REGRESIÓN LOGÍSTICA MULTINOMIAL

V. Pando Fernández y R. San Martín Fernández

Departamento de Estadística e Investigación Operativa. E.T.S. de Ingenierías Agrarias. Universidad de Valladolid. Avda. de Madrid 44. 34004–PALENCIA (España). Correo electrónico: vpando@eio.uva.es

Resumen

En este artículo se presenta la regresión logística multinomial como extensión multivariante de la regresión logística binaria clásica, ampliamente utilizada en la investigación forestal. A partir de la formulación matemática del modelo estadístico se explica la estimulación de parámetros mediante el método de máxima verosimilitud y se establecen los test estadísticos adecuados para la significatividad global del modelo y el efecto de cada regresor. También se calculan los intervalos de confianza asintóticos para los parámetros y se mide la calidad del ajuste mediante los coeficientes de determinación (pseudo-R^2) más ampliamente utilizados. La calidad en la predicción puede medirse, al igual que en el análisis discriminante, mediante la tabla de clasificación observados-predichos o mediante validación externa si se dispone de una muestra alternativa para ese propósito.

Palabras clave: *Logística, Multinomial, Politónica, Polinomial, Pseudo-R^2*

INTRODUCCIÓN

La regresión logística multinomial (HOSMER & LEMESHOW, 1989) es utilizada en modelos con variable dependiente de tipo nominal con más de dos categorías (politómica) y es una extensión multivariante de la regresión logística binaria clásica. Las variables independientes pueden ser tanto continuas (regresores) como categóricas (factores).

Tradicionalmente las variables dependientes politómicas han sido modeladas mediante análisis discriminante pero, gracias al creciente desarrollo de las técnicas de cálculo, cada vez es más habitual el uso de modelos de regresión logística multinomial, ya implementados en paquetes estadísticos como S.A.S. (PROC CATMOD) o S.P.S.S. (NOMREG), debido a la mejor interpretabilidad de los resultados que proporciona.

En el ámbito forestal esta técnica se ha utilizado por diversos autores como por ejemplo HARIG & FAUSCH (2002), HERAJARVI (2002), NORTH & REYNOLDS (1996) o RIO et *al.* (2004).

Con esta comunicación se pretende presentar las bases teóricas de esta técnica estadística tanto en lo que se refiere a su formulación y metodología de ajuste como al análisis e interpretación de los resultados obtenidos utilizando la subrutina NOMREG del paquete estadístico S.P.S.S. Para ello se considerará un caso con dos regresores y una variable politómica con tres categorías.

Formulación del modelo

Consideramos una variable aleatoria dependiente Y categórica nominal politómica con Soporte(Y)={1,2,3} y con probabilidades p_1=p(Y=1), p_2=p(Y=2) y p_3=p(Y=3)=1-p_1-p_2. Supongamos que queremos analizar el efecto que ejercen dos variables explicativas continuas X_1, X_2 sobre las probabilidades p_1 y p_2 que caracterizan a la variable Y. Podemos redefinir a

ISSN: 1575-2410

la variable Y mediante un vector (Y_1,Y_2) construido de la siguiente forma:

$$(Y_1,Y_2)=\begin{cases}(1,0) & si \quad Y=1\\(0,1) & si \quad Y=2\\(0,0) & si \quad Y=3\end{cases}$$

Las variables Y_1 e Y_2 tienen una distribución de Bernouilli con $E(Y_1)=p_1$ y $E(Y_2)=p_2$, al igual que la variable dependiente en una regresión logística binaria clásica. Obviamente estas dos variables no son independientes ya que $Cov(Y_1,Y_2)=-p_1p_2$.

