ESTADÍSTICA II (2022/23). Grado en Matemáticas
SOLUCIONES del examen final, 16 de enero de 2023

Problema 1: a) $\mathbf{X} \sim N_2\left(\begin{pmatrix} 0 \\ 1 \end{pmatrix}, \mathbf{BB}' = \begin{pmatrix} 2 & 3 \\ 3 & 6 \end{pmatrix}\right)$

$X_2|X_1 \sim N_c(\mu_c, \sigma_c^2)$, con $\mu_c = 1 + \frac{3}{2}X_1$ y $\sigma_c^2 = \frac{3}{2}$

b)

$$Q = \frac{1}{3}(Y_1 - Y_2, Y_2 - Y_3, Y_3 - Y_1)\begin{pmatrix} Y_1 - Y_2 \\ Y_2 - Y_3 \\ Y_3 - Y_1 \end{pmatrix} = \frac{1}{3}\mathbf{Y}'\mathbf{C}'\mathbf{C}\mathbf{Y} = \mathbf{Y}'\mathbf{A}\mathbf{Y},$$

donde

$$\mathbf{C} = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ -1 & 0 & 1 \end{pmatrix} \qquad \text{y} \qquad \mathbf{A} = \frac{1}{3}\mathbf{C}'\mathbf{C} = \frac{1}{3}\begin{pmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

Es obvio que $\mathbf{A}$ es simétrica y es sencillo comprobar que es idempotente. Por tanto, $Q \sim \chi_p^2$, donde $p = \text{tr}(\mathbf{A}) = 2$.

Problema 2: a) Definimos las variables X = "Causa de la muerte" e Y = "Sexo". El número de clases de X e Y es 4 y 2 respectivamente. Queremos hacer el contraste

$$\begin{aligned} H_0: &\quad X \text{ e } Y \text{ son independientes} \\ H_1: &\quad X \text{ e } Y \text{ son dependientes} \end{aligned}$$

Frecuencias observadas:

	Hombre	Mujer	
Enfermedades infecciosas y parasitarias	$o_{11} = 25728$	$o_{12} = 20273$	$o_{1\cdot} = 46001$
Tumores	$o_{21} = 67844$	$o_{22} = 45818$	$o_{2\cdot} = 113662$
Enfermedades del sistema circulatorio	$o_{31} = 55905$	$o_{32} = 63291$	$o_{3\cdot} = 119196$
Causas externas de mortalidad	$o_{41} = 10689$	$o_{42} = 6142$	$o_{4\cdot} = 16831$
	$o_{\cdot 1} = 160166$	$o_{\cdot 2} = 135524$	$n = 295690$

Estimación $\hat{e}_{ij}$ de las frecuencias esperadas bajo H_0:

	Hombre	Mujer
Enfermedades infecciosas y parasitarias	24917.299	21083.701
Tumores	61567.141	52094.859
Enfermedades del sistema circulatorio	64564.735	54631.265
Causas externas de mortalidad	9116.825	7714.175

La región de rechazo del contraste es $R = \{\chi^2 > \chi^2_{(4-1)(2-1);0.01} = \chi^2_{3;0.01} = 11.34\}$, siendo el estadístico del contraste

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{2} \frac{o_{ij}^2}{\hat{e}_{ij}} - n = 300269.5 - 295690 = 4579.5.$$

Por tanto, rechazamos H_0 a nivel $\alpha = 0.01$. Con las tablas de las distribuciones lo único que podemos decir del p-valor es que es mucho menor que 0.005, porque el valor del estadístico del contraste es mucho mayor que $\chi^2_{3;0.005} = 12.84$. El p-valor exacto sería $\mathbb{P}\{\chi^2_3 > 4579.5\}$.

b) El código correcto para hacer el contraste de (a) es el segundo. Se puede justificar porque el estadístico del contraste de (a) y el de este código (dado por `X-squared`) coinciden. También se puede justificar porque, para hacer un contraste χ^2 de homogeneidad o independencia, el argumento de `chisq.test` debe ser una matriz de frecuencias observadas y en el primer código el argumento es un vector.

Respecto al primer código, lo que hace es un contraste de bondad de ajuste de las frecuencias observadas a una distribución uniforme sobre las 8 clases.

