Qüestió - V. 10, n.º 2 (juny 1986) pp. 83-92

DISTANCIA ENTRE MODELOS LINEALES NORMALES

M. RIOS Y C. M. CUADRAS
UNIVERSIDAD DE BARCELONA

Este trabajo aborda el problema de comparar modelos lineales normales desde una perspectiva geométrica. A tal fin, se define una distancia geométrica informativa entre dos modelos lineales normales. La distancia propuesta es estudiada para diferentes condiciones experimentales. Se hallan además extensiones al modelo lineal normal multivariante. Finalmente, se deducen pruebas de significación para las distancias.

Keywords: INFORMATION METRIC; GEODESIC DISTANCE; MULTIVARIATE ANALYSIS OF VARIANCE; COMPARISON OF REGRESSIONS.

1. INTRODUCCION.

La forma más simple de comparación de modelos lineales, es la comparación de rectas de regresión. En sus diferentes facetas (test de paralelismo, test de coincidencia, etc.), es un problema clásico bien resuelto (véase /23/). También se ha estudiado la comparación de rectas de regresión a lo largo de un intervalo (/25/) y bajo la hipótesis de normalidad. Recientemente se ha estudiado un test a libre distribución para determinar si una recta de regresión $y = \alpha_1 + \beta_2 x$, es sistemáticamente mayor que otra recta de regresión $y = \alpha_2 + \beta_2 x$, es decir, decidir si $\alpha_1 + \beta_1 x > \alpha_2 + \beta_2 x$, para todo x de un intervalo I (/13/).

La extensión natural consiste en la comparación de curvas de regresión polinómica, problema que ha sido abordado por Cuadras y Sanchez (/12/), pero que queda como un caso particular de la comparación de modelos lineales más generales (/18/).

En este trabajo se aborda el problema de comparar dos modelos lineales

$$Y_1 = X_1 \beta_1 + e_1 \qquad Y_2 = X_2 \beta_2 + e_2 \qquad (1)$$

pero utilizando un enfoque geométrico consistente en definir una distancia entre las funciones de densidad paramétricas, asociadas a cada modelo.

Consideremos una clase de funciones de densidad paramétricas

$$S = \{p(x, \theta)\} \qquad (2)$$

donde $\theta = (\theta^1, \ldots, \theta^m)$ es un vector paramétrico perteneciente a un subconjunto H de R^m. Podemos introducir en S una estructura de variedad V y definir un tensor covariante de segundo orden (g_{ij}) que dote a V de estructura de espacio de Riemann o variedad riemanniana. Fue Rao (/21/) el primero en notar la importancia de este punto de vista geométrico-diferencial. Rao (/21/), alegando razones heurísticas, propone que el tensor métrico debe venir definido a través de la matriz de información de Fisher (suponiendo verificadas las convenientes condiciones de regularidad).

$$g_{ij} = E\left[(D_i \ln p)(D_j \ln p)\right] \qquad (3)$$

siendo

$$D_i \ln p = \frac{\partial}{\partial \theta_i} \ln p(x, \theta)$$

La distancia entre dos puntos de la variedad, es decir, entre dos densidades p_1, p_2, se calcula a través del elemento de arco, ds definido a través de:

- M. Ríos i C.M. Cuadras - Universitat de Barcelona - Facultat de Biologia - Dep. de Bioestadística Av. Diagonal, 637 - Barcelona.
- Article rebut el juliol de 1986.

$$ds^2 = \sum_{i,j} g_{ij}\, d\theta^i\, d\theta^j$$

integrando convenientemente a lo largo de una curva geodésica que una ambos puntos.

Diferentes autores han aplicado esta fecunda idea para obtener distancias entre distribuciones (véase /2/; /4/; /17/; /1/). Se puede probar, además, que imponiendo diferentes condiciones a una distancia, se llega por vía axiomática a la distancia propuesta por Rao (véase /19/).

Desde este punto de vista la distancia entre los dos modelos (1), suponiendo las condiciones usuales en los modelos lineales normales, es la distancia entre las distribuciones de probabilidad

$$N(X_1\,\beta_1,\ \sigma_1^2\, I) \qquad N(X_2\,\beta_2,\ \sigma_2^2\, I)$$

Obtendremos esta distancia para diferentes situaciones sobre la matriz de diseño, la varianza, etc. También desarrollaremos algunos contrastes de hipótesis relacionados con las mismas.

2. DISTANCIA ENTRE MODELOS LINEALES NORMALES.

Consideremos el modelo lineal normal $Y \sim \tilde{N}(\tilde{X}\beta, \sigma^2 I)$, donde X (nxm) es la matriz de diseño, $\beta = (\beta_1,\ldots,\beta_m)'$ es el vector de parámetros de regresión, $Y = (y_1,\ldots,y_n)'$ es el vector de observaciones independientes de la variable observable, σ^2 es la varianza del modelo.