Formulamos el modelo multivariante definido por las siguientes ecuaciones:

$$\left.\begin{aligned}p_1(X_1,X_2)=p_1=E(Y_1)=\frac{\exp(Z_1)}{1+\exp(Z_1)+\exp(Z_2)}\\p_2(X_1,X_2)=p_2=E(Y_2)=\frac{\exp(Z_2)}{1+\exp(Z_1)+\exp(Z_2)}\end{aligned}\right\}$$

donde $Z_1=\beta_{01}+\beta_{11}\cdot X_1+\beta_{21}\cdot X_2$ y $Z_2=\beta_{02}+\beta_{12}\cdot X_1+\beta_{22}\cdot X_2$, siendo β_{01}, β_{11}, β_{21}, β_{02}, β_{12}, β_{22}, parámetros que deseamos estimar.
(Observar que

$$p_3(X_1,X_2)=p_3=1-p_1-p_2=\frac{1}{1+\exp(Z_1)+\exp(Z_2)}).$$

Con el propósito de interpretar mejor los parámetros que aparecen en el modelo, podríamos reescribir éste de la siguiente forma:

$$\left.\begin{aligned}\frac{p_1}{p_3}=\exp(Z_1)=\exp(\beta_{01})\cdot(\exp(\beta_{11}))^{X_1}\cdot(\exp(\beta_{21}))^{X_2}\\\frac{p_2}{p_3}=\exp(Z_2)=\exp(\beta_{02})\cdot(\exp(\beta_{12}))^{X_1}\cdot(\exp(\beta_{22}))^{X_2}\end{aligned}\right\}$$

Al cociente p_1/p_3 se le denomina '*odds*' de la categoría 1 respecto de la categoría 3 y se le representa por $O_1(X_1, X_2) = O_1$ (idem. para O_2). De este modo puede observarse fácilmente que la razón de cambio en O_1 cuando X_1 se incrementa en una unidad manteniéndose constante X_2 viene dada por $\frac{O_1(X_1+1,X_2)}{O_1(X_1,X_2)}=\exp(\beta_{11})$, que recibe el nombre de '*odds-ratio*' de la categoría 1 respecto de la variable X_1 y se representa por $OR_1(X_1)$ (idem. para $OR_1(X_2)$, $OR_2(X_1)$ y $OR_2(X_2)$).

Es interesante observar que estas 'odds-ratio' dependen de las unidades en que vengan medidas las variables regresoras (si multiplicamos X_1 por 10, $OR_1(X_1)$ pasaría a ser $\sqrt[10]{\exp(\beta_{11})}$). Por tanto la importancia de cada variable regresora en el modelo debería medirse por el valor de la odds-ratio suponiendo estandarizada dicha variable. Este es el motivo por el que se habla de las '*odds-ratio*' *estandarizadas en las variables regresoras*. Por ejemplo $OR_1(X_1^*)=\exp(\beta_{11}\cdot S_{x1})$ siendo S_{x1} la desviación típica muestral de la variable X_1 (idem. para $OR_1(X_2^*)$, $OR_2(X_1^*)$ y $OR_2(X_2^*)$). Cuanto más grande sea este valor más relevante es la variable dentro del modelo.

También interesa definir las *proporciones de cambio en las* '*odds*' con respecto a cada variable regresora que, por ejemplo, para O_1 con respecto a X_1, viene dada por:

$$\frac{O_1(X_1+1,X_2)-O_1(X_1,X_2)}{O_1(X_1,X_2)}=OR_1(X_1)-1=\exp(\beta_{11})-1$$

y que representaremos por: $OC_1(X_1)$ (idem. para $OC_1(X_2)$, $OC_2(X_1)$ y $OC_2(X_2)$).

Otra formulación alternativa, y quizás más conocida, se obtiene tomando logaritmos en ambas ecuaciones del modelo:

$$\left.\begin{aligned}\ln\left(\frac{p_1}{p_3}\right)=Z_1=\beta_{01}+\beta_{11}\cdot X_1+\beta_{21}\cdot X_2\\\ln\left(\frac{p_2}{p_3}\right)=Z_2=\beta_{02}+\beta_{12}\cdot X_1+\beta_{22}\cdot X_2\end{aligned}\right\}$$

donde las expresiones del miembro izquierdo se denominan '*logits*' (al igual que en la regresión logística binaria) y los parámetros representan las *tasas de cambio en los* '*logits*' cuando una de las variables explicativas se incrementa en una unidad manteniéndose constante la otra.