Problema 3: a)

$$\text{RSS} = \sum_{i=1}^{n}(y_i - \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i})^2 = \sum_{i=1}^{n_1}(y_i - \beta_1 - \beta_3 x_{i3})^2 + \sum_{i=n_1+1}^{n}(y_i - \beta_2 - \beta_3 x_{i3})^2$$

b) Los estimadores de mínimos cuadrados (e.m.c.) de β_1, β_2 y β_3 se obtienen de resolver las ecuaciones:

$$0 = \frac{\partial \text{RSS}}{\partial \beta_1} \qquad 0 = \frac{\partial \text{RSS}}{\partial \beta_2} \qquad 0 = \frac{\partial \text{RSS}}{\partial \beta_3}.$$

Denotamos $\bar{y}_M = \sum_{i=1}^{n_1} y_i/n_1$ la media muestral de la respuesta en el grupo de mujeres. Análogamente denotamos $\bar{x}_{3,M} = \sum_{i=1}^{n_1} x_{i3}/n_1$. Y de la misma manera se definen $\bar{y}_H$ y $\bar{x}_{3,H}$.

$$\frac{\partial \text{RSS}}{\partial \beta_1} = -2\sum_{i=1}^{n_1}(y_i - \beta_1 - \beta_3 x_{i3}) = -2n_1\left(\bar{y}_M - \beta_1 - \beta_3 \bar{x}_{3,M}\right) = 0 \Rightarrow \beta_1 = \bar{y}_M - \beta_3 \bar{x}_{3,M}$$

$$\frac{\partial \text{RSS}}{\partial \beta_2} = -2\sum_{i=n_1+1}^{n}(y_i - \beta_1 - \beta_3 x_{i3}) = -2n_2\left(\bar{y}_H - \beta_2 - \beta_3 \bar{x}_{3,H}\right) = 0 \Rightarrow \beta_2 = \bar{y}_H - \beta_3 \bar{x}_{3,H}$$

$$\begin{aligned}
\frac{\partial \text{RSS}}{\partial \beta_3} &= -2\left(\sum_{i=1}^{n_1}(y_i - \beta_1 - \beta_3 x_{i3})x_{i3} + \sum_{i=n_1+1}^{n}(y_i - \beta_2 - \beta_3 x_{i3})x_{i3}\right) \\
&= -2\left(n_1\left(\frac{\sum_{i=1}^{n_1} x_{i3}y_i}{n_1} - \beta_1\bar{x}_{3,M} - \beta_3\frac{\sum_{i=1}^{n_1} x_{i3}^2}{n_1}\right) + n_2\left(\frac{\sum_{i=n_1}^{n} x_{i3}y_i}{n_2} - \beta_2\bar{x}_{3,H} - \beta_3\frac{\sum_{i=n_1}^{n} x_{i3}^2}{n_2}\right)\right)
\end{aligned}$$

Juntando las tres ecuaciones tenemos:

$$\begin{aligned}
0 &= n_1\left(\frac{\sum_{i=1}^{n_1} x_{i3}y_i}{n_1} - (\bar{y}_M - \beta_3\bar{x}_{3,M})\bar{x}_{3,M} - \beta_3\frac{\sum_{i=1}^{n_1} x_{i3}^2}{n_1}\right) \\
&\quad + n_2\left(\frac{\sum_{i=n_1}^{n} x_{i3}y_i}{n_2} - (\bar{y}_H - \beta_3\bar{x}_{3,H})\bar{x}_{3,H} - \beta_3\frac{\sum_{i=n_1}^{n} x_{i3}^2}{n_2}\right) \\
&= n_1\left(\frac{\sum_{i=1}^{n_1} x_{i3}y_i}{n_1} - \bar{x}_{3,M}\bar{y}_M\right) + n_2\left(\frac{\sum_{i=n_1}^{n} x_{i3}y_i}{n_2} - \bar{x}_{3,H}\bar{y}_H\right) \\
&\quad + \beta_3\left(n_1\left(\bar{x}_{3,M}^2 - \frac{\sum_{i=1}^{n_1} x_{i3}^2}{n_1}\right) + n_2\left(\bar{x}_{3,H}^2 - \frac{\sum_{i=n_1}^{n} x_{i3}^2}{n_2}\right)\right) \\
&= n_1\text{cov}_M(x_3, y) + n_2\text{cov}_H(x_3, y) - \beta_3(n_1 v_{3,M} + n_2 v_{3,H}),
\end{aligned}$$

siendo $v_{3,M}$ la varianza muestral de x_3 y $\text{cov}_M(x_3, y)$ la covarianza muestral entre x_3 y la respuesta en el grupo de mujeres y análogamente se define para el grupo de hombres. De la anterior ecuación despejamos el e.m.c. de β_3:

