Revista Colombiana de Estadística
Nº 9 - 1984

EL MODELO LINEAL EN EXPERIMENTOS CON VARIABLES CATEGORIZADAS

ESTUDIO DE UN CASO

Luis Alberto López Pérez
Depto. Mat. y Estadística
Universidad Nacional

Bernardo Chaves Córdoba
Instituto Colombiano
Agropecuario - ICA.

Resumen. Se pretende presentar y aplicar la metodología estadística propuesta por Grizzle, Starmer and Koch (1969), basada en el modelo lineal, la cual es de gran utilidad para el análisis de experimentos donde la respuesta a medir hace referencia al número de "unidades" que caen dentro de cada una de las t celdas (categorías); es decir, experimentos en los cuales la variable de interés sigue una distribución multinomial. Para dar una aplicación a esta metodología, se tuvo en cuenta los resultados de un experimento

donde se evalúa el grado de pudrición de mazorcas en la variedad de maíz sogamoseño al ataque del hongo *Fusarium, s.p.*

Introducción.

El uso de variables categorizadas es frecuente en estudios fitopatológicos, médico-veterinarios, epidemiológicos, de mejoramiento genético, etc., siendo frecuente el empleo de la estadística Ji-cuadrada para probar hipótesis de independencia. Sin embargo cuando se desea probar hipótesis acerca de combinaciones lineales de efectos de tratamientos donde las respuestas son de tipo categórico, el uso del estadístico Ji-cuadrado tradicional es limitado. Este hecho conduce a presentar esta metodología estadística basada en la teoría de los modelos lineales, aplicándola al caso particular del grado de pudrición de mazorca en la variedad de maíz sogamoseño.

Revisión de antecedentes.

En el estudio de datos categóricos, frecuentemente se encuentran distribuciones estadísticas como la Ji-cuadrado, que permite realizar pruebas de hipótesis de independencia entre dos

o mas variables medidas en una escala al menos nominal.

El modelo probabilístico que sigue a una variable categorizada es c :ocido en la teoría estadística como distribución multinomial con parámetros $(N, \pi_j; j = 1,2,\ldots,k)$ siendo π_j la probabilidad (la cual se supone constante) de que una unidad particular pertenezca a la j-ésima categoría. Es frecuente que dentro de las clasificaciones originales, se presentan subcategorías dando esto origen a una estructura multinomial multivariada, Johnson y Kotz (1969).

Generalmente se tiene interés en establecer que tipo de relación existe entre dos o más variables categorizadas, la prueba estadística comunmente utilizada es la Ji-cuadrada, cuya expresión es:

$$x^2 = \sum \frac{(O-E)^2}{E} \qquad (1)$$

Siendo O la frecuencia observada -número de unidades que pertenecen a cada celda- y E la frecuencia esperada. Es frecuente también hacer uso del estadístico Ji-cuadrado para pruebas de bondad de ajuste; tiene su base en que para muestras grandes la forma cuadrática

$$Q = (Y - \mu)' \sum_y^{-1} (Y - \mu) \qquad (2)$$

tiene una distribución Ji-cuadrado con $(r-1)$ grados de libertad (Chernoff 1956), citado por Lindeman et al (1976), en donde μ es el vector de medias en la distribución multinomial y $\sum_y$, la matriz de varianzas y covarianzas asociada a esta distribución.

Hay varios métodos que permiten estimar la probabilidad de que una observación particular pertenezca a cada una de las r-categorías, algunos de ellos permiten encontrar estimadores que minimizan la x^2 modificada, Bhapkar (1966). El principio de estimación mas utilizado es el de máxima verosimilitud propuesto por R.A. Fisher (1912). (Citado por Neyman J. (1949)), en él Neyman considera los estimadores que minimizan la expresión

$$X^2_{(N)} = \sum_{j=1}^{r} \frac{(X_j - n\,\pi_j)^2}{X_j} \qquad (3)$$

sujetos a la restricción impuesta por la hipótesis nula $F_k(\pi) = 0$. Bajo la hipótesis nula (3) sigue una distribución limite Ji-cuadrado con d grados de libertad, siendo d el número de parámetros restringidos por la hipótesis.

Grizzle, Starmer and Koch (1969), asumieron s distribuciones multinomiales con r categorías de respuesta, definieron k funciones de los

π_{ij}, los cuales tienen derivadas hasta de segundo grado, con matriz de primeras derivadas de rango k.

Método estadístico.

A continuación se presentan los resultados teóricos; soporte de la metodología estadística de los modelos lineales con datos categóricos desarrollada por Grizzle, Starmer y Koch (1969).

