ESTADÍSTICA II
Grado en Matemáticas (2022/23)

Tema 4: CLASIFICACIÓN Y REGRESIÓN LOGÍSTICA

4.1. Considera dos conjuntos de datos bidimensionales correspondientes a dos poblaciones P_0 y P_1:

$$\mathbb{X}_0 = \begin{pmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{pmatrix}, \quad \mathbb{X}_1 = \begin{pmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{pmatrix}.$$

a) Estima, a partir de estos datos, la función lineal discriminante de Fisher.
b) Clasifica la observación $\mathbf{x} = (2, 7)'$ utilizando la regla obtenida en el apartado anterior.

4.2. Considera los datos del fichero `infartos.RData`, sobre enfermedades coronarias en Sudáfrica. Se obtuvieron en una encuesta llevada a cabo en el marco del Coronary Risk-Factor Study (CORIS) en tres zonas rurales de Western Cape (Sudáfrica).[1] El objetivo del estudio era establecer la intensidad de factores de riesgo para enfermedades coronarias en esa región de alta incidencia.
Los individuos eran varones de raza blanca con edades entre 15 y 64 años. La etiqueta `clase` consignaba la presencia (`clase`=1) o ausencia (`clase`=0) de infarto de miocardio en el momento de la encuesta. Las características medidas fueron:

Nombre variable	Descripción
`sbp`	Tensión sanguínea sistólica
`tobacco`	Consumo de tabaco
`ldl`	Colesterol
`adiposity`	Medida de adiposidad
`typea`	Comportamiento "tipo A"
`obesity`	Medida de la obesidad
`alcohol`	Consumo de alcohol
`age`	Edad

Calcula la función lineal discriminante de Fisher para clasificar entre sano (`clase=0`) o enfermo (`clase=1`) a un individuo en función de las 8 variables regresoras contenidas el fichero. Compara los coeficientes de las variables con los correspondientes a la regla de clasificación basada en regresión logística. ¿Son muy diferentes?

4.3. Se toma una muestra (con tamaño $n_i = 10$, para $i = 0, 1$) de un vector bidimensional $\mathbf{X}$ en cada una de dos poblaciones, P_0 y P_1. Los vectores de medias y matrices de covarianzas muestrales son

$$\bar{\mathbf{x}}_0 = \begin{pmatrix} -0.11 \\ 0.17 \end{pmatrix} \quad \bar{\mathbf{x}}_1 = \begin{pmatrix} 1.19 \\ 0.80 \end{pmatrix} \quad \mathbf{S}_0 = \begin{pmatrix} 1.11 & 0.16 \\ 0.16 & 0.38 \end{pmatrix} \quad \mathbf{S}_1 = \begin{pmatrix} 0.49 & 0.15 \\ 0.15 & 0.43 \end{pmatrix}$$

Clasifica la observación $\mathbf{x} = (1, 1)'$ utilizando la regla discriminante lineal de Fisher.

[1] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Jooste, P., Ferreira, J. (1983). Coronary risk factor screening in three rural communities, *South African Medical Journal* 64, 430–436.

4.4. Para 100 lirios, 50 de ellos correspondientes a la especie *Versicolor* (`Y = 1`) y otros 50 correspondientes a la especie *Virginica* (`Y = 0`) se ha medido la longitud (`Long`) y la anchura (`Anch`) del pétalo en milímetros. Con los datos resultantes se ha ajustado un modelo de regresión logística con el objetivo de clasificar en alguna de las dos especies un lirio cuya especie se desconoce a partir de las medidas de su pétalo. A continuación se muestra un resumen de los resultados (algunos valores han sido suprimidos o sustituidos por letras):

```
Call:
glm(formula = y ~ Long + Anch, family = binomial)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.8965923 -0.0227388  0.0001139  0.0474898  1.7375172

Coefficients:
           Estimate Std. Error z value Pr(>|z|)
(Intercept)  45.272    13.610  3.327  0.00088
Long         -5.755     2.306  ***** BBBB
Anch        -10.447     3.755 -2.782  0.00540
---

    Null deviance: 138.629 on 99 degrees of freedom
Residual deviance: AAAA  on 97 degrees of freedom
AIC: 26.564

Number of Fisher Scoring iterations: 8
```

