La multidimensionalità

E' infrequente che una indagine statistica esamini un solo fattore. Quasi sempre le elaborazioni tabellari e grafiche sono multidimensionali

Numero di abitazioni nei comuni turistici lombardi. Indagine comunale ICI (anno1999)

Area	Numero abitazioni 1999			% abitazioni secondarie sul totale 1999	% di risposta alla scheda di rilevamento
	Abitazioni principali	Abitazioni secondarie	Totale		
Area Varesina	6004	2640	8644	*30,54*	*17,40*
Area Alto Lario occidentale, Alpi Lepontine	3189	4074	7263	*56,09*	*44,20*
Area Triangolo Lariano, Lario Intelvese	2601	2470	5071	*48,71*	*36,20*
Area Valchiavenna, Valtellina di Morbegno e Sondrio	2586	6525	9111	*71,62*	*23,80*
Area Valtellina di Tirano e Alta Valtellina	5616	8069	16887	*47,78*	*25,00*
Area Val Brembana, Valle Imagna	4563	15335	22191	*69,10*	*53,00*
Area Val Seriana, Val Seriana Superiore, Valle di Scalve	15652	29083	46285	*62,83*	*63,90*
Area Monte Bronzone, Valle Cavallina, Alto Sebino	9016	9352	18368	*50,91*	*61,90*
Area Val Camonica	3907	7816	11723	*66,67*	*68,20*
Area Lago d'Iseo orientale	5261	2209	7470	*29,57*	*60,00*
Area Garda bresciano, Benaco	5629	7661	13290	*57,64*	*44,50*
Area Val Trompia, Valle Sabbia	929	962	3235	*29,74*	*46,20*
Area Oltrepo' pavese e Lomellina	1668	3160	4828	*65,45*	*26,70*
Area Valsassina, Valvarrone	3960	17904	21864	*81,89*	*44,50*
Area Lario orientale, Valle San Martino	0	0	0	–	*18,20*
Area Pianura Padana	**55170**	**46530**	**101700**	*45,75*	*31,60*
Totale aree comuni turistici	**125751**	**163790**	**297930**	*54,98*	*40,90*

Fonte: nostre elaborazioni su scheda di rilevamento ICI inviate ai comuni turistici lombardi.

Raccogliere e classificare i dati non significa molto: è necessario elaborarli ed analizzarli (con uno scopo).

l'algebra matriciale è molto utile per trattare i dati multidimensionali.

il linguaggio delle matrici

La matematica è spesso considerata difficile e misteriosa a causa dei simboli che impiega.

Niente è più incomprensibile di un simbolismo che non si conosce. Pensate ai termini tecnici propri di alcuni mestieri o professioni.(*)

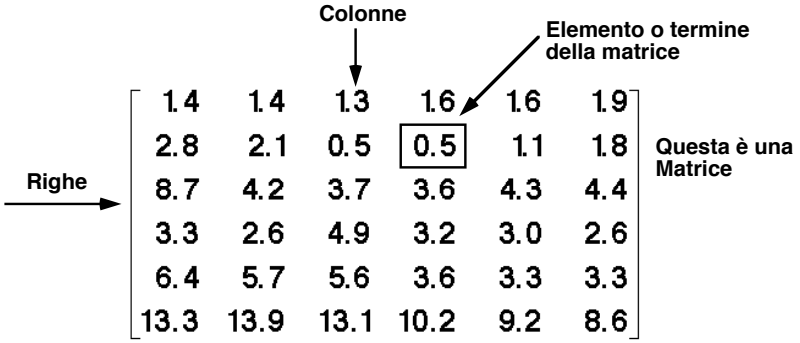
I simboli non sono difficili di per sé, anzi si introducono per facilitare

Il simbolismo matriciale e l'algebra delle matrici comportano una grande semplificazione:

Rendono possibile il trattamento aggregato di dati che altrimenti debbono essere analizzati singolarmente perdendo quindi di vista il loro significato globale, forse l'unico interessante (**)

Descrizione di una matrice

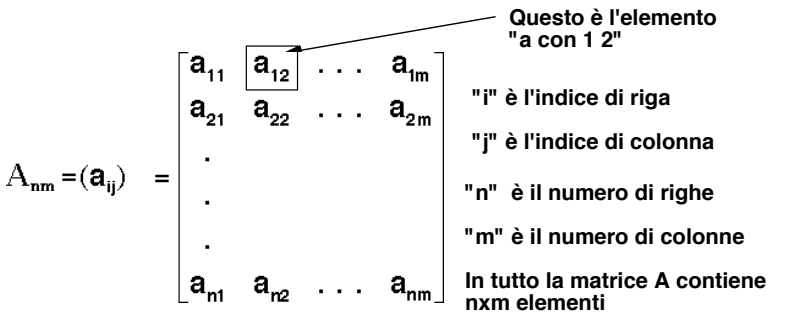
La tabella dell'esempio precedente può essere semplificata come segue



La matrice è una organizzazione di dati righe per colonne che è trattata come un blocco unico

Le dimensioni della matrice

La matrice è una notazione compatta in cui si racchiudono vari elementi disposti su un certo numero di righe e di colonne:



Le dimensioni di una matrice sono semplicemente il numero delle sue righe ed il numero delle sue colonne

Notazione matriciale

Le vendite di dentifricio sono largamente dipendenti dalla spesa in pubblicità

Marche	X Spesa	Y Vendite
Biancodent	2	5
Lucente	4	7
Newfresh	3	6
Perla	1	2
Aguablanca	3.5	6.2

Osservati

$$y_i = a + bx_i + u_i$$

$$= \hat{y}_i + u_i$$

Teorici

Matrice dei regressori

$$Y = \begin{bmatrix} 5 \\ 7 \\ 6 \\ 2 \\ 6.2 \end{bmatrix}; X = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 1 \\ 1 & 3.5 \end{bmatrix}; \beta = \begin{bmatrix} a \\ b \end{bmatrix}; u = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} \qquad Y = X\beta + u$$

Metodo dei minimi quadrati (Ordinari)

Fra i molti possibili criteri quello più usato è il metodo dei minimi quadrati già introdotto per il caso univariato

Tale metodo determina i parametri incogniti in modo da rendere minima la somma dei quadrati degli scarti fra valori osservati e valori teorici

$$s(\beta) \qquad y = X\beta + u$$

$$s = u^t u = (y - X\beta)^t (y - X\beta)$$

$$= y^t y - y^t X\beta - \beta^t X^t y + \beta^t X^t X\beta$$

$$= y^t y - 2\beta^t X^t y + \beta^t X^t X\beta$$

questo è uno scalare ed perciò sempre uguale al suo trasposto

La minimizzazione rispetto a "β" implica la derivazione di "s' rispetto a "β" ponendo poi le derivate uguali a zero.

Metodo dei M.Q.O./2

$$\frac{\partial s}{\partial \beta} = \frac{\partial\left(y^t y - 2\beta^t X^t y + \beta^t X^t X\beta\right)}{\partial \beta} = -2X^t y + 2(X^t X)\beta$$

Ponendo $-2X^t y + 2(X^t X)\beta = 0$

otteniamo il sistema $(X^t X)\hat{\beta} = X^t y$ ← **Sistema "normale"**

i "$\hat{\beta}$" sono i valori STIMATI, in base ai dati, dei valori VERI "β" che sono e rimarrannno comunque incogniti

Premoltiplicando per l'inversa si ha infine $\hat{\beta} = (X^t X)^{-1} X^t y$

Esempio_1

Spesa in pubblicità e vendita di dentifrici

$$Y = \begin{bmatrix} 5 \\ 7 \\ 6 \\ 2 \\ 6.2 \end{bmatrix}; X = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 1 \\ 1 & 3.5 \end{bmatrix}; \quad X^t y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 3 & 1 & 3.5 \end{bmatrix} * \begin{bmatrix} 5 \\ 7 \\ 6 \\ 2 \\ 6.2 \end{bmatrix} = \begin{bmatrix} 26.2 \\ 79.7 \end{bmatrix}$$

$$X^t X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 3 & 1 & 3.5 \end{bmatrix} * \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 1 \\ 1 & 3.5 \end{bmatrix} = \begin{bmatrix} 5 & 13.5 \\ 13.5 & 42.25 \end{bmatrix}$$

Pubblicità e dentifrici/2

$$\left(X^tX\right)^{-1} = \frac{1}{211.25 - 182.25}\begin{bmatrix} 42.25 & -13.5 \\ -13.5 & 5 \end{bmatrix}$$

$$\hat{\beta} = \left(X^tX\right)^{-1}X^ty = \frac{1}{29}\begin{bmatrix} 42.25 & -13.5 \\ -13.5 & 5 \end{bmatrix}\begin{bmatrix} 26.2 \\ 79.7 \end{bmatrix}$$

$$= \begin{bmatrix} 31/29 \\ 44.8/29 \end{bmatrix} = \begin{bmatrix} 1.07 \\ 1.55 \end{bmatrix}$$

Valori stimati

Il vettore dei parametri stimati "$\hat{\beta}$" serve per calcolare i valori TEORICI della variabile dipendente

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

In notazione matriciale: (Matrice *Hat*)

$$\hat{y} = X\hat{\beta} = X\left(X^tX\right)^{-1}X^ty = Hy \quad \textbf{con} \quad H = X\left(X^tX\right)^{-1}X^t$$

La matrice H è SIMMETRICA e IDEMPOTENTE. E' anche nota come "matrice cappello" perchè trasforma le y in y cappello.

$$H^t = \left[X\left(X^tX\right)^{-1}X^t\right]^t = X\left(X^tX\right)^{-1}X^t = H$$

$$H * H = X\left(X^tX\right)^{-1}X^t * X\left(X^tX\right)^{-1}X^t =$$

$$X\left(X^tX\right)^{-1}\left(X^t * X\right)\left(X^tX\right)^{-1}X^t = X\left(X^tX\right)^{-1}X = H$$

Pubblicità e dentifrici/3

$$Hy = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 1 \\ 1 & 3.5 \end{bmatrix}\frac{1}{29}\begin{bmatrix} 42.25 & -13.5 \\ -13.5 & 5 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 4 & 3 & 1 & 3.5 \end{bmatrix}\begin{bmatrix} 5 \\ 7 \\ 6 \\ 2 \\ 6.2 \end{bmatrix} = \hat{y}$$

$$\hat{y} = X\hat{\beta} = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 1 \\ 1 & 3.5 \end{bmatrix}\begin{bmatrix} 1.07 \\ 1.55 \end{bmatrix} = \begin{bmatrix} 4.17 \\ 7.27 \\ 5.72 \\ 2.62 \\ 6.50 \end{bmatrix}$$

Pubblicità e dentifrici/4

$$SSE = y^ty - \hat{\beta}^tX\,y$$

$$= \begin{bmatrix} 5 & 7 & 6 & 2 & 6.2 \end{bmatrix}\begin{bmatrix} 5 \\ 7 \\ 6 \\ 2 \\ 6.2 \end{bmatrix} - \begin{bmatrix} 1.07 & 1.55 \end{bmatrix}\begin{bmatrix} 26.2 \\ 79.7 \end{bmatrix}$$

$$= 152.44 - 151.569 = 0.871$$

Regressione Multipla

L'uso di modelli di regressione con più di una variabile esplicativa è una naturale estensione di ciò si è già fatto. L'equazione del modello è

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_m x_{im} + u_i \quad i = 1, 2, .., n$$

Da notare che ora le " x" hanno due pedici: il primo indica l'osservazione ed il secondo la variabile

In matrici, avremo $y = X\beta + u$

dove:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1m} \\ 1 & x_{21} & x_{22} & \ldots & x_{2m} \\ : & : & : & : & : \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix}; \quad y = \begin{bmatrix} y_1 \\ y_2 \\ : \\ y_n \end{bmatrix}; \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ : \\ \beta_m \end{bmatrix}$$

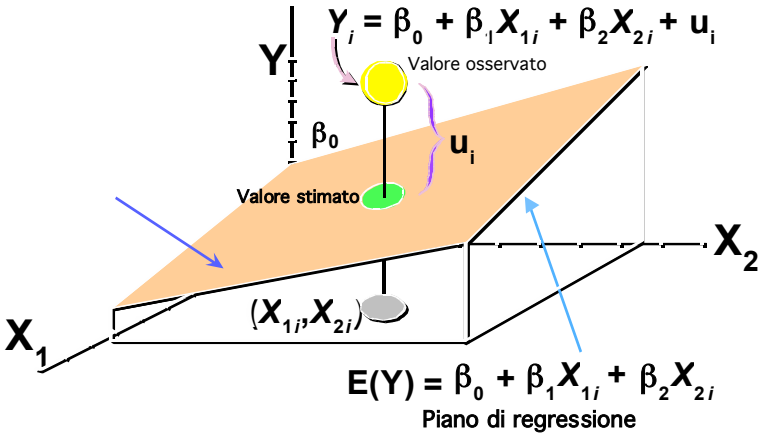
Questa colonna non sempre è presente

Le variabili indipendenti X sono anche dette REGRESSORI

Rappresentazione 3D

La variabile dipendente (anche detta risposta) è una funzione lineare dei regressori.

