DATA MINING PER IL MARKETING (63 ore)

Marco Riani

mriani@unipr.it

Sito web del corso

http://www.riani.it/DMM

Studio della distribuzione di $\hat{\beta}$

$$E(\hat{\beta}) = \beta$$

$$\text{var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

$$\hat{\beta} \sim N\left[\beta, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right]$$

Teorema di Gauss Markov (efficienza degli stimatori OLS p. 192)

$$
\begin{aligned}
var\left[\tilde{\hat{\beta}}\right] &= \sigma^2 DD' + \sigma^2 \left(X'X\right)^{-1} \\
&= var\left[\hat{\beta}\right] + \sigma^2 DD'.
\end{aligned}
$$

Stima di σ^2

- $E(s^2)$?

$$E(e'e) = \sigma^2(n-k)$$

- Qual è la distribuzione di s^2 (somma dei quadrati dei residui diviso i gradi di libertà)

Caratteristiche delle devianze

- Dev residua

$$\sum_{i=1}^{n} e'e = \epsilon' M \epsilon.$$

- Dev totale

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = y' A y = \epsilon' A \epsilon$$

$$A = (I - ii'/n)$$

- Dev regressione

$$\epsilon'(A - M)\epsilon$$

Come si distribuiscono le forme quadratiche idempotenti?

Come si distribuiscono le forme quadratiche idempotenti?

- Premessa: numero di autovalori diversi da zero di una matrice = rango della matrice (p. 294)
- Gli autovalori di una matrice idempotente sono 0 o 1(p. 288)
- La somma degli autovalori è uguale alla traccia (p.294)

- ➔ rango e traccia della matrice idempotente coincidono

Distribuzione delle forme quadratiche nella regressione

- Devianza residua

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = \frac{e'e}{\sigma^2} = (n-k)\frac{s^2}{\sigma^2} \sim \chi^2(n-k)$$

Distribuzione delle forme quadratiche nella regressione

- Devianza residua

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = \frac{e'e}{\sigma^2} = (n-k)\frac{s^2}{\sigma^2} \sim \chi^2(n-k)$$

Distribuzione della devianza residua e'e

- e'e=ε' M ε
- Scomposizione spettrale di M
- M= PΛP'
- e'e = ε' P ΛP' ε Ponendo P' ε=v
- e'e= v' Λ v $v \sim N(0, \sigma^2 I_n)$

Distribuzione della devianza residua e'e

- e'e = v' Λ v $v \sim N(0, \sigma^2 I_n)$

$$\sum_{i=1}^{n} e_i^2 = e'e = v' \begin{pmatrix} I_{n-k} & 0 \\ 0 & 0_k \end{pmatrix} v$$

$$\sum_{i=1}^{n} e_i^2/\sigma^2 = \sum_{i=1}^{n-k} v_i^2/\sigma^2$$

Distribuzione della devianza residua e'e

$$\frac{\sum_{i=1}^{n} e_i^2}{\sigma^2} = \frac{e'e}{\sigma^2} = (n-k)\frac{s^2}{\sigma^2} \sim \chi^2(n-k).$$

Distribuzione della devianza totale

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = y'Ay = \epsilon' A\epsilon$$

$$A = (I - ii'/n)$$

- Scomposizione spettrale di A
- A= PΛP’
- y’Ay=ε’ P ΛP’ ε Ponendo P’ ε=v
- y’Ay= v’ Λ v v~N(0, $\sigma^2 I_n$)

Distribuzione della devianza totale

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = y'Ay = \epsilon' A\epsilon$$

- y'Ay= v' Λ v $v \sim N(0, \sigma^2 I_n)$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = y'Ay = v' \begin{pmatrix} I_{n-1} & 0 \\ 0 & 0_1 \end{pmatrix} v = \sum_{i=1}^{n-1} v_i^2$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 \sim \sigma^2 \chi^2(n-1)$$

Affermazioni equivalenti (p. 197)

$$\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{\sigma^2} \sim \chi^2(n-1)$$

