APLICACION DEL ANALISIS DE CORRESPONDENCIA EN LA EVALUACION BLUP DE REPRODUCTORES PARA CARACTERES DISCRETOS

E.O. ROMANO (*)

Recibido:12-04-89
Aceptado:27-07-89

RESUMEN

La evaluación del mérito genético de toros para caracteres discretos usando metodología BLUP presenta la limitante que el carácter no se distribuye normalmente. La asignación de valores sucesivos a cada categoría del carácter discreto conlleva gran arbitrariedad. Se presenta un método para trabajar con caracteres no cuantitativos basado en el Análisis de Correspondencia. Mediante el análisis de correspondencia se propone una asignación menos arbitraria. La asignación de valores es tal que se maximiza la relación entre el carácter de interés y otro carácter complementario. La menor arbitrariedad se asegura si el carácter complementario muestra una conocida y estrecha relación con el carácter de interés. Se discuten las ventajas y limitaciones del método. Se presenta un ejemplo en el cual se asignan valores numéricos a cinco categorías del carácter dificultad de parto.

Palabras clave: Mérito genético, caracteres discretos, dificultad de parto, análisis de correspondencia.

USE OF THE ANALYSIS OF CORRESPONDANCE IN THE SIRE BLUP EVALUATION FOR CATEGORICAL TRAITS

SUMMARY

The categorical trait sire evaluation by BLUP has the limitation that the trait does not follow a normal distribution. The assignation of succesive values to each level of the categorical trait presents a great arbitrarity. A method for dealing with nonquantitative traits is presented which is based on the Analysis of Correspondance. The idea is to assign numeric values to each category of a categorical trait in a less arbitrary way than the usually adopted. The assignation of values would maximize the relationship with the trait of interest. The advantages and limitations of the method are discussed. An example is presented where numeric values are assigned to each of five categories of calving difficulty.

Key words: Genetic value, categorical trait, analysis of correspondance, calving difficulty.

INTRODUCCION

La estimación del mérito genético de toros mediante la aplicación de metodologías **BLUP** "Best Linear Unbiased Predictor" (Mejor Predicción Lineal Insesgada) ha recibido una aceptación generalizada en todo el mundo, en especial a través de la solución de modelos mixtos de Henderson. Tal metodología brindó la posibilidad de lograr ordenamientos ("rankings") de animales

(*)Departamento de Zootecnia, Facultad de Agronomía. UBA. Avda. San Martín 4453 (1417) Buenos Aires - Argentina -

de acuerdo a los respectivos valores genéticos, y de una forma mucho más precisa que lo que otras técnicas permitían.

La extraordinaria utilidad de la metodología **BLUP** ha hecho que los investigadores y técnicos desearan utilizar su esquema de análisis aún en aquellos casos en los cuales los supuestos en los que se basa no sean absolutamente satisfechos. Los estimadores que son obtenidos con la metodología de Henderson son **BLUP** si se cumple entre otros el supesto que el o los caracteres por los cuales se tiene interés evaluar siguen una distribución normal.

Es claro que aquellos caracteres que siguen una distribución discreta no satisfacen este supuesto. El deseo de tratar estos caracteres con la metodología **BLUP** ha sido considerado de diferentes maneras. El procedimiento más usado ha sido el de asignar arbitrariamente un valor a cada categoría para luego emplear esos valores como observaciones (Schaeffer y Wilton, 1976). La inconveniencia de este proceder se visualiza rápidamente cuando el carácter discreto no es ordinal. Esto es: cuando no se puede asignar ningún orden de preferencia para cada nivel del carácter.

En caracteres discretos ordinales como dificultad al parto, o estado corporal, resulta muy simple establecer el orden de preferencia de cada nivel. Sin embargo subsiste el problema de la asignación de los valores. Suponiendo cinco niveles de dificultad al parto, ¿ deberían asignarse los valores correlativamente del 1 al 5 ?, ¿ del 10 al 50 ?, ¿ del 342.000 al millón ?, ¿ los niveles deberían estar separados a igual distancia ?, ¿ a diferente ?. De la decisión va a depender la estimación de los parámetros poblacionales. Al no existir garantía que la asignación sucesiva de valores (1 al 5 en este caso) sea la correcta la evaluación de reproductores mediante su utilización es absolutamente arbitraria, con la excepción del respeto por la ordinalidad del carácter. Gianola (1980) resume los inconvenientes de todos estos procedimientos, entre ellos: la estimación de la heredabilidad depende de la asignación de valores, de forma tal que un conjunto de valores puede resultar "más heredable" que otro. Quartermain y Freeman (1967) propusieron un escalamiento que maximiza la heredabilidad de las diferencias genéticas.

