

Metodi per l'Analisi dei Dati Sperimentali

AA2009/2010



RIPARAMETRIZZAZIONE - INTERVALLI DI FIDUCIA

Gaetano D'Avino

Esercitazioni

Lezione 8

15/12/09



Riparametr. – Intervalli di fiducia

Esercitazioni
Lezione 8
15/12/09

Sommario

- **Riparametrizzazione**
- **Intervalli di fiducia: introduzione**
- **Intervalli di fiducia: esempi**
 - **Gaussiana con varianza nota**
 - Intervallo di fiducia per la media
 - **Gaussiana con varianza non nota**
 - Intervallo di fiducia per la media
 - Intervallo di fiducia per la varianza
 - **Regressione lineare**
 - Intervallo di fiducia per i parametri
 - Intervalli di fiducia per y al variare di x
 - Regioni di fiducia

*La teoria si trova nel file "**Lezione09.pdf**" sul sito del Prof. Maffettone*



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- Quando si ha a che fare con modelli lineari o multilineari, è possibile procedere alla RIPARAMETRIZZAZIONE
- La riparametrizzazione consiste nel trasformare il modello lineare/multilineare in un modello analogo più facile da risolvere (algebricamente e numericamente)
- In teoria si è visto il caso lineare:

$$y = a + bx \quad \longrightarrow \quad y = a + bx + b\bar{x} - b\bar{x} \quad \longrightarrow \quad y = a + b\bar{x} + b(x - \bar{x})$$



$$y = p_0 + p_1(x - \bar{x})$$

$$\begin{cases} p_0 = a + b\bar{x} \\ p_1 = b \end{cases}$$

- Quindi si ottiene un nuovo modello nei parametri p_0 e p_1



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- Il vantaggio della riparametrizzazione è che il calcolo della stima dei nuovi parametri è più agevole essendo le equazioni disaccoppiate (vedi teoria)
- L'estensione al caso multilineare è immediata

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + \dots$$



$$y = p_0 + p_1(x_1 - \bar{x}_1) + p_2(x_2 - \bar{x}_2) + p_3(x_3 - \bar{x}_3) + \dots$$

dove:
$$\begin{cases} p_0 = b_0 + b_1 \bar{x}_1 + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \dots \\ p_1 = b_1 \\ p_2 = b_2 \\ p_3 = b_3 \\ \dots \end{cases}$$



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- Nel caso multilineare, le equazioni da risolvere non sono disaccoppiate eccetto quella relativa all'intercetta
- Tuttavia è sempre conveniente riparametrizzare in quanto si hanno miglioramenti (numerici) della matrice caratteristica da costruire per risolvere il sistema lineare
- ESEMPIO:

Si consideri l'esempio già visto per il caso multilineare

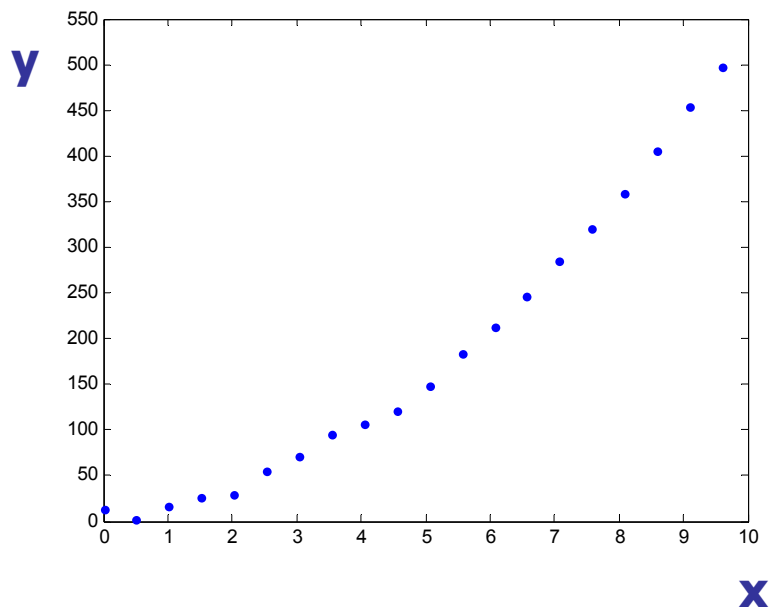
Sono state effettuate 20 prove (indipendenti) al variare delle condizioni sperimentali. I dati sono nel file "datiMultiL.txt"



