ITEA (2002), Vol. **98A N.° 2,** 256-263

ANÁLISIS DE VARIABLES CATEGÓRICAS MEDIANTE EL PROCEDIMIENTO CATMOD DE SAS®: APLICACIÓN A DATOS DE CRUZAMIENTO INDUSTRIAL EN BOVINO

Beatriz Silva, Javier Cañón

Dpto. Producción Animal, Facultad de Veterinaria, UCM, Madrid, España

Introducción

En producción animal es frecuente la expresión como variable categórica de caracteres de interés económico cuantitativos, entendidos como el resultado de la influencia de numerosos efectos o factores. El análisis de estas variables discretas mediante la utilización del modelo lineal, aunque no presente dificultades desde el punto de vista de la estimación de los parámetros del modelo, sí existe difícil justificación teórica cuando se trata de establecer pruebas de hipótesis (véase la revisión sobre caracteres discretos en CAÑÓN, 1986).

El paquete estadístico SAS presenta diversos módulos de análisis para tratar este tipo de variables, entre ellos el más utilizado puede ser el denominado CATMOD que permite, mediante el análisis de regresión logística, extender las técnicas del análisis de regresión múltiple al estudio de modelos en los que la variable dependiente no es continua, sino discreta.

Una presentación en forma de tabla de los datos correspondientes a variables categóricas sería la siguiente:

Poblaciones (combinaciones de niveles)	Categoría de respuesta 1	2	...	r	Total
1	n_{11}	n_{12}	...	n_{1r}	$n_{1\cdot}$
2	n_{21}	n_{22}	...	n_{2r}	$n_{2\cdot}$
:	:	:	...	:	:
s	n_{s1}	n_{s2}	...	n_{sr}	$n_{s\cdot}$

Lo que pretendemos con este trabajo es detallar la información que el procedimiento CATMOD de SAS proporciona, sobre todo desde la perspectiva de establecer funciones de los parámetros para dar respuesta a preguntas de interés en producción animal.

Material utilizado

Hemos utilizado la información proporcionada por unos 19000 terneros resultado del cruzamiento industrial entre hembras frisonas y machos de aptitud carnicera de diferentes razas: Blanco Azul Belga (BBB), Limousin (Li), Asturiana de Valles (Av) y Asturiana de Montaña (Am).

Como variables de interés recogidas en la base de datos figuran, además de la *raza* del padre: el *número de parto* de la vaca (con 3 niveles), *sexo* del ternero, *época de parto* (4 niveles), *dificultad al parto* (4 categorías) y la *conformación* del ternero (4 categorías). Las variables *dificultad al parto* y *conformación* del ternero son las variables de trabajo.

Modelos

Variable respuesta dicotómica

Sean $X_1,...,X_v$ el conjunto de variables explicativas, por simplicidad suponemos que Y es una variable que toma valores 1 y 0 con $\pi = P(Y = 1 | X_1,...,X_v)$ y por tanto $P(Y = 0 | X_1,...,X_v) = 1-\pi$.

Si trabajamos con la variable *dificultad al parto* con dos categorías correspondientes a la necesidad o no de llevar a cabo una cesárea y como variables explicativas consideramos el *sexo* del ternero (macho o hembra) y el *número de parto* de la vaca (en tres niveles: primeriza, 2° parto y 3° o más), tenemos 2 x 3 = 6 subpoblaciones determinadas por las categorías de las variables explicativas.