E' rappresentata dal piano di regressione nell'ipotesi che la forma lineare del legame sia appropriata e che tutti i regressori rilevanti vi siano stati inclusi senza sovrapporsi.



Regressione Multipla/2

Il sistema di equazioni per la stima dei parametri (sistema "normale") è pure quello di prima

$$\left(X^t X\right)\beta = X^t y \longrightarrow \hat{\beta} = \left(X^t X\right)^{-1} X^t y$$

$$X^t X = \begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \ldots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i2}x_{i1} & \ldots & \sum x_{im}x_{i1} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 & \ldots & \sum x_{im}x_{i2} \\ : & : & : & : & : \\ \sum x_{im} & \sum x_{i1}x_{im} & \sum x_{i2}x_{im} & \ldots & \sum x_{im}^2 \end{bmatrix}$$

$$X^t y = \begin{bmatrix} \sum y_i \\ \sum y_i x_{i1} \\ \sum y_i x_{i2} \\ : \\ \sum y_i x_{im} \end{bmatrix}$$

Somma dei quadrati degli errori

La somma degli scarti tra valori OSSERVATI e valori TEORICI della variabile dipendente è

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

Poiché H è simmetrica ed idempotente lo è anche (I-H)

Con le matrici abbiamo:

$$SSE = \left(\mathbf{y} - \hat{\mathbf{y}}\right)'\left(\mathbf{y} - \hat{\mathbf{y}}\right)$$

$$= \left(\mathbf{y} - \mathbf{H}\mathbf{y}\right)'\left(\mathbf{y} - \mathbf{H}\mathbf{y}\right) = \left[(\mathbf{I} - \mathbf{H})\mathbf{y}\right]'\left[(\mathbf{I} - \mathbf{H})\mathbf{y}\right] = \mathbf{y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{y}$$

$$= \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}\mathbf{y}$$

$$= \mathbf{y}'\mathbf{y} - \hat{\beta}^t\mathbf{X}\mathbf{y}$$

Questo è il vettore dei parametri stimati

Esempio

y	X1	X2
62	2	6
60	9	10
57	6	4
48	3	13
23	5	2

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i \quad \textbf{per } i = 1, 2, \ldots, 5$$

$$y = \begin{bmatrix} 62 \\ 60 \\ 57 \\ 48 \\ 23 \end{bmatrix}; X = \begin{bmatrix} 1 & 2 & 6 \\ 1 & 9 & 10 \\ 1 & 6 & 4 \\ 1 & 3 & 13 \\ 1 & 5 & 2 \end{bmatrix}; X^t y = \begin{bmatrix} 250 \\ 1265 \\ 1870 \end{bmatrix} \quad (X^t X) = \begin{bmatrix} 5 & 25 & 35 \\ 25 & 155 & 175 \\ 35 & 175 & 325 \end{bmatrix}$$

Continuazione esempio

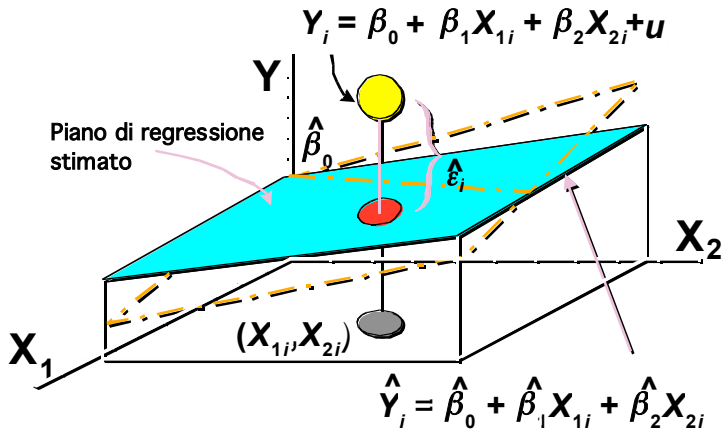
$$(X^t X)^{-1} = \frac{1}{480}\begin{bmatrix} 790 & -80 & -42 \\ -80 & 16 & 0 \\ -42 & 0 & 6 \end{bmatrix}$$

$$\hat{\beta} = (X^t X)^{-1} X^t y = \frac{1}{480}\begin{bmatrix} 790 & -80 & -42 \\ -80 & 16 & 0 \\ -42 & 0 & 6 \end{bmatrix}\begin{bmatrix} 250 \\ 1265 \\ 1870 \end{bmatrix} = \begin{bmatrix} 37 \\ 0.5 \\ 1.5 \end{bmatrix}$$

Il modello di regressione stimato è quindi

$$\hat{y}_i = 37 + 0.5 x_{i1} + 1.5 x_{i2}$$

Grafico della soluzione



La stima dei parametri determina un piano di regressione approssimato rispetto a quello vero, ma incognito.

Esempio_2

Flat.csv

In una società immobiliare si studia il legame tra prezzo di vendita di un appartamento ed alcuni indicatori della sua qualità.

Per stimare i parametri del modello si utilizza un campione casuale di n=20 appartamenti negoziati dalla società.

Una volta stimati i parametri si controllerà sia la coesione interna del modello che la sua utilità pratica.

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

$y = Prezzo\ \ di\ \ vendita$

$X_1 = Valore\ \ catastale\ \ della\ \ proprietà$

$X_1 = Migliorie$

$X_3 = Superficie\ \ calpestabile$

Apprtm	Prezzo	Valocat	Miglior	Superf
1	68900	5960	44967	1873
2	48500	9000	27860	928
3	55500	9500	31439	1126
4	62000	10000	39592	1265
5	116500	18000	72827	2214
6	45000	8500	27317	912
7	38000	8000	29856	899
8	83000	23000	47752	1803
9	59000	8100	39117	1204
10	47500	9000	29349	1725
11	40500	7300	40166	1080
12	40000	8000	31679	1529
13	97000	20000	58510	2455
14	45500	8000	23454	1151
15	40900	8000	20897	1173
16	80000	10500	56248	1960
17	56000	4000	20859	1344
18	37000	4500	22610	988
19	50000	3400	35948	1076
20	22400	1500	5779	962

```
Reg<-read.table(file="Flat.csv",header=TRUE,sep=",",dec=".")
names(Reg)
Mure<-lm(Prezzo~ValoCat+Miglior+Superf,data=Reg)
summary(Mure)
```

```
> Mure<-lm(Prezzo~ValoCat+Miglior+Superf,data=Reg)
> summary(Mure)

Call:
lm(formula = Prezzo ~ ValoCat + Miglior + Superf, data = Reg)

Residuals:
   Min     1Q Median     3Q    Max
-14662  -2155   1027   2722  15976

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1470.2759  5746.3246   0.256  0.80132
ValoCat        0.8145     0.5122   1.590  0.13137
Miglior        0.8204     0.2112   3.885  0.00131 **
Superf        13.5286     6.5857   2.054  0.05666 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7919 on 16 degrees of freedom
Multiple R-squared: 0.8974, Adjusted R-squared: 0.8782
F-statistic: 46.66 on 3 and 16 DF,  p-value: 3.9e-08
```

Proprietà della matrice scarto

- La matrice cappello $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t$ è al centro dei calcoli
- La matrice S=(I-H) è simmetrica e idempotente
- Gli elementi sulla diagonale di H verificano la relazione $\frac{1}{n} \le h_i \le 1$
- Il prodotto di "S" per la matrice X è la matrice nulla: $SX = X^tS = 0$
- La somma di riga di S è nulla: $u^tS = u^t(I-H) = u^t - u^tH = u^t - u^t = 0$

  (questo dipende dalla presenza di una colonna di "1" nella matrice X)
- La matrice di centramento $C = (I - \overline{U}) = \left(I - \frac{1}{n}u^tu\right)$

  è un caso particolare di matrice scarto con X=u

Riscrittura delle forme quadratiche

Si era già stabilito che la devianza dei dati osservati può scriversi come

$$S = \sum_{i=1}^{n}(y_i - \overline{y})^2 = y^t(I - \overline{U})y = y^tCy \qquad \widehat{\mathbf{U}} = \left(\frac{1}{n}\right)\mathbf{u}\mathbf{u}^t$$

M.A. osservati

dove $C = (I - \overline{U})$ **è la matrice di centramento**

Analogamente, la devianza dei dati teorici diventa

$$\sum_{i=1}^{n}\left(\hat{y}_i - \overline{\hat{y}}\right)^2 = \hat{y}^t C\hat{y} = y^tHCHy$$

M.A. teorici

N.B.

$$\hat{y} = Hy$$

Da notare che, poiché $u^tH=u^t$ la M.A. dei teorici è pari alla M.A. degli osservati

$$u^t\hat{y} = u^tHy = u^ty$$

Esempio esplicativo

$$y = \begin{bmatrix}3\\4\\8\\1\end{bmatrix}; \overline{y} = \frac{\sum_{i=1}^{4} y_i}{4} = \frac{1}{4}\begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}\begin{bmatrix}3\\4\\8\\1\end{bmatrix} = \frac{1}{4}u^ty = 4$$

$$\sum_{i=1}^{4}(y_i - \overline{y})^2 = (3-4)^2 + (4-4)^2 + (8-4)^2 + (1-4)^2 = 1 + 0 + 16 + 9 = 26$$

Per esprimere questo calcolo con le matrici dobbiamo avere un vettore colonna formato dalla media aritmetica

$$\overline{y} = \begin{bmatrix}\overline{y}\\ \overline{y}\\ \overline{y}\\ \overline{y}\end{bmatrix} = \begin{bmatrix}\frac{1}{4}u^ty\\ \frac{1}{4}u^ty\\ \frac{1}{4}u^ty\\ \frac{1}{4}u^ty\end{bmatrix} = \frac{1}{4}\begin{bmatrix}u^t\\ u^t\\ u^t\\ u^t\end{bmatrix}y = \frac{1}{4}Uy = \overline{U}y$$

Questa è una matrice somma

Esempio esplicativo (continua)