a) Escribe la fórmula de lo que en la salida de `R` se llama `Deviance residuals` y calcula la suma de estos residuos al cuadrado.
b) Calcula la desviación residual `AAAA` y contrasta, usando el método de razón de verosimilitudes, la hipótesis de que ninguna de las dos medidas influye en la variable respuesta: $H_0 : \beta_1 = \beta_2 = 0$.
c) Calcula el p-valor `BBBB` y contrasta a nivel $\alpha = 0.05$ la hipótesis nula de que la longitud del pétalo no es significativa para explicar la respuesta.
d) Para un lirio se sabe que la longitud del pétalo es de 4.9 mm y la anchura es 1.5 mm. ¿En cuál de las dos especies se debe clasificar?

4.5. En un experimento descrito en Prentice (1976) se expuso una muestra de escarabajos a cierto pesticida. Tras cinco horas de exposición a distintos niveles de concentración del pesticida algunos de los escarabajos murieron y otros sobrevivieron. Los resultados para cada dosis aparecen en la tabla siguiente:

Dosis ($\log_{10} CS_2 mgl^{-1}$)	N. insectos	N. muertos
1.6907	59	6
1.7242	60	13
1.7552	62	18
1.7842	56	28
1.8113	63	52
1.8369	59	53
1.8610	62	61
1.8839	60	60

Formula un modelo de regresión logística para analizar estos datos y estima la probabilidad de que muera un escarabajo expuesto durante cinco horas a una dosis de concentración 1.8.

4.6. Para tratar la meningitis bacteriana es vital aplicar con urgencia un tratamiento con antibióticos. Por ello, es importante distinguir lo más rápidamente posible este tipo de meningitis de la meningitis vírica. Con el fin de resolver este problema se ajustó con `R` un modelo de regresión logística a las siguientes variables medidas en 164 pacientes del *Duke University Medical Center*:

Nombre variable	Descripción
`age`	Edad en años
`bloodgl`	Concentración de glucosa en la sangre
`gl`	Concentración de glucosa en el líquido cefalorraquídeo
`pr`	Concentración de proteína en el líquido cefalorraquídeo
`whites`	Leucocitos por mm^3 de líquido cefalorraquídeo
`polys`	Porcentaje de leucocitos que son leucocitos polimorfonucleares
`abm`	Tipo de meningitis: bacteriana (`abm=1`) o vírica (`abm=0`)

El resultado del ajuste se muestra a continuación (algunos valores se han sustituido por letras):

```
Call:
glm(formula = abm ~ age + bloodgl + gl + pr + whites + polys,
    family = "binomial")

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.6433113 -0.2515780 -0.0426214  0.0009792  3.3999069

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.7729088 2.4465149 -3.995 6.48e-05 ***
age         -0.0745558 0.0254888 -2.925 0.003444 **
bloodgl      0.0495798 0.0137182  3.614 0.000301 ***
gl          -0.0566176 0.0186024 -3.044 0.002338 **
pr           0.0506505 0.0133574  3.792 0.000149 ***
whites       0.0007971 0.0005108    B   0.118660
polys        0.0453840 0.0145852  3.112 0.001860 **

    Null deviance:  A     on 163 degrees of freedom
Residual deviance: 51.539 on 157 degrees of freedom
AIC: 65.539
```

a) Calcula el valor de `A` en la salida anterior sabiendo que hay 68 pacientes con meningitis bacteriana en la muestra.
b) Calcula el valor de `B` en la salida anterior. A nivel $\alpha = 0.1$, ¿puede afirmarse que al aumentar la cantidad de leucocitos en el líquido cefalorraquídeo disminuye la probabilidad de que la meningitis sea de tipo vírico?
c) En un análisis realizado a un paciente de 15 años se han determinado los siguientes valores para el resto de variables:

`bloodgl`	`gl`	`pr`	`whites`	`polys`
119	72	53	262	41

¿En cuál de los dos tipos de meningitis debe clasificarse este paciente?

