BetaReg: pacchetto R

Marta Rotari - Idriss Riouak

Università degli studi di Udine
Dipartimento di matematica e informatica
Applied Statistic and Data Analysis
idriss.riouak@spes.uniud.it marta.rotari@spes.uniud.it

16 febbraio 2018

Sommario

La regressione è un metodo statistico che permette l'analisi delle relazioni che intercorrono tra due variabili che possono assumere valori nel continuo o nel discreto. Lo scopo di questa relazione è quello di studiare e analizzare un modello di regressione nel quale il dominio delle variabili di risposta possono assumere valori nell'intervallo limitato $(0, 1)$. *Il modello analizzato è chiamato modello di regressione con variabili di risposta Beta, introdotto per la prima volta nel 2004 da Cribari-Neto e Ferrari [1]. In particolare andremo ad analizzare l'implementazione in* R *del modello, evidenziandone i pregi e difetti.*

I. Introduzione

Un modello di regressione è un modello statistico, il cui scopo è sia quello di studiare ed analizzare le relazioni tra una una variabile *dipendente*, detta variabile di risposta, e una o più variabili *indipendenti*, dette variabili esplicative, che di effettuare predizioni dato un nuovo valore per la variabile esplicativa.

Il *modello di regressione lineare semplice* ha la seguente forma

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{1}$$

dove la componente casuale ε_i è normalmente distribuita con media zero e varianza σ^2. Tale modello è ampiamente utilizzato in svariate applicazione, tuttavia non è appropriato per situazioni dove la variabile risposta è limitata ad assumere valori in un intervallo $(0, 1)$, in quanto, i valori stimati potrebbero eccedere tale intervallo [1].

Prima dell'avvento del modello di regressione con variabili Beta, per effettuare un'analisi in cui la variabile di risposta (v.r.) y assumeva valori in (0,1), era consuetudine effettuare delle trasformazioni di y. Dunque si considerava $\tilde{y} = log(\frac{y}{1-y})$ alla quale veniva applicato il modello di regressione lineare semplice. Tale approccio presentava le seguenti problematiche:

I Disugaglianza di Jensen: ovvero i parametri dovevano essere interpretati rispetto il valore atteso di $\tilde{y}$ anziché rispetto quello di y.

II Eteroschedasticità: la varianza aumentava all'avvicinarsi della media e decresceva spostandosi versoi limiti dell'intervallo.

III Asimmetria: in generale la distribuzione di tassi e di proporzioni è asimmetrica e dunque la stima degli intervalli per il test dell'ipotesi basate su approssimazioni Gaussiane potrebbero essere imprecise per campioni di piccole dimensioni.

Nel 2004, Cribari-Neto e Ferrari, con l'articolo *"Beta Regression for Modelling Rates and Proportions"* [1], descrivono come il modello di regressione con variabili Beta sia il migliore per trattare proporzioni e tassi. Successivamente nel 2016, nell'articolo *"Beta Regression in R"* [2], i due autori forniscono un'implementazione in R di tale modello.

[1]Un esempio classico è «Teaching Program»[3](pg. 67)

II. Beta distribuzione

Come già anticipato, il modello di regressione con variabili Beta, d'ora in avanti BetaReg, si presta perfettamente per modellare situazioni in cui la variabile di risposta y assuma valori nell'intervallo aperto (0,1)[2].

BetaReg è basato su un'alternativa parametrizzazione della funzione di densità della distribuzione beta; la funzione di densità di una variabile casuale (v.c.) beta è data nel seguente modo

$$f(y;p,q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1}, \quad 0 < y < 1$$

Dove $p > 0$ e $q > 0$ e $\Gamma(\cdot)$ è la funzione gamma.

Ferrari e Cribari-Neto ne hanno proposto una parametrizzazione differente:

$$f(y,\mu,\phi) = \frac{\Gamma\phi}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi^{-1}}(1-y)^{(1-\mu)\phi-1}$$

con

$$\mu = \frac{p}{p+q}, \qquad \phi = p+q$$

dove $0 < \mu < 1$, $\phi > 0$ e $0 < y < 1$.