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 \sim \sigma^2\chi^2(n-1)$$

Distribuzione delle forme quadratiche nella regressione

- Devianza di regressione

$$
\begin{aligned}
SSR &= SST - SSE \\
&= y'Ay - y'My \\
&= y'(A - M)y.
\end{aligned}
$$

Riassunto finale

- Le forme quadratiche idempotenti hanno una distribuzione chi quadrato (dato che gli autovalori sono 0 e 1)
- Il numero di gradi di libertà è dato dal numero di autovalori uguali ad 1 (traccia ossia rango della matrice idempotente)

Scomposizione della devianza totale e distribuzione delle forme quadratiche (p. 197)

forme quadratiche

Fonte di variabilità	Espressione	Distribuzione
Dovuta ai residui (Devianza residua)	$y'My = \epsilon'M\epsilon$	$\sigma^2\chi^2(n-k)$
Totale (Devianza totale)	$y'Ay = \epsilon'A\epsilon$	$\sigma^2\chi^2(n-1)$
Dovuta al modello (Devianza di regressione)	$y'(A-M)y = \epsilon'(A-M)\epsilon$	$\sigma^2\chi^2(k-1)$ sotto H_0 : $\beta_1 = ..., \beta_{k-1} = 0$ ossia sotto $y = i\alpha + \epsilon$

$A = I - ii'/n$, $M = I - H = I - X(X'X)^{-1}X'$

Inferenza su un generico coeff. di regressione parziale (p. 197)

cioè ogni singolo $\hat{\beta}_j$ ha distribuzione normale

$$\hat{\beta}_j \sim N\left[\beta_j, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}_{jj}\right]$$

Se indichiamo con S^{jj} l'elemento $j-$esimo della matrice $(\mathbf{X}'\mathbf{X})^{-1}$ allora la variabile

$$z_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 S^{jj}}} \qquad (1.22)$$

si distribuisce come una normale standardizzata.

Inferenza su un generico coeff. di regressione parziale

si distribuisce come una normale standardizzata. Il problema è che generalmente σ^2 è ignoto. Una sua stima corretta è data da:

$$s^2 = \frac{\mathbf{e}'\mathbf{e}}{n-k} \tag{1.23}$$

dove $\mathbf{e}$ è il vettore dei residui. Lo *standard error di regressione* è dato da $s = \sqrt{s^2}$.

H0: $\beta_j=0$

- Analisi della distribuzione del test t_j

$$
\begin{aligned}
t_j &= \frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \\
&= \frac{\hat{\beta}_j - \beta_j}{\sqrt{s^2 S^{jj}}} \\
&= \frac{\left(\hat{\beta}_j - \beta_j\right) / \sqrt{\sigma^2 S^{jj}}}{\sqrt{\left[(n-k)\, s^2/\sigma^2\right] / (n-k)}} \\
&= \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-k)}{n-k}}}
\end{aligned}
$$

t_j presenta una distribuzione T di Student con n-k gradi di libertà

Intervallo di conf. di un generico coeff. di regressione parziale

$$\Pr\left(\hat{\beta}_j - t_\gamma s_{\hat{\beta}_j} \leq \beta_j \leq \hat{\beta}_j + t_\gamma s_{\hat{\beta}_j}\right) = 1 - \gamma$$

ove

$$s_{\hat{\beta}_j} = \sqrt{s^2 S^{jj}}$$

è lo *standard error* di $\hat{\beta}_j$. $\perp\!\!\!\perp$

Analisi della bontà di adattamento

- R2 nei modelli di regressione lineare multipla

Analisi della varianza e coeff. di correlazione lineare multipla (modelli senza intercetta)

$$e'e = y'y - \widehat{y}'\widehat{y}$$

- Indice di bontà di adattamento

$$\frac{\sum_{i=1}^{n} \hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2}.$$

Modelli con intercetta

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \widehat{y})^2 + \sum_{i=1}^{n} e_i^2$$