Sin embargo, para abordar el problema con mayor generalidad, utilizaremos la matriz de diseño reducida X (/7/ y /9/). Designamos por X la matriz kxm, que contiene las k filas no repetidas de $\tilde{X}$. El número k representa las condiciones experimentales diferentes bajo las cuales se obtienen réplicas de la variable observable. Designamos por

$$D = \text{diag}\,(n_1,\ n_2,\ \ldots,\ n_k) \qquad (4)$$

la matriz diagonal que contiene el número de réplicas n_i obtenidas bajo la i-ésima condición experimental. Finalmente,

$$\bar{Y} = (\bar{y}_1,\ \bar{y}_2,\ \ldots,\ \bar{y}_k)' \qquad (5)$$

representa el vector columna con las medidas muestrales calculadas para cada condición experimental. Por ejemplo, dado el modelo lineal

$$y_1 = \beta_1 + \beta_2 + e_1$$
$$y_2 = \beta_1 + \beta_2 + e_2$$
$$y_3 = \beta_1 + \beta_2 + e_3$$
$$y_4 = \beta_1 - \beta_2 + e_4$$

tenemos n = 4, m = 2, k = 2 y además

$$\tilde{X} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

$$D = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \qquad \bar{Y} = \left((y_1+y_2+y_3)/3,\ y_4\right)'$$

Volviendo al caso general, la estimación por mínimos cuadrados de los parámetros β son

$$\hat{\beta} = (\tilde{X}'\tilde{X})^{-}\tilde{X}'Y = (X'DX)^{-}X'D\bar{Y} \qquad (6)$$

y la suma de cuadrados residual es

$$R_o^2 = Y'Y - \hat{\beta}'\tilde{X}'Y = Y'Y - \hat{\beta}'X'D\bar{Y} =$$
$$= Y'Y - \bar{Y}'DX\,(X'DX)^{-}X'D\bar{Y} \qquad (7)$$

donde $(X'DX)^{-}$ es una g-inversa de X'DX. Es fácil ver que podemos expresar también el modelo lineal como $\bar{Y} \sim N(X\beta, \sigma^2 D^{-1})$.

Por otra parte, si consideramos la variedad de las distribuciones normales n-variantes de vector de medias $\mu = X\beta$, matriz de covarianzas $\sigma^2 D^{-1}$, e introducimos la métrica informacional, obtenemos el tensor métrico

$$G = \sigma^{-2} D \qquad (8)$$

(véase /14/, /3/). Como el tensor métrico es constante, resulta que la geometría inducida en la variedad paramétrica es Euclídea.

Consideremos ahora dos modelos lineales identificados con las distribuciones $N(X\beta_1, \sigma^2 D^{-1})$, $N(X\beta_2, \sigma^2 D^{-1})$. Por todo lo dicho, la distancia entre ambos modelos es la longitud de la curva geodésica que une los vectores paramétricos β_1, β_2, que en este caso es una recta. Luego la distancia (al cuadrado) es

$$\delta^2(\beta_1,\beta_2) = (\mu_1-\mu_2)' G (\mu_1-\mu_2)$$

$$= (\beta_1-\beta_2)' X'DX (\beta_1-\beta_2)/\sigma^2 \quad (9)$$

La estimación de la distancia puede conseguirse tras sustituir en (9) los parámetros β por sus estimaciones (6). Entonces, si σ^2 es conocida, obtenemos

$$\hat{\delta}_1^2(\beta_1,\beta_2) =$$

$$= (\bar{Y}_1-\bar{Y}_2)'DX (X'DX)^- (X'DX) (X'DX)X'D (\bar{Y}_1-\bar{Y}_2)$$

$$= (\bar{Y}_1-\bar{Y}_2)'DX (X'DX)^- X'D (\bar{Y}_1-\bar{Y}_2) / \sigma^2 \quad (10)$$

Si la varianza común σ^2 es desconocida, consideraremos entonces los dos modelos conjuntamente a fin de calcular la suma de cuadrados residual

$$\sum_{i=1}^{2} (Y_i' Y_i - \hat{\beta}_i' X' D \bar{Y}_i) \quad (11)$$

que tiene 2(n-r) g.l, donde n es el número de observaciones y r es el rango de la matriz de diseño. Y_i es el vector de observaciones, $\bar{Y}_i$ es el vector de medias (i=1,2). Sustituyendo σ^2 por el correspondiente estimador insesgado, obtenemos

$$\hat{\delta}_2^2(\beta_1,\beta_2) =$$

$$= \left[2(n-r) (\bar{Y}_1-\bar{Y}_2)'DX (X'D)^- X'D (\bar{Y}_1-\bar{Y}_2)\right] /$$

$$\sum_{i=1}^{2} (Y_i' Y_i - \hat{\beta}_i X' D \bar{Y}_i) \quad (12)$$

Supongamos ahora que ambos modelos están asociados a una misma matriz de diseño reducida X, pero que el número de réplicas por condición experimental es diferente. Indiquemos

$$D_i = \text{diag}(n_{i1}, n_{i2}, \ldots, n_{ik})$$

$$n_{i.} = \sum_{h=1}^{k} n_{ih} \qquad i = 1, 2 \quad (13)$$

Entonces $\bar{Y}_1 \sim N(X\beta_1, \sigma^2 D_1^{-1})$, $\bar{Y}_2 \sim N(X\beta_2, \sigma^2 D_2^{-1})$, es decir, ambos modelos tienen la misma estructura en cuanto a parámetros de regresión (condición indispensable para que tenga sentido su comparación), pero distinta matriz de covarianzas. En este caso no es inmediato definir una distancia porque no existe una transformación contínua que una dos puntos de la variedad correspondiente (salvo que D_1 sea proporcional a D_2), pues en realidad estamos en dos variedades paramétricas distintas. Sin embargo, por razones que justificamos en la sección siguiente, vamos a definir una distancia. Sean