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- Il diagramma dei dati è:



- Si suppone che un modello parabolico sia adeguato:

$$y = a + bx + cx^2$$



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- Vedemmo che, risolvendo il sistema lineare:

$$(X^T \cdot X) \hat{\theta} = X^T \cdot y$$

dove:

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_N & x_N^2 \end{bmatrix}$$

si avevano le stime dei parametri:

$$\hat{a} = 7.6237$$

$$\hat{b} = 4.6601$$

$$\hat{c} = 4.8487$$



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- La "bontà" numerica di una matrice contenente i coefficienti di un sistema lineare si stima attraverso il **numero di condizionamento (cond)**
- Tralasciando la sua definizione rigorosa, diremo solo che:
 - "cond" è compreso nell'intervallo $[1, +\infty]$
 - se "cond" $\rightarrow 1$, la matrice è numericamente "buona" (*problema ben condizionato*)
 - se "cond" $\rightarrow \infty$, la matrice è vicina alla singolarità (*problema mal condizionato*)
- Se la matrice è mal condizionata, NON ESISTONO algoritmi numerici in grado di risolvere accuratamente il sistema lineare
- Quindi, il numero di condizionamento dovrebbe essere il più piccolo possibile



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- In Matlab, il numero di condizionamento di una matrice si calcola col comando:

`cond(A)`

dove A è la matrice del sistema lineare

- Calcoliamo allora il numero di condizionamento per la matrice nel nostro caso multilineare:

$$\text{cond}(X^T \cdot X) \cong 16230$$

- Che succede al numero di condizionamento se ripeto la stima dei parametri con il modello riparametrizzato?



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- Riparametrizziamo quindi il modello:

$$y = a + bx + cx^2 + b\bar{x} - b\bar{x} + c\bar{x}^2 - c\bar{x}^2$$



$$y = p_0 + p_1(x - \bar{x}) + p_2(x^2 - \bar{x}^2)$$

$$\begin{cases} p_0 = a + b\bar{x} + c\bar{x}^2 \\ p_1 = b \\ p_2 = c \end{cases}$$

- La nuova matrice **X** sarà:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 - \bar{x} & x_1^2 - \bar{x}^2 \\ 1 & x_2 - \bar{x} & x_2^2 - \bar{x}^2 \\ \dots & \dots & \dots \\ 1 & x_N - \bar{x} & x_N^2 - \bar{x}^2 \end{bmatrix}$$

da cui si ricava: $\text{cond}(\mathbf{X}^T \cdot \mathbf{X}) \cong 1478$



Riparametrizzazione

Esercitazioni
Lezione 8
15/12/09

- Ovviamente, la stima dei parametri "a", "b", "c", ottenuta ritrasformando le stime di " p_0 ", " p_1 ", " p_2 ", è identica alla precedente senza riparametrizzazione
- La differenza consiste nella migliore qualità numerica della matrice $\mathbf{X}^T \mathbf{X}$ che riduce errori numerici per la risoluzione del sistema di equazioni lineari



Intervalli di fiducia: introduzione

Esercitazioni
Lezione 8
15/12/09

- Finora abbiamo considerato stime PUNTUALI dei parametri (di una VA Gaussiana, di una regressione, ecc.)
- Comunque non è possibile conoscere in maniera esatta il valore di un parametro
- Ciò ci motiva ad aggiungere una **STIMA DI UN INTERVALLO** in cui, con sufficiente garanzia, cade il valore del parametro
- Ovviamente non è possibile conoscere al 100% un intervallo in cui cade un parametro (o meglio, tale intervallo è $[-\infty, +\infty]$)
- Questi intervalli si chiamano **INTERVALLI DI FIDUCIA** (oppure intervalli di confidenza)