La matrice di centramento "C" è simmetrica e idempotente e questo deriva dalle proprietà della matrice somma. Sfruttando queste proprietà abbiamo

$$(y-\bar{y})^t(y-\bar{y}) = (y-\bar{U}y)^t(y-\bar{U}y) = \left[(I-\bar{U})y\right]^t\left[(I-\bar{U})y\right]$$
$$= (Cy)^tCy = y^tC^tCy = y^tCy$$

$$y^tCy = \begin{bmatrix}3 & 4 & 8 & 1\end{bmatrix}\left\{\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{bmatrix} - \frac{1}{4}\begin{bmatrix}1&1&1&1\\1&1&1&1\\1&1&1&1\\1&1&1&1\end{bmatrix}\right\}\begin{bmatrix}3\\4\\8\\1\end{bmatrix}$$

$$= \begin{bmatrix}3 & 4 & 8 & 1\end{bmatrix}\begin{bmatrix}\frac{3}{4}&\frac{-1}{4}&\frac{-1}{4}&\frac{-1}{4}\\\frac{-1}{4}&\frac{3}{4}&\frac{-1}{4}&\frac{-1}{4}\\\frac{-1}{4}&\frac{-1}{4}&\frac{3}{4}&\frac{-1}{4}\\\frac{-1}{4}&\frac{-1}{4}&\frac{-1}{4}&\frac{3}{4}\end{bmatrix}\begin{bmatrix}3\\4\\8\\1\end{bmatrix} = \begin{bmatrix}3 & 4 & 8 & 1\end{bmatrix}\begin{bmatrix}-1\\0\\4\\-3\end{bmatrix} = -3+0+32-3 = 26$$

Misura della bontà di adattamento

Per accertare se il modello di regressione sia adatto ai dati esistono varie misure. Ad esempio, il COEFFICIENTE DI CORRELAZIONE MULTIPLA

$$R\_multiplo = \frac{\left[\sum_{i=1}^{n}(y_i-\bar{y})\left(\hat{y}_i-\hat{\bar{y}}\right)\right]^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2\sum_{i=1}^{n}\left(\hat{y}_i-\hat{\bar{y}}\right)^2}; \qquad \hat{\bar{y}} = \bar{y}$$

La media delle osservate e la media delle stimate coincidono nei minimi quadrati

che è dato dal quadrato del COEFFICIENTE DI CORRELAZIONE LINEARE tra i valori osservati ed i valori teorici.

Per costruzione, tale misura è compresa tra zero ed uno.

Tende ad assumere valori elevanti anche in presenza di adattamenti solo sufficienti

Scomposizione della devianza totale

La devianza complessiva dei dati osservati (SST) si scompone come segue:

$$S = SST = \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n}\left[\left(y_i-\hat{y}_i\right)+\left(\hat{y}_i-\bar{y}\right)\right]^2$$
$$= \sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2 + \sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2 + 2\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)\left(\hat{y}_i-\bar{y}\right)$$

Con le matrici avremo:

$$y^tCy = y^t(I-H)y + \left(\hat{y}-\bar{U}y\right)^t\left(\hat{y}-\bar{U}y\right) + 2\left(y-\hat{y}\right)\left(\hat{y}-\bar{U}y\right)$$
$$= y^t(I-H)y + y^t(H-\bar{U})(H-\bar{U})y + 2y^t(I-H)(H-\bar{U})y$$
$$= y^tSy + y^tCHy$$

$H\bar{U} = \bar{U}$

$$(I-H)*(H-\bar{U}) = H-\bar{U}-H^2+H\bar{U} = H-\bar{U}-H+\bar{U} = 0$$

$$(H-\bar{U})*(H-\bar{U}) = H^2-H\bar{U}-\bar{U}H+\bar{U}^2$$
$$= H-\bar{U}-\bar{U}+\bar{U} = H-\bar{U} = H-H\bar{U} = CH$$

Definizione dell'R^2

E' la misura più nota di adattamento. Si definisce a partire dalla relazione:

Devianza totale SST — **Devianza residua SSE** — **Devianza spiegata SSR**

$$y^tCy = y^t(I-H)y + y^tCHy$$

R^2 è il rapporto tra devianza spiegata e devianza totale

$$R^2 = \frac{Dev.Spieg.}{Dev.Tot.} = \frac{y^tCHy}{y^tCy} = \frac{SSR}{SST}$$

Esprime la parte di variabilità che è colta dal modello di regressione

Inoltre, per complemento: $R^2 = 1-\frac{Dev.Res.}{Dev.Tot.} = 1-\frac{y^t(I-H)y}{y^tCy} = 1-\frac{SSE}{SST}$

Esempi

1) Pubblicità e dentifrici

$$R^2 = \frac{\sum_{i=1}^{5}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{5}\left(y_i - \bar{y}\right)^2} = \frac{13.948}{15.152} = 0.92$$

2) dall'esercizio_3

$$SSE = y^t y - \hat{\beta}^t X^t y = 13526 - \begin{bmatrix}37 & 0.5 & 1.5\end{bmatrix}\begin{bmatrix}250\\1265\\1870\end{bmatrix}$$

$$= 13526 - 12687.5 = 838.54$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{838.5}{1026} = 0.183$$

Esempio esplicativo

In questo caso riteniamo illogica la presenza di una termine fisso ovvero se tutti i regressori sono nulli lo deve essere anche la dipendente.

(Ad esempio quando sia la y che le x sono degli scarti da valori fissi).

$$y = \begin{bmatrix}62\\60\\57\\48\\23\end{bmatrix}; X = \begin{bmatrix}2 & 6\\9 & 10\\6 & 4\\3 & 13\\5 & 2\end{bmatrix}; (X^tX) = \begin{bmatrix}155 & 175\\175 & 325\end{bmatrix}; X^ty = \begin{bmatrix}1265\\1870\end{bmatrix}$$

$$\hat{\beta} = \left(X^tX\right)^{-1}X^ty = \frac{1}{790}\begin{bmatrix}3355\\2739\end{bmatrix} = \begin{bmatrix}4.25\\3.47\end{bmatrix}$$

Il modello è ora $\hat{y}_i = 4.25x_{i1} + 3.47x_{i2}$

Le stime dei parametri sono cambiate data l'assenza della intercetta.

Il modello senza intercetta

Consideriamo il modello di regressione lineare multipla

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + u_i$$

Il significato del termine "β_0" è chiaro: rappresenta il livello raggiunto dalla dipendente, al netto dell'errore "u", allorchè tutti i regressori siano nulli.

Talvolta è appropriato escludere tale termine dalla procedura di stima per lavorare sul modello SENZA INTERCETTA

$$y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + u_i$$

Nella matrice dei regressori non c'è più la colonna degli uno e la (X^tX) è la stessa tranne per la scomparsa della prima riga e prima colonna

R^2 nel modello senza intercetta

In questo caso la definizione prescinde dalla media delle osservate e si adotta la scomposizione

$$SST = y^ty = \sum_{i=1}^{n} y_i^2$$

$$SSE = y^ty - \hat{\beta}^tX^ty$$

$$SSR = \hat{\beta}^tX^ty$$

Ne consegue che $\longrightarrow$ $R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}^tX^ty}{y^ty}$

Da notare che, a causa di errori di programmazione, alcuni packages danno valori negativi. Questo è dovuto all'uso della formula:

$$R^2 = \frac{\hat{\beta}^tX^ty}{y^ty - \overline{n\bar{y}^2}}$$

che è valida solo per il modello con intercetta. Se è senza intercetta il termine cerchiato non deve essere considerato (è nullo per costruzione)

Esempio

Modello per il consumo di Benzina. Serie storica 1947-1974

$$Cosdef = \beta_0 + \beta_1 Permed + \beta_2 Numaut + \beta_3 Kilit + \beta_4 Pop + u$$

$Cosdef$ = *Prezzo deflazionato benzina*

$Permed$ = *percorrenza media per auto*

$Numaut$ = *numero auto circolanti*

Km *percorsi con un litro*

Popolazione presente

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5860.0716  1763.2006   3.324  0.00296 **
Numaut         1.2796     0.2488   5.144 3.27e-05 ***
Kilit         88.8666   125.7643   0.707  0.48690
Cosdef         1.2083     9.7564   0.124  0.90251
Pop          -12.5187    12.7393  -0.983  0.33599
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 199.9 on 23 degrees of freedom
Multiple R-squared: 0.9584, Adjusted R-squared: 0.9511
F-statistic: 132.3 on 4 and 23 DF,  p-value: 1.599e-15
```

I cambiamenti ci sono e sono consistenti

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
Numaut   0.7757     0.2349   3.302   0.0030 **
Kilit  304.6181   128.2932   2.374   0.0259 *
Cosdef  18.0817     9.9231   1.822   0.0809 .
Pop     13.8878    11.8605   1.171   0.2531
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 238.1 on 24 degrees of freedom
Multiple R-squared: 0.9995, Adjusted R-squared: 0.9995
F-statistic: 1.311e+04 on 4 and 24 DF,  p-value: < 2.2e-16
```

R^2 corretto

il denominatore di R^2 non dipende dal numero di regressori. Il numeratore aumenta al loro aumentare perché cresce comunque la capacità esplicativa del modello

Ad esempio per ottenere R^2=1 con "n" osservazioni basta adattare un modello polinomiale di grado "n-1"

$$y_i = \sum_{j=0}^{n-1} \beta_i X_i^j + e_i$$

dove "x" è un regressore QUALSIASI (anche i vostri numeri di matricola)

Per ovviare a questo problema si usa R^2 corretto.

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

Per m=1 le due formule coincidono e la correzione non ha praticamente effetto se $R^2 \geq 0.98$.

Se poi risulta $R^2 \leq \frac{m-1}{n-1}$ **allora** $\overline{R}^2 \leq 0$

Esempio

Consumi interni di latte e derivati e di formaggi. Valori in miliardi di lire a prezzi correnti

Anno	Latte	Formaggi	Generi alim.
1970	565	976	15528
1971	632	1087	16188
1972	722	1255	17546
1973	758	1395	20395
1974	960	1507	24878
1975	1134	1747	28569
1976	1385	2376	34752
1977	1796	3229	41334
1978	1974	3800	47573
1979	2277	4604	55216
1980	2693	5032	65705
1981	3195	5658	77217
1982	3497	7066	91688
1983	4573	8414	104696
1984	4977	9699	115141

a) Stimare il modello:

$$G.A. = \beta_0 + \beta_1 Latte + \beta_2 Formaggi + errore$$

```
Model 1: Generi.alim. ~ Latte + Formaggi
Model 2: Generi.alim. ~ -1 + Latte + Formaggi
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1     12  79536211
2     13 122968991 -1 -43432780 6.5529 0.02501 *
```

b) Calcolare il coefficiente R^2

c) Se si omette il termine costante β_0, cosa succede all'R^2?

Campioni e popolazione

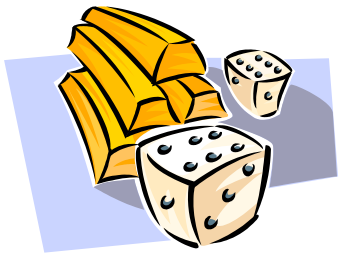
RIcordiamo che i valori con cui operiamo sono valori campionari e quindi sono quelli, ma potevano essere altri.

Ogni campione può dare una sola stima del modello (fermo restando l'ampiezza campionaria)

$$y = \hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i X_i \longrightarrow y = \beta_0 + \sum_{i=1}^{m} \beta_i X_i$$

Tale calcolo è una delle tante stime che si sarebbero poute ottenere dai campioni provenienti da una data popolazione.

Poiché i campioni variano, variano anche le stime.



Disponendo un solo campione ci dobbiamo basare su delle ipotesi concernenti la popolazione e sulle proprietà statistiche che ne conseguono

Un problema più grande

Per risolvere un problema conviene inserirlo in un problema più ampio al quale si devono dare risposte più semplici (non necessariamente più facili).

