STATISTICA 1, metodi matematici e statistici

Introduzione al linguaggio R

Esercitazione 7: 20-05-2004

Luca Monno

Università degli studi di Pavia

luca.monno@unipv.it
http://www.lucamonno.it

Test del rapporto delle massime verosimiglianze

Consideriamo ora un sistema generale di ipotesi come

$$H_0 : \theta \in \Omega_0, \quad \theta \in \Omega$$

una zona critica intuitiva è del tipo

$$\{x : V(x) \leq \xi\}$$

dove

$$V(x) = \frac{\sup_{\theta \in \Omega_0} L(\theta)}{\sup_{\theta \in \Omega} L(\theta)}$$

è il cosiddetto rapporto delle verosimiglianze massimizzate. Il valore $\xi \in [0, 1]$ determina l'ampiezza del test tramite la relazione

$$\sup_{\theta \in \Omega_0} P\{x : V(x) \leq \xi\} = \alpha$$

.

Supponiamo che su un campione casuale estratto da una normale con varianza σ^2 vogliamo verificare il seguente sistema di ipotesi

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0.$$

Allora

$$V(x) = \frac{L(\theta_0)}{L(\hat{\theta})} = \ldots \text{(conti)} \ldots = \exp\left[-\frac{n}{2\sigma^2}(\bar{x} - \theta_0)^2\right]$$

per cui la regione critica è data da

$$\left\{ x : \left(\frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}} \right)^2 \geq -2 \log \xi \right\}$$

Sotto l'ipotesi nulla la statistica

$$\left(\frac{\bar{x} - \theta_0}{\sigma/\sqrt{n}}\right)^2 \quad (1)$$

ha distribuzione $Chi^2(1)$. Pertanto per avere un test di ampiezza α basta calcolare il quantile di livello $1 - \alpha$ da un $Chi^2(1)$ e vedere se la statistica test (1) supera o meno tale valore.

Dato

```
> set.seed(5)
> y <- rnorm(10, mean = 1, sd = 1)
> mean(y)

[1] 0.9211485
```

Verifichiamo il sistema di ipotesi

$$H_0 : \theta = 0 \quad H_1 : \theta \neq 0.$$

La statistica test (1) è pari a

```
> t <- ((mean(y) - 0)/(1/sqrt(10)))^2
> t

[1] 8.485145
```

il quantile di livello $1 - \alpha$ da un $Chi^2(1)$ è

```
> q <- qchisq(0.95, 1)
> q

[1] 3.841459

> t > q

[1] TRUE
```

Un importante aspetto applicativo è che se Ω_0 e Ω sono intervalli rispettivamente di dimensione k_0 e k, la distribuzione campionaria di $G^2 = -2 \log V(X_1, \ldots, X_n)$ sotto la condizione $\theta \in \Omega_0$ è del tipo $Chi^2(k - k_0)$.

ESERCIZIO: verifificare che per campioni provenienti da una distribuzione esponenziale con media θ pari a 1 e numerosità 10, la statistica test G^2 per la verifica dell'ipotesi nulla $H_0 : \theta = 1$ rispetto all'alternativa $H_1 : \theta \neq 1$ è asintoticamente $Chi^2(1)$.

Ma se la varianza è incognita?
Il test del rapporto delle massime verosimiglianze porta alla zona di rifiuto:

$$R = \left\{ x : \left(\frac{\bar{x} - \theta_0}{S/\sqrt{n}} \right)^2 \geq k \right\}$$

dove sotto l'ipotesi H_0 si ha

$$t = \frac{\bar{x} - \theta_0}{S/\sqrt{n}} \sim t_{n-1}$$

Questo test è chiamato test-t e può essere calcolato con R utilizzando la funzione `t.test`:

```
> t.test(y, mu = 0)
```

L'argomento `mu=0` serve ad indicare che l'ipotesi nulla è $\theta_0 = 0$.
Nell'output ci sono varie quantità interessanti:

1. `t` (t-value) è il valore della statistica test
2. `df` sono i gradi. di libertà della t di student ($n - 1$).
3. `p-value` è, in questo caso, $P(|t_{n-1}| > |t_{oss}|)$. Se il `p-value` è maggiore di α rifiuto H_0, altrimenti accetto.
4. l'intervallo di confidenza per la media

Notiamo che il test è di default bilaterale e di ampiezza 0.05.

Con la stessa funzione è possibile anche fare un test sulla differenza delle medie di due differenti campioni.

Modelli Lineari

Ricordiamo che per un modello lineare del tipo

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

sotto le ipotesi di *omoschedasticità* minimizzando la quantità:

$$D(\boldsymbol{\beta}) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (X\boldsymbol{\beta})_i]^2$$

si ottengono le stime per i parametri

$$\hat{\boldsymbol{\beta}} = \left(X^t X\right)^{-1} X^t \mathbf{Y} \qquad s^2 = \frac{D(\hat{\boldsymbol{\beta}})}{n-p}$$

Esempio

Consideriamo il dataset `auto.dat`: in cui sono presenti i dati realtivi a 203 modelli di automobili in circolo negli USA nel 1985. I dati si riferiscono a 27 caratteristiche.

```
> auto = read.table("auto.dat", header = T)
> attach(auto)
> names(auto)
```

vogliamo studiare il comportamento della percorrenza urbana (l'inverso del consumo) in funzione della cilindrata.

```
> y = percorr.urbana
> x = cilindrata
> formula = y ~ x
> plot(formula)
```

Per stimare i parametri del nostro modello dobbiamo costruire la matrice X. In realtà c'è un comando che permette di farlo automaticamente avendo la formula del modello:

```
> X = model.matrix(formula)
```

A questo punto per stimare i parametri del modello utilizziamo la formula, ricordando che la funzione `solve` restituisce la matrice inversa, `t` la trasposta e l'operatore `%*%` fa il prodotto matriciale

```
> B = solve(t(X) %*% X) %*% t(X) %*% y
> sigma2 = sum((y - X %*% B)^2)/(203 - 2)
> sigma = sqrt(sigma2)
```

Ora possiamo sovrappore al grafico precedente la retta stimata:

```
> abline(B, col = 2, lty = 2)
```

La funzione `lm` unita alla funzione `summary` fa questo e molto altro:

```
> mod = lm(y ~ x)
> mod
> summary(mod)
```

Nel `summary` sono contenuti numerose quantità interessanti:

1. alcune statistiche sui residui: e_i
2. le stime dei parametri (`Estimate`)
3. gli errori standard ovvero la radice dei valori sulla diagonale della matrice delle varianze e covarienze di $\hat{\beta}$ (`Std. Error`)
4. il valore della statistica test t per provare l'ipotesi $H_0 : \beta = 0$ (`t value`)
5. il corrispondente valore-p: (`Pr(>|t|)`)
6. la stima di σ (`Residual standard error`) con i relativi gradi di libertà
7. il valore R^2 (`Multiple R-Squared`)

Il valore-p è una quantità molto interessante: se è un valore molto piccolo vuol dire che alta probabilità non sbagliamo rifiutando l'ipotesi che $H_0 : \beta = 0$, quindi $\beta \neq 0$ che implica che la componete di cui β è il coefficiente è significativa nella detreminazione di y

Il valore R^2 è sempre compreso tra 0 e 1 e indica la bontà di adattamento del modello ai dati