Denoteremo con $y \sim \mathcal{B}(\mu,\phi)$ se la v.c. y segue una beta distribuzione con parametri μ e ϕ. Si noti che $p = \mu\phi$ e $q = \phi(1-\mu)$, da cui segue che $E(y) = \mu$ e che $VAR(y) = \frac{V(\mu)}{1+\phi} = \frac{\mu(1-\mu)}{1+\phi}$.

Il parametro ϕ è anche chiamato *parametro di precisione*, in quanto per un fissato μ, all'aumentare di ϕ diminuisce il valore della varianza.

III. Il modello di regressione Beta

Sia $y_1, y_2, ..., y_n$ un campione casuale tale che $\forall_{i=1}^{n} : y_i \sim \mathcal{B}(\mu_i,\phi)$. Il modello di regressione

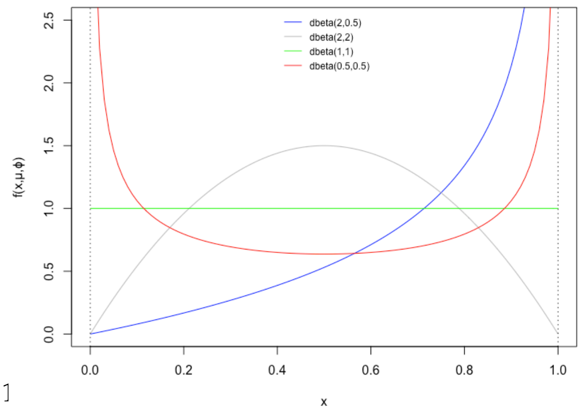

Figura 1: *Rappresentazione grafica della distribuzione Beta, utilizzando il comando R:* dbeta.

Beta è definito nel seguente modo

$$g(\mu_i) = x_i^t\beta = \eta_i \tag{2}$$

dove $\beta = (\beta_1,\beta_2,...,\beta_k)^t$, con $k < n$, è un vettore $k \times 1$, $x_i = (x_{i1}, x_{i2}, ..., x_{ik})^t$ è un vettore di k variabili esplicative mentre $\eta_i = \beta_1 x_{i1} ... \beta_k x_{ik}$[3] è un predittore lineare. Infine $g(\cdot) : (0,1) \to \mathbb{R} \in \mathcal{C}^2$ è una funzione di collegamento avente derivata seconda costante. Le funzioni di collegamento più utilizzate sono:

- **logit:** $g(\mu) = log(\frac{\mu}{(1-\mu)})$
- **probit:** $g(\mu) = \Phi^{-1}(\mu)$, dove $\Phi(\cdot)$ è la funzioni di distribuzione normale.
- **log-log complementare:** $g(\mu) = \log(-\log(1-\mu))$
- **log-log:**$g(\mu) = \log(-\log(\mu))$
- **Cauchy:** $g(\mu) = \tan(\pi(\mu - 0.5))$

Denotiamo con $l(\beta,\phi) = \sum_{i=1}^{n} l_i(\mu_i,\phi)$ la funzione di verso somiglianza, dove

$$\begin{aligned} l_i(\mu_i,\phi) = &\log\Gamma(\phi) - \log(\mu_i\phi) \\ &- \log\Gamma((1-\mu_i)\phi) + (\mu_i\phi - 1)\log y_i \\ &+ \{(1-\mu_i)\phi - 1\}\log(1-y_i) \end{aligned}$$

con μ_i definito come nell'equazione (2) ovvero $\mu_i = g^{-1}(x_i^t\beta)$.

[2] Si noti che se la variabile y dovesse assumere valori nell'intervallo (a,b), dove $a < b$ e sia a che b sono valori noti, allora è possibile modellare $\frac{y-a}{b-a}$ al posto di y. Mentre se la variabile y dovesse assumere come valori in [0,1], una possibile trasformazione potrebbe essere $\frac{y\cdot(n-1)\cdot 0.5}{n}$ dove n è la grandezza del campione.