Un programa SAS tan sencillo como:

```
PROC CATMOD;
MODEL dificultad = sexo n.° parto;
RUN;
```

proporciona la siguiente información:

The CATMOD Procedure

Response	dificultad	Response Levels	2
Weight Variable	None	Populations	6
Data Set	Mis_datos	Total Frequency	19.073
Frequency Missing	0	Observations	19.073

Population Profiles				Response Frequencies		
				Response Number		
Sample	sexo	n.° parto	Sample Size	Sample	1	2
1	H	Primeriza	607 = $n_{11.}$	1	606 = n_{111}	1 = n_{112}
2	H	2.° parto	1.595 = $n_{12.}$	2	1.593 = n_{121}	2 = n_{122}
3	H	≥ 3	6.000 = $n_{13.}$	3	5.998 = n_{131}	2 = n_{132}
4	M	Primeriza	694 = $n_{21.}$	4	690 = n_{211}	4 = n_{212}
5	M	2.° parto	1.985 = $n_{22.}$	5	1.975 = n_{221}	10 = n_{222}
6	M	≥ 3	8.192 = $n_{23.}$	6	8.171 = n_{231}	21 = n_{232}

Obsérvese que en algunas celdas tenemos un bajo número de observaciones, lo que implica que las aproximaciones asintóticas tipo Chi-cuadrado deberían interpretarse con reservas y comprobarse mediante tests exactos.

Response Profiles
Response dificultad

1. Sin cesárea
2. Con cesárea

Como $0 < \pi < 1, 0 < 1-\pi < 1 \Rightarrow 0 < \frac{\pi}{1-\pi} < \infty \Rightarrow -\infty < \ln\left(\frac{\pi}{1-\pi}\right) < \infty,$

aplicamos la transformación *logit* para extender el modelo de regresión lineal

$$Y = \alpha + \sum_{j=1}^{v} \beta_j X_j \text{ a}$$

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \sum_{j=1}^{v} \beta_j X_j \Rightarrow \pi = \frac{e^{\alpha+\sum_{j=1}^{v}\beta_j X_j}}{1+e^{\alpha+\sum_{j=1}^{v}\beta_j X_j}} \text{ y así } 1-\pi = \frac{1}{1+e^{\alpha+\sum_{j=1}^{v}\beta_j X_j}}$$

Los parámetros a estimar son α y los coeficientes de regresión logística (β_j), para ello consideramos la función de verosimilitud:

$$L = \prod_{i=1}^{n} P(Y_i / X_{i1},\ldots,X_{iv}) = \prod_{i=1}^{n} \left[\left(\frac{e^{\alpha+\sum_{j=1}^{n}\beta_j X_i}}{1+e^{\alpha+\sum_{j=1}^{n}\beta_j X_i}} \right)^{Y_i} \times \left(\frac{1}{1+e^{\alpha+\sum_{j=1}^{n}\beta_j X_i}} \right)^{1-Y_i} \right]$$

que nos proporcionará las estimaciones máximo verosímiles mediante un proceso iterativo.

Maximum Likelihood Analysis

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion	Parameter Estimates 1	2	3	4
0	0	25595.153	1.0000	0	0	0	0
1	0	4884.2397	0.8092	1.9886	0.005566	-0.003753	-0.001734
2	0	1863.0173	0.6186	3.0985	0.0179	-0.0120	-0.005571
. . . . .							
9	0	546.64596	1.4025E-6	6.2631	0.8562	-0.3472	-0.1966
10	0	546.64596	5.263E-11	6.2632	0.8562	-0.3472	-0.1966

Maximum likelihood computations converged.

Para diferenciar entre los parámetros y las estimaciones a éstas últimas las denotaremos a y bj, en este caso a = 6,263, b_1 = 0,856 (es el cambio diferencial para el sexo hembra [para el macho será $-b_1$], b_2= -0,347 (corresponde a las vacas primerizas) y b_3 = -0,196 (es el cambio diferencial para las vacas en su 2° parto) con lo que para las vacas en su parto de 3° o mayor orden tendremos el coeficiente de regresión logística: $-b_2-b_3$.

El ajuste del modelo completo se comprueba mediante el contraste de hipótesis de que todos los coeficientes de regresión logística son 0; es decir, H_0: $\beta_j = 0\ \forall j$.

Comparamos el modelo dado con el restringido

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha$$

mediante un estadístico Chi-cuadrado. Así también vemos qué variables explicativas son significativas y la conveniencia o no de eliminarlas del modelo.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	1	577.60	<.0001
sexo	1	12.81	0.0003
n.° parto	2	6.07	0.0480

De acuerdo con estos resultados el sexo es muy significativo en el grado de dificultad al parto y también el número de parto puede considerarse que influye al nivel de significación usual de 0,05.

