Análisis Multivariante.
Febrero 2020, $1^{\underline{a}}$ semana

23 de diciembre de 2020

Índice

Ejercicio 1

Suponga que un muestreo aleatorio simple de tamaño 40, sobre una población $N_3(\mu, V)$, se han obtenido los siguientes resúmenes estadísticos:

$$\bar{x} = \begin{pmatrix} 3 \\ 5 \\ 4 \end{pmatrix}, S = \begin{pmatrix} 6 & 1 & -2 \\ 1 & 3 & 2 \\ -2 & 2 & 5 \end{pmatrix}$$

(a) Contraste la hipótesis de que todas las componentes de la media son igual a 4.

(b) ¿Qué decidiría sobre la hipótesis de esfericidad?

Solución:

apartado (a)

Para contrastar si todas las componentes de la media son igual a 4 planteamos el contraste de hipótesis:

$$H_0 : \mu = (4,4,4)' = \mu_0, V = cualquiera$$

$$H_1 : \mu \neq (4,4,4) = \mu_0, V = cualquiera$$

y según el método de la razón de verosimilitudes, para una muestra grande ($n > 30$), el estadístico de contraste es:

$$\lambda = n\, log\left(1 + \frac{T^2}{n-1}\right) \approx T^2 \sim \chi^2_g$$

donde $T^2 = (n-1)(\bar{x}-\mu_0)'S^{-1}(\bar{x}-\mu_0)$ y $g = dim\,\Omega_1 - dim\,\Omega_0$. Si operamos teniendo en cuenta que $n = 40$, $\bar{x} = (3,5,4)'$ y $\mu_0 = (4,4,4)'$ tenemos:

$$T^2 = 39\,(-1,1,0)\frac{1}{41}\begin{pmatrix} 11 & -9 & -8 \\ -9 & 26 & -14 \\ 8 & -14 & 17 \end{pmatrix}(-1,1,0)' = \frac{39}{41}\,7 \approx 6{,}65$$

Por otro lado $g = \frac{p(p+1)}{2} + p - \frac{p(p+1)}{2} = p = 3$, por tanto, fijado α, rechazamos H_0 si:

$$p_{valor} = 1 - P\{T^2 > \chi^2_3\} = 1 - P\{6{,}65 > \chi^2_3\} = P\{6{,}65 < \chi^2_3\} < \alpha$$

A modo de intuición, la esperanza de una χ^2 son sus grados de libertad y su varianza el doble de los grados de libertad, y como $T^2 = 6{,}65$

$$E[\chi^2_3] + \sqrt{Var(\chi^2_3)} < 6{,}65 < E[\chi^2_3] + 2\sqrt{Var(\chi^2_3)}$$

por lo que 6,65 está entre una y dos desviaciones típicas de la media de la distribución, por lo que es un valor relativamente valor alto y deberíamos esperar un pvalor relativamente bajo pero que no nos permita rechazar H_0.[1]

apartado (b)

[1] $P\{6{,}65 < \chi^2_3\} \approx 0{,}08$ que es un pvalor bajo, pero ligeramente más alto que los α usuales, así que no rechazaríamos la hipótesis nula.

La hipótesis de esfericidad significa que asumimos que la matriz de covarianzas poblacional es de la forma $V = \sigma^2 I$. Esta hipótesis se llama de esfericidad por que las curvas de nivel de la función de densidad de una normal $N_p(\mu, \sigma^2 I)$ son esferas, y por tanto, la distribución es simétrica en todas las direcciones del espacio.

Para comprobar la hipótesis de esfericidad, planteamos el contraste:

$$H_0 : V = \sigma^2 I,\ \mu = cualquiera$$

$$H_1 : V = cualquiera,\ \mu = cualquiera$$

El estimador de máxima verosimilitud de σ^2 es:

$$\hat{\sigma}^2 = \frac{tr(S)}{p} = \frac{1}{p}\sum_{i=1}^{p} s_{ii}$$

es decir, la varianza muestral media de todas las componentes. Por otro lado, el estadístico de contraste según el método de la razón de verosimilitudes es:

$$\lambda = np\, log\, \hat{\sigma}^2 - n\, log|S|$$

En nuestro problema tenemos:

$$\hat{\sigma}^2 = \frac{1}{3}(6+3+5) = \frac{14}{3}$$

$$\lambda = 40\, 3\, log\frac{14}{3} - 40\, log\, 41 \approx 36{,}31$$

Finalmente, como $\lambda \sim \chi^2_g$, debemos encontrar los grados de libertad g que vienen dados por la diferencia entre la dimensión de los espacios en que se mueven los parámetros bajo ambas hipótesis, es decir:

$$g = dim(\Omega_1) - dim(\Omega_0) = p + \frac{p(p+1)}{2} - p - 1 = \frac{(p+2)(p-1)}{2} = \frac{(3+2)(3-1)}{2} = 5$$