Quass y Van Vleck (1980) desarrollaron una metodología que obtiene estimadores **BLUP** evitando la asignación de valores. Gianola y Foulley (1983) y Harville y Mee (1984), trabajando separadamente, presentaron un nuevo enfoque que piensa al carácter discreto como una manifestación de un carácter contínuo subyacente. Esta idea en si no es nueva (Snell, 1964; Cox, 1974), pero se encuadra dentro del esquema bayesiano. Este enfoque, si bien está teniendo una creciente aceptación en Estados Unidos y Alemania (Hoeschele, I., comunicación personal) presenta un incremento de los requerimientos computacionales con respecto a los métodos tradicionales, a la vez que ciertos supuestos no siempre fáciles de validar (Fuolley **et al**, 1987).

El objetivo del presente trabajo es el presentar un método de tratar los caracteres discretos que, si bien basado en una asignación de valores,haga que la misma sea menos arbitraria que la usual, presentando a su vez un esquema de trabajo que disminuya las necesidades computacionales del "thereshold model" y que eluda la postulación de ciertos supuestos. La idea central es la de utilizar el análisis de correspondencia como una forma de escalamiento (Hill, 1974; Gómez Riera, 1985) de forma tal que los nuevos valores asignados maximicen la relación entre el carácter en cuestión y algún otro carácter determinado.

La menor arbitrariedad se aseguraría de ser posible afirmar la existencia de una correlación genética no nula entre ambos caracteres. Se presenta un ejemplo demostrativo.

DESCRIPCION DEL METODO

El primer paso en el análisis consiste en la elaboración de una tabla de contingencia que especifique las frecuencias de ocurrencia simultánea de los dos caracteres, el de interés y el secundario, estrechamente relacionados en forma genética.

Sea el vector $x = (x1,...,xn)$ conformado por cada uno de los n valores que corresponden a uno de los dos caracteres en cuestión. Sea el vector $y = (y_1,...,y_m)$ el correspondiente al otro carácter. Sea $A = (n_{ij})$ la matriz m x n conformada por las frecuencias de ocurrencia de las m x n celdas correspondientes.

Se pueden definir además dos funciones: $f(i)=x$ y $g(j)=y$. El análisis consiste en hallar a partir de los datos, es decir a partir de la matriz A, los valores x_i e y_j; para i de 1 a n, y j de 1 a m, que maximicen la correlación entre f(i) y g(j).

El análisis de correspondencia cumple con estas condiciones. Se puede establecer que el triplete (δ, x, y) es una solución de orden cero del análisis de correspondencia de la matriz A, Co (A), si sólo si:

$$(F^{1/2}\, x) = (F^{-1/2}\, A\, C^{-1/2})\,(C^{1/2}\, y) \qquad (1)$$

$$(C^{1/2}\, y) = (F^{-1/2}\, A\, C^{-1/2})'\,(F^{1/2}\, x) \qquad (2)$$

Siendo:

$F^{1/2} = \text{Diag}(\sqrt{n_{i.}})$, es decir la matriz diagonal compuesta por las raíces cuadradas de los totales de fila.

$C^{1/2} = \text{Diag}(\sqrt{n_{.j}})$, la matriz diagonal compuesta por las raíces cuadradas de los totales de cada columna.

δ^2 = Autovalor de la matriz

$$(F^{-1/2}\, A\, C^{-1/2})\,(F^{-1/2}\, A\, C^{-1/2})'$$ y de la matriz

$$(F^{-1/2}\, A\, C^{-1/2})'\,(F^{-1/2}\, A\, C^{-1/2}).$$

Esto se ve a partir de (1) y (2), ya que:

$$\delta^2(F^{1/2}x) = (F^{-1/2}AC^{-1/2})\,(F^{-1/2}AC^{-1/2})'\,(F^{1/2}x) \qquad (3)$$

$$\delta^2(C^{1/2}y) = (F^{-1/2}AC^{-1/2})'\,(F^{-1/2}AC^{-1/2})\,(C^{1/2}y) \qquad (4)$$

Si se reescribieran (3) y (4):

$$\delta^2\, x^* = BB'\, x^* \qquad (5)$$

$$\delta^2\, y^* = B'B\, y^* \qquad (6)$$

Se visualiza más claramente que es un autovalor de BB' o bien de B'B, siendo x^*, y^* sus autovectores, respectivamente.