Estimación de parámetros

Dada una muestra de datos $(Y_{1i}, Y_{2i}, X_{1i}, X_{2i})$ con i=1,2....,n podemos definir, en función de los parámetros del modelo, las funciones Z_{1i}, Z_{2i}, p_{1i}, p_{2i} y abordar el problema de la estimación de los mismos mediante el método de máxima verosimilitud, como se muestra a continuación.

Con el modelo planteado, la función de verosimilitud de la muestra viene dada por la siguiente expresión:

$$L = \prod_{i=1}^{n} \left(p_{1i}^{Y_{1i}} \cdot p_{2i}^{Y_{2i}} \cdot p_{3i}^{1-Y_{1i}-Y_{2i}} \right) = \prod_{i=1}^{n} \left(\left(\frac{p_{1i}}{p_{3i}} \right)^{Y_{1i}} \cdot \left(\frac{p_{2i}}{p_{3i}} \right)^{Y_{2i}} \cdot p_{3i} \right)$$

En vez de trabajar con esta expresión se utiliza la función auxiliar:

$$\Lambda = -2 \cdot \ln(L) = -2 \cdot \sum_{i=1}^{n} \left(Y_{1i} \cdot \ln\left(\frac{p_{1i}}{p_{3i}} \right) + Y_{2i} \cdot \ln\left(\frac{p_{2i}}{p_{3i}} \right) + \ln(p_{3i}) \right) =$$

$$= 2 \cdot \sum_{i=1}^{n} \left(\ln\left(1 + \exp(Z_{1i}) + \exp(Z_{2i})\right) - Y_{1i} \cdot Z_{1i} - Y_{2i} \cdot Z_{2i} \right)$$

El problema de maximizar la verosimilitud equivale al de minimizar la función auxiliar Λ_Λ y puede resolverse por métodos numéricos de forma iterativa partiendo de la estimación inicial $\beta_{11}=\beta_{21}=\beta_{12}=\beta_{22}=0$, $\beta_{01}=\ln(n_1)-\ln(n-n_1-n_2)$ y $\beta_{02}=\ln(n_2)-\ln(n-n_1-n_2)$ siendo n_1 y n_2 el número de observaciones en las categorías 1 y 2 respectivamente. Estos estimadores iniciales se obtienen suponiendo que no hay una influencia de las variables regresoras en el modelo planteado y para ellos el valor inicial de la función auxiliar que debemos de minimizar es:

$$\Lambda_0 = -2 \cdot \left(n_1 \cdot \ln\left(\frac{n_1}{n} \right) + n_2 \cdot \ln\left(\frac{n_2}{n} \right) + (n - n_1 - n_2) \cdot \ln\left(\frac{n - n_1 - n_2}{n} \right) \right)$$

Una vez alcanzada la convergencia del método iterativo, designaremos por Λ_φ al mínimo obtenido y por $\hat{\beta}_{01}, \hat{\beta}_{11}, \hat{\beta}_{21}, \hat{\beta}_{02}, \hat{\beta}_{12}, \hat{\beta}_{22}$ a los valores estimados de los parámetros del modelo.

Significatividad global del modelo

Podemos contrastar la hipótesis de no existencia de un efecto significativo global de las variables regresoras teniendo en cuenta que la diferencia entre el valor inicial y el valor final de la función auxiliar Λ tiene una distribución χ^2 con 4 grados de libertad (en general, número de regresores multiplicado por número de categorías menos una). El p-valor del test para la hipótesis nula de que no existe efecto de las variables regresoras ($\beta_{11}=\beta_{21}=\beta_{12}=\beta_{22}=0$) vendrá dado por $p(\chi^2_4>\Lambda_o-\Lambda_f)$.

