Modelli per dati binari

Vers. 1.0.0

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

Dati Binomiali I

Ipotizziamo che i dati y_i provengano da una Binomiale di parametro (π_i, n_i).

$$g : [0,1] \to \mathbb{R}$$
$$\mu_i \to \eta_i$$

La densità è quindi

$$(*) \quad \binom{n_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{n_i - y_i}$$

Siamo interessati a modellizzare π_i e utilizziamo la funzione link logistica:

$$\eta_i = g(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) \Rightarrow \pi_i = \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i\boldsymbol{\beta})}$$

Abbiamo che

$$\boxed{log\left(\frac{\pi_i}{1-\pi_i}\right) = \text{logit}(\pi_i) = \mathbf{x}_i\boldsymbol{\beta}}$$

Dati Binomiali II

Per interpretare i parametri possiamo innanzitutto vedere che

$$1 - \pi_i = (1 + \exp(\mathbf{x}_i \boldsymbol{\beta}))^{-1} \qquad (*)$$

che mostra come π_i sia monotona in ogni variabile x, con il segno del β che ne decide il "verso".

Il valore di β_j si può interpretare calcolando

$$\frac{1}{\underbrace{1+\exp(x_i\beta)}_{1-\pi_i}} \cdot \underbrace{\frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}}_{\pi_i}$$

$$\frac{\partial \pi_i}{\partial x_{ij}} = \beta_j \frac{\exp(\mathbf{x}_i \boldsymbol{\beta})}{(1 + \exp(\mathbf{x}_i \boldsymbol{\beta}))^2} = \beta_j \pi_i (1 - \pi_i)$$

β_j determina la pendenza della derivata, ed è massima a $\pi = 0.5$, con valore $\beta_j/4$.

Dati Binomiali III

Il metodo precedente funzione meglio con le variabili quantitative, per le qualitative possiamo considerare l'odds-ratio.

Consideriamo un modello in cui abbiamo una variabile (~~quantitativa~~ qualitativa) x_{ij} che assume valore 1 se si appartiene ad un gruppo specifico, e 0 altrimenti. Ipotizziamo anche che l'osservazione i ha $\underline{x_{i2} = 1}$ e la h ha $\underline{x_{h2} = 0}$. Consideriamo

$$\text{OR} = \frac{\pi_i/(1-\pi_i)}{\pi_h/(1-\pi_h)} = \frac{\exp\left(\cancel{\beta_1} + \beta_2 + \cancel{\sum_{j=3}^{p} x_{ij}\beta_j}\right)}{\exp\left(\cancel{\beta_1} + \cancel{\sum_{j=3}^{p} x_{hj}\beta_j}\right)} = \exp(\beta_2)$$

ODDS = exp(logit) !

e il suo logaritmo

$$\log(\text{OR}) = \text{logit}(\pi_i) - \text{logit}(\pi_h) = \beta_2$$

Valori positivi di β_2 ci dicono che passando da $x_{h2} = 0$ a $x_{h2} = 1$, la prob cresce.

Dati Binomiali IV

Possiamo calcolare la statistica U come

$$U_j = \sum_{i=1}^{n} \underbrace{\frac{(y_i - \mu_i)x_{ij}}{\text{Var}(y_i)}} \underbrace{\frac{\partial \mu_i}{\partial \eta_i}}$$

ricordando che

- $E(y_i) = n\pi_i = \mu_i$ (*)
- $\text{Var}(y_i) = n_i\pi_i(1-\pi_i)$

e dato che $\eta_i = \log\left(\frac{\mu_i/n_i}{1-\mu_i/n_i}\right)$, abbiamo che

$\eta_i = \text{logit}(\pi_i)$

DA (*) $\to$ $\pi_i = \frac{\mu_i}{n}$

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{n_i\pi_i(1-\pi_i)}$$

e quindi

$$\to \quad U_j = \sum_{i=1}^{n} \frac{(y_i - n_i\pi_i)x_{ij}}{\cancel{n_i\pi_i(1-\pi_i)}} \cancel{n_i\pi_i(1-\pi_i)} = \sum_{i=1}^{n} (y_i - n_i\pi_i)x_{ij}$$

β E' QUI!

Dati Binomiali V

Ricordando che

$$\mathbf{J} = \mathbf{X}^T \mathbf{W} \mathbf{X}$$

con

$$W_{ii} = \frac{1}{\text{Var}(y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \longrightarrow \frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{Var(y_i)}$$

e che $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \mathbf{J}^{-1})$, possiamo calcolare $\underline{\hat{\mathbf{J}}}$ che è

$$\boxed{\mathbf{X}^T \text{diag}(n_i \hat{\pi}_i (1 - \hat{\pi}_i)) \mathbf{X}}$$

Che ci permettono di fare test di ipotesi su $\boldsymbol{\beta}$.