Intervalli di fiducia: definizione

Esercitazioni
Lezione 8
15/12/09

DEFINIZIONE: Sia (Y_1, \dots, Y_N) un campione casuale proveniente da una popolazione di cui interessa stimare il parametro θ . Siano $G_1 = g_1(Y_1, \dots, Y_N)$ ed $G_2 = g_2(Y_1, \dots, Y_N)$ due statistiche tali che $G_1 \leq G_2$. Se si può scrivere $P(G_1 < \theta < G_2) = 1 - \alpha$ allora l'intervallo aleatorio (G_1, G_2) è un intervallo fiduciario per θ al $(1 - \alpha)100\%$.

- La quantità $(1 - \alpha)$ prende il nome di coefficiente fiduciario e G_1 e G_2 sono i limiti fiduciari inferiore e superiore
- Si noti che α è un parametro da scegliere (in genere $\alpha = 0.01, 0.05, 0.1$) mentre G_1 e G_2 sono i parametri da stimare
- Al variare di α , l'intervallo cambia la dimensione dell'intervallo: al diminuire di α , aumenta la dimensione dell'intervallo
 - Per $\alpha = 0 \rightarrow G_1 = -\infty$ e $G_2 = +\infty$



Intervalli di fiducia

Esercitazioni
Lezione 8
15/12/09

- A seconda del problema in esame, si possono avere diverse procedure per la stima degli intervalli di fiducia (CI)
- In questa esercitazione si forniranno solo le "ricette" per la stima di intervalli di fiducia. Per i dettagli si veda la teoria
- Distingueremo i seguenti casi:
 1. Stima di CI per la media di una VA Gaussiana con varianza nota
 2. Stima di CI per la media e la varianza di una VA Gaussiana entrambe non note
 3. Stima di CI per i parametri di una regressione lineare
 4. Stima della regione di fiducia per i parametri di una regressione lineare
- In tutti i casi ipotizzeremo **errori Gaussiani** e **misure indipendenti**



CI: Gaussiana con σ^2 nota

Esercitazioni
Lezione 8
15/12/09

- Effettuiamo N prove sperimentali (y_1, \dots, y_N) e la VA Y è di tipo Gaussiano di media μ e varianza σ^2 : $Y = N(\mu, \sigma^2)$
- Supponiamo σ^2 nota e vogliamo stimare un intervallo di fiducia per la media μ
- Procedura:
 1. scegliamo $\gamma = 1 - \alpha$
 2. calcoliamo il valore c tale che $F(c) - F(-c) = \gamma$ dove F è la CDF di una Gaussiana standard $N(0, 1)$
 3. calcoliamo la costante: $k = \frac{c \cdot \sigma}{\sqrt{N}}$
 4. L'intervallo cercato è: $[\bar{y} - k \leq \mu \leq \bar{y} + k]$



CI: Gaussiana con σ^2 nota

Esercitazioni
Lezione 8
15/12/09

Dati

y_i

0.5703
0.6282
0.5337
0.4195
0.4875
0.6827
0.4995
0.5723
0.5048
0.5663
0.4488
0.3514
0.2994
0.4559
0.3132
0.6280
0.5662
0.5471
0.6377
0.5847

- Esempio: si considerino i dati riportati a destra e si supponga di conoscere la varianza dello strumento di misura, $\sigma^2 = 0.1$
- Si vuole stimare l'intervallo di confidenza per la media μ_Y , con $\gamma = 0.9$
- I dati si trovano nel file "datiCI1.txt" e la procedura precedentemente descritta è implementata nell'M-file "CIeserc1.m"
- L'intervallo di fiducia è [0.3986, 0.6312]
- Osservando la formula $k = \frac{c \cdot \sigma}{\sqrt{N}}$ si nota che, per dimezzare il valore di k (avendo un intervallo più preciso a parità di γ), occorre quadruplicare il numeri di prove sperimentali