Significatividad del efecto de cada variable regresora

Si llamamos Λ_{-1} al mínimo de la función auxiliar que se obtendría eliminando del modelo la variable X_1 ($\beta_{11}=\beta_{12}=0$) se verifica que la diferencia entre los mínimos de la función auxiliar en el modelo reducido y en el modelo completo tiene una distribución χ^2 con 2 grados de libertad (en general, número de regresores menos uno multiplicado por número de categorías menos una). Por tanto el p-valor del test para la hipótesis nula de que no existe efecto de la variable X_1 ($\beta_{11}=\beta_{12}=0$) vendrá dado por $p(\chi^2_2>\Lambda_{-1}-\Lambda_\varphi)$. De modo similar podríamos calcular Λ_{-0} (mínimo de la función auxiliar eliminando β_{01} y β_{02} del modelo) y Λ_{-2} (mínimo de la función auxiliar eliminando del modelo la variable X_2) y construir tests de hipótesis para $\beta_{01}=\beta_{02}=0$ y $\beta_{21}=\beta_{22}=0$, respectivamente.

Significatividad de cada parámetro

Teniendo en cuenta que el cuadrado de cada estimador dividido por su error estándar tiene una distribución χ^2 con 1 grado de libertad podemos construir test de hipótesis para la igualdad de cada parámetro a cero y podremos saber qué estimadores de los parámetros del modelo son significativamente distintos de cero. Por ejemplo, para el test de hipótesis $\beta_{11}=0$ el p-valor sería $p\left(\chi_1^2 > \left(\frac{\hat{\beta}_{11}}{s.e.(\hat{\beta}_{11})} \right)^2 \right)$, siendo $s.e.(\hat{\beta}_{11})$ el valor correspondiente al error estándar del estimador del parámetro β_{11}.

Intervalos de confianza para los parámetros

Basándonos en la normalidad asintótica de los estimadores máximo verosímiles podemos construir, utilizando la distribución normal, intervalos de confianza asintóticos para cada uno de los parámetros del modelo y, mediante

las transformaciones correspondientes, intervalos de confianza (I.C.) para las *OR* y las *OC*. Por ejemplo, para el parámetro β_{11}, y utilizando un grado de confianza de 1–α, tendríamos:

I. C. para β_{11}:

$$\left(\hat{\beta}_{11} - z_{\alpha/2} \cdot s.e.(\hat{\beta}_{11}),\ \hat{\beta}_{11} + z_{\alpha/2} \cdot s.e.(\hat{\beta}_{11})\right)$$

I. C. para $OR_1(X_1)$:

$$\left(\exp\left(\hat{\beta}_{11} - z_{\alpha/2} \cdot s.e.(\hat{\beta}_{11})\right), \exp\left(\hat{\beta}_{11} + z_{\alpha/2} \cdot s.e.(\hat{\beta}_{11})\right)\right)$$

I: C: para $OR_1(X_1^*)$:

$$\left(\exp\left(S_{X_1} \cdot \left(\hat{\beta}_{11} - z_{\alpha/2} \cdot s.e.(\hat{\beta}_{11})\right)\right), \exp\left(S_{X_1} \cdot \left(\hat{\beta}_{11} + z_{\alpha/2} \cdot s.e.(\hat{\beta}_{11})\right)\right)\right)$$

I. C. para $OC_1(X_1)$:

$$\left(\exp\left(\hat{\beta}_{11} - z_{\alpha/2} \cdot s.e.(\hat{\beta}_{11})\right) - 1, \exp\left(\hat{\beta}_{11} + z_{\alpha/2} \cdot s.e.(\hat{\beta}_{11})\right) - 1\right)$$

siendo $z_{\alpha/2}$ el valor que, en una distribución normal (0,1), verifica $p(Z>z_{\alpha/2})=\alpha/2$

Calidad del ajuste

Al igual que en la regresión logística binaria, la calidad del ajuste en la regresión logística multinomial se mide mediante coeficientes de determinación conocidos como Pseudo-R^2. De entre todos ellos comentaremos los más clásicos, que son los que proporciona el paquete estadístico S.P.S.S.

El primero se basa en la función auxiliar Λ utilizada en el ajuste, se conoce como pseudo-R^2 de Mc-Fadden y viene dado por: $R^2_{MF} = 1 - \frac{\Lambda_f}{\Lambda_0}$.