Consideriamo una combinazione lineare dei parametri incogniti

$$c^t\beta = \sum_{i=0}^{m} c_i\beta_i$$

Risultati soddisfacenti e agevoli da trattare si ottengono spesso con stimatori ottenuti come una…

<u>funzione lineare dei dati osservati nella dipendente y</u>

$$\gamma^t y = \sum_{i=1}^{n} \gamma_i y_i$$

Le costanti c sono note. Le incognite sono i parametri γ

Soluzione

Due dei requisiti richiesti ad uno stimatore sono:

Essere corretto

$$E\left(\gamma^t y\right) = c^t\beta \quad ovvero \quad E\left(\gamma^t y\right) = \gamma^t E(y) = \gamma^t X\beta = c^t\beta \Rightarrow \gamma^t X = c^t$$

Avere varianza minima (fra quelli corretti e funzioni lineari delle y)

$$Var\left(\gamma^t y\right) = \gamma^t V\gamma \quad dove \quad Var(y) = V$$

La V è una matrice di varianze-covarianze, cioè ogni entrata sulla diagonale è una varianza ed ogni elemento fuori diagonale è una covarianza.

Quindi V è una matrice simmetrica

$$Var(y) = \begin{bmatrix} Var(y_1) & Cov(y_2y_1) & Cov(y_3y_1) & \cdots & Cov(y_ny_1) \\ Cov(y_1y_2) & Var(y_2) & Cov(y_3y_2) & \cdots & Cov(y_ny_2) \\ Cov(y_1y_3) & Cov(y_2y_3) & Var(y_3) & & \vdots \\ \vdots & \vdots & & \ddots & Cov(y_ny_{n-1}) \\ Cov(y_1y_n) & Cov(y_2y_n) & \cdots & Cov(y_{n-1,n}y_2) & Var(y_n) \end{bmatrix}$$

Dobbiamo minimizzare $\gamma^t V\gamma$ tenuto conto del vincolo sulla correttezza $\gamma^t X = c^t$.

Soluzione/2

Usiamo 2θ come vettore (mx1) dei moltiplicatori di Lagrange.

Il problema di minimo diventa

$$\underset{(\gamma,\theta)}{Min}\left\{ w = \gamma^t V\gamma - 2\left(\gamma^t X - c^t\right)\theta \right\}$$

Le derivate parziali rispetto a γ e θ comportano

$$\frac{\partial w}{\partial \theta} = \gamma^t X - c^t = 0 \Rightarrow \gamma^t X = c^t$$

$$\frac{\partial w}{\partial \gamma} = 2V\gamma - 2X\theta = 0 \Rightarrow V\gamma = X\theta \Rightarrow \gamma = V^{-1}X\theta$$

A questo punto possiamo determinare i moltiplicatori

$$\gamma^t X = c^t \Rightarrow \theta^t X^t V^{-1} X = c^t \Rightarrow \theta^t = c^t\left[X^t V^{-1} X\right]^{-1}$$

B.L.U.E.

Restano da determinare i pesi della combinazione

$$\gamma^t = \theta^t X^t V^{-1} \Rightarrow c^t\left[X^t V^{-1} X\right]^{-1} X^t V^{-1}$$

Quindi il "miglior stimatore corretto funzione lineare delle osservazioni" cioè *Best Linear Unbiased estimator* (BLUE) della combinazione $c^t\beta$ è

$$\gamma^t y = c^t\left[X^t V^{-1} X\right]^{-1} X^t V^{-1} y$$

Con matrice di varianze-covarianze $Var\left(\gamma^t y\right) = c^t\left[X^t V^{-1} X\right]^{-1} c$

Poiché V^{-1} esiste, quella ottenuta è l'unica soluzione possibile del minimo vincolato e quindi $\gamma^t y$ è l'unico BLUE di $c^t\beta$.

Questo è vero per ogni vettore di costanti c.

Una analisi specifica

Definiamo c come l'i-esima riga e_i della matrice identità I_m. Ne consegue che

$$e_i^t\beta = \beta_i \;\; (i-esimo \;\; parametro)$$

Quindi, il BLUE di β_i è

$$e_i^t\left[X^tV^{-1}X\right]^{-1}X^tV^{-1} \;\; (i-esima \;\; colonna \;\; della \;\; matrice)$$

$$con \;\; varianza \;\; e_i^t\left[X^tV^{-1}X\right]^{-1}e_i$$

Ripetendo le operazioni per ogni parametro si arriva a

$$BLUE \;\; di \;\; \beta = \tilde{\beta} = \left[X^tV^{-1}X\right]^{-1}X^tV^{-1}y$$

$$Var\left(\tilde{\beta}\right) = \left[X^tV^{-1}X\right]^{-1}$$

La matrice V è considerata nota. Se fosse incognita e si decidesse di stimarla occorrerebbe valutare n(n+1)/2 parametri.

Un caso particolare

Le osservazioni sulla variabile dipendente sono considerate incorrelate ed a varianza omogenea (omoschedastiche).

Queste due ipotesi implicano che la matrice di varianze-covarianze sia

$$V = \sigma^2 I_n \qquad V = \begin{bmatrix} \sigma^2 & 0 & & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots 0 & 0 \\ 0 & 0 & \sigma^2 & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \sigma^2 \end{bmatrix}$$

Dove $0< \sigma^2 < \infty$ è la varianza comune delle y.

Si ottiene di conseguenza $\hat{\beta} = \left(X^tX\right)^{-1}X^ty; \qquad Var\left(\hat{\beta}\right) = \sigma^2\left(X^tX\right)^{-1}$

Lo stimatore dei minimi quadrati ordinari è il BLUE di β sotto le ipotesi di incorrelazione e omoschedasticità delle osservazioni.

Quindi gli OLS danno uno stimatore non distorto che ha la varianza minima tra quelli definiti come funzioni lineari delle osservazioni sulla dipendente

N.B. Varianza minima non significa varianza piccola.

Stima della varianza per gli OLS

Consideriamo la forma quadratica

$$x^tAx = \sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}x_ix_j$$

Ipotizziamo che E(x)=*μ* e che Var(x)=V. Poiché

$$E\left(xx^t\right) = V + \mu\mu^t$$

avremo

$$E\left(x^tAx\right) = E\left[Tr\left(x^tAx\right)\right] = Tr\left[E\left(Ax^tx\right)\right] = Tr\left[AE\left(x^tx\right)\right]$$

E quindi

$$E\left(x^tAx\right) = Tr\left[AV + A\mu\mu^t\right] = Tr\left(AV\right) + \mu^tA\mu$$

Stima della varianza per gli OLS/2

La devianza dei residui SSE è espressa da una forma quadratica

$$SSE = y^t\left(I-H\right)y$$

Dove H è la matrice cappello simmetrica e idempotente.

$$E\left(SSE\right) = E\left[y^t\left(I-H\right)y\right] = Tr\left[\left(I-H\right)I\sigma^2\right] + \beta^tX^t\left(I-H\right)X\beta$$
$$= \sigma^2Tr\left[\left(I-H\right)\right] = \sigma^2\left[n-ran\left(X\right)\right]$$

Dove ran(X) è il rango della matrice dei regressori.

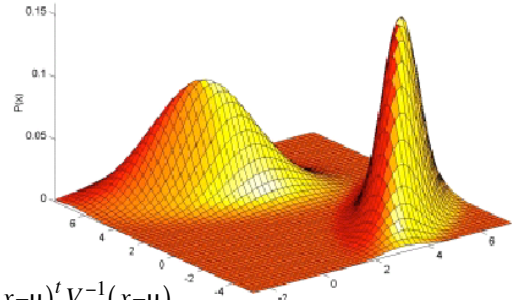
Uno stimatore corretto della varianza degli errori (e delle dipendenti) è quindi

$$\hat{\sigma}^2 = \frac{SSE}{n-ran\left(x\right)};\; se \;\; X \;\; ha \;\; rango \;\; pieno \;\; allora \;\; \hat{\sigma}^2 = \frac{SSE}{n-m-1}$$

La gaussiana multivariata

Un vettore di variabili casuali ha distribuzione gaussiana e multivariata con edia μ e matrice di varianze-covarianze V

$$x \sim N(\mu, V)$$



se la sua funzione di densità congiunta è data da

$$f(x_1, x_2, \ldots, x_m) = \frac{e^{-0.5(x-\mu)^t V^{-1}(x-\mu)}}{(2\pi)^{0.5n}|V|^{0.5}}$$

<u>Proprietà importante</u>

Se z=Ax è una trasformazione del vettore delle x allora anche z avrà distribuzione gaussiana

$$z \sim N(A\mu, A^t V A)$$

Conseguenze

La gaussianità degli errori u si estende alle osservazioni sulla y

$$y \sim N(X\beta, \sigma^2 I_n)$$

Anche gli stimatori dei parametri hanno distribuzione gaussiana

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

Si ottengono inoltre diversi altri risultati collaterali che saranno indicati di volta in volta

Regressione ed inferenza

L'ipotesi che la var(y) e quindi var(u) sia finita è sufficiente per assicurare che il metodo dei minimi quadrati produca uno stimatore BLUE.

Questo però non basta per condurre ragionamenti probabilistici efficaci.

Per espletare l'inferenza nel modello di regressione lineare di solito si considera una delle due ipotesi alternative seguenti

Gli errori del sono indipendenti ed il numero di casi n è grande. Grazie alla versione multivariata del teorema limite centrale si ha:

$$u \sim N(0, \sigma^2 I_n)$$

Gli errori del modello hanno distribuzione gaussiana multivariata

$$u \sim N(0, \sigma^2 I_n)$$

Il primo è un risultato asintotico basato sulle ipotesi; il secondo è una vera e propria congettura.

t di Student

L'efficacia di un regressore ai fini della determinazione di y può essere misurata verificando l'ipotesi

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases}$$

La statistica test che si utilizza è data dal rapporto tra lo stimatore MQO del parametro e la sua deviazione standard

$$t_i = \frac{\hat{\beta}_i}{std(\hat{\beta}_i)} = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{v_{ii}}}$$

V_{ii} è un elemento della diagonale di $(X^tX)^{-1}$.

Tali statistiche hanno distribuzione t-Student con n-m-1 gradi di libertà. Se n è grande si può utilizzare la gaussiana.

p-value

Indica la probabilità che valori della statistica test -inferiori o uguali a quello osservato- siano sopravvenuti solo per effetto della sorte.

Quindi, il p-value misura la probabilità di sbagliare, nelle condizioni date, se si rifiuta l'ipotesi nulla

$$Ipotesi\ \ nulla\quad H_0:\beta_0=0,\ \ p-value=0.0019$$

Il modello senza intercetta potrebbe essere migliorativo solo in 2 casi su 1000 (circa). E' bene rifiutare H_0

$$Ipotesi\ \ nulla\quad H_0:\beta_0=0,\ \ p-value=0.3483$$

Il modello senza intercetta è migliorativo una volta su tre. Non è consigliabile rifiutare H_0.

p-Value/2

Dipende sia dalla distribuzione della statistica test che dal tipo di alternativa.

Nel caso della gaussiana si ha:

$$\text{Se } H_1:\theta>\theta_0 \Rightarrow p-\text{value}=P\left(\hat{\theta}\geq\theta_c\right)=1-\phi\left(Z_c\right)$$
$$\text{Se } H_1:\theta<\theta_0 \Rightarrow p-\text{value}=P\left(\hat{\theta}\leq\theta_c\right)=1-\phi\left(Z_c\right)$$
$$\text{Se } H_1:\theta\neq\theta_0 \Rightarrow p-\text{value}=P\left(\left|\hat{\theta}\right|\geq\theta_c\right)=2\left[1-\phi\left(|Z_c|\right)\right]$$

Formule analoghe possono essere determinate per la t-Student e per le altre distribuzioni coinvolte nella verifica di ipotesi (F-Fisher, etc.)