Ahora bien, para el análisis de correspondencia, la primera solución es trivial, cumpliendo solamente una función de centrado. Esto indica que el primer autovalor hallado siempre será igual a 1.

Definiendo cual es el sistema de interés, si (5) ó (6), obteniendo el autovector correspondiente al primer autovalor no trivial, y transformándolo nuevamente, se hallará el escalamiento buscado.

Es posible que el escalamiento hallado consista de cifras poco usuales y por lo tanto poco manejables y de más dificultosa percepción. Ante esto, Gómez Riera (1985) propone una trasformación simple del vector de códigos hallado, consistente en multiplicarlo por un escalar que redondee las cifras y facilite su visualización. El autor considera que este tipo de transformación puede hacer perder la propiedad original de máxima correlación entre ambos caracteres, aunque concede que su empleo facilitaría la aceptación del método por clientes potenciales, siendo de todas formas un escalamiento menos arbitrario que el usual. Las posibles consecuencias de estas transformaciones, así como sus ventajas y desventajas requerirían de una futura investigación.

En síntesis:

1. Defínase cuál es su carácter de interés, si x ó y.

2. Si el mismo resultó **x**, calcúlese **BB'**. Si fuese **y**, calcúlese **B'B**.

3. Obténgase los autovalores, que serán los mismos en ambos casos. Una verificación del proceso hasta este momento es que el primer autovalor sea igual a 1.

4. Eliminando el autovector correspondiente al autovalor trivial identifique al siguiente autovector, el cual será x^* si se está trabajando con (5) o bien y^* si se está con (6).

5. Para obtener el escalamiento x deseado, premultipique x^* por $F^{-\frac{1}{2}}$, ya que $x^* = F^{\frac{1}{2}} x$. Si se desea escalar y, premultiplique y^* por $C^{-\frac{1}{2}}$, ya que $y^* = C^{\frac{1}{2}} y$.

6. De ser necesario, redondee las cifras obtenidas.

En función aclaratoria, se incluye un ejemplo, el cual fue procesado utilizando el **PROC IML** del paquete **SAS** instalado en el Departamento de Estadística del INTA en Castelar.

DISCUSION SOBRE LA IMPLEMENTACION DEL METODO

Ventajas y Limitaciones

Resulta evidente la ventaja de poder usar un escalonamiento menos arbitrario que el usual. La cuestión remanente es la practicidad de esta metodología.

Dos son las posibles formas de implementación más generales. Una es directamente antes de cada evaluación, para cada conjunto de datos. La otra podría efectuarse una única vez por algún organismo oficial, para que luego los interesados utilicen los códigos calculados. La discusión de estas dos implementaciones dará idea de los alcances del método así como de las futuras investigaciones para su ajuste.

El efectuar una codificación inicial requiere para su éxito que su resultado sea confiable por los usuarios futuros. Para ello debería contar con una gran cantidad de datos, lo que obviamente mejoraría la precisión de la evaluación con respecto a las individuales. Este sistema además no elevaría la complejidad de la usual estimación por parte de los futuros usuarios, dado que la misma se desarrollaría siguiendo los esquemas tradicionales.

Aunque la implementación individual de este método no cuenta con las ventajas de precisión por la cantidad de datos y requiere una mayor complejidad computacional a cada uno de los usuarios, a su vez no cuenta con la necesidad de estudiar los siguientes factores a tener en cuenta cuando se hace una codificación general: acción de diferentes razas, edades de las madres, zonas, establecimientos, estación, etc. Muy posiblemente se podrían efectuar diferentes subdivisiones de los datos y realizar codificaciones para diferentes zonas, razas, etc. Futuras investigaciones aclararían la necesidad de hacer estas subdivisiones, así como la de incluir o no diversos factores.

De todas formas, la posibilidad de realizar evaluaciones utilizando codificaciones menos arbitrarias que las usuales merecerían ser tenidas en cuenta, al menos para aquellos usuarios que ulicen una implementación del tipo individual.

