ESTADÍSTICA II **Curso 2016/2017**
Grado en Matemáticas

Relación 4 de problemas

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

$$X_0 = \begin{pmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{pmatrix}, \quad X_1 = \begin{pmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{pmatrix}.$$

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

2. Considera los datos sobre enfermedades coronarias en Sudáfrica (`infartos.RData`). 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?

3. 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. 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.

5. 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?

6. Supongamos que la distribución de X condicionada a $Y = 1$ es normal con vector de medias μ_1 y matriz de covarianzas Σ, mientras que la distribución de X condicionada a $Y = 0$ es normal con vector de medias μ_0 y la misma matriz de covarianzas Σ (caso homocedástico). Demuestra que el error de la regla Bayes (error Bayes) del correspondiente problema de clasificación es:

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

donde $\Delta^2 = (\mu_0 - \mu_1)'\Sigma^{-1}(\mu_0 - \mu_1)$ 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, $\pi_0 = \pi_1 = 1/2$).