$$A_i = D_i X (X'D_i X)^- X'D_i \qquad i = 1,2$$

$$A = (D_1+D_2) X |X'(D_1+D_2) X|^- X' (D_1+D_2)$$

$$\hat{\beta}_i = (X' D_i X)^- X' D \bar{Y}_i \qquad i = 1,2$$

$$\bar{Y} = (D_1+D_2)^{-1} (D_1 \bar{Y}_1 + D_2 \bar{Y}_2)$$

$$\mu = (D_1+D_2)^{-1} (D_1 X \beta_1 + D_2 X \beta_2)$$

$$n.. = n_{1.} + n_{2.} \quad (14)$$

Por definición, la distancia (al cuadrado), entre

$N(X\beta_1, \sigma^2 D_1^{-1})$ y $N(X\beta_2, \sigma^2 D_2^{-1})$ es (15)

$$\delta_3^2(\beta_1,\beta_2) = 2 \left(\sum_{i=1}^{2} \beta_i' X'D_i X\beta_i - \mu' A \mu\right)/\sigma^2$$

Con algo de esfuerzo se puede comprobar que si $D_1 = D_2$, entonces $\delta^2(\beta_1,\beta_2) = \delta_3^2(\beta_1,\beta_2)$. La estimación de esta distancia (al cuadrado) es

$$\delta_3^2(\beta_1,\beta_2) = 2(n...-2r)\left(\sum_{i=1}^{2} \bar{Y}_i' A_i \bar{Y}_i - \bar{Y}'A\bar{Y}\right) /$$

$$\left(\sum_{i=1}^{2} Y_i' Y_i - \bar{Y}_i' A_i \bar{Y}_i\right) \quad (16)$$

Por otra parte, el caso de observaciones faltantes, es decir, valores 0 en la diagonal de D_1 ó D_2, se puede abordar también modificando las expresiones anteriores convenientemente (poner 0 en la observación correspondiente, sustituir D_i^{-1} por D_i^-, etc.). (Véase /10/).

Finalmente, busquemos la distancia entre $N(X\beta_1, \sigma_1^2 I)$ y $N(X\beta_2, \sigma_2^2 I)$. Con las mismas notaciones de antes y suponiendo ahora $D_1 = D_2 = D$, consideremos las distribuciones $N(X\beta_1, D^{-1}\sigma_1^2)$, $N\beta_2, D^{-1}\sigma_2^2)$. Ahora las varianzas son distintas. Este caso es bastante más complejo, resultando que el tensor métrico no es constante, luego la variedad no es Euclídea. La distancia (al cuadrado) que se obtiene (ver /20/), es

$$\delta_4^2(\beta_1,\beta_2) = 2n \log^2 \frac{1+\sqrt{\Delta_{12}}}{1-\sqrt{\Delta_{12}}} \quad (17)$$

siendo

$$\Delta_{12} = \frac{\alpha_{12}^2 + 2\,(\sigma_2 - \sigma_1)^2}{\alpha_{12}^2 + 2\,(\sigma_2 + \sigma_1)^2}$$

$$\alpha_{12}^2 = (\beta_1-\beta_2)'\, X'DX\, (\beta_1-\beta_2) \,/\, n$$

Debido a que la variedad paramétrica en el caso $\sigma_1^2 = \sigma_2^2$ está incluida en la variedad paramétrica para $\sigma_1^2 \neq \sigma_2^2$, la distancia obtenida en el primer caso será mayor, es decir,

$$\delta^2\,(\beta_1,\beta_2) \geq \delta_4^2\,(\beta_1,\beta_2)$$

3. COMPARACION ENTRE MODELOS LINEALES NORMALES.

Sean los modelos lineales normales $N(X_1\beta_1, \sigma_1^2\, I)$, $N(X_2\beta_2, \sigma_2^2\, I)$, con la misma matriz de diseño reducida X. Diremos que ambos modelos son iguales si se verifica

$$X\beta_1 = X\beta_2 \qquad \sigma_1^2 = \sigma_2^2 \tag{18}$$

Supongamos que una distancia $d(\beta_1,\beta_2)$ es cero si se verifica (18), y positiva en caso contrario. Es razonable entonces plantear el contraste de hipótesis

$$H_o:\ d^2\,(\beta_1,\beta_2) = 0, \quad H_1:\ d^2\,(\beta_1,\beta_2) > 0 \tag{19}$$

La decisión de aceptar H_o para un cierto nivel de significación, significa aceptar (18), es decir, aceptar que la distribución de los vectores de observaciones Y_1, Y_2 es idéntica y por lo tanto tenemos el mismo modelo para ambos conjuntos de datos. Estudiaremos el contraste (19) bajo diferentes situaciones, donde d será alguna de las distancias definidas en (9), (15) ó (17).

3.1. MATRICES DE DISEÑO IDENTICAS.

Supongamos $\tilde{X}_1 = \tilde{X}_2 = X$, es decir $D_1 = D_2 = D$ y también que $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Expresando los dos modelos conjuntamente tendremos

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} \tilde{X} & 0 \\ 0 & \tilde{X} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} \tag{20}$$

Entonces la suma de cuadrados residual será

$$R_o^2 = \sum_{i=1}^{2} (Y_i'\, Y_i - \bar{Y}_i'\, DX\, (X'DX)^{-}\, X'\, D\, \bar{Y}_i) \tag{21}$$

con 2(n-r) g.l., donde X es la matriz de diseño reducida y r=ran(X).