[3] Per convenzione $x_{i1} = 1$. In tal modo ogni modello ha l'intercetta (null-model) [3].

i. Determinizzazione degli stimatori.

Siano
$$y_t^* = log(\frac{y_t}{1-y_t})$$
e
$$\mu_t^* = \psi(\mu_t \phi) - \psi((1-\mu_t)\phi),$$
dove $\psi(x) = \frac{\partial \log \Gamma(x)}{\partial x}$ con $x > 0$ è detta funzione *digamma.* Denotiamo con
$$\nabla(\beta, \phi) = \begin{pmatrix} U_\beta(\beta,\phi) \\ U_\phi(\beta,\phi) \end{pmatrix}$$
la funzione *score*, ottenuta differenziando la funzione di log-verosimiglianza rispetto i due parametri sconosciuti. Dunque
$$U_\beta(\beta,\phi) = \frac{\partial l(\beta,\phi)}{\partial \beta} = \phi X^t T(y^* - u^*),$$
dove X è la matrice del modello di dimensione $n \times k$, T è una matrice diagonale la cui dimensione $n \times n$ definita come $T = diag\{g'(\mu)_1^{-1}, ..., g'(\mu_i)^{-1}\}$, $y^* = (y_1^*, ..., y_n^*)$ e $\mu^* = (\mu_1^*, ..., \mu_n^*)$. Mentre
$$\begin{aligned} U_\phi(\beta,\phi) = \sum_{t=1}^{n} \{ & \mu_t(y_t^* - \mu_t^*) + log(1-y_t) \\ & - \phi((1-\mu_t)\psi) + \phi(\psi)\} \end{aligned}$$
Possiamo dunque concludere che gli stimatori di massima verosimiglianza (MLEs) per β e ϕ sono ottenibili ponendo rispettivamente $U_\beta(\beta,\phi)$ e $U_\phi(\beta,\phi)$ uguali a zero. Tale tipo di equazioni non sono risolvibili analiticamente, ma il risultato può essere approssimato attraverso un algoritmo numerico quale l'algoritmo di *Newton*. Tali algoritmi necessitano di un punto di partenza (β_0, ϕ_0), che nel caso di β utilizzando il metodo dei minimi quadrati è
$$\beta_0 = (X^t X)^{-1} X^t z,$$
dove $z = (g(y_1), ..., g(y_n))^t$. Mentre per ϕ, *Ferrari e Cribari-Neto* in [1] suggeriscono come punto di partenza
$$\phi_0 = \frac{1}{n}\sum_{t=1}^{n} \frac{\breve{\mu}_t(1 - l\breve{\mu}_t)}{\breve{\sigma}_t^2},$$
dove $\breve{\mu}_t$ è ottenuto applicando la funzione $g^{-1}(\cdot)$ al *t-esimo* valore stimato dal modello di regressione lineare di $g(y_1), ..., g(y_n)$ su X:
$$\breve{\mu}_t = g^{-1}(x_t^t (X^t X)^{-1} X^t z)$$
e
$$\breve{\sigma}_t^2 = \frac{\breve{e}^t \breve{e}}{(n-k)[g'(\breve{\mu})_t]^2}$$
dove $\breve{e} = z - X(X^t X)^{-1} X^t z$.

Consideriamo ora la matrice d'informazione di *Fisher*, che servirà per poter approssimare l'errore standard degli stimatori $\hat{\beta}$ e $\hat{\phi}$. Poniamo prima $W = diag\{w_1, ..., w_n\}$, con
$$w_t = \phi\{\psi'(\mu_t\phi) + \psi'((1-\mu_t)\phi)\}\frac{1}{\{g'(\mu_t)\}^2},$$
$c = (c_1, ..., c_n)^t$, dove
$$c_t = \phi\{\psi'(\mu_t\phi)\mu_t - \phi'((1-\mu_t)\phi)(1-\mu_t)\}$$
e $\psi'(\cdot)$ è la funzione *trigamma*, definita come segue
$$\psi'(x) = \frac{\partial^2}{\partial z^2} \log \Gamma(x).$$
Sia dunque K la matrice d'informazione di *Fisher*:
$$K = K(\beta,\phi) = \begin{pmatrix} K_{\beta\beta} & K_{\beta\phi} \\ K_{\phi\beta} & K_{\phi\phi} \end{pmatrix}, \qquad (3)$$
dove

- $K_{\beta\beta} = \phi X^t W X$,
- $K_{\beta\phi} = K_{\phi_\beta}^t = X^t T c$,
- $K_{\phi\phi} = tr(D)$.