Finalmente se brinda al lector un ejemplo de codificación seguido de un ejemplo de evaluación del valor de cría siguiendo una codificación usual y la aquí presentada.

AGRADECIMIENTO

El autor agradece al Departamento de Estadística del INTA de Castelar por su apoyo en el "software" empleado, así como al Ing. Eduardo Manfredi por sus aclaraciones y comentarios.

UN EJEMPLO DE ESCALAMIENTO

Supóngase que se desea dar valores a cinco categorías de dificultad al parto, con el objeto de utilizarlas posteriormente en la estimación **BLUP** del valor de cría de reproductores para este carácter. Se cuenta con una subdivisión del carácter en cinco niveles, deseándose en consecuencia codificar cada nivel para obtener una estimación menos sesgada del valor de cría.

Conociendo la estrecha relación entre dificultad al parto y peso al nacer, se decidió la construcción de una tabla de incidencia que involucre a ambos caracteres. Dado que se tomó el criterio de dividir los pesos también en cinco grupos, la matriz **A** será cuadrada y de orden 5x5.

Sea:

$$A = \begin{bmatrix} 0 & 2 & 2 & 3 & 40 \\ 17 & 15 & 15 & 17 & 18 \\ 18 & 30 & 18 & 15 & 10 \\ 35 & 27 & 23 & 18 & 13 \\ 40 & 20 & 18 & 6 & 15 \end{bmatrix}$$

con sus totales de fila: **totfila** = (47 82 91 116 99)'

y de columna: **totcol** = (110 94 76 59 96)

Así $F^{\frac{1}{2}}$ = diag. (6,86 9,05 9,54 10,77 9,95)

$C^{\frac{1}{2}}$ = diag. (10,49 9,69 8,72 7,68 9,80)

Supóngase ahora que el carácter de interés, dificultad al parto, se encuentra representando en el vector **y**. Debido a esto, va a interesar obtener la matriz **B'B**.

Así B'B es:

$$\begin{bmatrix} 0,3073345 & 0,2485223 & 0,2283920 & 0,1780860 & 0,1527099 \\ 0,2485223 & 0,2451490 & 0,2100382 & 0,1824087 & 0,1510366 \\ 0,2283920 & 0,2100382 & 0,1871384 & 0,1622445 & 0,1437389 \\ 0,1780860 & 0,1824087 & 0,1622445 & 0,1583923 & 0,1442951 \\ 0,1527099 & 0,1510366 & 0,1437389 & 0,1442951 & 0,4460656 \end{bmatrix}$$

Los autovalores de B'B son:

1,0000000 0,2928260 0,0415981 0,0096556 0,0000001

Los autovectores:

0,5028654	0,3357323	-0,709997	0,2241825	-0,282944
0,5658569	0,2590482	0,3072643	-0,753310	-0,234337
0,4179864	0,2038313	0,0682686	0,0825010	0,8787907
0,3682827	0,1013848	0,6230145	0,6109488	-0,304440
0,4697762	-0,876557	-0,093198	-0,046912	-0,008486

De acuerdo a que el primer autovalor es trivial, se considera el segundo, con su correspondiente autovector. Este será el y^*. Para obtener el **y** buscado, se premultiplica y^* por $C^{-\frac{1}{2}}$ obteniéndose:

$$y+ = (0,0320108 \quad 0,0267188 \quad 0,0233811 \quad 0,0131992 \quad -0,089463)'$$

Este **y+** sería el vector con los escalamientos sin redondear. Si se deseara contar con cifras más concisas, un posible escalamiento sería:

$$y = (32 \quad 27 \quad 23 \quad 13 \quad -8)$$

UN EJEMPLO DE OBTENCION DE ESTIMADORES BLUP CON EL ESCALAMIENTO USUAL Y EL AQUI PRESENTADO

Supóngase que se desea obtener estimadores **BLUP** del valor de cría de animales, utilizando el siguiente modelo muy sencillo:

$$y_{ij} = u + s_i + e_{ij}$$

donde: **u** es una constante desconocida, s_i es una variable aleatoria no observable (efecto toro), e_{ij} es el error residual, $E(y_{ij}) = u$, $E(s_i) = 0$, $E(e_{ij}) = 0$, $V(e) = 1\sigma^2 e$, $V(s) = 1\sigma^2 s$, $Cov(e,s) = 0$, $h^2 = a\sigma^2 s / (\sigma^2 s + \sigma^2 e) \Rightarrow \sigma^2 e = 15\sigma^2 s$.