Su rango teórico de valores es $0 \leq R^2_{MF} \leq 1$, pero muy raramente su valor se aproxima a 1. Suele considerarse una buena calidad del ajuste cuando $0.2 \leq R^2_{MF} \leq 0.4$ y excelente para valores superiores.

Otros autores prefieren coeficientes basados directamente en la verosimilitud *L*, y no en la función auxiliar Λ. El más conocido es el pseudo-R^2 de Cox-Snell, definido como

$$R^2_{CS} = 1 - \frac{\left(\sqrt[n]{L_0}\right)^2}{\left(\sqrt[n]{L_f}\right)^2} = 1 - \exp\left(\frac{\Lambda_f - \Lambda_0}{n}\right)$$, siendo

$L_0=\exp(-\Lambda_0/2)$ y $L_f=\exp(-\Lambda_f/2)$. El rango teórico de valores para este coeficiente es

$0 \leq R^2_{CS} \leq 1 - \left(\sqrt[n]{L_0}\right)^2$, lo que le hace poco interpretable al depender de L_0. Por este motivo es preferible el pseudo-R2 de Nagelkerke, que se define como

$$R^2_N = \frac{R^2_{CS}}{1 - \left(\sqrt[n]{L_0}\right)^2} = \frac{1 - \exp\left(\frac{\Lambda_f - \Lambda_0}{n}\right)}{1 - \exp\left(-\frac{\Lambda_0}{n}\right)}$$

y su rango de valores es $0 \leq R^2_N \leq 1$ por lo que puede interpretarse del mismo modo que el coeficiente de determinación de la regresión lineal clásica, aunque es más difícil que alcance valores próximos a 1.

Para comparar modelos de regresión logística multinomial con diferente número de variables regresoras suelen introducirse coeficientes Pseudo-R^2 ajustados. El más conocido es el de Mc-Fadden, definido como

$Adj - R^2_{MF} = 1 - \frac{0.5 \cdot \Lambda_f + k + 1}{0.5 \cdot \Lambda_0 + 1}$, siendo *k* el número de regresores.

Calidad en la predicción

Si, a partir del modelo ajustado, clasificamos cada observación en la categoría más probable, podemos construir una matriz de clasificación observados-predichos y utilizar el porcentaje de clasificaciones correctas como una medida de la calidad de predicción, del mismo modo que se hace en el análisis discriminante.

BIBLIOGRAFÍA

AGRESTI, A.; 1990. *Categorical Data Analysis.* John Wiley and Sons. New York.

COX, D.R. & SNELL, E.J.; 1989. *The Analysis of Binary Data.* Chapman and Hall. London.

HOSMER, D.W. & LEMESHOW, S; 1989. *Applied Logistic Regression.* Wiley Interscience. New York.

HARIG, A.L. & FAUSCH, K.D.; 2002. Minimum habitat requirements for establishing translocated cutthroat trout populations. *Ecol. Appl.* 12(2): 535-551.

HERAJARVI, H.; 2002. Internal knottiness with respect to sawing patterns in Betula pendula and B. bubescens. *Baltic Forestry* 8(1): 42-50.

MCFADDEN, D.; 1979. Quantitative methods for analysing travel behaviour of individuals: some recent developments. *In*: B*ehavioural travel modelling*: 279-318. Croom Helm. London.

MENARD, S.; 2000. Coefficients of Determination for Multiple Logistic Regression Analysis. *The American Statistician* 54: 17-24.

NAGELKERKE, N.J.; 1991. A note on a general definition of the coefficient of determination. *Biometrika* 78: 691-692.

NORTH, M.P. & REYNOLDS, J.H.; 1996. Microhabitat analysis using radiotelemetry locations and polytomous logistic regression. *J. Wild. Manage.* 60(3): 639-653.

RÍO, M. DEL; BRAVO, F.; PANDO, V.; SANZ, G. & SIERRA, R.; 2004. Influence of individual tree and stand attributes in stem straightness in Pinus pinaster Ait. Stands. *Ann. Sci. For.* 61(2): 141-148.