Dati Binomiali VI

Possiamo anche calcolare in maniera esplicita la devianza notando che

$$\hat{y}_i = n_i \hat{\pi}_i \quad !!!$$

e che il valore stimato di $\hat{\pi}$ nel modello saturo è y_i/n_i, o, detta diversamente, la previsione è $\hat{y}_i = y_i$.
Abbiamo che

$$\underbrace{D = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right)}_{} \sim \chi^2_{n-p}$$

Dati Binomiali VII

Se vogliamo confrontare modelli annidati con l'ipotesi

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \qquad H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1$$

dove $\boldsymbol{\beta}_0$ ha q elementi (modello M_0) e $\boldsymbol{\beta}_1$ ha p elementi (modello M_1), con $q < p < n$,
PICCOLO GRANDE
possiamo usare Possiamo calcolare

$$\Delta D = D_0 - D_1 = 2\left(L(\mathbf{y}; \hat{\boldsymbol{\beta}}_1) - L(\mathbf{y}; \hat{\boldsymbol{\beta}}_0)\right)$$

$$\Delta D = 2\sum_{i=1}^{n}\left(y_i \log\left(\frac{\hat{y}_i^1}{\hat{y}_i^0}\right) + (n_i - y_i)\log\left(\frac{n_i - \hat{y}_i^1}{n_i - \hat{y}_i^0}\right)\right) \sim \chi_{p-q,\nu_0-\nu_1}$$

dove

- $\hat{y}_i^j = n_i \hat{\pi}_i^j$ è la previsione sotto il modello $j-$esimo con stima $\hat{\pi}_i^j$;
- $\underline{\nu_0 = 2\left(L(\boldsymbol{\beta}_{max}; \mathbf{y}) - L(\boldsymbol{\beta}_0; \mathbf{y})\right)}$;
- $\underline{\nu_1 = 2\left(L(\boldsymbol{\beta}_{max}; \mathbf{y}) - L(\boldsymbol{\beta}_1; \mathbf{y})\right)}$;

Dati Binomiali VIII

$$\hat{\mu}_i = n_i \cdot \hat{\pi}_i$$
$$Var(y_i) = n_i \hat{\pi}_i (1-\hat{\pi}_i)$$

Definiamo i residui

- Residui di Pearson

$$e_i = \underbrace{\frac{y_i - \hat{\mu}_i}{\sqrt{\mathsf{Var}(\underset{y}{\cancel{\hat{\mu}}}_i)}}}_{} = \underbrace{\frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}}_{}$$

- deviance residuals

$$d_i = \mathsf{sign}(y_i - n_i\hat{\pi}_i)\sqrt{y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)}$$

Dati Binomiali IX

La somma dei residui di Pearson al quadrato è usato come un indice di model fitting:

$$Z^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

La distribuzione approssimata la si può trovare calcolando l'espansione di Taylor della funzione $s \log \frac{s}{t}$ nel punto $s = t$

$$g(s) = s \log \frac{s}{t} = (s-t) + \frac{1}{2}\frac{(s-t)^2}{t} + \dots$$

applicata a D, ottenendo

$$D = 2\sum_{i=1}^{n}\left((y_i - n_i\hat{\pi}_i) + \frac{1}{2}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i} + ((n_i - y_i) - (n_i - n_i\hat{\pi}_i)) + \right.$$
$$\left. \frac{1}{2}\frac{((n_i - y_i) - (n_i - n_i\hat{\pi}_i))^2}{n_i - n_i\hat{\pi}_i} + \dots \right)$$

Facendo tutti i calcoli si vede che

$$Z^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)} \approx D \sim \chi^2_{n-p}$$

Generalmente Z^2 performa meglio di D.

Modelli per dati Multinomiali

Gianluca Mastrantonio

gianluca.mastrantonio@polito.it

Dati MUltinomiali I

La distribuzione può essere considerata della famiglia esponenziale perchè è la distribuzione di variabili Poisson, condizionatamente alla loro somma n. Ipotizziamo di avere J variabili, ognuna da $Y_j \sim \text{Pois}(\lambda_j)$. La loro densità congiunta è

$$\prod_{j=1}^{J} \frac{\lambda_j^{y_j} e^{-\lambda_j}}{y_j!}$$

Siamo interessati a trovare

$$f(\mathbf{y}|n) = \frac{f(\mathbf{y}, n)}{f(n)} = \boxed{\frac{f(\mathbf{y})}{f(n)}}$$

dove $\mathbf{y}$ è il vettore delle y_j osservate

y_i	y_2	y_3	y_4	y_5

$n = \sum_{i=1}^{5} y_i$

Dati MUltinomiali II

Ipotizziamo che $J = 2$, la distribuzione di n si trova come

$n = Y_1 + Y_2$

- $P(n = k) = P(Y_1 + Y_2 = k) = \sum_{i=0}^{k} P(Y_2 = k - i)P(Y_1 = i)$

che è uguale a

$\frac{e^{-\lambda_1}\lambda_1^i}{i!} \cdot \frac{e^{-\lambda_2}\lambda_2^{k-i}}{(k-i)!}$

$$\frac{e^{-(\lambda_1+\lambda_2)}}{k!} \sum_{i=0}^{k} \frac{k!\lambda_1^i \lambda_2^{k-i}}{i!(k-i)!}$$