En el caso particular de una repetición con r categorías de respuesta y s tratamientos, se presenta el siguiente arreglo:

	Categorías de respuesta			
Tratamiento	1	2 ...	r	total
1	n_{11}	n_{12}	n_{1r}	n_1
2	n_{21}	n_{22}	n_{2r}	n_2
⋮	⋮	⋮	⋮	⋮
s	n_{s1}	n_{s2}	n_{sr}	n_s

en donde los $n_{ij}(i = 1,\ldots,s,\ j = 1,2,\ldots,r)$ son frecuencias observadas asociadas a la estructura multinomial anterior. Si π_{ij} representa la probabilidad de que una observación pertenezca a la población i de la categoría j, entonces se define el vector $(\pi_i') = (\pi_{i1},\pi_{i2},\ldots,\pi_{ir})$ como el

vector de probabilidades asociados a las r categorías de respuesta.

Si $P_{ij} = \frac{n_{ij}}{n_i}$ es el estimador máximo verosímil de π_{ij}, el vector de estimadores de π_i es de la forma ${}_1P'_{ir} = [P_{i1}, P_{i2}, \ldots, P_{ir}]$, Johnson y Kotz (1969).

La varianza estimada asociada a una población específica (en una repetición del experimento) esta dada por:

$$\underset{r\times r}{V(P_i)} = \frac{1}{n_i}\left(D_{P_i} - P_i\, P_i'\right)$$

con D_{P_i} una matriz diagonal con elementos P_i.

La matriz de varianzas y covarianzas estimada de $V(\pi)$ para los s tratamientos es diagonal por bloques de la forma:

$$\underset{rs\times rs}{V(P)} = \begin{bmatrix} V(P_1) & & & \emptyset \\ & V(P_2) & & \\ & & \ddots & \\ \emptyset & & & V(P_s) \end{bmatrix}$$

El ajuste de datos categóricos al modelo lineal se hace por el método de mínimos cuadrados generalizados (M.C.G.), en donde la variable dependiente es una función lineal de las estimaciones de la probabilidad π_{ij}.

El modelo propuesto es $Y = X\theta + e$, con

$$\theta' = [\mu, \tau_1, \tau_2, \ldots, \tau_{\delta-1}; \beta_1, \beta_2, \ldots, \beta_{k-1}]$$

$$\tau_\delta = -\sum_{i=1}^{\delta-1} \tau_i \qquad \beta_k = -\sum_{j=1}^{\delta-1} \beta_j$$

$Y = F(\pi) = A\pi$;

A : matriz de orden (n, w), en donde n es el número de observaciones.

$w = n \times \delta \times k$ siendo k el número de repeticiones.

X : la matriz de diseño reparametrizada de orden (n, r) asociada a una estructura de bloques al azar

θ : es el vector de parámetros

e : el error aleatorio no observable, el cual tiene distribución normal con media 0 y matriz de varianzas y covarianzas $AV(\pi)A'$.

El resultado de AP, con P estimador máximo verosímil de π es el porcentaje promedio ponderado sobre las r categorías. Las ponderaciones se asignan en orden ascendente de acuerdo al número de categorías estudiadas. La estructura de A con r categorías de respuesta es:

$$A = \begin{bmatrix} 1 & 2 & 3 & \ldots & r & 0 & 0 & 0 & \ldots & 0 & 0 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 & 1 & 2 & 3 & \ldots & r & 0 & 0 & 0 & \ldots & 0 \\ \cdot & \cdot & \cdots & & & & & & \cdots & & & & & \cdots & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 1 & 2 & 3 & \ldots & r \end{bmatrix}$$

Como el interes del investigador es estimar los parámetros en el modelo y probar hipótesis acerca de la diferencia entre tratamientos, se tiene entonces que el estimador de M.C.G. para θ es:

$$\hat{\theta} = (X'S^{-1}X)^{-1}XS^{-1}y \quad \text{con} \quad S = AV(P)A' \ ;$$

matriz de varianzas covarianzas estimada para $A\pi$.