Finalmente, fijado un nivel de significación α, rechazamos H_0 cuando se cumpla:

$$p_{valor} = 1 - P\{\lambda > \chi^2_g\} = P\{36{,}31 < \chi^2_5\} < \alpha$$

Sin conocer los valores de una χ^2_6 podemos notar que el valor del estadístico es demasiado alto como para ser comparado con una χ^2_5, ya que esta tiene por esperanza sus grados de libertad que son 5 y varianza el doble de los grados de libertad que es 10. Por tanto, 36,31 está a más de tres desviaciones típicas de la media ($5 + 3\sqrt{10} < 36{,}31$) y es de esperar que podamos rechazar H_0 con gran seguridad ya que $P\{36{,}31 < \chi^2_5\} \approx 0$.

Ejercicio 2

¿Cuál es el planteamiento teórico del problema de clasificación entre dos poblaciones? ¿En qué se traduce la solución cuando las poblaciones son normales? Aplique la metodología al caso:

$$\mu_1 = \begin{pmatrix} 2 \\ 3 \end{pmatrix},\ \mu_2 = \begin{pmatrix} 6 \\ 1 \end{pmatrix},\ S = \begin{pmatrix} 3 & 2 \\ 2 & 2 \end{pmatrix}$$

Solución:

Supongamos que tenemos dos poblaciones P_1 y P_2 con funciones de densidad f_1 y f_2 y queremos clasificar en una de ellas una nueva observación x_0. Sean π_1 y π_2 las probabilidades *a priori* de pertenecer a una de las dos poblaciones, su función de probabilidad es una mixtura de la forma:

$$f(x) = \pi_1 f_1(x) + \pi_2 f_2(x)$$

y la probabilidad *a posteriori* de pertenecer a cada una vendrá dada por el teorema de Bayes:

$$P(1|x_0) = \frac{\pi_1 f_1(x_0)}{f(x_0)}$$

$$P(2|x_0) = \frac{\pi_2 f_2(x_0)}{f(x_0)}$$

Por tanto, deberemos clasificar x_0 en P_2 si se cumple:

$$P(2|x_0) > P(1|x_0) \Rightarrow \pi_2 f_2(x_0) > \pi_1 f_1(x_0) \tag{1}$$

Si además las poblaciones son $p-$normales con la misma matriz de covarianzas, entonces:

$$f_i(x) = \frac{1}{(2\pi)^{p/2}|V|^{1/2}} exp\{-\frac{1}{2}(x-\mu_i)'V^{-1}(x-\mu_i)\}$$

sustituyendo en (1) y tomando logaritmos a ambos lados de la desigualdad tenemos

$$log\,\pi_2 + (x_0-\mu_2)'V^{-1}(x_0-\mu_2) > log\,\pi_1 + (x_0-\mu_1)'V^{-1}(x_0-\mu_1)$$

$$log\,\pi_2 + D_2^2 < log\,\pi_1 + D_1^2 \tag{2}$$

donde D_i es la distancia de Mahalanobis de x_0 a la media poblacional μ_i. Esto reduce el problema a clasificar x_0 en la población cuyo centro esté más cerca a x_0 con la distancia de Mahalanobis.