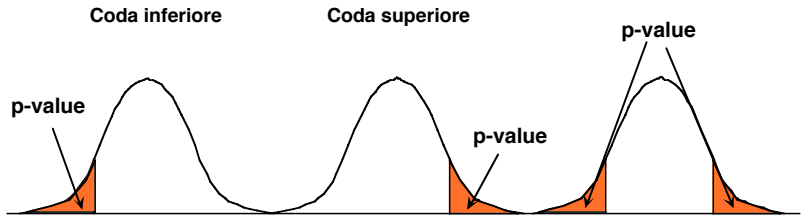
Un parametro associato ad p-value molto piccolo si dice "significativo". Questo vuol dire che ritenendo non nullo parametro si commetterà un errore con una probabilità molto bassa

Precisazioni

Rispetto all'ipotesi che il parametro abbia un valore prefissato ci sono tre casi:

$$\begin{cases}H_0:\beta_i=0\\H_1:\beta_i<0\end{cases};\qquad \begin{cases}H_0:\beta_i=0\\H_1:\beta_i>0\end{cases};\qquad \begin{cases}H_0:\beta_i=0\\H_1:\beta_i\neq 0\end{cases}$$

Nei primi due il test è unidirezionale (o ad una coda), nel terzo è bidirezionale (o a due code).



Linee guida

Se p -value ≤ 1%.	***Aldilà di ogni ragionevole dubbio si può rifiutare H_0***
Se 1% ≤ p -value ≤ 5%.	***Ci sono buone ragioni per rifiutare H_0***
Se 5% ≤ p -value ≤ 10%.	***Ci sono ragioni per rifiutare H_0, ma non sono del tutto convincenti***
Se p -value > 10%.	***E' consigliabile non rifiutare H_0***

I valori sono solo apparentemente bassi.

Le condizioni di applicabilità dei test (ad esempio la distribuzione gaussiana) sono valide solo in parte).

Di conseguenza, solo una forte evidenza può convincere a rifiutare l'ipotesi nulla (angolatura conservatrice)

Esempio

Data set discusso in Ruppert and Carroll (1980). Campione di n=28 misurazioni della salinità del mare presso North Carolina's Pamlico Sound.

a) Stimare i parametri del modello di regressione (Salinity dipendente)

b) Valutarne la significatività

Salinity	Lagged Salinity	Trend	Water Discharge
7.60	8.20	4	23.005
7.70	7.60	5	23.873
4.30	4.60	0	26.417
5.90	4.30	1	24.868
5.00	5.90	2	29.895
6.50	5.00	3	24.200
8.30	6.50	4	23.215
8.20	8.30	5	21.862
13.20	10.10	0	22.274
12.60	13.20	1	23.830
10.40	12.60	2	25.144
10.80	10.40	3	22.430
13.10	10.80	4	21.785
12.30	13.10	5	22.380
10.40	13.30	0	23.927
10.50	10.40	1	33.443
7.70	10.50	2	24.859
9.50	7.70	3	22.686
12.00	10.00	0	21.789
12.60	12.00	1	22.041
13.60	12.10	4	21.033
14.10	13.60	5	21.005
13.50	15.00	0	25.865
11.50	13.50	1	26.290
12.00	11.50	2	22.932
13.00	12.00	3	21.313
14.10	13.00	4	20.769
15.10	14.10	5	21.393

Test-F

Consideriamo il modello di regressione multipla

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + u_i$$

L'adattamento può essere visto da una diversa angolature:

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \ldots = \beta_m = 0 \\ H_1 : \beta_i \neq 0 \ \ per \ \ almeno \ \ un \ \ "i" \end{cases}$$

Non esiste alcuna relazione Tra regressori e dipendente

Qualcuno dei regressori ha un certo impatto sulla "y"

Se l'ipotesi nulla non può essere rifiutata allora il modello è del tutto INADATTO ed occorre cambiare i dati o cambiare modello o entrambi

La prova di questa ipotesi si basa sulla statistica test F -Fisher

$$F_c = \frac{SSR}{SSE} * \frac{N-(m+1)}{m+1} = \frac{\frac{y^t H y}{m+1}}{\frac{y^t S y}{N-(m+1)}}$$

Esempio

$$y = \begin{bmatrix} 10 \\ 20 \\ 17 \\ 12 \\ 11 \end{bmatrix}; \ X = \begin{bmatrix} 1 & 6 & 28 \\ 1 & 12 & 40 \\ 1 & 10 & 32 \\ 1 & 8 & 36 \\ 1 & 9 & 34 \end{bmatrix}; \ \hat{\beta} = \frac{1}{24} \begin{bmatrix} 56 \\ 50 \\ -5 \end{bmatrix}$$

$$SSE = 11.5; \quad SSR = 1038.24; \ m = 2; \ N = 5$$

$$F = \frac{\frac{1038.24}{3}}{\frac{11.5}{5-3}} = 60.1878$$

=FDIST(60.1878,3,2)=0.0164=1.6%

Quindi il modello è almeno contestabile. Ci vuole un approfondimento sui singoli regressori

Da notare che l'adattamento è invece elevato

$$R^2 = \frac{SSR}{SST} = \frac{1038.24}{1054} = 0.9851$$

Relazione tra l'R^2 ed il test F

E' intuitivo che una relazione ci sia dato che misurano lo stesso aspetto: l'adattamento generale

$$R^2 = \frac{SSR}{SST}; \ 1 - R^2 = \frac{SSE}{SST} \Rightarrow \frac{R^2}{1-R^2} = \frac{\frac{SSR}{SST}}{\frac{SSE}{SST}} = \frac{SSR}{SSE}$$

$$F = \left(\frac{SSE}{SSR}\right)\left(\frac{n-m-1}{m+1}\right) \Rightarrow F = \frac{R^2}{1-R^2}\left(\frac{n-m-1}{m+1}\right)$$

Quindi valori elevati di F corrispondono a valori elevati dell'R^2 e viceversa.

Questa relazione è simile a quella che lega la t-Student del coefficiente angolare al test-F nel modello di regressione lineare semplice.

Esempio

Dati regionali al 1991: Tasso di attività in funzione della scolarità d'obbligo e secondaria in rapporto alla popolazione residente

```
Call:
lm(formula = TASAT ~ SCOLOBB + SCUSECS)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8730 -1.4364 -0.3738  2.2938  4.9198

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.35099   31.23439   0.139    0.891
SCOLOBB      0.35258    0.29838   1.182    0.253
SCUSECS      0.02226    0.08473   0.263    0.796

Residual standard error: 2.927 on 18 degrees of freedom
Multiple R-squared: 0.07569,    Adjusted R-squared: -0.02701
F-statistic: 0.737 on 2 and 18 DF,  p-value: 0.4925
```

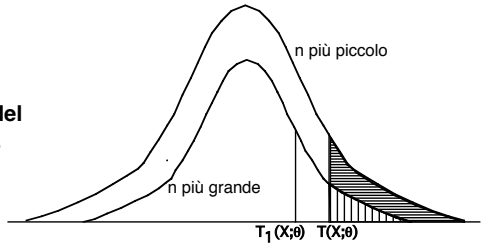
Il modello è pessimo perché il p-value dell'F è al 49% e perché nessuno dei parametri ha un p-value inferiore all'1%

Ampiezza del campione e p-value

La statistica test è, in genere, uno stimatore consistente del parametro sotto ipotesi.

Quindi, all'aumentare dell'ampiezza del campione, la sua variabilità si riduce.

Questo implica che le code della distribuzione della statistica test diventano più sottili.



A parità di p-value, la corrispondente statistica test è inferiore.

Ovvero, la stessa statistica test può avere un p-value più piccolo perché il campione è più grande.

ATTENZIONE!
Campioni molto grandi possono rendere valori della statistica test poco rilevanti dal punto di vista pratico.

Esempio

Johann Tobias Mayers used measurements of the location of the crater Manilius on the moon's surface (a point always observable from earth) to locate the moon's equator and its angle of inclination to the earth.
The data set comprises n=27 observations.

a) Stimare i parametri del modello di regressione
b) Valutarne la significatività singolarmente
c) Valutarne la significatività congiuntamente

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34907 -0.09447 -0.01684  0.09244  0.32220