La hipótesis H_o, que implica $X\beta_1 = X\beta_2$, podemos plantearla en términos del análisis de la varianza como la hipótesis lineal

$$H_o:\ (X, -X) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = 0 \tag{22}$$

sobre el modelo conjunto (20). Vemos que (22) es una hipótesis lineal de rango r que es equivalente a $\delta^2(\beta_1,\beta_2) = 0$ (ver (9)). Sea entonces $\bar{Y} = (\bar{Y}_1 + \bar{Y}_2)/2$. Bajo H_o, la distribución de $\bar{Y}$ es $N(X\beta, (2D)^{-1}\sigma^2)$ y la suma de cuadrados residual es

$$R_1^2 = Y_1'Y_1 + Y_2'Y_2 - \tfrac{1}{2}(\bar{Y}_1+\bar{Y}_2)'\,2DX(X'2DX)^{-}X'2D(\bar{Y}_1+\bar{Y}_2)/2$$

luego

$$R_1^2 - R_o^2 = (\bar{Y}_1-\bar{Y}_2)'\, DX\, (X'DX)^{-}\, X'D\, (\bar{Y}_1-\bar{Y}_2)/2$$

y por lo tanto, bajo la hipótesis nula

$$F = \frac{(R_1^2 - R_o^2)\,/\,r}{R_o^2\,/\,2(n-r)} \tag{23}$$

sigue la distribución F de Fisher-Snedecor con r y 2(n-r) g.l. En consecuencia, la significación de la distancia $\delta^2(\beta_1,\beta_2)$ se decide a través del estadístico

$$F = \frac{1}{2r}\,\hat{\delta}_2^2\,(\beta_1,\beta_2) \tag{24}$$

con distribución F con r y 2(n-r) g.l., cuando H_o es cierta. Obsérvese que se llega al mismo resultado obtenido por Cuadras (/8/), pero ahora a través de un enfoque geométrico.

3.2. MATRICES DE DISEÑO DISTINTAS

Supongamos $D_1 \neq D_2$, luego $\tilde{X}_1 \neq \tilde{X}_2$, aunque la matriz de diseño reducida X sea la misma para ambos modelos. Supongamos también $\sigma_1^2 = \sigma_2^2 = \sigma^2$. Planteamos ahora el contraste (19) pero utilizando la distancia (15). Empleando las notaciones (14), la suma de cuadrados residual es

$$R_o^2 = \sum_{i=1}^{2} (Y_i' Y_i - \bar{Y}_i' A_i \bar{Y}_i)$$

con (n..-2r) g.l.

Relacionemos ahora (22), es decir, la hipótesis $X\beta_1 = X\beta_2$, con $\delta_3^2(\beta_1,\beta_2)$. Indiquemos este valor común por $X\beta$. Según (14) resulta entonces

$$\mu = (D_1 + D_2)^{-1} (D_1 + D_2) X\beta = X\beta$$

$$\mu' A \mu = \beta' X (D_1+D_2) X (X' (D_1+D_2) X)^{-} X' (D_1+D_2) X\beta$$

$$= \beta' X (D_1+D_2) X\beta = \sum_{i=1}^{2} \beta' X D_i X\beta$$

Luego $X\beta_1 = X\beta_2 - X\beta$ implica $\delta_3^2(\beta_1,\beta_2) = 0$. Por otra parte, bajo esta hipótesis, la distribución de $\bar{Y} = (D_1+D_2)^{-1} (D_1 Y_1 + D_2 Y_2)$ es $N(X\beta, (D_1+D_2)^{-1}\sigma^2)$ y la suma de cuadrados residual es

$$R_1^2 = \sum_{i=1}^{2} Y_i' Y_i - \bar{Y}' A \bar{Y}$$

Tenemos entonces que

$$R_1^2 - R_o^2 = \sum_{i=1}^{2} \bar{Y}_i' A_i \bar{Y}_i - \bar{Y}' A \bar{Y}$$

y, bajo la hipótesis nula $\delta_3^2(\beta_1,\beta_2) = 0$, resulta que

$$F = \frac{(R_1^2 - R_o^2) / r}{R_o^2 / (n..-2r)} \qquad (25)$$

sigue la distribución F con r y (n.. -2r). Relacionando (16) con (25), podemos utilizar el estadístico

$$F = \frac{1}{2r} \hat{\delta}_3^2 (\beta_1,\beta_2) \qquad (26)$$

para decidir si aceptamos o rechazamos la hipótesis nula. Esta propiedad añade otra justificación a la definición de $\delta_3^2(\beta_1,\beta_2)$.