$$\begin{aligned}
\hat{\beta}_3 &= \frac{n_1\text{cov}_M(x_3, y) + n_2\text{cov}_H(x_3, y)}{n_1 v_{3,M} + n_2 v_{3,H}} = \frac{n_1\text{cov}_M(x_3, y)}{n_1 v_{3,M} + n_2 v_{3,H}} + \frac{n_2\text{cov}_H(x_3, y)}{n_1 v_{3,M} + n_2 v_{3,H}} \\
&= \frac{\text{cov}_M(x_3, y)}{v_{3,M}}\frac{n_1 v_{3,M}}{n_1 v_{3,M} + n_2 v_{3,H}} + \frac{n_2\text{cov}_H(x_3, y)}{v_{3,H}}\frac{n_2 v_{3,H}}{n_1 v_{3,M} + n_2 v_{3,H}} \\
&= \hat{\beta}_{3,M} w_M + \hat{\beta}_{3,H} w_H,
\end{aligned}$$

donde $\hat{\beta}_{3,M}$ y $\hat{\beta}_{3,H}$ son las pendientes de las rectas de mínimos cuadrados de y sobre x_3 en el grupo de mujeres y en el de hombres respectivamente y

$$w_M = \frac{n_1 v_{3,M}}{n_1 v_{3,M} + n_2 v_{3,H}} \qquad \text{y} \qquad w_H = \frac{n_2 v_{3,H}}{n_1 v_{3,M} + n_2 v_{3,H}}$$

son pesos ($w_M, w_H \geq 0$, $w_M + w_H = 1$). Por tanto, $\hat{\beta}_3$ es la media ponderada de las pendientes $\hat{\beta}_{3,M}$ y $\hat{\beta}_{3,H}$. Por otro lado, $\hat{\beta}_1 = \bar{y}_M - \hat{\beta}_3 \bar{x}_{3,M}$ y $\hat{\beta}_2 = \bar{y}_H - \hat{\beta}_3 \bar{x}_{3,H}$ son términos muy parecidos a los términos independientes de las rectas de mínimos cuadrados de y sobre x_3 en el grupo de mujeres y en el de hombres respectivamente:

$$\hat{\beta}_{0,M} = \bar{y}_M - \hat{\beta}_{3,M}\bar{x}_{3,M} \qquad \text{y} \qquad \hat{\beta}_{0,H} = \bar{y}_H - \hat{\beta}_{3,H}\bar{x}_{3,H}.$$

De hecho, en el límite, cuando la población de mujeres o de hombres sea mayoritaria y tienda al 100 %,

$$\hat{\beta}_1 \xrightarrow{w_M \to 1} \hat{\beta}_{0,M} \qquad \text{y} \qquad \hat{\beta}_1 \xrightarrow{w_H \to 1} \hat{\beta}_{0,H}.$$

Problema 4: **a)** El modelo logístico es $\mathbb{P}\{Y = 1|\text{LI}\} = 1/(1+\exp(-(\beta_0+\beta_1\text{LI})))$. La probabilidad de remisión estimada en un paciente con `LI = 28` es

$$\hat{\mathbb{P}}\{Y = 1|\text{LI} = 28\} = 1/(1 + \exp(-(-3.77714 + 0.14486 \cdot 28))) = 0.569.$$

Como esta probabilidad estimada es mayor que 0.5, asignamos al paciente al grupo 1 (Remisión sí).

b) El contraste pedido es

$$\begin{aligned} H_0 &: \quad \beta_1 \leq 0 \\ H_1 &: \quad \beta_1 > 0. \end{aligned}$$

La región de rechazo es $R = \{z > z_{0.05} = 1.645\}$, siendo el estadístico del contraste

$$z = \texttt{A} = \frac{0.14486}{0.05934} = 2.441.$$

Por tanto, al nivel de significación del 5 % hay suficiente evidencia para rechazar la hipótesis nula.

c)

$$\hat{O}(\text{LI} = 34) = \frac{\hat{\mathbb{P}}\{Y = 1|\text{LI} = 34\}}{\hat{\mathbb{P}}\{Y = 0|\text{LI} = 34\}} = 3.15.$$

Representa cuánto más probable es que remita el cáncer que la no remisión en el paciente con LI=34. En este caso, es más del triple de probable que remita.