El procedimiento de cálculo más sencillo se deduce de simplificar (2) usando:

$$D_i^2 = (x_0 - \mu_i)' V^{-1} (x_0 - \mu_i) =$$
$$= x_0' V^{-1} x_0 - 2x_0' V^{-1} \mu_i + \mu_i' V^{-1} \mu_i$$

y sustituyendo en (2) tenemos[2]:

$$log\, \pi_2 + x_0' V^{-1} x_0 - 2x_0' V^{-1} \mu_2 + \mu_2' V^{-1} \mu_2 <$$
$$< log\, \pi_1 + x_0' V^{-1} x_0 - 2x_0' V^{-1} \mu_1 + \mu_1' V^{-1} \mu_1$$
$$log\, \pi_2 - 2x_0' V^{-1} \mu_2 + \mu_2' V^{-1} \mu_2 < log\, \pi_1 - 2x_0' V^{-1} \mu_1 + \mu_1' V^{-1} \mu_1$$
$$log\, \pi_2 + 2x_0' V^{-1} (\mu_1 - \mu_2) + \mu_2' V^{-1} \mu_2 - \mu_1' V^{-1} \mu_1 < log\, \pi_1$$

finalmente, como $\mu_2' V^{-1} \mu_2 - \mu_1' V^{-1} \mu_1 = (\mu_2 - \mu_1)' V^{-1} (\mu_1 + \mu_2)$ obtenemos:

$$log\, \pi_2 + 2x_0' V^{-1} (\mu_1 - \mu_2) + (\mu_2 - \mu_1)' V^{-1} (\mu_1 + \mu_2) < log\, \pi_1$$
$$log\, \pi_2 + 2x_0' V^{-1} (\mu_1 - \mu_2) - (\mu_1 + \mu_2)' V^{-1} (\mu_1 - \mu_2) < log\, \pi_1$$
$$log\, \pi_2 + (2x_0 - (\mu_1 + \mu_2))' V^{-1} (\mu_1 - \mu_2) < log\, \pi_1$$

llamando $w = V^{-1}(\mu_2 - \mu_1)$ el problema se reduce a clasificar en P_2 si:

$$log\, \pi_2 + ((\mu_1 + \mu_2) - 2x_0)' w < log\, \pi_1$$
$$log\, \pi_2 + (\mu_1 + \mu_2)' w < log\, \pi_1 + 2x_0' w \qquad (3)$$

NOTA IMPORTANTE: Si las probabilidades a priori son iguales entonces a la función $L(x) = x_0' w - \left(\frac{\mu_1 + \mu_2}{2}\right)' w$ se la conoce como función discriminante lineal de Fisher, de manera que clasifica en P_2 si $L(x) > 0$ y en P_1 si $L(x) < 0$.

Aplicando esto a nuestro problema:

$$w = \begin{pmatrix} 3 & 2 \\ 2 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 1 & -1 \\ -1 & 3/2 \end{pmatrix} \begin{pmatrix} 4 \\ -2 \end{pmatrix} = \begin{pmatrix} 6 \\ -7 \end{pmatrix}$$

[2]Este procedimiento no viene en el libro y puede ser de interés, se ve en el Cuadras.

Ahora, clasificamos en P_2 si:

$$log\,\pi_2 + \begin{pmatrix} 6 & -7 \end{pmatrix} x_0 > log\,\pi_1 + \begin{pmatrix} 6 & -7 \end{pmatrix} \begin{pmatrix} 8 \\ 4 \end{pmatrix} = log\,\pi_1 + 20$$

Es decir, la regla de decisión sería:

$$d(x_0) = \begin{cases} P_2 \text{ si } log\,\pi_2 + 6x_{01} - 7x_{02} > log\,\pi_1 + 20 \\ \\ P_1 \text{ si } log\,\pi_2 + 6x_{01} - 7x_{02} \leq log\,\pi_1 + 20 \end{cases}$$

A modo de comprobación, se ve que si $x_0 \approx \mu_2$ entonces, x_0 está en P_2 si $log\,\pi_2 + 29 > log\,\pi_1 + 20$ y si las probabilidades a priori fueran iguales, quedaría clasificada en P_2, como es lógico, ya que x_0 estaría muy cerca del centro de P_2.

Ejercicio 3

Describa, sin demostraciones, los conceptos y/o métodos aludidos en las siguiente expresiones:

(a) Distribuciones elípticas.

(b) Criterios para la selección de modelos.

Solución:

apartado (a)

Decimos que una variable x sigue una distribución esférica si su función de densidad depende de la variable solo por la distancia euclidiana $x'x = \sum x^2$.

Sean A una matriz cuadrada de dimensión $p \times p$, m un vector columna $p \times 1$ y x una variable esférica $p-$dimensional, decimos que y es elíptica si es de la forma:

$$y = m + Ax$$

Su esperanza es cero, como las esféricas, y su matriz de covarianzas es $V_y = \sigma^2 AA'$, que viene del cambio de variable en la distribución esférica de x, que tiene por matriz de covarianzas $V_x = \sigma^2 I$.