Coeff. correlazione lineare multipla

Se nel modello di regressione è presente l'intercetta, l'indice di determinazione è esprimibile come il quadrato della correlazione tra i valori osservati ed i valori stimati dall'equazione di regressione. In simboli

$$R^2 = r^2_{y,\hat{y}} = \left[\frac{cov(y,\hat{y})}{var(y)var(\hat{y})}\right]^2 = \frac{\left[\sum_{i=1}^{n}(y_i-\overline{y})(\widehat{y}_i-\overline{y})\right]^2}{\left[\sum_{i=1}^{n}(y_i-\overline{y})^2\right]\left[\sum_{i=1}^{n}(\widehat{y}_i-\overline{y})^2\right]} \qquad (1.27)$$

Criteri per confrontare i modelli

- In assenza di relazione lineare tra X e y qual è il valore atteso di R^2

$$E(R^2) \approx \frac{k-1}{n-1}$$

Criteri per confrontare i modelli

E' necessario, perciò correggere R^2 da questa componente casuale.

$$\overline{R}^2 = R^2 - \frac{k-1}{n-1}$$

$$R^2 = 1 \qquad \overline{R}^2 = 1 - \frac{k-1}{n-1} = \frac{n-k}{n-1}$$

Criteri per confrontare i modelli

$$\overline{R}^2 = \left(R^2 - \frac{k-1}{n-1} \right) \frac{n-1}{n-k}$$

$\overline{R}^2$

- tende a 0 in assenza di dipendenza lineare e tende a 1 in presenza di dipendenza lineare perfetta.

Criteri per confrontare i modelli

- Dopo semplici passaggi

$$\overline{R}^2 = 1 - \frac{\text{DEV.Residua } / (n-k)}{\text{DEV.Totale}/ (n-1)}$$

Ripasso sullle v.c

- Normale (standardizzata)
- chi^2 (forme quadratiche idempotenti)
- T di Student
- F (rapporto tra forme quadratiche idempotenti indipendenti)

Test di verifica di ipotesi su combinazioni lineari dei coefficienti

- Esempi

$$\beta_2 = \beta_3, \ \beta_2 + \beta_3 + \beta_4 = 1,$$

Test di verifica di ipotesi su combinazioni lineari dei coefficienti

- Se vogliano testare simultaneamente q ipotesi la forma generale è
- $R\beta=r$
- dove R (q × k) di costanti note
- r= vettore noto di q elementi

Test di verifica di ipotesi su combinazioni lineari dei coefficienti

Se R è un vettore riga ($q = 1$) che presenta tutti valori nulli tranne nella posizione i-esima dove assume valore 1 ($R = (0 \ldots 0\ 1\ 0 \ldots 0)$) e r è uno scalare che assume valore nullo ($r = 0$), dato che

$$\begin{aligned} R\beta &= (0 \ldots 0\ 1\ 0 \ldots 0) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_i \\ \ldots \\ \beta_k \end{pmatrix} \\ &= \beta_i \end{aligned}$$

si ottiene $H_0 : \beta_i = 0$. Pertanto, questa specificazione di R e r rappre-

Test di verifica di ipotesi su combinazioni lineari dei coefficienti

2. $R = (0\ 1\ \ \text{-}1\ \ 0 \ldots 0)$ e $r = 0$ rappresenta l'ipotesi

$$\beta_2 - \beta_3 = 0$$

ossia $\beta_2 = \beta_3$.