Poisson con $\lambda = \lambda_1 + \lambda_2$

$$= \boxed{\frac{(\lambda_1 + \lambda_2)^k e^{-(\lambda_1+\lambda_2)}}{k!}} \boxed{\sum_{i=0}^{k} \frac{k!}{i!(k-i)!}^{(*)} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^i \left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{k-i}} = 1$$

i termini dentro la sommatoria sono la densità di una binomiale con k estrazioni e prob $\frac{\lambda_1}{\lambda_1+\lambda_2}$. La sommatoria è quindi pari a 1 (somma su tutti i possibili valori) e questo dimostra che la somma n è Poisson. Per inmduzione si può estendere a qualsiasi J.

$(*)\ \binom{k}{i}$

Dati MUltinomiali III

Abbiamo quindi che

$n = y_1 + y_2 + \dots y_J$

$$f(\mathbf{y}|n) = \frac{\prod_{j=1}^{J} \frac{\lambda_j^{y_j} \cancel{e^{-\lambda_j}}}{y_j!}}{\frac{(\sum_{j=1}^{J} \lambda_j)^n \cancel{e^{-\sum_{j=1}^{J} \lambda_j}}}{n!}} = \underbrace{\left(\frac{\lambda_1}{\sum_{j=1}^{J} \lambda_j}\right)^{y_1} \cdots \left(\frac{\lambda_J}{\sum_{j=1}^{J} \lambda_j}\right)^{y_J}} \frac{n!}{y_1! \dots y_J!}$$

e definendo

$$\pi_j = \frac{\lambda_j}{\sum_{j=1}^{J} \lambda_j}$$

abbiamo la multinomiale

$$\left(\sum_{j=1}^{J} \lambda_j\right)^n = \left(\sum_{j=1}^{J} \lambda_j\right)^{y_1} \cdot \left(\sum_{j=1}^{J} \lambda_j\right)^{y_2} \cdot \dots$$

Dati MUltinomiali IV

La multinomiale si modella come sequenze di Binomiali.

Per esempio si può considerare la **logistica nominale** in cui si modella

$$\text{logit}(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = \mathbf{X}_j\boldsymbol{\beta}_j$$

con classe di riferimento la prima (potrebbe essere qualsiasi altra).

Il valore stimato di π_j, esclusa la classe di riferimento, è

$$\boxed{\hat{\pi}_j} = \hat{\pi}_1 \exp\left(\mathbf{X}_j\hat{\boldsymbol{\beta}}_j\right) = \frac{\exp\left(\mathbf{X}_j\hat{\boldsymbol{\beta}}_j\right)}{\underbrace{1 + \sum_{j=2}^{J} \exp\left(\mathbf{X}_j\hat{\boldsymbol{\beta}}_j\right)}}$$

con

$$\hat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp\left(\mathbf{X}_j\hat{\boldsymbol{\beta}}_j\right)}$$

Dati MUltinomiali V

Altri approcci sono

$$P_2 = \log\left(\frac{\pi_1+\pi_2}{\pi_3+\pi_4+\dots}\right)$$

- **Cumulativa**: si modella

$$log\left(\frac{\pi_1+\dots+\pi_j}{\pi_{j+1}+\dots+\pi_J}\right) = \mathbf{X}_j\boldsymbol{\beta}_j$$

- **Categorie adiacenti**: si modella

$$\log\left(\frac{\pi_1}{\pi_2}\right), \log\left(\frac{\pi_2}{\pi_3}\right), \log\left(\frac{\pi_{J-1}}{\pi_J}\right)$$

  con $\log\left(\frac{\pi_j}{\pi_{j+1}}\right) = \mathbf{X}_j\boldsymbol{\beta}_j$

- **Continuous ratio**: si modella

$$log\left(\frac{\pi_j}{\pi_{j+1}+\dots+\pi_J}\right) = \mathbf{X}_j\boldsymbol{\beta}_j$$

- ...

Dati MUltinomiali VI

Tutta l'inferenza sul modello:

- Interpretazione dei parametri;
- test sui parametri;
- test sul modello;
- residui;
- ...

si fa sui singoli rapporti di probabilità