Revista Colombiana de Matemáticas
Volumen VIII (1974), págs. 29-33

PRUEBA PARA EL ERROR DE AJUSTE DE UN MODELO MULTIVARIANTE

por

Luis H. RODRÍGUEZ

RESUMEN

En el ajuste de un modelo a una serie de observaciones se presenta el interesante problema de decidir sobre lo adecuado del modelo para describir tales observaciones. Una prueba para esta clase de decisión se denomina "Error de ajuste". No conocíamos una tal prueba para el caso de un modelo multivariante (cada observación es un vector), por lo que este artículo hacemos una extensión de la técnica de "error de ajuste" utilizada en el análisis univariante al caso multivariante, y se produce una prueba de hipótesis basada en el máximo valor propio de una matriz aleatoria.

§ 1. *Introducción.* Cuando se ajusta un modelo a una serie de datos, es posible comparar los valores observados con aquellos calculados mediante el uso del modelo que se ajusta; lo anterior produce una serie de diferencias $(y_{ij} - \hat{y}_{ij})$ entre y_{ij}, el valor observado para el j-ésimo $(j = 1, \ldots, b)$ dato correspondiente al i-ésimo $(i = 1, \ldots, t)$ tratamiento, y el valor $\hat{y}_{ij}$, de este dato calculado mediante el modelo. Sea n_R el número total de tales diferencias menos el número p de parámetros en el modelo ajustado. Entonces la cantidad $\sum_i \sum_j (y_{ij} - \hat{y}_{ij})^2$, se llama la *suma residual de cuadrados* (SRC). Esta suma de cuadrados se halla compuesta por dos factores, a saber : la *suma de cuadrados de error puro* (SEP), causado por la variación

y_{ij} de las observaciones con relación a su media, y *la suma de error de ajuste* (SEA), causado por el efecto del modelo que se ajusta. En el caso de llevar a cabo un submuestreo para unidades sometidas a un mismo tratamiento, la hipótesis referente a la significación de SEA puede someterse a prueba mediante el cociente entre el error puro y el error de ajuste, o sea :

$$(1.1) \qquad \frac{\sum_i \sum_j (y_{ij}-y_{i\cdot})^2 / n_E}{[\sum_i \sum_j (y_{ij}-\hat{y}_{ij})^2/n - \sum_i \sum_j (y_{ij}-y_{i\cdot})^2]/(n_R-n_E)}$$

donde $n_E = t(b-1)$, lo cual conduce a una prueba de hipótesis que utiliza la distribución F de Snedecor para la región crítica [1].

Para el caso del análisis multivariante, si el submuestreo se ha llevado a cabo, esta prueba no puede efectuarse con una estadística similar a (1.1), puesto que ahora se trata de ajustar no un único modelo sino más bien un modelo para cada una de las componentes del vector observado. La hipótesis para el error de ajuste estará compuesta de tantas subhipótesis como modelos se ajusten. Para reducir el problema multivariante a una forma tratable se puede utilizar la técnica de prueba de hipótesis conocida como el *principio de unión - intersección debida a Roy* [2].

§ 2. *Método de prueba de hipótesis.* La situación del caso multivariante puede resumirse en la notación siguiente : sea,

$$(2.1) \qquad \vec{y}' = (y_{ij1}, y_{ij2}, \ldots, y_{ijp}), \qquad \begin{matrix} i = 1, \ldots, b \\ j = 1, \ldots, t \\ k = 1, \ldots, p \end{matrix}$$

el vector $(1 \times p)$ cuya k-ésima componente es la observación hecha en la j-ésima unidad de la i-ésima "población".

Sea A una matriz cuya i-ésima fila es el vector $\vec{a}'_i$. Entonces el modelo mul-

tivariante puede ser representado así :

$$Y = A^{*} \Xi + E \tag{2.2}$$

donde Y es una matriz de observaciones $(bt \times p)$; A^{*} es una matriz de diseño $A^{*} = (A \otimes \vec{j})$ donde $\vec{y} = (1, 1, \ldots, 1)$; el rango de A es $m < bt$; Ξ es una matriz $(r \times p)$ de parámetros, donde r es el número de parámetros en el modelo ; y E es una matriz $(bt \times p)$ de errores aleatorios con valor esperado 0 y matriz de covarianza Σ .