3.3. PROBLEMA DE BEHRENS-FISHER

Supongamos ahora $\sigma_1^2 \neq \sigma_2^2$. Nos limitaremos entonces al caso $\tilde{X}_1 = \tilde{X}_2$. Deseamos comparar los dos modelos, pero sin suposición alguna acerca de las varianzas. En realidad, vamos a plantear una comparación conjunta entre $(X\beta_1,\sigma_1^2)$ y $(X\beta_2\sigma_2^2)$ utilizando la distancia δ_4^2. En efecto, obsérvese que se verifican las igualdades $X\beta_1 = X\beta_2, \sigma_1^2 = \sigma_2^2$ si y sólo si

$$\delta_4^2 (\beta_1,\beta_2) = 0.$$

El contraste (19) tomando la distancia δ_4, podemos resolverlo a partir de un resultado obtenido por Oller /15/. Los estimadores máximo verosímiles de β_1,σ_i^2 son

$$\hat{\beta}_i = (X'DX)^{-} X' Y_i$$

$$\hat{\sigma}_i^2 = R_o^2 (i) / n =$$

$$= (Y_i' Y_i - \bar{Y}_i' DX (X'DX)^{-} X' D \bar{Y}_i) / n$$

Teniendo en cuenta (17) y sustituyendo parámetros por estimaciones

$$\hat{\alpha}_{12}^2 = (\hat{\beta}_1 - \hat{\beta}_2)' X' DX (\hat{\beta}_1 - \hat{\beta}_2) / n$$

$$= (\bar{Y}_1 - \bar{Y}_2)' DX (X'DX)^{-} X'D (\bar{Y}_1 - \bar{Y}_2)$$

$$\hat{\Delta}_{12} = \frac{\hat{\alpha}_{12}^2 + 2 (\hat{\sigma}_2 - \hat{\sigma}_1)^2}{\hat{\alpha}_{12}^2 + 2 (\hat{\sigma}_2 + \hat{\sigma}_1)^2}$$

la estimación de la distancia (al cuadrado) es

$$\hat{\delta}_4^2 (\beta_1,\beta_2) = 2n \log^2 \frac{1 + \sqrt{\hat{\Delta}_{12}}}{1 - \sqrt{\hat{\Delta}_{12}}}$$

Para decidir si esta distancia difiere significativamente de 0, aplicaremos la propiedad (/14/) de que bajo la hipótesis nula, la distribución de

$$U = \frac{n}{2} \hat{\delta}^2 (\beta_1,\beta_2)$$

puede aproximarse a una ji-cuadrado con (r+1) g.l., si el número de réplicas en todas las condiciones experimentales es suficientemente elevado.

4. DISTANCIA ENTRE MODELOS MULTIVARIANTES:

Consideremos ahora el modelo

$$Y = \tilde{X} B + E \qquad (27)$$

donde Y es una matriz de datos nxp, $\tilde{X}$ es una matriz de diseño nxm, B es una matriz de parámetros mxp, E tiene el mismo orden que Y. Las filas de Y contienen n observaciones p-dimensionales estocásticamente independientes. Cada observación p-dimensional se asimila a un vector aleatorio con distribución normal de

matriz de covarianzas constante Σ. Por otra parte, E contiene n filas estocásticamente independientes, cada una de ellas con distribución N $(0,\Sigma)$. En definitiva (27) representa el modelo general lineal del análisis multivariante de la varianza (MANOVA). Véase -- /24/ y /9/ .

Consideremos ahora dos modelos

$$Y_1 = \tilde{X}B_1 + E_1 \qquad Y_2 = \tilde{X}B_2 + E_2 \tag{28}$$

donde las matrices son todas del mismo orden. Ambos tienen la misma matriz de diseño. Definimos la siguiente distancia (al cuadrado) - entre los modelos (28) como sigue

$$L^2 = \text{tr} \quad (\Sigma^{-1} (B_1-B_2)' \tilde{X}'\tilde{X} (B_1-B_2)) \tag{29}$$

A continuación vamos a justificar esta definición. Para cada variable observable del modelos (27) tenemos el modelo univariante

$$y_i = \tilde{X} \beta_i + e_i \qquad 1 \leq i \leq p \tag{30}$$

donde y_i es la columna i-ésima de Y, etc. Consideremos entonces los siguientes vectores columna

$$y = (y_1',\ldots,y_p')', \qquad \beta = (\beta_1',\ldots, \beta_p')',$$

$$e = (e_1',\ldots,e_p')',$$

donde y, e son de orden np, β es de ordem mp. El modelo (27) puede expresarse ahora como

$$y = A \beta + e \tag{31}$$

siendo

$$A = I_p \otimes \tilde{X} \tag{32}$$

el producto de Kronecker de la matriz identidad I_p con la matriz $\tilde{X}$. Por otra parte, el -- vector aleatorio y sigue la distribución normal multivariante N $(A \beta, \Sigma_o)$ siendo

$$\Sigma_o = \Sigma \otimes I_n \tag{33}$$

Los dos modelos (28) pueden entonces asociarse a sendas distribuciones $N(A \beta_1,\Sigma_o)$, $N(A \beta_2,\Sigma_o)$. Si consideramos la variedad paramétrica de las distribuciones $N(\mu,\Sigma_o)$, con Σ_o constante, el tensor métrico es entonces

$$G = \Sigma_o^{-1} = \Sigma^{-1} \otimes I_n \tag{34}$$

Luego, procediendo análogamente a la definición (9), la distancia entre ambos modelos - es

$$\delta^2 (B_1,B_2) = (\beta_1-\beta_2)'A' (\Sigma^{-1} \otimes I_n) A (\beta_1-\beta_2) \tag{35}$$

Ahora bien, por las propiedades del producto de Kronecker

$$A'(\Sigma^{-1} \otimes I_n) A = (I_p \otimes \tilde{X}') (\Sigma^{-1} \otimes I_n) (I_p \otimes \tilde{X}) =$$
$$= (\Sigma^{-1} \otimes \tilde{X}') (I_p \otimes \tilde{X}) = (\Sigma^{-1} \otimes \tilde{X}'\tilde{X})$$