CI: Gaussiana con σ^2 non nota

Esercitazioni
Lezione 8
15/12/09

- Effettuiamo N prove sperimentali (y_1, \dots, y_N) e la VA Y è di tipo Gaussiano di media μ e varianza σ^2 : $Y = N(\mu, \sigma^2)$
- Supponiamo σ^2 non nota e vogliamo stimare un intervallo di fiducia sia per la media μ che per la varianza σ^2
- Procedura per CI per la media:
 1. scegliamo $\gamma = 1 - \alpha$
 2. calcoliamo il valore c tale che $F(c) = 0.5(1 + \gamma)$ dove F è la CDF di una T di Student ad $N - 1$ gradi di libertà
 3. calcoliamo la costante: $k = \frac{c \cdot s}{\sqrt{N}}$ dove $s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2}$
 4. L'intervallo cercato è: $[\bar{y} - k \leq \mu \leq \bar{y} + k]$



CI: Gaussiana con σ^2 non nota

Esercitazioni
Lezione 8
15/12/09

- Procedura per CI per la varianza:

1. scegliamo $\gamma = 1 - \alpha$

2. calcoliamo i valori c_1 e c_2 tali che:

– $F(c_1) = 0.5(1 - \gamma)$

– $F(c_2) = 0.5(1 + \gamma)$

dove F è la CDF di una χ^2_{n-1}

3. calcoliamo le costanti:

– $k_1 = \frac{(N-1)s^2}{c_1}$

– $k_2 = \frac{(N-1)s^2}{c_2}$

dove: $s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{y})^2}$

4. L'intervallo cercato è: $[k_2 \leq \sigma^2 \leq k_1]$



CI: Gaussiana con σ^2 non nota

Esercitazioni
Lezione 8
15/12/09

Dati

y_i

0.5703
0.6282
0.5337
0.4195
0.4875
0.6827
0.4995
0.5723
0.5048
0.5663
0.4488
0.3514
0.2994
0.4559
0.3132
0.6280
0.5662
0.5471
0.6377
0.5847

- Esempio: si considerino i dati riportati a destra (uguali a quelli dell'esempio precedente)
- Si vuole stimare l'intervallo di confidenza per la media μ_Y e per la varianza σ^2_Y con $\gamma = 0.9$
- I dati si trovano nel file "datiCI1.txt" e le procedure sono implementate nell'M-file "CIeserc2.m"
- Gli intervalli di fiducia sono:
 - [0.4734, 0.5563] per la media
 - [0.0073, 0.0216] per la varianza
- Si noti che al crescere del numero N di prove sperimentali, la T-Student tende ad una Gaussiana



CI: regressione lineare/1

Esercitazioni
Lezione 8
15/12/09

- Effettuiamo N prove sperimentali (y_1, \dots, y_N) al variare delle condizioni sperimentali (x_1, \dots, x_N)
- Supponiamo di aver stimato i parametri a e b di una regressione lineare: $y = a + bx$ e vogliamo calcolare i rispettivi CI
- Procedura per il calcolo dei CI:
 1. scegliamo $\gamma = 1 - \alpha$
 2. calcoliamo il valore c tale che $F(c) = 0.5(1 + \gamma)$ dove F è la CDF di una T di Student ad $N - 2$ gradi di libertà
 3. calcoliamo le quantità:

$$MSE = \frac{1}{N-2} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$S_{xx} = \sum_{i=1}^N (x_i - \bar{x})^2$$

*MSE: errore
quadratico
medio*

*Sxx: somma cor-
retta dei
quadrati
delle x*



CI: regressione lineare/1

Esercitazioni
Lezione 8
15/12/09

4. Definiamo le costanti:

$$k_a = c \sqrt{MSE \left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

$$k_b = c \sqrt{\frac{MSE}{S_{xx}}}$$

5. Gli intervalli cercati sono:

$$[\hat{a} - k_a \leq a \leq \hat{a} + k_a]$$

$$[\hat{b} - k_b \leq b \leq \hat{b} + k_b]$$



CI: regressione lineare/1

Esercitazioni
Lezione 8
15/12/09

Dati

x_i	y_i
1.0	7.60333
2.0	8.37930
3.0	10.8113
4.0	8.21049
5.0	15.1662
6.0	18.9210
7.0	18.6537
8.0	21.5874
9.0	23.3174
10.0	23.6398

