Regresión Logística Multinomial

Martin Blanco, Alicia Giménez

1 de noviembre de 2018

Resumen: el presente trabajo presenta un esbozo de las características del Modelo de Regresión Logística Multinomial la cual generaliza el Método de Regresión Logística para problemas multiclase, es decir con más de dos resultados posibles discretos, prediciendo las probabilidades de los diferentes resultados posibles de una distribución categórica como variables independientes. Además de la presentación de definiciones y conceptos generales, se presenta un ejemplo de aplicación de la técnica mediante el programa de procesamiento estadístico Rstudio, la redacción del material se realizó con Rmarkdowm y para el desarrollo se utilizaron los paquetes .

Introducción al Modelo de Regresión Logística Multinomial

El modelo de regresión logística multinomial o también conocido como modelo con respuesta politómica, es una generalización del modelo de regresión logístico binomial (Mc Cullagh, 1989) en el que se desea estimar la probabilidad de que el individuo presente o no un evento específico, dado un conjunto de variables que explican características particulares de los individuos. En el caso del modelo multinomial, la variable endógena tiene más de dos alternativas a considerar como posibles respuestas, por lo cual la distribución de probabilidad adecuada para modelar este fenímeno es la distribución multinomial.

Se debe tener en cuenta que la regresión logística multinomial difiere de la regresión logística condicional y ordinal, pues en la regresión condicional las variables explicativas hacen referencia a atributos de las alternativas, variando sus valores para cada una de ellas, mientras que pueden o no variar para cada individuo, además solo se estima un vector de parámetros, a diferencia de el caso multinomial en el que hay tantos vectores como categorías menos una. Por otra parte en la regresión ordinal, las estimaciones no se pueden realizar con los modelos mencionados anteriormente, ya que esta tiene como particularidad el uso de la información adicional suministrado por el orden de las categorías de la variable respuesta.

Ahora bien, la variable respuesta del modelo de regresión logística multinomial es una variable aleatoria con distribución multinomial, que se puede considerar como el número de éxitos en cada una de las g categorías que se presentan en n ensayos independientes, y su funci?n de distribución viene dada por (Feller, 1967).

$$P(Y_1 = n_1, n_2, ..., Y_g = n_g; n_1, p_1, p_2, ..., p_g) =$$

$$P\binom{n}{n_1 n_2 ... n_g} p^{n_1} p^{n_2} ... p^{n_g}$$

donde $\sum_{j=1}^{g} n_j = n$ y $\sum_{j=1}^{g} p_j = 1$. En esencia, la regresión logística multinomial estima las probabilidades de esta distribución para cada individuo i, teniendo en cuenta un conjunto de variables explicativas (Mc Fadden, 1974).

Para realizar las estimaciones de los parámetros en el modelo logístico multinomial, se utiliza la funcion canónica de enlace de la distribución multinomial a la familia exponencial, la cual es llamada transformación logit:

$$EY_i = log(\frac{p_i j}{p_i g})$$

Donde ij es la probabilidad del individuo i pertenezca a la categoría j, ig p corresponde a la probabilidad del mismo individuo en la categoría g, la cual es definida como la categoría de referencia, de la variable con distribución multinomial Y.

Al realizar una revisión de la literatura sobre el enlace canónico multinomial, se encuentra poca información de los antecedentes teóricos de esta función, debido a que el modelo multinomial puede verse como una extensión del modelo de regresión logístico binomial, del cual muchos autores realizan el desarrollo formal y luego generalizan al caso politomico, mostrando directamente los resultados, como Hosmer y Lemeshow (2000), McCullagh y Nelder (1989), Jobson (1991), Silva (1995) entre otros. Dado esta situación este trabajo pretende mostrar formalmente el planteamiento del modelo de regresión logístico multinomial partiendo de dicha función, y las expresiones a través de las cuales se estiman las probabilidades de los individuos.

La Función Canónica de Enlace

Como se menciona anteriormente para la estimación de los parámetros del modelo se utiliza el hecho de que la distribución multinomial pertenece a la familia exponencial, lo cual se puede demostrar de forma sencilla. Supongase que ; Fy y,θ es una función de distribución dependiendo de único parámetro θ. Se dice que pertenece a la familia exponencial si su función de densidad (o de masa, en el caso discreto) puede expresarse de la forma:

fy,$y;\theta = expa(\theta)b(Y) + C(\theta) + d(y)$

En el caso de distribuciones que dependen de k parámetros, por ejemplo θ=θ_1, θ_2, . . . ,θ_k , ,. . . , k , la definición anterior se generaliza de la siguiente manera (Tusell, 2003):

fy,$y;\theta = exp\sum_{i=1}^{k} a_i(\theta)b(Y) + C(\theta) + d(y)$

Manipulando algebraicamente y utilizando algunas propiedades de las funciones logarítmicas y exponenciales se obtiene finalmente:

$P = exp(\sum_{j=1}^{g-1} n_j log(\frac{p_i j}{p_i g}) + nlog(p_i g) + logA)$

Lo cual puede ser expresado como una función de la familia exponencial, mostrada en donde:

b_i(y)=n_j

c(θ)=n.logP_ij

a_i(θ)=log($\frac{P_i j}{P_i g}$)=1

EY_i

Que es el parámetro natural de la familia y también llamada la función de enlace canónico. Así queda demostrado que la distribución Multinomial pertenece a la familia exponencial.