Sólo se consideran efectos aditivos.

Supóngase además que se cuentan con los datos que figuran en el Cuadro N° 1. Son 4 toros (**A, B, C, D**), con información sobre sus respectivos números de hijos y de su distribución según la dificultad de parto de su progenie.

Cuadro N° 1

TORO	NUMERO DE HIJOS	CLASE DE DIFICULTAD AL PARTO				
		I	II	III	IV	V
A	10	5	2	1	2	0
B	25	9	5	5	3	2
C	50	12	10	10	8	10
D	100	50	25	15	7	3

En consecuencia las ecuaciones normales se establecerán de la siguiente forma:

$$\begin{bmatrix} 185 & 10 & 25 & 50 & 100 \\ 10 & 25 & 0 & 0 & 0 \\ 25 & 0 & 40 & 0 & 0 \\ 50 & 0 & 0 & 65 & 0 \\ 100 & 0 & 0 & 0 & 115 \end{bmatrix} \begin{bmatrix} u \\ sA \\ sB \\ sC \\ sD \end{bmatrix} = \begin{bmatrix} 411 & 4337 \\ 20 & 119 \\ 59 & ,661 \\ 144 & 898 \\ 188 & 2659 \end{bmatrix}$$

Los dos vectores correspondientes al lado derecho de las ecuaciones representan los obtenidos mediante un escalonamiento usual (**1 a 5**) de la dificultad al parto, y mediante el escalonamiento obtenido del ejemplo anterior, respectivamente.

La solución de ambos sistemas de ecuaciones da por resultado a:

b1' = (2,299 -0,120 0,038 0,447 -0,365)
b2' = (21,856 -3,982 2,864 -0,2998 4,116)

Por lo tanto, si se clasificarán los animales por su capacidad para transmitir una menor dificultad al parto, los "rankings" serían:

Escalonamiento Usual: **D A B C**
Escalonamiento Nuevo: **C A B D**

Lo que aparece en consecuencia como un distinto ordenamiento de los animales. Esta es una situación previsible y que debería llevar a la reflexión acerca de como estimar valores de cría de caracteres discretos de la forma menos arbitraria posible.

BIBLIOGRAFIA

1) COX, N.R. (1974). Estimation of the correlation between a continous and discrete variable. **Biometrics,** 30:171-178.

2) FOULLEY, J.L.; D. GIANOLA and I. HOSCHELE, (1987). Empirical Bayes estimation of parameters for n polygenic binary traits. **Génét. Sél. Evol.**, 19(2):197-224.

3) GIANOLA, D. (1980). A method of sire evaluation for dichotomies. **J. of An. Sci.**, 51:1266-1271.

4) GIANOLA, D. and J.L. FOULLEY, (1983). Sire evaluation for ordered categorical data with a threshold model. **Génét. Sél. Evol.**, 15:201-224.

5) GOMEZ RIERA, P. (1985). Análisis de correspondencia: Su uso para la cuatificación de variables categóricas. **INTA Reg. Cuyo.** s/n. 7p.

6) HARVILLE, D.A. and R.W. MEE. (1984). A mixed model procedure for analyzing ordered categorical data. **Biometrics,** 40:393-401.

7) HILL, M.O. (1974). Correspondence Analysis: A Neglected Multivariate Method. **Appl. Statist.**, 23:340-354.

8) QUAAS, R.L. and L.D. VAN VLECK, (1980). Categorical trait sire evaluation by Best Linear Unbiased Prediction of future progeny category frequencies. **Biometrics,** 36:117-122.

9) QUARTERMAIN, A.R. and A.E. FREEMAN, (1967). Some transformetions of scale and the estimation of genetic parameters from daughter-dam regression. **Biometrics,** 23:823-833.

10) SHAEFFER, L. and J.W. WILTON. (1976). Methods of sire evaluation for calving ease. **J. of Dairy Sciencie,** 59:541-544.

11) SNELL, E.J. (1964). A scaling procedure for ordered categorical data. **Biometrics,** 20:592-607.