A continuación encontramos un estadístico Chi-cuadrado de cociente de verosimilitudes para comprobar la bondad de ajuste del modelo (la cercanía de los valores predichos por el modelo a los observados) que en nuestro caso tendrá la forma:

$$Q_L = \sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{2} 2n_{ijk}\ln\left(\frac{n_{ijk}}{m_{ijk}}\right)$$

donde n_{ijk} es el número de observaciones en los niveles i y j de las variables explicativas sexo y n.° de parto para la categoría k de la variable respuesta y m_{ijk} son los valores esperados, es decir:

$$m_{ijk} = \begin{cases} n_{ij\cdot}\cdot\theta_{ij} & si\ k=1 \\ n_{ij\cdot}\cdot(1-\theta_{ij}) & si\ k=2 \end{cases}$$

(θ_{ij} estimación de la probabilidad de que se produzca la primera respuesta [que no sea necesario realizar una cesárea] en los niveles i y j de las correspondientes variables explicativas:

$$\theta_{11}=\frac{e^{a+b_1+b_2}}{1+e^{a+b_1+b_2}},\theta_{12}=\frac{e^{a+b_1+b_3}}{1+e^{a+b_1+b_3}},$$

$$\theta_{13}=\frac{e^{a+b_1-b_2-b_3}}{1+e^{a+b_1-b_2-b_3}},\theta_{21}=\frac{e^{a-b_1+b_2}}{1+e^{a-b_1+b_2}},$$

$$\theta_{22}=\frac{e^{a-b_1+b_3}}{1+e^{a-b_1+b_3}},\theta_{23}=\frac{e^{a-b_1-b_2-b_3}}{1+e^{a-b_1-b_2-b_3}})$$

Obtenemos:

	DF	Chi-Square	Pr > ChiSq
Likelihood Ratio	2	0.55	0.7585

Con un p-valor del 0,7585 no rechazamos la hipótesis nula de que el modelo es acertado.

También contrastamos la hipótesis de si los distintos coeficientes de regresión logística son significativos o no (H_0: $\beta_j = 0$), mediante el estadístico $z = b_j/s_j$, donde s_j es el error estándar (la raíz cuadrada de la cuasivarianza muestral) de b_j. SAS utiliza z^2, el estadístico de Wald, que sigue una distribución Chi-cuadrado con un grado de libertad (χ_1^2). Con los datos con los que estamos trabajando, vemos que en la fila correspondiente al sexo tenemos

$$\left(\frac{0,8562}{0,2392}\right)^2 = 12,81$$

(que como sólo tiene 2 niveles y por tanto un sólo coeficiente, coincide con el valor de la Chi-cuadrado en la tabla superior).

Analysis of Maximum Likelihood Estimates

Effect	Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	6.2632	0.2606	577.60	<.0001
sexo	2	0.8562	0.2392	12.81	0.0003
n.º parto	3	-0.3472	0.3218	1.16	0.2806
	4	-0.1966	0.2538	0.60	0.4385

Una vez ajustado el modelo, podemos utilizar las "odds ratio" para dar respuesta a preguntas de interés como, por ejemplo, ¿cuánto más probable es tener que practicar una cesárea si el ternero que nace es macho?:

$$\frac{e^{a+b_1+b_2}}{e^{a-b_1+b_2}} = \frac{e^{a+b_1+b_3}}{e^{a-b_1+b_3}} = \frac{e^{a+b_1-b_2-b_3}}{e^{a-b_1-b_2-b_3}} =$$

$$= e^{2b_1} = e^{2*0.8562} \approx 5,5$$

Obtenemos que, aproximadamente, la probabilidad de que no sea necesaria una cesárea si el ternero que nace es hembra es unas 5,5 veces más alta que en el caso en que el ternero es macho.