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.55824    0.03540 411.268   <2e-16 ***
X1           1.50580    0.04173  36.082   <2e-16 ***
X2          -0.07192    0.05083  -1.415     0.17
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.154 on 24 degrees of freedom
Multiple R-squared: 0.9823, Adjusted R-squared: 0.9808
F-statistic:   666 on 2 and 24 DF,  p-value: < 2.2e-16
```

Coefficienti ed unità di misura

Il confronto tra i coefficienti non sempre è lecito in quanto riflettono l'unità di misura del loro regressore

Esempio:

$$y_i = 5 + 2x_{i1} + 200x_{i2}$$

Unità di misura:
y=Litri
x1=centilitri
x2=litri

Sebbene $\hat{\beta}_2$ sia molto più grande di $\hat{\beta}_1$ l'effetto su "y" è uguale: l'incremento di un litro in uno dei regressori (fermo l'altro) induce la stessa variazione in "y"

$$\text{varia } x_1\text{: } y_{nuovo} - y_{vecchio} = 5 + 2*100 + 200*x_2 - 5 - 2*0 - 200*x_2 = 200$$

$$\text{varia } x_2\text{: } y_{nuovo} - y_{vecchio} = 5 + 2*x_1 + 200*1 - 5 - 2*x_1 - 200*0 = 200$$

Per cogliere l' importanza dei regressori è opportuno lavorare con regressori **e dipendente** ***ADIMENSIONALI privi cioè di unità di misura***

Variabili standardizzate

La tecnica usuale di uniformare le unità di misura è quella di standardizzare le variabili (regressori + risposta)

La standardizzazione di una variabile si ottiene sottraendo la media aritmetica e dividendo per lo scarto quadratico medio

$$x_{ij}^{*} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad \text{con } i = 1,2,\ldots,n;\ j = 1,2,\ldots,m;$$

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}; \quad s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}$$

Per costruzione, le x* hanno media zero e SQM uno

Quindi, invece di operare sul modello

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{im} + u_i$$

si usa

$$y_i^{*} = \beta_1^{*} X_{i1} + \beta_2^{*} X_{i2} + \ldots + \beta_m^{*} X_{im} + u_i$$

E' sparita l'intercetta e sono cambiati i coefficienti. Perché?

Regressori in unità standard

Per ottenere una matrice in cui i regressori siano standardizzati è utile la matrice centrante

$$C = \left(I - \frac{1}{n}U\right)$$

Cosicchè le colonne della matrice CX_2 sono formate da deviazioni dalla rispettiva media.

$$CX_2 = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1m} \\ x_{21} & x_{22} & \ldots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nm} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_m \\ \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_m \\ \vdots & \vdots & & \vdots \\ \bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_m \end{bmatrix} = \begin{bmatrix} (x_{11}-\bar{x}_1) & (x_{12}-\bar{x}_2) & \ldots & (x_{1m}-\bar{x}_m) \\ (x_{21}-\bar{x}_1) & (x_{22}-\bar{x}_2) & \ldots & (x_{2m}-\bar{x}_m) \\ \vdots & \vdots & & \vdots \\ (x_{n1}-\bar{x}_1) & (x_{n2}-\bar{x}_2) & \ldots & (x_{nm}-\bar{x}_m) \end{bmatrix}$$

Per completare l'operazione occorre coinvolgere gli SQM. Questo si ottiene postmoltiplicando la matrice per una matrice diagonale così formata

$$D = \begin{bmatrix} \frac{1}{s_1} & 0 & \ldots & 0 \\ 0 & \frac{1}{s_2} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \frac{1}{s_m} \end{bmatrix}$$

Regressori in unità standard/2

$$CX_2D = \begin{bmatrix} \frac{x_{11}-\bar{x}_1}{s_1} & \frac{x_{12}-\bar{x}_2}{s_2} & \ldots & \frac{x_{1m}-\bar{x}_m}{s_m} \\ \frac{x_{21}-\bar{x}_1}{s_1} & \frac{x_{22}-\bar{x}_2}{s_2} & \ldots & \frac{x_{2m}-\bar{x}_m}{s_m} \\ \vdots & \vdots & & \vdots \\ \frac{x_{n1}-\bar{x}_1}{s_1} & \frac{x_{n2}-\bar{x}_2}{s_2} & \ldots & \frac{x_{nm}-\bar{x}_m}{s_m} \end{bmatrix} = \begin{bmatrix} x_{11}^{*} & x_{12}^{*} & \ldots & x_{1m}^{*} \\ x_{21}^{*} & x_{22}^{*} & \ldots & x_{2m}^{*} \\ \vdots & \vdots & & \vdots \\ x_{n1}^{*} & x_{n2}^{*} & \ldots & x_{nm}^{*} \end{bmatrix}$$

La matrice dei regressori standardizzati, si ottiene se si pre-moltiplica la "X2" per la matrice centrante e pos-tmoltiplica per la diagonale degli SQM

Qual'è l'effetto sui parametri?

$$\beta_2^{*} = \left(X_2^{*t} C X_2^{*}\right)^{-1} X_2^{*t} C y^{*} \qquad \text{con } X_2^{*} = CX_2D; y^{*} = s_y^{-1} Cy;$$

$$\beta_2^{*} = D^{-1}\left(X_2^{t} C X_2\right)^{-1} D^{-1} D X_2^{t} C C s_y^{-1} y = s_y^{-1} D^{-1} \beta_2$$

SQM delle osservate

Questi parametri si chiamano COEFFICIENTI BETA

Riduzione dei problemi di calcolo

La stima dei minimi quadrati è molto sensibile agli errori di arrotondamento dovuti al calcolo della matrice inversa $(X^tX)^{-1}$

L'uso delle variabili standardizzate riduce il problema.

$$\frac{1}{n-1}\left(X^{*t}X^{*}\right)_{ij} = (n-1)\sum_{k=1}^{n} x_{ki}^{*} x_{kj}^{*} = \sum_{i=1}^{n}\left(\frac{x_{ki}-\bar{x}_i}{s_i}\right)\left(\frac{x_{kj}-\bar{x}_j}{s_j}\right) = (n-1)r_{ij}$$

Che è proporzionale al coefficiente di correlazione lineare tra x_i ed x_j.

Poiché $-1 \le r_{ij} \le 1$ i valori numerici sono più controllabili

Esempio

Nel caso di due regressori avremo:

$$\left(X^{*t}X^{*}\right) = R = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \Rightarrow R^{-1} = \frac{1}{1-r_{12}^2}\begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}$$

I coefficienti BETA

Cerchiamo di leggere bene il risultato ottenuto

Da notare che ora i coefficienti includono una indicazione che proviene dalla dipendente: Sy

$$\beta_2^* = s_y^{-1} D^{-1} \beta_2 = \begin{bmatrix} \frac{s_1}{s_y} & 0 & \dots & 0 \\ 0 & \frac{s_2}{s_y} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{s_m}{s_y} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} \Rightarrow \beta_i^* = \frac{s_i}{s_y} \beta_i$$

I coefficienti BETA si ottengono moltiplicando le stime ordinarie per il rapporto tra l'sqm del regressore e l'sqm delle osservate

Ma che succede all'intercetta? Poiché vale la relazione:

$$\beta_0^* = \bar{y}^* - \bar{x}^{*t} \beta_2^* = 0 - 0 * \beta_2^* = 0$$

E' identicamente nulla!

E' per questo che scompare nei modelli con variabili standardizzate

I coefficienti BETA/2

Esempio_5

L'ammissione all'UNICAL avviene per merito e per reddito, ma quale dei due è più rilevante? Ecco dei dati relativi ad un campione di studenti

Media(+)	Punti merito	Punti reddito
22	36	34
24	38	39
19	30	31
21	32	32
25	40	36
27	39	37
28	38	33

(+) media su 3 esami specifici di 1^anno

$$M_i = \beta_0 + \beta_1 PM_i + \beta_2 PR_i + u_i$$

$$M_i = -1.15 + 0.8 PM_i - 0.119 PR_i$$

$$\beta_1^* = 0.926; \beta_2^* = -0.105 \qquad R^2 = 0.867$$

In questo caso non c'è molto scarto tra le stime classiche ed i coefficienti Beta. Infatti, pure questi evidenziano una maggiore importanza per i punti merito

L'uso dei beta è legittimo solo nei casi in cui le variabili sono espresse su scale comparabili

I coefficienti Beta riflettono la variazione (in unità standard) nella risposta per unità standard di variazione nei regressori.

In prima approssimazione si può dire che maggiore è il coefficiente Beta, maggiore è l'importanza del regressore nello spiegare la risposta

La variabilità dei coefficienti Beta

Si pone per i coefficienti Beta il problema della precisione? Vediamo

$$V^* = \left(X^{*t} X^*\right)^{-1} = \left(D X_2^t C X_2 D\right)^{-1} = D^{-1} \left(X_2^t C X_2\right)^{-1} D^{-1} = R^{-1}$$

Quindi, per i parametri diversi dall'intercetta, abbiamo

$$\text{errore standard: } s\left(\beta_i^*\right) = \hat{\sigma} \sqrt{v_{ii}^*}; \qquad t - \text{student: } t_i^* = \frac{\beta_i^*}{s\left(\beta_i^*\right)}$$

Poiché, per costruzione si ha

$$\beta_i^* = \frac{s_y}{s_i} \beta_i; \quad \mathbf{e} \quad v_{ii}^* = \frac{s_y}{s_i} v_{ii} \quad \Rightarrow \quad t_i^* = t_i$$

La precisione sarà la stessa sia per i parametri naturali che per i coefficienti BETA.

Ne consegue che, essendo i t-student più informativi, si tende a trascurare il calcolo dei coefficienti BETA

L'indice di presenza

Si è visto che una misura della rilevanza di un regressore nel modello

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_m X_{im} + u_i$$

era data dal suo coefficiente Beta o dallo t-Student. Entrambe però risentono molto delle interrelazioni tra regressori

Una misura più efficace è l'indice di presenza

$$p_i = \frac{|\beta_i| * \|x_i\|}{\|y\|} \qquad \text{dove:} \begin{cases} \|x\| \text{ indica la norma di un vettore: } \sqrt{\sum_{i=1}^{n} x_i^2} \\ |\beta_i| \text{ valore assoluto dei parametri stimati nel modello con risposta e regressori in scala naturale} \end{cases}$$

Il numeratore rappresenta la parte di y attribuibile al regressore

E' quindi un indicatore di forza o debolezza del regressore rispetto alla formulazione del modello.

Se l'importanza è superiore a 0.5 il regressore deve essere presente anche se le altre diagnostiche tenderebbero ad escluderlo

Valore teorici (previsioni)

$$E(y) = X\beta$$

Se x_0 è una osservazione su tutti i regressori, allora il valore atteso della previsione è

$$E(y_0) = x_0{}^t\beta$$

Tuttavia β è incognito e quindi dobbiamo scegliere una via alternativa adoperando la stima di β ottenuta con gli OLS

$$E(y_0) = x_0{}^t\hat{\beta} = \hat{y}_0$$

Poiché la stima è solo una delle possibili realizzazioni dello stimatore, c'è incertezza anche nella stima del valore atteso della Y

Le previsioni/2

il valore previsto può essere considerato una…

Previsione del valore atteso della dipendente cioè una stima puntuale del valore atteso della y dato che $x=x_0$

$$E(y|x_0) = x_0{}^t\hat{\beta}$$

Qui è un parametro

Previsione del valore della dipendente corrispondente ai regressori $x=x_0$

$$y|x_0 = x_0{}^t\hat{\beta} + u$$

Qui è l'osservazione di una variabile casuale

La varianza delle previsioni dipende dalla particolare angolatura adottata

Precisione ed attendibilità

Poiché la dipendente è una variabile casuale dobbiamo aspettarci uno scarto tra valore previsto e valore che si realizza

Possiamo tenere conto di questa variabilità usando gli intervalli di previsione.

Gli intervalli di previsione sono due valori (a loro volta variabili casuali) con le seguenti caratteristiche

PRECISIONE. Legata all'ampiezza dell'intervallo

ATTENDIBILITA'. Legata alla probabilità con il quale la procedura include il valore incognito corrispondente al valore dato dei regressori.

Precisione ed attendibilità dipendono dalla variabilità associata alle stime campionarie dei parametri

Leva dell'osservazione

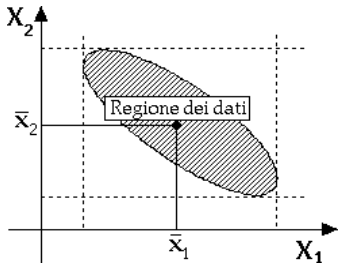
Un'indicazione della variabilità dei regressori si ha dalla diagonale della matrice cappello

$$H = X(X^tX)^{-1}X^t$$

Ricordiamo che, per gli elementi sulla diagonale di H: $h_i = x_i^t(X^tX)^{-1}x_i$

si ha $\frac{1}{n} \le h_i \le 1; \quad Tr(H) = \sum_{i=1}^{n} h_i = m + 1$

h_i è detto LEVA *(Leverage)* dell'osservazione i-esima ed è una misura della distanza tra l'osservazione ed il CENTRO dell'insieme dei dati (vettore delle medie dei regressori)



Infatti:

$$h_i = \frac{1}{n} + (x_i - \bar{x})^t(X^tCX)(x_i - \bar{x})$$

Varianza delle previsioni

Come parametro $$var\left[E\left(y|x_0\right)\right]=\hat{\sigma}^2 x_0^t\left(X^tX\right)^{-1}x_0$$

Come osservazione

$$var\left[y|x_0\right]=\hat{\sigma}^2 x_0^t\left(X^tX\right)^{-1}x_0+\hat{\sigma}^2=\hat{\sigma}^2\left[x_0^t\left(X^tX\right)^{-1}x_0+1\right]$$

Nel secondo caso c'è maggiore incertezza rispetto al primo e quindi la varianza è più grande, a parità di altre condizioni.

Da notare il ruolo della leva h0 nella misura della variabilità

Esempio

Dati su un campione di 5 persone

Persona	Reddito	Scolarità	Età
Cecco	10	6	28
Gisa	20	12	40
Debra	17	10	32
Rita	12	8	36
Peppe	11	9	34

$$\left(X^tX\right)=\begin{bmatrix}5 & 45 & 170\\ 45 & 125 & 1562\\ 170 & 1562 & 5860\end{bmatrix};\quad X^ty=\begin{bmatrix}70\\ 665\\ 2430\end{bmatrix}\quad \left(X^tX\right)^{-1}=\frac{1}{2880}\begin{bmatrix}50656 & 1840 & -1960\\ 1840 & 400 & -60\\ -1960 & -60 & 100\end{bmatrix}$$

$$\hat{\beta}=\frac{1}{24}\begin{bmatrix}56\\ 50\\ -5\end{bmatrix};\quad \hat{y}=X\hat{\beta}=\begin{bmatrix}9\\ 19\\ 16.5\\ 11.5\\ 14\end{bmatrix};\quad \hat{\sigma}^2=\frac{11.5}{5-3}=5.75$$

Per Mr. Tazio è noto che X_0= (1 11 24)

Il valore previsto del reddito è

$$\hat{y}_0=x_0\hat{\beta}=20.25\quad con\quad var\left(\hat{y}_0\right)=\begin{cases}\hat{\sigma}^2x_0^t\left(X^tX\right)^{-1}x_0=5.75*8.7=50.025\\ \hat{\sigma}^2\left[x_0^t\left(X^tX\right)^{-1}x_0+1\right]=5.75*9.7=55.775\end{cases}$$

Intervalli di confidenza (valore previsto)

Per il fissato valore dei regressori il valore previsto (come parametro) è

$$E\left(y_0\right)=x_0{}^t\hat{\beta}=\hat{y}_0$$

Poiché $\hat{\beta}$ **è uno stimatore cioè una variabile casuale, lo sarà anche** $\hat{y}_0$

Se gli stimatori MQO sono normali lo saranno anche i valori previsti in quanto ne sono una combinazione lineare.

$$\hat{y}_0\sim N\left[y_0,\sigma^2x_0^t\left(X^tX\right)^{-1}x_0\right]$$

Se sostituiamo σ^2 con la sua stima $\hat{\sigma}^2$ **otterremo una distribuzione t-Student per il parametro** $\hat{y}_0$

Intervalli di confidenza/2

La conoscenza della distribuzione ci consente di determinare i limiti di un intervallo di confidenza

$$P\left(L_n\le y_0\le U_n\right)=1-\alpha$$

(1-α) è detto livello di confidenza. E' una probabilità che misura il grado di attendibilità della procedura

I valori dei limiti si ottengono attraverso i quantili della t-Student

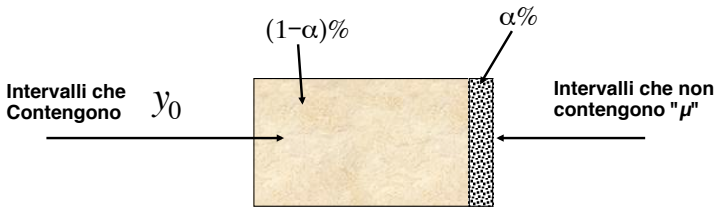
$$\hat{y}_0-t_{1-\frac{\alpha}{2},n-m-1}\hat{\sigma}\sqrt{x_0^t\left(X^tX\right)^{-1}x_0}<y_0<\hat{y}_0+t_{1-\frac{\alpha}{2},n-m-1}\hat{\sigma}\sqrt{x_0^t\left(X^tX\right)^{-1}x_0}$$

I due limiti sono due statistiche e quindi delle variabili casuali che includono il valore previsto con probabilità (1-α)

Interpretazione

GIUSTA: il 95% dI TUTTI I POSSIBILI intervalli, ognuno basato su di un campione diverso, costruiti con questo schema, conterrà il valore vero.

Tuttavia non è possibile essere sicuri che UN PARTICOLARE INTERVALLO contenga o no il valore vero



SBAGLIATA: L'intervallo contiene il valore vero con una probabilità del 95% (in effetti y_0 è un parametro che non può variare a piacimento: o è incluso nell'intervallo oppure no).

Intervalli di previsione

Si ci interessa un un intervallo di previsione per il possibile valore della y che corrisponde a x_0, dovremo modificare il tipo di intervalli

$$\hat{y}_0 - t_{1-\frac{\alpha}{2},n-m-1}\hat{\sigma}\sqrt{1+x_0^t\left(X^tX\right)^{-1}x_0} < y_0 < \hat{y}_0 + t_{1-\frac{\alpha}{2},n-m-1}\hat{\sigma}\sqrt{1+x_0^t\left(X^tX\right)^{-1}x_0}$$

Questi limiti racchiudono i valori potenziali della dipendente non solo la media che ci si aspetta che questi raggiungano.

L'intervallo di previsione è sempre più ampio del corrispondente intervallo di confidenza

Nel primo caso teniamo conto della variabilità dovuta alla stima dei parametri.

Nel secondo dobbiamo aggiungere la variabilità dovuta all'errore di stima intrinseco nel modello di regressione

Esempio

Riprendiamo i dati dell'esempio 7

Per Mr. Tazio è noto che X_0= (1 11 24)

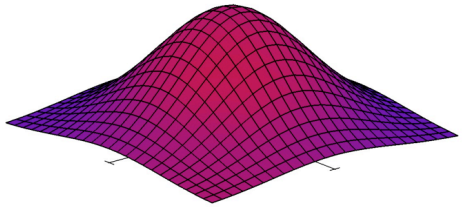
Il valore previsto medio del reddito per configurazioni dei regressori come quella di Mr. Tazio è

$$Confidenza\ \ al\ \ 99\%: 20.25 - 9.92\sqrt{50.025} < y_0 < 20.25 + 9.92\sqrt{50.025}$$
$$Previsione\ \ al\ \ 99\%: 20.25 - 9.92\sqrt{55.725} < y_0 < 20.25 + 9.92\sqrt{57.725}$$

$$Confidenza\ \ -49.912 < y_0 < 90.413$$
$$Previsione\ \ -53.802 < y_0 < 94.302$$

Gli intervalli di entrambi i tipi sono, in questo caso, inutilizzabili.

Con pochi dati e con un elevato grado di attendibilità, la precisione (cioè la lunghezza dell'intervallo) ne ha molto risentito



Esempio

Narula (1987). Un data set di n=31 osservazioni in cui le X provengono dalla gaussiana. Il valore vero dei parametri è (0,1,1).

a) Stimare I parametri e valutare la qualità generale del modello.

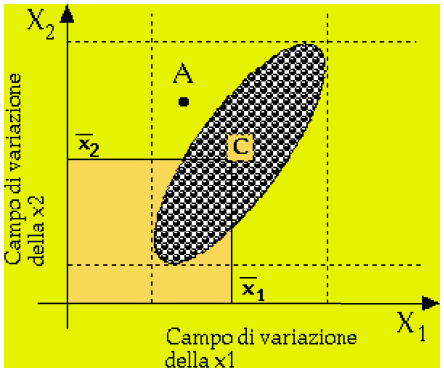
b) Produrre un intervallo di confidenza e di previsione per la combinazione (-2,2).