Test di verifica di ipotesi su combinazioni lineari dei coefficienti

3. $R = (0\ 0\ 1\ 1\ 0 \ldots 0)$ e $r = 2$ rappresenta l'ipotesi

$$\beta_3 + \beta_4 = 2$$

Esercizio

- Supponiamo che k=5. Determinare la matrice R ed il vettore r per testare simultaneamente le seguenti ipotesi
- $\beta_2+3\beta_4=1$
- $\beta_1-5\beta_5=0$
- $\beta_3=0$
- $\beta_3+\beta_4+\beta_5=2$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}$$

Esercizio

- $\beta_2+3\beta_4=1$
- $\beta_1-5\beta_5=0$
- $\beta_3=0$
- $\beta_3+\beta_4+\beta_5=2$

$$R=\begin{pmatrix}0&1&0&3&0\\1&0&0&0&-5\\0&0&1&0&0\\0&0&1&1&1\end{pmatrix}$$

$$\beta=\begin{pmatrix}\beta_1\\\beta_2\\\beta_3\\\beta_4\\\beta_5\end{pmatrix}$$

$$r=\begin{pmatrix}1\\0\\0\\2\end{pmatrix}$$

Test di verifica di ipotesi su combinazioni lineari dei coefficienti

4. Porre

$$R_{(k-1\times k)} = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}$$

e $r_{(k-1\times 1)} = (0, 0, \dots, 0)'$ equivale a testare l'ipotesi

$$\begin{pmatrix} \beta_2 \\ \beta_3 \\ \dots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}$$

Test di verifica di ipotesi su combinazioni lineari dei coefficienti

5. Porre $R = (0\ I_q)$ e $r = 0$ dove il simbolo 0 all'interno della matrice R denota una matrice nulla di ordine $q \times (k-q)$ e r è un vettore colonna di q elementi, rappresenta l'ipotesi che gli ultimi q elementi di β siano tutti nulli.

$$\beta_{k-q+1} = \beta_{k-q+2} = \cdots = \beta_k = 0$$

Esercizio

- Supponiamo che k=6. Determinare la matrice R ed il vettore r per testare simultaneamente le seguenti ipotesi
- $\beta_3=\beta_4=\beta_5=\beta_6=0$

Esercizio

- $\beta_3=\beta_4=\beta_5=\beta_6=0$

$$R=\begin{pmatrix}0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{pmatrix}$$

$$\beta=\begin{pmatrix}\beta_1\\\beta_2\\\beta_3\\\beta_4\\\beta_5\\\beta_6\end{pmatrix}$$

$$r=\begin{pmatrix}0\\0\\0\\0\end{pmatrix}$$

Statistica test

$$F = \frac{(r - R\hat{\beta})'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})/q}{e'e/(n-k)}$$

ha una distribuzione F di Fisher con q e $n-k$ gradi di libertà.

Dimostrazione

$$\text{Se } r = R\beta,\ r - R\hat{\beta} = R(\beta - \hat{\beta}) = -R(X'X)^{-1}X'\epsilon.$$

$$F = \frac{(r - R\hat{\beta})'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})/q}{e'e/(n-k)}$$

- Il numeratore si può scrivere ε'Q ε

$$Q = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$$

Q è simmetrica ed idempotente.

Devo dimostrare che QQ=Q

$$Q = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$$

$$X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X' \quad X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$$

$$X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1} \quad R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$$

$$X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1} \quad R(X'X)^{-1}X'$$

- ε'Q ε = forma quadratica idempotente

ε'Q ε ~ σ² chi^2

$$Q = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$$

$$\begin{aligned} trQ &= tr[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'X(X'X)^{-1}R' \\ &= tr\,I_q = q \end{aligned}$$

- chi^2(q) dove q è il numero di righe della matrice R (numero di vincoli)

Distribuzione del test F

$$F = \frac{(r - R\hat{\beta})'[R(X'X)^{-1}R']^{-1}(r - R\hat{\beta})/q}{e'e/(n-k)}$$

Numeratore ε'*Q* ε/*q*

$$Q = X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'$$

Denominatore ε'M ε/(n-k)

$$M = I - X(X'X)^{-1}X' = I - H$$

Esempio con Excel

- File regr-test.xlsx

Supponiamo di voler testare la seguente ipotesi congiunta:

$H_0 : \beta_2 = 0$ assenza di trend

$H_0 : \beta_3 = 1$ la propensione marginale ad investire è pari ad 1

$H_0 : \beta_4 + \beta_5 = 0$ gli investitori fanno riferimento al tasso di interesse in termini reali