Regresión logística

Marcelo Molinatti

Contents

1 Regresión logística **1**
1.1 Ejercicios 2
1.2 Soluciones 2
1.2.1 Ecuaciones de Newton-Raphson sobre $L(\beta)$. 2
1.2.2 Implementación del método de Newton-Raphson en R. 3

1 Regresión logística

Los métodos de regresión que hemos introducido en los capítulos anteriores, no alcanzan cuando la variable de respuesta es discreta. En estos casos el método de regresión logística es una alternativa.

Consideremos para ilustrar el caso en el que la variable de respuesta $Y \in \{0, 1\}$, podemos pensar en dos categorías, por ejemplo presencia de eventos de hipertensión o no. Consideramos en un primer caso la dependencia de Y con una única variable explicativa o covariable X. En el caso de hipertensión podríamos considerar el consumo de sal.

Consideramos las probabilidades $P(Y = 1|X = x) = p(x)$ y $P(Y = 0|X = x) = 1 - p(x)$, queremos realizar inferencia sobre la probabilidad p. Como Y es una variable aleatoria con distribución Bernoulli. Utilizaremos el método de máxima verosimilitud. Sea $y_1 \dots y_n$, $x_1 \dots x_n$ una muestra de la variable Y y X respectivamente. La función de verosimilitud está dada por:

$$V = \prod_{i=1}^{n} p(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

al tomar logaritmo y agrupar términos tenemos

$$\begin{aligned} L &= logV \\ &= \sum_{i=1}^{n} y_i \log\, p(x_i) + (1 - y_i) \log\, (1 - p(x_i)) \\ &= \sum_{i=1}^{n} y_i \log\, \left(\frac{p(x_i)}{1 - p(x_i)} \right) + \log\, (1 - p(x_i)) \end{aligned}$$

el término $g(x) = \log\, (p(x)/(1 - p(x)))$ es llamado transformación logit, si despejamos y escribimos $p(x)$ en términos de $g(x)$ obtenemos

$$p(x) = \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{1}{1 + e^{-g(x)}}$$

Si suponemos la relación lineal $g(x) = \beta_0 + \beta_1 x$ expresamos L como

$$L = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \log\,(1 + e^{\beta_0+\beta_1 x_i})$$
$$= n\beta_0\overline{y} + n\beta_1\overline{yx} - \sum_{i=1}^{n} \log\,(1 + e^{\beta_0+\beta_1 x_i})$$

La función L puede ser maximizada numéricamente para hallar los estimados $\hat{\beta}_0$, $\hat{\beta}_1$.

Estos cálculos pueden ser extendidos al caso cuando tenemos un vector de covariables $X = (X_1, \ldots, X_k)$. En este caso tomamos $g(x) = \langle \beta, x \rangle$ para realizaciones del vector aleatorio $X_j = x_j$, $j = 1, \ldots, m$ y vector de parámetros β_j. En efecto, la función de verosimilitud queda escrita como

$$L = \sum_{i=1}^{n} y_i(\langle \beta, x_{ij} \rangle) - \log\,(1 + e^{\langle \beta, x_i \rangle})$$

donde se usa $x_{i0} = 1$ para todo $i = 1, \ldots, n$. Del mismo modo, los estimados de β_j pueden ser obtenidos numéricamente.

1.1 Ejercicios

1. Desarrolle las ecuaciones del método de Newton para la función de log-verosimilitud en ambos casos.
2. Implemente estas ecuaciones en Octave o R.
3. Utilice por lo menos dos conjuntos de datos del capitulo 1 del libro *Applied Logistic Regression*, de Hosmer *et al.* para probar su algoritmo y compare con los estimados obtenidos con las rutinas implementadas de Octave o R.

1.2 Soluciones

1.2.1 Ecuaciones de Newton-Raphson sobre $L(\beta)$.

Para encontrar los estimados $\hat{\beta}$ se debe resolver $\partial L(\beta)/\partial\beta_j = 0$ para cada uno de los k parámetros. Si se supone que comenzamos con un estimador inicial *suficientemente bueno* de β, digamos $\beta_{(0)}$, y se expande el vector de derivadas $L'(\hat{\beta})$ alrededor de la solución $\hat{\beta}$, por expansión de Taylor:

$$L'(\hat{\beta}) \approx L'(\beta_{(0)}) - L^{(2)}(\beta_{(0)}) \left[\hat{\beta} - \beta_{(0)}\right]$$

Como la función $L'(\hat{\beta})$ es igual a cero en $\hat{\beta}$, entonces se resuelve la ultima ecuación para $\hat{\beta}$, llamando a la solución $\beta_{(1)}$:

$$\beta_{(1)} = \beta_{(0)} + \left[L^{(2)}(\beta_{(0)})\right]^{-1} L'(\beta_{(0)})$$

Cuando solo se tiene un regresor, la ecuación anterior se reduce a:

$$\beta_{(1)} = \beta_{(0)} + \frac{L'(\beta_{(0)})}{L^{(2)}(\beta_{(0)})}$$

En el caso de la ecuación (1), la primera derivada con respecto al coeficiente β_j viene dada por:

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} y_i x_{ij} - \frac{x_{ij}}{1 + e^{-\langle \beta, x_i \rangle}}$$

y la segunda derivada queda:

$$\frac{\partial^2 L}{\partial \beta_j^2} = -\sum_{i=1}^{n} \frac{x_{ij}^2 e^{-\langle \beta, x_i \rangle}}{(1 + e^{-\langle \beta, x_i \rangle})^2}$$

1.2.2 Implementación del método de Newton-Raphson en R.

Para encontrar la solución final $\hat{\beta}$, se procede de forma iterativa, reemplazando en cada iteración el valor de $\beta_{(0)}$ por el de $\beta_{(1)}$, para encontrar una nueva actualización, y así sucesivamente hasta que el algoritmo converge: se alcanza una actualización que no difiere de la actualización anterior (dentro de una tolerancia dada lo suficientemente pequeña).