```
Call:
lm(formula = y ~ X1 + X2, data = Reg)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9091 -0.5984 -0.2026  0.4869  4.0233

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1333     0.2105   0.633   0.5318
X1           -0.5231     0.2190  -2.388   0.0239 *
X2            0.4719     0.1014   4.653 7.16e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 1.141 on 28 degrees of freedom
Multiple R-squared: 0.4361, Adjusted R-squared: 0.3959
F-statistic: 10.83 on 2 and 28 DF,  p-value: 0.0003284
```

Validità della stima

L'interpolazione avviene per osservazioni già realizzate. L'estrapolazione può coinvogere dei loro valori ipotetici, magari futuri.

I minimi quadrati delineano un modello valido solo nella regione dei dati effettivamente osservati



La regione dei dati è definita dalle distanze di Mahalanobis dal centro dei dati

$$M_i = x_i^t C\left(X^t CX\right)X^t Cx_i = h_i - \frac{1}{n}$$

L'estrapolazione dovrebbe avvenire solo se h_e ricade nell'ellissoide contenente i dati osservati. In pratica si dovrebbe avere: $h_e \le h_{max}$.

N.B. il punto "A" è normale ai sensi dei due campi di variazione, ma è anomalo se si guarda alla sua distanza dal centro

Capacità estrapolativa del modello

E' possibile quantificare la capacità estrapolativa del modello con una misura basata sui residui "deleted"

L'idea è di adattare il modello ad (n-1) osservazioni e di utilizzarne una per valutare la qualità dell'estrapolazione (*jacknife*)

Supponiamo di cancellare la i-esima rilevazione. La stima dei parametri del modello applicato ai dati rimanenti è

$$\beta_{(i)} = \left(X_{(i)}^t X_{(i)}\right)^{-1} X_{(i)}^t y_{(i)}$$

L'errore di estrapolazione connesso all'i-esimo dato è

$$e_{(i)} = y_i - \hat{y}_{(i)} = y_i - x_i^t \beta_{(i)} = y_i - x_i^t \left(X_{(i)}^t X_{(i)}\right)^{-1} X_{(i)}^t y_{(i)}$$

Formalismi necessari

Poiché

$$\left(X_{(i)}^t X_{(i)}\right) = \left(X^t X - x_i x_i^t\right)$$

La nuova inversa può essere ottenuta con la formula Sherman-Morrison

$$\left(X_{(i)}^t X_{(i)}\right)^{-1} = \left(X^t X\right)^{-1} - \frac{\left(X^t X\right)^{-1} x_i x_i^t \left(X^t X\right)^{-1}}{x_i^t \left(X^t X\right)^{-1} x_i - 1} = W^{-1} + \frac{W^{-1} x_i x_i^t W^{-1}}{1 - h_i}$$

$$h_i = x_i^t (X^t X)^{-1} x_i$$

$$h_i^2 = h_i \ (\text{H è idempotente})$$

Tenuto conto che $X_{(i)}^t y_{(i)} = X^t y - x_i y_i$

Si ottiene

$$e_{(i)} = y_i - x_i^t \left[W^{-1} + \frac{W^{-1} x_i x_i^t W^{-1}}{1 - h_i}\right]\left(X^t y - x_i y_i\right) =$$

$$= y_i - x_i^t \hat{\beta} + h_i y_i - \frac{h_i x_i^t \hat{\beta}}{1 - h_i} + \frac{h_i^2 y_i}{1 - h_i} = \frac{y_i - x_i^t \hat{\beta}}{1 - h_i} = \frac{\hat{e}_i}{1 - h_i}$$

L'indice PRESS

Si è dimostrato che l'errore di estrapolazione è connesso al residuo del modello completo ed alla leva della osservazione

$$e_{(i)} = \frac{\hat{e}_i}{1 - h_i}$$

la misura della capacità estrapolativa del modello si ottiene dall'indice PRESS *(Prediction Error Sum of Squares)*

$$\text{Press} = \sum_{i=1}^{n} e_{(i)}^2 = \sum_{i=1}^{n} \left[\frac{\hat{e}_i}{1 - h_i}\right]^2$$

L'indice PRESS è una media quadratica ponderata dei residui del modello in cui il peso è determinato dalla leva: maggiore è la leva, maggiore il contributo di quell'errore e più alto sarà il PRESS

A parità di altre condizioni è sempre preferibile avere un modello con un PRESS basso

Esempio: PRESS per gli Hald Data

La stima del modello con due regressori produce

$$y_i = 52.58 + 1.468x_{i1} + 0.662x_{i2}$$

i	e_i	h_i	$[e_i/(1-h_i)]^2$
1	-1.574	0.251	4.418
2	-1.04	0.261	2.020
3	-1.547	0.118	2.955
4	-1.658	0.242	4.790
5	-1.392	0.083	2.309
6	4.047	0.115	20.922
7	-1.303	0.361	4.162
8	-2.075	0.241	7.480
9	1.824	0.171	4.940
10	1.362	0.550	9.168
11	3.264	0.184	16.003
12	0.862	0.196	1.153
13	-2.893	0.214	13.557
		Press=	93.883

Se invece si utilizzano tre regressori si ha

$$y_i = 71.65 + 1.452x_{i1} + 0.416x_{i2} - 0.273x_{i4}$$

con PRESS=85.3516

Occorre decidere se l'aggiunta di un regressore si possa ritenere bilanciata dalla diminuzione del PRESS.

Linearità del modello di regressione

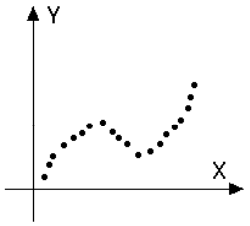
La linearità del modello di regressione dipende solo da come vi compaiono i parametri.

Il modello

$$\sqrt{y_i} = a + bx_i^3 - ce^{z_i} + u_i$$

è lineare dato che a, b e c compaiono potenza uno.

Spesso è lo scatterplot a suggerire funzioni particolari



In questo caso il modello lineare è inadatto; è più verosimile una cubica

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + u_i$$
$$= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

Anche in questo caso il modello è lineare

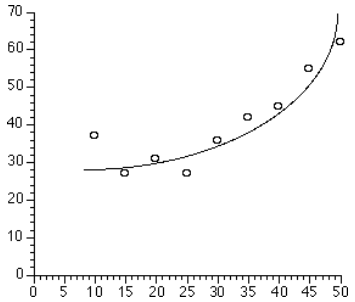
Esempio: La curva dei costi

I costi totali di una linea di produzione sono collegabili alla quantità prodotta in base ad un modello quadratico

$$C_i = \beta_0 + \beta_1 Q_i + \beta_2 Q_i^2 + u_i$$

Ponendo $X_1=Q$ e $X_2=Q^2$ la MATRICE DEI DATI cioè il vettore della dipendente e la MATRICE DEI REGRESSORI per il costo a vari livelli di produzione potrebbe essere la tabella

C_i	$X_{1i}=Q_i$	$X_{2i}=(Q_i)^2$
37	10	100
27	15	225
31	20	400
27	25	625
36	30	900
42	35	1225
45	40	1600
55	45	2025
62	50	2500



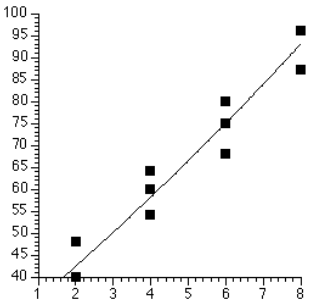
N.B. REGRESSORE ≠ VAR. INDIPENDENTE

Esempio: regressione quadratica

Per i dati seguenti

$$y^t = [60 \quad 75 \quad 96 \quad 40 \quad 54 \quad 68 \quad 87 \quad 64 \quad 48 \quad 80]$$
$$x^t = [4 \quad 6 \quad 8 \quad 2 \quad 4 \quad 6 \quad 8 \quad 4 \quad 2 \quad 6]$$

Adattare il modello $y_i = b_0 + b_1 x_i + b_2 x_i^2 + u_i$



$$X = \begin{bmatrix} 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 8 & 64 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \\ 1 & 6 & 36 \\ 1 & 8 & 64 \\ 1 & 4 & 16 \\ 1 & 2 & 4 \\ 1 & 6 & 36 \end{bmatrix};$$

L'inversa esiste anche se c'è relazione tra due colonne

$$(X^tX) = \begin{bmatrix} 10 & 50 & 292 \\ 50 & 292 & 1880 \\ 292 & 1880 & 12880 \end{bmatrix}$$

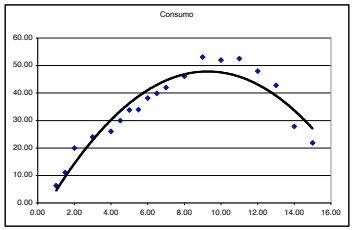
$$X^ty = \begin{bmatrix} 672 \\ 3690 \\ 22940 \end{bmatrix};$$

$$\hat{\beta} = \begin{bmatrix} 30.3 \\ 6.7 \\ 0.12 \end{bmatrix}$$

Esempio: una curva di domanda

Consumo pro-capite di zucchero in vari paesi secondo il livello dei prezzi

Lo scatterplot suggerisce quadratica