Las propiedades mas relevantes son:

1. Como la densidad de las esféricas depende de $x'x$, deducimos del cambio que la densidad de las elípticas depende de la distancia de Mahalanobis

$$(y-m)'V_y^{-1}(y-m)$$

2. Las curvas de nivel de la función de densidad son elipsoides centrados en m.

Los casos más conocidos de distribuciones elípticas son la distribución normal multivariante y la *distribución t multivariante*, que resulta de dividir cada componente de una distribución normal entre la raíz cuadrada de una χ^2 dividida entre sus grados de libertad.

apartado (b)

Los criterios mas usuales para seleccionar modelos son el criterio de Akaike (AIC) y el criterio Bayesiano (BIC).

El criterio de Akaike trata de seleccionar el modelo cuyas predicciones sean lo más precisas posibles. Considerando una variable respuesta y y un modelo M_k de parámetros $\theta_1, \ldots, \theta_k$, deberemos seleccionar un modelo tal que $f(y|M_i)$, con $M_i = (\theta_1, \ldots, \theta_i, 0, \ldots, 0)$, sea lo más próxima posible a la verdadera distribución $f(y)$.

Para esto, usamos la divergencia de *Kullback-Leibler*[3]:

$$KL(f(y|M_i), f(y)) = \int log \, \frac{f(y|M_i)}{f(y)} f(y)dy \tag{4}$$

y se puede probar (Akaike, 1985) que minimizar (4) es equivalente a minimizar:

$$AIC = -2L(M_i) + 2p_i$$

[3]Se usa mucho para medir como de distintas son dos funciones de densidad, en nuestro caso, queremos medir como de distintas son la verdadera densidad de y y la que obtenemos con el modelo.

con p_i el número de parámetros del modelo y $L(M_i)$ el soporte del modelo M_i. Se puede ver fácilmente que AIC es estrictamente positivo y que tiende a disminuir al introducir nuevas variables, efecto que tratamos de corregir sumando el doble de la cantidad de parámetros.

Por otro lado, con la misma notación que el criterio anterior, el criterio de Bayes propone considerar los modelos como hipótesis sobre los datos y seleccionar el modelo con mayor probabilidad *a posteriori*, que vienen dadas por:

$$P(M_j|X) = \frac{f(X|M_j)P(M_j)}{\sum f(X|M_i)P(M_i)}$$

Por tanto, aplicar este criterio, equivale a seleccionar el modelo M que cumpla:

$$M = \max_{j=1,\dots,m} f(X|M_j)P(M_j)$$

Suponiendo que tenemos una muestra grande y que la distribución del vector de parámetros $\hat{\theta}_j$ de un modelo M_j es asintóticamente normal, con p_j la dimensión del vector de parámetros, el criterio de Bayesiano se puede escribir de la forma (*Bayesian Information Criterion*):

$$BIC(M_j) = -2L(\hat{\theta}_j|X) + p_j log\, n \tag{5}$$

y seleccionaríamos el modelo que haga tal cantidad mínima. Al igual que el criterio de Akaike, cabe observar que este criterio siempre disminuye al aumentar la cantidad de parámetros, por lo que sumando la cantidad de parámetros por el logaritmo del tamaño de la muestra se intenta corregir esta sensibilidad.

El criterio BIC trata de seleccionar el modelo correcto, maximizando la probabilidad a posteriori, y se puede probar que es consistente, es decir, que para un tamaño muestral suficientemente grande selecciona con probabilidad 1 el verdadero modelo si este está entre los candidatos $M_1, \dots, M_m$. Sin embargo, el criterio de Akaike es más conformista, no pretende encontrar el verdadero modelo, que puede no estar entre los candidatos, si no que trata de encontrar el que aporte mejores predicciones. En general, el criterio de Bayes seleccionará modelos con pocos parámetros y el criterio de Akaike tiende a seleccionar modelos con más parámetros de los necesarios si las muestras no son suficientemente grandes.

Concluyendo, ambos criterios comparten la misma forma funcional:

$$C(M_j) = -2L(M_j) + p_j\, g(n)$$

y se distinguen en la función $g(n)$, que puede ser modificada para dar lugar a variaciones de los criterios anteriores.