Como los elementos de esta matriz son

$$\sigma^{ij} \tilde{X} \tilde{X}' \qquad i,j = 1,\ldots, p$$

siendo $\Sigma^{-1} = (\sigma^{ij})$, desarrollando (35) obtenemos

$$\delta^2 (B_1,B_2) =$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{p} (\sigma^{ij}) (\beta_i{}^1-\beta_j{}^2) \tilde{X}'\tilde{X} (\beta_i{}^1- \beta_j{}^2) \tag{36}$$

donde $\beta_i^h (h=1,2)$ es el i-ésimo vector columna de $B_h (h=1,2)$. Relacionando (29) con (36) vemos que

$$L^2 = \delta^2 (B_1, B_2) \tag{37}$$

Luego la distancia L queda justificada por tratarse de una distancia geodésica entre - dos distribuciones normales multivariantes con matriz de covarianzas Σ_o común.

Se obtiene una estimación de L^2 sustituyendo en (29) los marámetros por sus estimaciones. Las estimaciones (en el sentido de los mínimos cuadrados) de B_1, B_2 son

$$\hat{B}_i = (\tilde{X}'\tilde{X})^- \tilde{X}' Y_i \qquad i = 1,2 \tag{38}$$

La matriz de dispersión residual para cada uno de los modelos (28) es

$$\mathbb{R}_o(i) = Y_i' (I-\tilde{X} (\tilde{X}'\tilde{X})^- \tilde{X}') Y_i \qquad i=1,2 \tag{39}$$

matriz que sigue la distribución de Wishart $W_p(\Sigma,n-r)$, siendo $r=\text{ran } (\tilde{X})$. La estimación conjunta de la matriz de covarianzas común Σ es

$$\hat{\Sigma} = \mathbb{R}_o /2(n-r) \tag{40}$$

siendo

$$\mathbb{R}_o = \mathbb{R}_o(1) + \mathbb{R}_o(2) \qquad (41)$$

una matriz que sigue la distribución de Wishart $W_p(\Sigma, 2(n-r))$. Sustituyendo en (29) - obtenemos

$$\hat{L}^2 = 2(n-r(tr[\mathbb{R}_o^{-1} (Y_1-Y_2)' \tilde{X} (\tilde{X}'\tilde{X})^- \tilde{X}' (Y_1-Y_2)] \qquad (42)$$

Podemos también utilizar la matriz de diseño reducida para obtener otra expresión equivalente a (42) que generalice (12). Indiquemos ahora por X la matriz de diseño reducida que contiene las k filas no repetidas de $\tilde{X}$, por $\bar{Y}$ la matriz kxp que contiene los p vectores columnas con las medidas muestrales (véase (5)), y sea de nuevo D la matriz (4). Entonces

$$\hat{L}^2 = 2(n-r)tr[\mathbb{R}_o^{-1} (\bar{Y}_1-\bar{Y}_2)'DX (X'DX)^- X'D (\bar{Y}_1-\bar{Y}_2)] \qquad (43)$$

Consideremos ahora el caso de que los dos modelos multivariantes (28) tengan matrices - de diseño distintas $\tilde{X}_1, \tilde{X}_2$, pero misma matriz de diseño reducida X (condición indispensable para que tenga sentido comparar ambos modelos). Entonces $D_1 \neq D_2$. Como en el caso - univariante, no existe una distancia geodésica entre ambos modelos, pues las distribuciones de Y_1, Y_2, están asociadas a variedades paramétricas distintas. Definamos entonces las matrices A_i, A como en (14), y las matrices

$$\bar{Y} = (D_1+D_2)^{-1} (D_1 \bar{Y}_1 + D_2 \bar{Y}_2)$$
$$M = (D_1+D_2)^{-1} (D_1XB_1 + D_2XB_2) \qquad (44)$$

Por definición, tomaremos como distancia - (al cuadrado), entre los modelos $Y_1 = \tilde{X}_1B_1+E_1$, $Y_2 = \tilde{X}_2B_2+E_2$, a la siguiente cantidad

$$L_1^2 = 2tr\left\{\Sigma^{-1} \left[\left(\sum_{i=1}^{2} B'_i X D_i X B_i\right)-M' A M\right]\right\} \qquad (45)$$

donde Σ es la matriz de covarianzas común a ambos modelos. Con algo de esfuerzo se puede probar que si $\tilde{X}_1 = \tilde{X}_2$, entonces $L = L_1$, y que $XB_1 = XB_2$ (ambos modelos son iguales) implica $L_1 = 0$. Por otra parte, la estimación de (45) es

$$\hat{L}_1^2 = 2(n..-2r)\ tr\left\{\mathbb{R}_o^{-1} \left[\left(\sum_{i=1}^{2} \bar{Y}_i' A_i \bar{Y}_i\right)-\bar{Y}'A\bar{Y}\right]\right\} \qquad (46)$$

Finalmente, la distancia entre dos modelos normales multivariantes con distinta matriz de covarianzas, es un problema actualmente no resuelto, por la sencilla razón de que - todavía no se conoce la distancia geodésica entre dos distribuciones normales $N(\mu_1,\Sigma_1)$, $N(\mu_2,\Sigma_2)$, con $\Sigma_1 \neq \Sigma_2$, aunque se han obtenido algunos resultados parciales (/16/, /3/, /6/). En otras palabras, no estamos en condiciones de generalizar (17) al caso multivariante.