O para comparar la dificultad al parto en función de la paridad de la vaca:

$$\frac{e^{a+b_1-b_2-b_3}}{e^{a+b_1+b_2}} = \frac{e^{a-b_1-b_2-b_3}}{e^{a-b_1+b_2}} =$$

$$= e^{-2b_2-2b_3} = e^{-2*(-0.3472)-(-0.1966)} \approx 1,65$$

es decir, la probabilidad de no tener que realizar cesárea si la paridad de la vaca es superior a 2 es un 65% más elevada que si la vaca fuera primeriza.

Variable respuesta con más de dos niveles

Consideremos ahora el tratamiento de una variable categórica ordenada, con un número de respuestas superior a 2, por ejemplo la variable *conformación* del ternero con notas 1, 2, 3 ó 4 que agruparemos en tres niveles: el primero es la mejor nota, el 1, el segundo nivel se refiere a nota 2 y el tercero agrupa las notas 3 y 4 y como variables explicativas la *raza* (Limousin, Asturiana de Valles, Asturiana de Montaña y Blanco Azul Belga) y el *número de parto* de la vaca. En este caso trabajamos con un tamaño de muestra de 19.073 cabezas que se dividen en 4 × 3 = 12 subpoblaciones:

The CATMOD Procedure

Data Summary

Response	conform	Response Levels	3
Weight Variable	None	Populations	12
Data Set	Mis_datos	Total Frequency	19073
Frequency Missing	0	Observations	19073

Population Profiles				Response Frequencies Response Number			
Sample raza	n.° parto	Sample	Size	Sample	1	2	3
1	Li	Primeriza	114	1	8	64	42
2	Li	2° parto	104	2	6	78	20
3	Li	≥ 3	319	3	23	245	51
4	Av	Primeriza	336	4	13	205	118
5	Av	2° parto	1662	5	137	1187	338
6	Av	≥ 3	5698	6	486	4060	1152
7	Am	Primeriza	730	7	4	319	407
8	Am	2° parto	66	8	0	34	32
9	Am	≥ 3	107	9	1	55	51
10	BBB	Primeriza	121	10	9	89	23
11	BBB	2° parto	1748	11	254	1221	273
12	BBB	≥ 3	8068	12	1051	5869	1148

Response Profiles

Response	conform
1	Mejor nota
2	Nota 2
3	Notas 3 o 4

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	2	1089.54	<.0001
raza	6	343.41	<.0001
n.° parto	4	63.95	<.0001
Likelihood Ratio	12	18.12	0.1120

Ambas variables explicativas son significativas y al nivel 0,05 (< 0,112) se trata de un buen modelo.

Analysis of Maximum Likelihood Estimates

Effect	Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.7583	0.1256	196.02	<.0001
	2	0.9235	0.0347	708.11	<.0001

Analysis of Maximum Likelihood Estimates (continuación)

Effect	Parameter	Estimate	Standard Error	Chi-Square	Pr > ChiSq
raza	3	0.5140	0.1827	7.91	0.0049
	4	0.2433	0.0831	8.58	0.0034
	5	0.5300	0.1325	15.99	<.0001
	6	0.1166	0.0435	7.17	0.0074
	7	-2.3757	0.3523	45.47	<.0001
	8	-0.8340	0.0729	130.79	<.0001
n.° parto	9	-0.6923	0.1306	28.11	<.0001
	10	-0.3854	0.0568	45.98	<.0001
	11	0.3470	0.0775	20.03	<.0001
	12	0.1573	0.0402	15.31	<.0001