La hipótesis lineal general a probar es H_o: $C\theta = 0$ en donde C es una matriz cuya estructura es de la forma:

$$_{s-1}C_r = \begin{bmatrix} \overbrace{0\ 1\ 0\ \ldots\ 0}^{s-1} & \overbrace{0\ \ldots\ 0}^{k-1} \\ 0\ 0\ 1\ \ldots\ 0 & 0\ \ldots\ 0 \\ \cdot\ \cdot\ \cdot\ \cdots & \vdots \\ 0\ 0\ 0\ \ldots\ 1 & 0\ \ldots\ 0 \end{bmatrix}$$

la naturaleza de C, depende de las hipótesis que se planteen con respecto a contrastes entre tratamiendos. La suma de cuadrados debida a la hipótesis H_o: $C\theta = 0$ es:

$$SC(C\hat{\theta}) = (C\hat{\theta})' \ [C(X'S^{-1}X)^{-1}C']^{-1}(C\hat{\theta})$$

bajo H_o: $SC(C\hat{\theta}) \overset{a}{\sim} x^2(r(C))$

en donde: $r(C)$ es el rango de C; si $SC(C\hat{\theta}) > x^2 r(C)$ se rechaza H_o.

La suma de cuadrados debida al error, se define como:

$$SC(error) = Y'[S^{-1} - \hat{\theta}' (X'S^{-1}\hat{\theta})^{-1}]Y$$

Bajo H_o: los datos se ajustan al modelo; se tiene que:

$$SC(error) \sim x^2(n-r(c)): \text{ si } SC(error) > x^2 = (n-r(c))(\alpha)$$

se rechaza la hipótesis nula.

Cuando los efectos de los bloques tienen interpretación práctica (razas, especies, camadas, etc.) se construye una matriz C_1 tal que la hipótesis $H_o: C_1\theta = 0$ se puede probar.

Material experimental.

Para presentar una aplicación del método, se hizo uso de los resultados de un experimento realizado en el Centro Nacional de Investigaciones Agropecuarias del ICA, cuyo objetivo fué evaluar del grado de pudrición de mazorcas, causada por el hongo *Fusarium*, S.P. en la variedad de maíz sogamoseño.

Los materiales evaluados fueron:

- Sogamoseño V.O. variedad original τ_1 ;

- Ciclo I Sogamoseño (M.P.)I, selección masal por prolificidad τ_2;
- Ciclo II Sogamoseño (M.P.)II, selección masal por prolificidad τ_3;
- Ciclo III Sogamoseño (M.P.)III, selección masal por prolificidad τ_4;
- Cruzamiento de Sogamoseño V.O. con MB510 (M.P) VIII, material prolífico τ_5;
- Cruzamiento Sogamoseño V.O. con MB513 (M.N.P) VIII, material no prolífico τ_6;
- MB510 (M.P)VIII (testigo) τ_1;
- MB513 (M.N.P)VIII (testigo) τ_8;
- ICAV506, variedad comercial (testigo) τ_9.

En el arreglo del material experimental evaluado, se utilizó el diseño de bloques al azar con 6 repeticiones.

La evaluación del material experimental se efectuó teniendo en cuenta la siguiente escala.

Grado de pudrición	*% tejido enfermo*
G0	mazorcas aparentemente sanas
G1	1 - 25
G2	26 - 50

Para efecto de ésta aplicación se incluyeron tres categorías equivalentes al 50% o menos del tejido atacado, debido a que por encima de

de este porcentaje el número de mazorcas afectadas no eran considerables.

Resultados.

El arreglo del material experimental, se presenta en la Tabla 1 para el análisis de la información se hizo uso del procedimiento FUNCAT, implementado en el paquete SAS.

El valor del estadístico Ji-cuadrado para probar la hipótesis de igualdad entre materiales evaluados (tratamientos) fue de 171.27, con $\hat{\alpha} = 0.0001$, de encontrar el valor mayor al calculado, lo cual muestra suficiente evidencia de diferencia al grado de pudrición entre los materiales.

La prueba de bondad de ajuste al modelo resultó altamente significativa, el valor del estadístico Ji-cuadrado fue 158.82. Para dar una recomendación de la variedad "mas" resistente a la pudrición, se plantearon las siguientes comparaciones:

- H_o: $3(\tau_2+\tau_3+\tau_4+\tau_5)-4(\tau_7+\tau_8+\tau_9) = 0$. Este contraste compara la resistencia a la pudrición del material prolífico contra los materiales testigos, el valor del estadístico x^2 fue de 133.35 con un nivel de significancia estimado

de 0.0001, lo cual indica que los materiales testigo son más resistentes a la pudrición que el material prolífico.

- $H_o: \tau_1-\tau_6 = 0$. Compara la variedad regional contra el material prolífico, no se encontró diferencia significativa en la resistencia a la pudrición; el valor del estadístico Ji-cuadrado fue de 2.59 con $\hat{\alpha} = 0.1075$.