5. COMPARACION ENTRE MODELOS MULTIVARIANTES.

Diremos que dos modelos lineales normales - multivariantes

$$Y_1 = \tilde{X}_1 B_1 + E_1 \qquad Y_2 = \tilde{X}_2 B_2 + E_2 \qquad (47)$$

con la misma matriz de diseño reducida X, - son iguales si

$$XB_1 = XB_2 \qquad (48)$$

Como en el caso univariante, plantearemos - el contraste de hipótesis

$$H_o: d^2(B_1, B_2) = 0 \qquad H_1: d^2(B_1, B_2) \neq 0 \qquad (49)$$

donde $d(B_1, B_2)$ es una distancia que vale - cero si se verifica (48). Estudiemos primero el caso $\tilde{X}_1 = \tilde{X}_2$. La matriz de dispersión residual conjunta para los modelos (47) es

$$\mathbb{R}_o = \sum_{i=1}^{2} (Y'_i Y_i - \bar{Y}_i' DX (X'DX)^- X'D \bar{Y}_i) \qquad (50)$$

cuya distribución es Wishart $W_p(\Sigma, 2(n-r))$. Si se verifica la hipótesis nula L = 0, es decir, si se verifica (48), entonces la matriz de dispersión residual es

$$\mathbb{R}_1 = \left(\sum_{i=1}^{2} Y'_i Y_i\right) - \frac{1}{2} (\bar{Y}_1+\bar{Y}_2)' DX (X'DX)^- X'D (\bar{Y}_1+\bar{Y}_2)$$

y por lo tanto, la matriz de la desviación de la hipótesis es

$$\mathbb{R}_1 - \mathbb{R}_o = (Y_1-Y_2)' DX (X'DX)^- X'D (Y_1-Y_2) / 2$$

Sean ahora $\lambda_1, \ldots, \lambda_n$ los valores propios de $\mathbb{R}_o$ respecto $\mathbb{R}_1$. Entonces, para decidir si se verifica (48), podemos utilizar el estadístico

$$V = tr[\mathbb{R}_o^{-1} (\mathbb{R}_1-\mathbb{R}_o)] = \sum_{i=1}^{n} (1-\lambda_i) / \lambda_i \qquad (51)$$

que se conoce como criterio de Lawley-Hotelling (véase /9/).

Comparando (51) con (43) obtenemos

$$V = \hat{L}^2 / 4 (n-r) \qquad (52)$$

La prueba de significación de L^2 se hace a través del criterio de Lawley-Hotelling, que en casos especiales es equivalente a la distribución F.

Finalmente, consideremos el caso $\tilde{X}_1 \neq \tilde{X}_2$, es decir, $D_1 \neq D_2$. Entonces, como en el caso univariante, se comprueba que la matriz de dispersión correspondiente a la desviación de la hipótesis es

$$R_1 - R_o = \sum_{i=1}^{2} \bar{Y}_i' A_i \bar{Y}_i - \bar{Y}' A \bar{Y} \qquad (53)$$

Relacionando (53) con (46) vemos que

$$V = \hat{L}_1^2 / 2 (n..-2r) \qquad (54)$$

y la significación de L_1 se comprueba también a través del criterio de Lawley-Hotelling. Para realizar esta prueba de significación se puede utilizar la aproximación (Seber /24/)

$$V \cong c F$$

donde F sigue una distribución $F_{a,b}$ siendo en este caso

$$a = dr \qquad b = 4+(a+2)/(B-1)$$
$$c = a(b-2)\ b\ (n..-2r-d-1)$$
$$B = (n..-r-d-1)(n..-2r)/(n..-2r-d-3)(n..2r-d)$$

6. CONCLUSIONES Y APLICACIONES.

La distancia entre modelos lineales normales se puede obtener enfocando el problema como distancia geodésica en una variedad paramétrica de distribuciones normales. La distancia resultante es doblemente interesante, pues es invariante, tanto por transformaciones admisibles de los parámetros, como de las variables aleatorias. La distancia ha sido estudiada bajo diferentes situaciones: idéntica matriz de diseño, número de réplicas distintas, varianzas diferemtes. En los dos primeros casos, la prueba de significación de las distancias obtenidas se resuelven mediante el test F. En el último caso, bajo ciertas condiciones, a través de una aproximación asintótica a la distribución jicuadrado.

La generalización al caso multivariante no presenta dificultad en los dos primeros casos. La prueba de significación nos lleva entonces al criterio de Lawley-Hottelling. Sin embargo, no está resuelto actualmente el caso de matrices de covarianzas distintas.

Las distancias obtenidas pueden aplicarse a la siguiente situación: sean $P_1, P_2, \ldots, P_q$ poblaciones estadísticas. Supongamos que los datos de una cierta variable observable se asimilan a una distribución $N(\tilde{X}_i\beta_i, \sigma_i^2 I)$ para la población p_i. Calculando entonces la matriz Δ de interdistancias entre las q poblaciones, podemos aplicar las técnicas usuales de análisis de datos multidimensionales: análisis de coordenadas principales, análisis de proximidades, análisis "cluster", etc., a fin de representar y clasificar las q poblaciones.