- Esempio: si considerino i dati riportati a destra
- Effettuando una regressione lineare, si ottengo le seguenti stime dei parametri a e b:

$$\hat{a} = 4.367 \quad \hat{b} = 2.04764$$

- Si vogliono calcolare i CI per tali stime con $\gamma = 0.95$
- I dati si trovano nel file "datiCI3.txt"
- Gli intervalli di fiducia sono:
 - [1.4044, 7.3296] per il parametro a
 - [1.5702, 2.5251] per il parametro b

Suggerimento: si scriva una M-file con i passaggi per il calcolo degli intervalli di fiducia per la regressione lineare



CI: regressione lineare/2

Esercitazioni
Lezione 8
15/12/09

- Effettuiamo N prove sperimentali (y_1, \dots, y_N) al variare delle condizioni sperimentali (x_1, \dots, x_N)
- Supponiamo di aver stimato i parametri a e b di una regressione lineare: $y = a + bx$
- Una volta fissato un valore x_0 della condizione sperimentale, il corrispondente y_0 sarà dato da: $y_0 = \hat{a} + \hat{b}x_0$
- Essendo y_0 una stima di y in x_0 , è anch'essa una VA
- Quindi possiamo calcolare l'intervallo di fiducia di y per un fissato valore di x_0



CI: regressione lineare/2

- Procedura per il calcolo dei CI:

- scegliamo $\gamma = 1 - \alpha$
- calcoliamo il valore c tale che $F(c) = 0.5(1 + \gamma)$ dove F è la CDF di una T di Student ad $N - 2$ gradi di libertà
- calcoliamo le quantità:

$$MSE = \frac{1}{N-2} \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

$$S_{xx} = \sum_{i=1}^N (x_i - \bar{x})^2$$

- calcoliamo la costante:

$$k(x) = c \sqrt{MSE \left(\frac{1}{N} + \frac{(x - \bar{x})^2}{S_{xx}} \right)}$$

- l'intervallo cercato è: $[\hat{y}_0 - k(x_0) \leq y_0 \leq \hat{y}_0 + k(x_0)]$



CI: regressione lineare/2

Esercitazioni
Lezione 8
15/12/09

Dati

x_i	y_i
1.0	7.60333
2.0	8.37930
3.0	10.8113
4.0	8.21049
5.0	15.1662
6.0	18.9210
7.0	18.6537
8.0	21.5874
9.0	23.3174
10.0	23.6398

- Esempio: si considerino i dati riportati a destra (che sono gli stessi del caso precedente)

- La regressione lineare ci fornisce:

$$\hat{a} = 4.367 \qquad \hat{b} = 2.04764$$

- Si vogliono calcolare i CI per y al variare di x_i , per $\gamma = 0.95$
- I dati si trovano nel file "datiCI3.txt"
- Un diagramma dei risultati consente di visualizzare graficamente gli intervalli di fiducia appena calcolati

Suggerimento: si scriva una M-file con i passaggi per il calcolo degli intervalli di fiducia per la regressione lineare che consenta di diagrammare gli intervalli

1. Gaussiana con varianza non nota

Sono state effettuate delle misure di un processo e i dati sono nel file "datiEserc1.txt".

Si stimino gli intervalli di confidenza per la media e la varianza quando $\gamma = 0.9, 0.95$ e 0.99

2. Regressione lineare

I dati nel file "datiEserc2.txt" sono modellabili attraverso un modello lineare. Si chiede di:

- Stimare i parametri della regressione
- Stimare gli intervalli di confidenza per i parametri ($\gamma = 0.9$)
- Calcolare e diagrammare la regione di confidenza ($\gamma = 0.9$)