Sotto le condizioni di normalità, d'indipendenza e di omogeneità di varianza delle variabili, quando la grandezza del campione è grande, vale che
$$\begin{pmatrix} \hat{\beta} \\ \hat{\phi} \end{pmatrix} \sim \mathcal{N}_{k+1}\left(\begin{pmatrix} \beta \\ \phi \end{pmatrix}, K^{-1} \right).$$
Denoteremo con $SE(\hat{\beta}_j)$ l'errore standard asintottico del MLE $\hat{\beta}_j$, che si ottiene dall'inversa della matrice di *Fisher* (3) valutata in $\hat{\beta}_j$ e in $\hat{\phi}$.

ii. Intervallo di confidenza

E' possibile determinare un intervallo di confidenza $(1-\alpha)100\%$[4] per i coefficienti $\hat{\beta}_j$, con $j = 1, ..., k$. Tale intervallo è:

$$\left[\hat{\beta}_j \pm \Phi^{-1}\left(1 - \frac{\alpha}{2} SE(\hat{\beta}_j)\right)\right],$$

dove $\Phi(\cdot)$ è la funzione di distribuzione cumulativa di una variabile casuale normale.

Analogamente un intervallo di confidenza $(1-\alpha)100\%$ per il parametro $\hat{\phi}$ è il seguente

$$\left[\hat{\phi} \pm \Phi^{-1}\left(1 - \frac{\alpha}{2} SE(\hat{\phi})\right)\right]$$

dove

$$\begin{aligned} SE(\hat{\phi}) &= \sqrt{tr(D) - \phi^{-1} c^t T^t X (X^t W X)^{-1} X^t T c} \\ &= \sqrt{\hat{\gamma}}. \end{aligned}$$

In fine è possibile determinare un intervallo di confidenza $(1-\alpha)100\%$ per il valore atteso della variabile risposta μ per un dato vettore d'osservazioni delle variabili regressori $x_0 = [1, x_{01}, x_{02}, ..., x_{0k}]$:

$$[Lim_{sx}, Lim_{dx}]^5$$

dove

$$Lim_{sx} = \left[g^{-1}\left(\hat{\eta} - \Phi^{-1}\left(\frac{1-\alpha}{2}\right) SE(\hat{\eta})\right)\right]$$

mentre

$$Lim_{dx} = \left[g^{-1}\left(\hat{\eta} + \Phi^{-1}\left(\frac{1-\alpha}{2}\right) SE(\hat{\eta})\right)\right],$$

con $\hat{\eta} = x_0^t \hat{\beta}$ e $SE(\hat{\eta}) = \sqrt{x_0^t \widehat{\text{cov}}(\hat{\beta}) x_0}$ dove $\widehat{\text{cov}}(\hat{\beta})$ è ottenuto dall'inversa della matrice di *Fisher* (3) valutata negli MLEs escludendo la riga e la colonna relative al parametro di precisione $\hat{\phi}$.

[4] con $\alpha \in (0, \frac{1}{2})$.

[5] Tale intervallo è valido solo per funzioni di collegamento $g(\cdot)$ strettamente crescenti.

Indice

Riferimenti bibliografici

[1] **Beta Regression for Modelling Rates and Proportions.**, Ferrari SLP, Cribari-Neto Francisco (2004). Journal of Applied Statistics, 31(7), 799815.

[2] **Beta Regression in R**, Francisco Cribari-Neto, Achim Zeileis.

[3] **Towards multiple linear regression and logistic regression**, Paolo Vidoni, 2017-2018. Lecture 5. Applied Statistics and Data Analysis.