- $H_o: 3\tau_1-\tau_7-\tau_6-\tau_9 = 0$ contrasta la variedad regional contra los materiales testigos. El valor del estadístico $x^2 = 40.78$ y $\hat{\alpha} = 0.0001$ conducen claramente a establecer diferencias. En este caso los materiales testigos se mostraron más resistentes a la pudrición que la variedad regional.

- $H_o: \tau_1-\tau_9$ compara la variedad regional sogamoseño contra la variedad comercial ICAV506, se encontró significativa al grado de pudrición, siendo más resistente la variedad comercial, como lo muestra el valor de la $x^2 = 27.07$ con $\hat{\alpha} = 0.0001$.

Conclusiones.

Cuando se presenta el problema de la falta de ajuste del modelo, se presenta una sobre-

estimación (o subestimación), así también se sobreestima la hipótesis de diferencia entre tratamientos como también las comparaciones.

Este hecho conduce al investigador a encontrar otras funciones de la forma $Y = AP = F(P)$, no necesariamente lineales. Las transformaciones más frecuentes son log-lineales de la forma $F(P) = K[\log(AP)]$, donde k es una matriz de orden $(v \times 1)$; funciones exponenciales $F(P) = Q\{\exp k(\log(AP))\}$, donde Q de orden $q \times v$ y funciones logarítmicas compuestas.

$F(P) = L\{\log/Q(\exp\{K/\log(AP)/\}\}$ con L de orden $1 \times q$.

La selección masal por prolificidad, no contribuye al mejoramiento por resistencia con respecto a la variedad regional, y estos a su vez son menos resistentes que las variedades MB510(M.P)VIII, MB513(M.N.P)VIII y la variedad comercial ICAV506. Por otra parte no se observa tolerancia de pudrición el cruzar las variedades MB510(M.P)VIII, pues no se encontraron diferencias significativas.

Los materiales que mostraron mayor resistencia a la pudrición fueron el MB510(M.P)VIII, MB513(M.N.P)VIII y la variedad comercial ICAV-560.

TABLA 1. Número de mazorcas clasificadas en cada grado de pudrición y porcentaje ponderado y = AP.

	β_1									β_2								
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9
G0	17	14	14	26	38	13	52	60	65	9	16	19	18	27	21	67	65	63
G1	20	20	29	34	35	47	48	45	37	24	30	35	42	41	34	34	66	33
G2	6	8	17	7	14	12	12	12	7	6	3	2	9	8	9	4	6	5
y=AP	1.7	1.9	2.0	1.7	1.7	2.0	1.6	1.6	1.5	1.9	1.7	1.9	1.8	1.7	1.8	1.4	1.5	1.4
	β_3									β_4								
G0	11	17	22	23	25	19	67	58	42	12	22	9	12	47	15	42	53	42
G1	21	40	30	32	55	54	34	27	51	28	27	42	40	37	43	40	44	50
G2	5	2	5	5	8	2	4	4	7	5	0	4	6	2	3	8	6	7
y=AP	1.8	1.7	1.7	1.7	1.8	1.8	1.4	1.4	1.6	1.8	1.5	1.9	1.9	1.5	1.8	1.6	1.5	1.6
	β_5									β_6								
G0	13	11	5	23	17	23	41	45	47	15	5	4	11	14	55	48	45	5
G1	18	0	29	26	46	28	68	47	40	38	30	31	36	42	25	49	57	14
G2	9	6	12	9	11	6	7	8	11	5	9	2	7	5	5	7	2	3
y=AP	1.9	1.7	2.1	1.7	1.9	1.7	1.7	1.6	1.6	1.8	2.0	1.9	1.9	1.8	1.4	1.6	1.6	1.9

Fuente: Departamento de fitopatología ICA - Tibaitatá.

BIBLIOGRAFIA

Bhapkar, V.P., A note on the equivalence of two test criterio for hypotheses in categorical data. J. Amer. Statist. ASSOC 61, p. 228-235, 1960.

Grizzle, J.E. Starmer, C.F., Koch, G.G., Analysis of categorical data by linear models. Biometrics 25. pp.489-504, 1969.

Johnson, N.I. Kotz, S., Distribution in statistics. Discrete distributions. Houghton Milflin Company-Boston, 1969.

Lindeman, R.J., Miranda, P. Gold, R.Z., Introduction to bivariate and multivariate analysis. Scott, Foresman and Company, 1976.

Neyman, J., Contribution to theory of test. proceedings of the Berkeley Symposium on Mathematical Statistics and probability, University of California Berkeley. pp. 239-273, 1949.

S.A.S. User's guide, 1979.

* *