4.7. Supongamos que la distribución de $\mathbf{X}$ condicionada a $Y = 1$ es normal con vector de medias $\boldsymbol{\mu}_1$ y matriz de covarianzas $\boldsymbol{\Sigma}$, mientras que la distribución de $\mathbf{X}$ condicionada a $Y = 0$ es normal con vector de medias $\boldsymbol{\mu}_0$ y la misma matriz de covarianzas $\boldsymbol{\Sigma}$ (caso homocedástico). Demuestra que el error de la regla Bayes (error Bayes) del correspondiente problema de clasificación es:

$$L^* = 1 - \Phi(\Delta/2),$$

donde $\Delta^2 = (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)$ es el cuadrado de la distancia de Mahalanobis entre los dos vectores de medias y Φ es la función de distribución de una v.a. normal estándar. (Se supone que las probabilidades a priori de ambas poblaciones son iguales, $\pi_0 = \pi_1 = 1/2$).

4.8. El conjunto de datos `DatosNOSL` contiene información de 100 pacientes normales (grupo `NO`, $Y = 0$) y de 150 pacientes con espondilolistesis (grupo `SL`, $Y = 1$), una afección de la columna vertebral en la que una vértebra se desliza hacia adelante sobre la que está debajo. En cada paciente se miden seis características biomecánicas obtenidas a partir de la forma y orientación de la pelvis y de la columna lumbar[2]:

`V1`	Incidencia pélvica
`V2`	Inclinación pélvica
`V3`	Angulo de lordosis lumbar
`V4`	Pendiente del sacro
`V5`	Radio pélvico
`V6`	Grado de espondilolistesis
`V7`	Clase (`SL` o `NO`)

Se ofrece a continuación una vista parcial de los datos:

```
head(DatosNOSL)
```

```
   V1    V2    V3    V4     V5    V6 V7
74.38 32.05 78.77 42.32 143.56 56.13 SL
89.68 32.70 83.13 56.98 129.96 92.03 SL
44.53  9.43 52.00 35.10 134.71 29.11 SL
77.69 21.38 64.43 56.31 114.82 26.93 SL
76.15 21.94 82.96 54.21 123.93 10.43 SL
83.93 41.29 62.00 42.65 115.01 26.59 SL
```

Se ajusta con R un modelo de regresión logística a `DatosNOSL` obteniéndose la siguiente salida, en la que se han sustituido algunos valores por letras.

```
reglog = glm(V7~V1+V2+V3+V4+V5+V6,data=DatosNOSL,family="binomial")

summary(reglog)

Call:
glm(formula = V7 ~ V1 + V2 + V3 + V4 + V5 + V6, family = "binomial",
    data = DatosNOSL)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.87275 -0.10493  0.00014  0.03487  2.52323

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

[2]Fuente de los datos: UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml`

```
(Intercept) -4.541516  4.860944 -0.934   0.350
V1          34.682723 73.264591  0.473   0.636
V2         -34.591437 73.276658   AAAA    BBBB
V3          -0.009126  0.045632 -0.200   0.841
V4         -34.624532 73.269332 -0.473   0.637
V5          -0.028484  0.028235 -1.009   0.313
V6           0.294571  0.054590  5.396 6.81e-08 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 336.506 on 249 degrees of freedom
Residual deviance: 45.721 on 243 degrees of freedom
AIC: CCCC

Number of Fisher Scoring iterations: 9
```

a) Calcula el valor de `AAAA`. ¿A qué contraste de hipótesis corresponde este estadístico? A nivel $\alpha = 0.1$, ¿cuál es la conclusión acerca del contraste?

b) Calcula el valor de `CCCC`. A nivel $\alpha = 0.01$, usar el método de razón de verosimilitudes para contrastar si ninguna de las características influye conjuntamente sobre la clase Y.

c) En un análisis biomecánico a un nuevo paciente se han observado las siguientes características:

V1	V2	V3	V4	V5	V6
80	30	70	60	90	50

Estima la probabilidad de que el paciente padezca espondilolistesis. ¿En cual de los dos grupos clasificaríamos al paciente?