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

10 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

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

dove la componente casuale ε_i è normalmente distribuita con media zero e varianza σ^2.

Prima dell'avvento del modello di regressione con variabili Beta, per effettuare un'analisi di regressione 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.

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)[1].

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, 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 $0 < y < 1$.

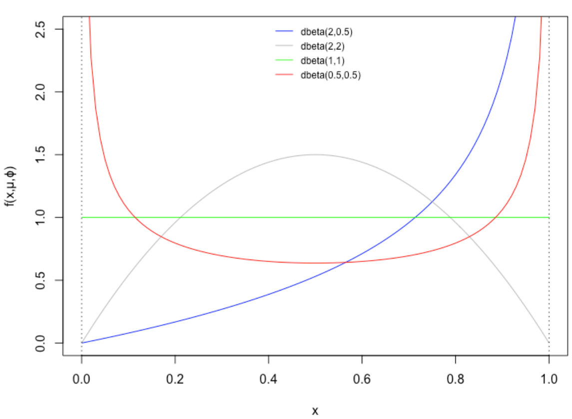

Figura 1: *Rappresentazione grafica della distribuzione Beta, utilizzando il comando R:* dbeta.

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 regessione Beta è definito nel seguente modo

$$g(\mu_i) = x_i^t\beta = \eta_i$$

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 $\eta_i = \beta_1 x_{i1} ... \beta_k x_{ik}$[2] 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:** $g(\mu) = log(\frac{\mu}{(1-\mu)})$
- **porbit:** $g(\mu) = \Phi^{-1}(\mu)$, 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

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

[1]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.

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

Indice

List of Algorithms

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.