Ahora bien, empleando el método de estimación por máxima verosimilitud que no establece restricción alguna respecto de las características de las variables predictoras, se estiman los parámetros $\beta_i j$_ij de la regresión lineal, la función de verosimilitud se define como sigue:

NO ME SALE LA PARTE DEL SUB-INDICE Y SUPERINDICE————————-

L=Πp_i1^{yi1,p_i2}yi2. . . ,p_igˆig

Al aplicar el logaritmo y remplazando los resultados anteriores, maximizando la función de verosimilitud se obtiene un sistema de ecuaciones que requiere de los métodos numéricos para encontrar la mejor solución.

Ln L=$\sum_{i=1}^{n}(\sum_{j=1}^{g-1} y_i j.ln\mathrm{e}^{x'i}\beta_j - \sum_{j=1}^{g-1} y_i j.ln(1 + \mathrm{e}^{x'i}\beta_j)$

EJEMPLO DE APLICACION

Sexo	*Edad*	*No/Poco*	*Importante*	*Muy Importante*	*Total*
H	18-23	25	12	7	42
H	24-40	9	21	15	45
H	>40	5	14	41	60
M	18-23	40	17	8	65
M	24-40	17	15	12	44
M	>40	8	15	18	41

	Sexo	Edad	Importancia	Observaciones
1	M	18-23	1	26.00
2	M	18-23	2	12.00
3	M	18-23	3	7.00
4	M	24-40	1	9.00
5	M	24-40	2	21.00
6	M	24-40	3	15.00
7	M	>40	1	5.00
8	M	>40	2	14.00
9	M	>40	3	41.00
10	H	18-23	1	40.00
11	H	18-23	2	17.00
12	H	18-23	3	8.00
13	H	24-40	1	17.00
14	H	24-40	2	15.00
15	H	24-40	3	12.00
16	H	>40	1	8.00
17	H	>40	2	15.00
18	H	>40	3	18.00

```
test1 <-glm(Importancia ~ Sexo + Edad,weights = Observaciones, family=binomial(link = logit), data=Nuev
test1
```

```
##
## Call:  glm(formula = Importancia ~ Sexo + Edad, family = binomial(link = logit),
##     data = Nuevo, weights = Observaciones)
##
## Coefficients:
## (Intercept)        SexoM    Edad18-23    Edad24-40
##      1.6020       0.5723      -2.2490      -0.9893
##
## Degrees of Freedom: 17 Total (i.e. Null);  14 Residual
## Null Deviance:       388.5
## Residual Deviance: 328.6     AIC: 336.6
```

```
summary(test1)
```

```
##
## Call:
## glm(formula = Importancia ~ Sexo + Edad, family = binomial(link = logit),
##     data = Nuevo, weights = Observaciones)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -5.963  -5.028   2.700   3.316   6.026
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6020     0.3267   4.904 9.39e-07 ***
## SexoM         0.5723     0.2708   2.114  0.03456 *
## Edad18-23    -2.2490     0.3577  -6.287 3.23e-10 ***
## Edad24-40    -0.9893     0.3801  -2.603  0.00925 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 388.47  on 17  degrees of freedom
## Residual deviance: 328.64  on 14  degrees of freedom
## AIC: 336.64
##
## Number of Fisher Scoring iterations: 5
```

```
coef(test1)
```

```
## (Intercept)       SexoM   Edad18-23   Edad24-40
##   1.6020444   0.5723414  -2.2489967  -0.9893373
```

```
fitted(test1)
```

```
##         1         2         3         4         5         6         7
## 0.4813559 0.4813559 0.4813559 0.7658543 0.7658543 0.7658543 0.8979256
##         8         9        10        11        12        13        14
## 0.8979256 0.8979256 0.3436767 0.3436767 0.3436767 0.6485581 0.6485581
##        15        16        17        18
## 0.6485581 0.8323039 0.8323039 0.8323039
```

En la salida anterior, lo primero que vemos es la llamada, esta es R, que nos recuerda qué modelo ejecutamos, qué opciones especificamos, etc.

A continuación vemos los residuos de desviación, que son una medida del ajuste del modelo. Esta parte de la salida muestra la distribución de los residuos de desviación para los casos individuales utilizados en el modelo. A continuación, discutimos cómo usar los resúmenes de la estad??stica de desviación para evaluar el ajuste del modelo.

La siguiente parte de la salida muestra los coeficientes, sus errores estándar, la estad??stica z (a veces denominada estad??stica z de Wald) y los valores p asociados. Ambos Sexo y los dos niveles de edad son estad??sticamente significativos. Los coeficientes de regresión log??stica proporcionan el cambio en las probabilidades de registro del resultado para un aumento de una unidad en la variable predictiva.