Asumiendo que todas las unidades provenientes de una misma población tienen el mismo modelo (entiéndase por "población" el conjunto de observaciones correspondientes al i-ésimo tratamiento) se tiene que :

$$Y = \begin{bmatrix} \vec{y}'_{11} \\ \vec{y}'_{1t} \\ \vec{y}'_{b1} \\ \vec{y}'_{bt} \end{bmatrix} \tag{2.3}$$

donde cada uno de los vectores $\vec{y}_{ij}$ se distribuye como una normal multivariante con media $\vec{a}'_i \Xi$ y matriz de covarianza Σ , es decir, $\vec{y}_{ij} \sim MVN (a'_i \Xi, \Sigma)$ y además son independientes por razón del muestreo. La hipótesis para el error de ajuste se plantea en la forma siguiente :

$$\begin{aligned} H_o &: E(Y) = (A \otimes \vec{j}_t) \\ H_1 &: E(Y) = (I_b \otimes \vec{j}_t) D \end{aligned} \tag{2.4}$$

donde D es una matriz $(b \times p)$ de parámetros bajo la hipótesis alternativa, $\otimes$ es el producto directo de matrices, ie. , $A \otimes B = (Ab_{ij})$.

El principio de unión-intersección de Roy [2] , puede entonces ser aplicado y da como resultado las siguientes hipótesis, equivalentes a (2.4) :

$$\cap\ H_{o\vec{l}}: \quad \{E(Y\vec{l}) = (A \otimes \vec{j}_t)\ \vec{l}\}$$

(2.5)

$$\cup\ H_{1\vec{l}}: \cup\{E(Y\vec{l}) = (I_b \otimes \vec{j}_t)\, D\, \vec{l}\}\ ;$$

donde el vector $\vec{l} \neq \vec{0}$; la unión y la intersección se consideran con respecto a $\vec{l}$. Es claro que de ser aceptada la hipótesis $H_{o\vec{l}}$ de (2.5) para todo $\vec{l}$, también lo será H_o de (2.4). Si aquella no es aceptada para algún $\vec{l}$ entonces H_o de (2.4) se rechaza.

TEOREMA 2.1. Sea $\vec{l}' S_E \vec{l}$, *una forma cuadrática que corresponde a la matriz de covarianza de error puro. Sea* $\vec{l}' S_C \vec{l}$; *una forma cuadrática que corresponde a la matriz de covarianza de errores residuales. Entonces* $\vec{l}' S_H \vec{l}$, *la forma cuadrática de la matriz de covarianzas debida a errores de ajuste, se puede expresar como:*

$$\vec{l}' S_H \vec{l} = \vec{l}' Y' [(I_b - AA^-) \otimes (1/t)\, J_t]\, Y \vec{l}$$

donde A^- *denota el inverso de Penrose de la matriz* A *y* J_t *una matriz cuyos elementos son todos iguales a uno.*

TEOREMA 2.2. S_H *y* S_E , *se distribuyen independientes como distribuciones Wishart centrales, con grados de libertad respectivamente* $(b-m)$ *y* $b(t-1)$.

Una prueba de estos dos teoremas se halla en [3] .

Como quiera que la hipótesis nula de (2.5) es aceptada si para todo $\vec{l} \neq 0$ es aceptada, entonces el principio de unión-intersección de Roy[2] permite establecer como regla de rechazo para H_o de (2.4) la estadística :

$\underset{\vec{l}}{Sup}\ \frac{\vec{l}' S_H \vec{l}}{\vec{l}' S_E \vec{l}} = C_1$ = valor propio máximo de la matriz $S_H S_E^{-1}$. Tablas de la distribución del máximo valor propio de una matriz aleatoria se encuentran en Heck [4].

3. *Conclusiones.* En el análisis multivariante de la prueba de error de ajuste de un modelo, el principio de unión-intersección de Roy permite transformar la hipótesis original en forma tal que la estadística para llevar a cabo la prueba de hipótesis correspondiente es el máximo valor propio de una matriz aleatoria. La distribución de esta estadística es conocida y ha sido tabulada por varios autores. Aunque los resultados expuestos en este artículo son válidos para todo modelo que reuna las condiciones establecidas, la experiencia personal del autor le indica que el caso de mayor aplicación de esta prueba está en los casos de regresión múltiple, en donde el submuestreo a que se hace referencia es más utilizado.

BIBLIOGRAFÍA

1. *Draper, N. R. and Smith, H., Applied Regression Analysis,* John Wiley & Son Inc. New York, 1966.
2. *Roy, S. N., On a heuristic method of test construction and its use in multivariate analysis,* Ann. Math. Statist., 24 (1953), 220 - 238.
3. *Rodríguez, L. H., Moment generating function and moments of a Wishart matrix with related problems in Manova,* Ph.D. Dissertation, Kansas State University, 1973.
4. *Heck, D. L., Chart of some upper percentage points of the distribution of the largest characteristic root,* Ann. Math. Statist., 31 (1960), 625 - 642.

Departamento de Matemáticas y Estadística
Universidad Nacional de Colombia
Bogotá, D. E., Colombia, S. A.

(Recibido en octubre de 1973)