál es la curva *roc* del modelo final? ¿Cuál es su *auc*? ¿Cuál es el punto de corte óptimo para utilizar el modelo?
   - ¿Se puede aplicar el análisis discriminante a estos datos? ¿Qué técnica es mejor en este caso?

   En el documento que se entregue habrá que incluir el código utilizado.

rta	exp
1	48
1	49
1	51
1	56
1	67
1	67
1	67
1	70
1	70
1	72
1	76
1	78
1	81
1	82
1	82
1	84
1	89
1	99
1	126
1	136
0	40
0	40
0	46
0	47
0	48
0	48
0	49
0	49
0	50
0	50
0	50
0	50
0	50
0	52
0	52
0	55
0	55
0	56
0	59
0	62
0	62
0	63
0	65
0	66
0	71
0	75
0	76
0	78
0	83
0	95
0	98
0	102
0	187

Tabla 1: Datos observados del esquema rta ← exp

c	Sano	Enf	*ffp(c)*	*fvp(c)*	$\widehat{auc(c)}$
39.0		20	0	1.000	0.061
43.0		20	0.061	1.000	0.030
46.5		20	0.091	1.000	0.030
47.5		20	0.121	1.000	0.059
48.5		19	0.182	0.950	0.056
49.5		18	0.242	0.900	0.136
50.5		18	0.394	0.900	
51.5		17	0.394	0.850	
53.5		17	0.455	0.850	
55.5		17	0.515	0.850	
57.5	18		0.545	0.800	
60.5	19		0.576	0.800	
62.5	21		0.636	0.800	
64.0	22		0.667	0.800	
65.5	23		0.697	0.800	
66.5	24		0.727	0.800	
68.5	24		0.727	0.650	0.000
70.5	24		0.727	0.550	0.017
71.5	25		0.758	0.550	0.000
73.5	25		0.758	0.500	0.015
75.5	26	10	0.788	0.500	0.014
77.0	27	9		0.450	0.013
79.5	28	8		0.400	0.000
81.5	28	7		0.350	0.000
82.5	28	5		0.250	0.008
83.5	29	5		0.250	0.000
86.5	29	4		0.200	0.000
92.0	29	3		0.150	0.005
96.5	30	3		0.150	0.005
98.5	31	3		0.150	0.000
100.5	31	2		0.100	0.003
114.0	32	2	0.970		0.000
131.0	32	1	0.970		0.000
161.5	32	0	0.970		0.000
188.0	33	0	1.000		0.000

Tabla 2: Datos parciales para roc del esquema rta ← exp

	$X=1$	$X=0$	
$Y=1$	21	16	37
$Y=0$	8	31	39
	29	47	76

Tabla 3: Tabla con datos observados del esquema $D \leftarrow D$

	$X=1$	$X=0$	
$Y=1$	a	b	r_1
$Y=0$	c	d	r_0
	s_1	s_0	n

Tabla 4: Tabla con datos genéricos del esquema $D \leftarrow D$