Continuando con la misma notación para las estimaciones, en este caso es un poco más complicado pero sigue el mismo patrón. Tenemos que $a_1 = -1,76$ es la media para la nota de conformación 1 (1er nivel frente al 3er nivel), $a_2 = 0,92$ media para la nota 2 (2° nivel frente al 3er nivel de respuesta); $b_1 = 0,514$ es el cambio diferencial para la raza Limousin en el 1er nivel de respuesta frente al 3°, $b_2 = 0,243$ corresponde al 2° nivel *versus* el 3er nivel de respuesta, $b_3 = 0,53$ corresponde a la raza Asturiana de Valles para el 1er nivel *vs* el 3°, $b_4 = 0,117$ para el 2° nivel de respuesta frente al 3°, $b_5 = -2,376$ es el cambio diferencial para la raza Asturiana de Montaña en el 1er nivel frente al 3er nivel de respuesta, $b_6 = -0,834$ para el 2° nivel frente al 3° y finalmente $-b_1-b_3-b_5$ para la raza Blanco Azul Belga y el 1er nivel *vs* 3°, $-b_2-b_4-b_6$ para el 2° frente al 3°; $b_7 = -0,692$ es el cambio diferencial correspondiente a las vacas primerizas para el 1er nivel de respuesta frente al 3°, $b_8 = -0,385$ corresponde al 2° nivel frente al 3°, $b_9 = 0,347$ es el cambio diferencial para las vacas en su 2° parto en el 1er nivel de respuesta *vs* el 3°, $b_{10} = 0,157$ para el 2° nivel *vs* el 3er nivel, finalmente para las vacas en su tercer parto o parto de mayor orden tenemos $-b_7-b_9$ para el 1er nivel de respuesta frente al 3° y $-b_8-b_{10}$ corresponde al 2° nivel *vs* el 3°.

Así, si nuestro interés es comparar la probabilidad de obtener una nota de conformación del ternero en lugar de otra en función de la raza de su padre, la razón de probabilidades será, por ejemplo:

$$\frac{e^{a_2-b_2-b_4-b_6}}{e^{a_2+b_6}} = e^{-2b_6-b_2-b_4} = e^{1.308} \approx 3,7$$

es decir, es 3,7 veces más probable obtener nota de conformación de 2 en lugar de una conformación 3 ó 4 cuando se utiliza como raza paterna la Blanco Azul Belga en lugar de la Asturiana de Montaña.

Otra razón de probabilidades que puede tener resultar de interés es la siguiente:

$$\frac{e^{a_1-b_1-b_3-b_5}}{e^{a_1+b_1}} = e^{-2b_1-b_3-b_5} = e^{0.818} \approx 2,3$$

es decir, es 2,3 veces más probable obtener un ternero con una nota de conformación 1 en lugar de una nota 3 ó 4 cuando se utiliza la raza Blanco Azul Belga en lugar de la raza Limousin en cruzamiento industrial con vacas Frisonas.

Igualmente podríamos comparar la influencia de la paridad de la vaca sobre la nota de conformación del ternero mediante el cociente:

$$\frac{e^{a_1+b_7}}{e^{a_1-b_7-b_0}} = e^{-2b_7-b_0} = e^{2*(-0.692)+0.347} =$$

$$= e^{-1.037} \approx 0{,}35$$

lo que indica que es 2,82 veces más probable (1/0,35) obtener un ternero con la mejor conformación en lugar de un ternero con conformaciones 3 ó 4 si se trata de un tercer parto o superior que si fuera el primer parto de la vaca.

Agradecimientos

La información ha sido proporcionada por ASEAVA y ASEAMO. Agradecemos la ayuda prestada por M.ª del Carmen Bravo Llatas del Servicio Informático de Apoyo a Docencia e Investigación de la UCM.

Bibliografía

CAÑÓN J., 1986. Caracteres Discretos en Mejora Genética Animal. Investigación agraria, Producción y Sanidad Animales, 1 (3): 205-236.

GRIZZLE J.E., STARMER C.F., KOEH G.G., 1968. Analysis of catogorical data by linear models. Biometrics, September 1969, 489-504.

HOSMER D.W., LEMESHOW S., 1989. Applied Logistic Regression, New York: Wiley.

STOKES M.E., DAVIS C.S., KOCH G.G., 1995. Categorical Data Analysis using de SAS System, (Books by Users, SAS Institutet's Author Service).