/22/, /11/, exponen un estudio sobre la población de N niños sospechosos de padecer Diabetes Melitus. A cada niño se le mide la variable observable "glucosa" para distintos valores de una variable independiente t (tiempo). Entonces se calculó la curva de glucemia ajustando "glucosa" a un polinomio de grado 3 en la variable t. Se realizó exactamente lo mismo para la variable "insulina", y para cada niño se obtuvo entonces un modelo normal bivariante. Calculando a continuación la matriz de distancias entre los N niños, y efectuando un análisis de coordenadas principales, junto con la obtención de un dendograma, se obtuvieron 4 grupos principales, que permitieron establecer grupos clínicos que fueron de utilidad para el diagnóstico de la diabetes.

7. REFERENCIAS.

/1/ AMARI, S. : "Differential-Geometrical Methods in Statistics". Lect. Notes in Statistics, 29, Springer-Verlag, Berlin (1985).

/2/ ATKINSON, C. y MITCHEL, A.F.S. : "Rao's distance measure". Sankhya, 43, A, 345-365 (1981).

/3/ BURBEA, J. : "Informative Geometry of Probability Spaces". Techn. Rep. N°.84-52, Center for Multivariate Analysis, U. of Pittsburgh. (1984).

/4/ BURBEA, J. y RAO, C.R.: "Entropy differential metric, distance and divergence measures in probability spaces: a unified approach". J. Multivariate Analysis, 12, 575-596. (1982).

/5/ BURBEA, J. y RAO, C.R.: "Differential metrics in probability spaces". Probability Math. Statist., Vol. 3, Fase 2, pp. 241-258. (1984).

/6/ CALVO, M. y OLLER, J.M. :"Métodos numéricos aplicados al cálculo de geodésicas entre distribuciones normales multivariantes" (en preparación). (1985).

/7/ CUADRAS, C.M.: "Análisis discriminante de funciones paramétricas estimables". Trab. Estad. Inv. Oper., 25 (3), 3-31. (1974).

/8/ CUADRAS, C.M. : "Sobre la comparació estadística de corbes experimentals" QUESTIIO 3 (1), 1-10.

/9/ CUADRAS, C.M.: "Métodos de Análisis Multivariante". Eunibar. Barcelona (1981).

/10/ CUADRAS, C.M.: "Diseños no balanceados y con observaciones faltantes en MANOVA". Actas XIII Jorn. Est. I. Op. vol. II, secc. III, 32-36, Valladolid. (1983).

/11/ CUADRAS, C.M., OLLER, J.M., ARCAS, A.y RIOS, M.: "Métodos geométricos de la Estadística". QUESTIIO, 9 (4) 219-250. (1985).

/12/ CUADRAS, C.M. y SANCHEZ, M.: "Un método de comparación entre dos curvas experimentales: una aplicación al estudio de parámetros conductuales". Rev. Psicol. Gen. Aplic., 32 (146), 441-452. (1977).

/13/ HILL, N.J. y PADMANABHAN, A.R.: "Robust comparison of two regression lineas and biomedical applications". Biometrics 40 (4), 985-994. (1984).

/14/ OLLER, J.M. : "Utilización de métricas riemannianas en análisis de datos multidimensionales y su aplicación a la Biologia." Pub. Bioest. y Biomat., n°11, D. de Bioestadística, U. de Barcelona. (1983).

/15/ OLLER, J.M. y CUADRAS, C.M.: "Defined distances for some probability distributions". Proceed. II World Conf. Math. Serv. Man.(A. Ballester, D. Cardús, E. Trillas, eds). U. Pol. Las Palmas, 563-565. (1982).

/16/ OLLER, J.M. y CUADRAS, C.M.: "Sobre una distancia definida para la distribución normal multivariante". Actas XIII, Jorn. Est. I.Op., vol. II, secc.III, 168-173, Valladolid. (1983).

/17/ OLLER, J.M. y CUADRAS, C.M.: "Rao's distance for negative multinomial distributions". Sankhya, A, 47 (1), 75-83.(1985a).

/18/ OLLER, J.M.: "Information metric for extreme value and logistic probability distributions". Sankhya, (en prensa).(1986).

/19/ OLLER, J.M. y CUADRAS, C.M.: "Sobre ciertas condiciones que deben verificar las distancias entre espacios probabilísticos". Actas XV, Reunión Nac. Est.I Op., U. de Oviedo (en prensa). (1985b).

/20/ OLLER, J.M. y RIOS, M.: "The information metric for univariate linear normal models". Sankhya. (en prensa). (1985b).

/21/ RAO, C.R.: "Information and the accuracy attainable in the estimation of statistical parameters". Bull. Calcutta Math. Soc., 17, 81-91. (1945).

/22/ RIOS, M.: "Métricas entre modelos lineales y su aplicación al tratamiento de datos en Medicina". Pub. Bioest. Biomat., n° 16, D. de Bioestadística, U. de Barcelona (1985).

/23/ SEBER, G.A.F.: "Linear Regression Analysis". Wiley, New York. (1977).

/24/ SEBER, G.A.F.: "Multivariable Observations". Wiley, New York. (1984).

/25/ TSUTAKAWA, R.K. y HEWETT, J.E.: "Comparison of two regression lines over a finite interval". Biometrics, 34, 391-398. (1978).