```
Call:
lm(formula = y ~ ., data = Reg)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6255 -1.6041  0.0552  1.5903  5.0299

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.657950   2.957078   1.913   0.0750 .
x            3.565721   1.567124   2.275   0.0380 *
x2           0.655716   0.231517   2.832   0.0126 *
X3          -0.055274   0.009797  -5.642 4.68e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.587 on 15 degrees of freedom
Multiple R-squared: 0.9706, Adjusted R-squared: 0.9648
F-statistic: 165.2 on 3 and 15 DF,  p-value: 1.034e-11
```

Le stime confermano l'ipotesi. L’intercetta è forse sacrificabile.

Stima della curva di domanda

Cosa succede se invece utilizziamo una polinomiale di grado superiore?

Adattamento di una cubica:

Regression Statistics	
Multiple R	0.9852
R Square	0.9706
Adjusted R Square	0.9648
Standard Error	2.5872
Observations	19

	P-value
Intercept	0.0750
X Variable 1	0.0380
X Variable 2	0.0126
X Variable 3	0.0000

I parametri sono tutti significativi (β_0 non conta) anche se β_1 sembra meno significativo. Comunque β_3 ha un p-value molto elevato
In questo caso è difficile scegliere tra la cubica e la quadratica. La prima è però più "semplice"

Adattamento di una quartica:

Regression Statistics	
Multiple R	0.9861
R Square	0.9724
Adjusted R Square	0.9645
Standard Error	2.5956
Observations	19

	P-value
Intercept	0.5525
X Variable 1	0.0774
X Variable 2	0.9984
X Variable 3	0.8771
X Variable 4	0.3582

R^2 è aumentato perchè è cresciuto il numero di regressori, ma la stima è poco attendibile.

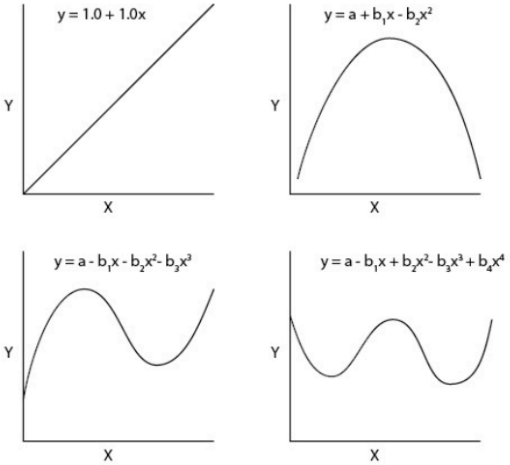
I valori alti del p-value sono dovuti ad un altro problema: la COLLINEARITA'

Regressione polinomiale

Se si ritiene che il legame di dipendenza tra la variabile dipendente ed una o più variabili esogene sia accertato per logica, ma si ignora la forza e la forma si può formulare il modello usando più regressori per la stessa variabile

L’idea è di aggiungere delle potenze successive della variabile esogena fino ad ottenere un adattamento soddisfacente.

Possono però insorgere problemi di overflow dato che le potenze poi si raddoppiano nel calcolo della covarianza; sono inolgtre possibili problemi di multicollinearità dato che potenze ravvicinate hanno andamenti abbastanza simili, almeno in alcuni intervalli.



Regressione polinomiale/2

In base al teorema di Taylor ogni funzione dotata di

Derivate prime continue nell’intervallo chiuso [a,b] fino all’ordine (n-1)

Derivata n-esima continua nell’intervallo aperto (a,b)

In [a,b] può essere espressa come

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2}f''(a) + \frac{(x-a)^3}{3!}f^{(3)}(a) + \\ + \frac{(x-a)^{n-1}}{(n-1)!}f^{(n)}(a) + \frac{(x-a)^n}{n!}f^{(n)}(\theta) \quad a < \theta < x$$

Se si pone a=0 e θ=a si ha (approssimativamente)

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_n x^n \quad dove \quad \beta_i = \frac{f^{(i)}(0)}{i!}$$

Regressione polinomiale/3

Se si ipotizza una funzione abbastanza liscia tra y ed x (cioè esistono le derivate fino ad un certo ordine) la f(x) che le lega può essere ben approssimata con un polinomio di grado adeguato.

Ci sono però delle difficoltà

Un grado elevato comporta problemi di rappresentazione numerica.

Se un regressore è nell'ordine di 10^4 la sua potenza quinta è nell'ordine di 10^{20}. Nella matrice (X^tX) ci troveremo termini dell'ordine di 10^{40} con perdita di cifre significative tanto maggiore quanto minore è la potenza del computer.

Un grado elevato comporta problemi di condizionamento nella matrice dei coefficienti

Le potenze elevate tendono a collocarsi , tra di esse, n un rapporto quasi lineare e questo determina problemi di dipendenza lineare (collinearità).

Esempio

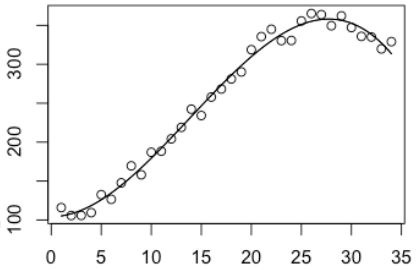
Numero indice della produzione industriale in una regione meridionale. Dati trimestrali destagionalizzati.

a) Individuate e stimate il tipo di trend polinomiale

b) Valutare la qualità del modello ottenuto.

c) Quali accorgimenti si possono adoperare per attenuare I problemi derivanti dall'suo di un polinomio di grado elevato?

```
Call:
lm(formula = NIPI ~ Trim + I(Trim^2) + I(Trim^3), data

Residuals:
    Min      1Q  Median      3Q     Max
-15.134  -7.894  -2.428   7.560  15.863

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 105.271473   7.205119  14.611 3.52e-15 ***
Trim         -0.296672   1.757063  -0.169    0.867
I(Trim^2)     1.008597   0.115754   8.713 1.02e-09 ***
I(Trim^3)    -0.024118   0.002176 -11.084 3.95e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.367 on 30 degrees of freedom
Multiple R-squared: 0.9905, Adjusted R-squared: 0.9895
F-statistic:  1039 on 3 and 30 DF,  p-value: < 2.2e-16
```



Si pu centrare la variabile su cui poi si calcolano le potenze.

Polinomi ortogonali

L'uso dei polinomi comporta il ricalcolo di ogni termine se una delle potenze del polinomio è ritenuta poco significativa e quindi cancellata ovvero si vuole includere un termine addizionale.

Per semplificare i calcoli si possono adoperare i polinomi ortogonali

$$z_0 = 1; \qquad z_1 = a_1 + b_1 x; \qquad z_2 = a_2 + b_2 x + c_2 x^2$$

$$z_3 = a_3 + b_3 x + c_3 x^2 + d_3 x^3 \qquad z_4 = a_4 + b_4 x + c_4 x^2 + d_4 x^3 + e_5 x^5$$

I coefficienti dei polinomi devono essere scelt in modo tale che

$$\mathbf{z}_i{}^t \mathbf{z}_j = 0 \quad per \quad i \neq j$$

I regressori z in questo caso non sono delle semplici potenze della variabile esplicativa X, ma polinomi separati in X vincolati ad essere ortogonali.

I vantaggi sono che ogni parametro di ogni polinomio in ogni potenza può essere calcolato autonomamente dagli altri

La variabilità spiegata da ogni regressore-polinomio è calcolabile separatamente ed esprime l'incremento dovuto all'aggiunta di un nuovo regressore

Polinomi ortogonali/2

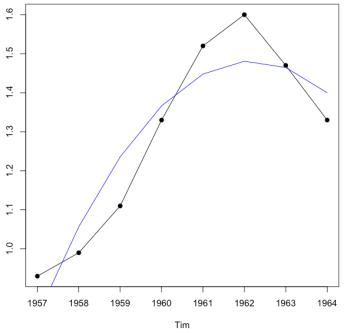
L'ortogonalità dei polinomi implica che

$$\mathbf{X}^t\mathbf{X} = diag\left(A_{00}, A_{11} \ldots, A_{rr}\right) \;\; con \;\; A_{jj} = \sum_{i=1}^{n} \left[z_i\left(x_i\right)\right]^2; A_{00} = n$$

Se i valori della variabile indipendente sono equispaziati allora i coefficienti dei polinomi sono più semplici da calcolare.

In caso contrario si ricorre ad algoritmi di trasformazione delle colonne della matrice dei regressori per ottenere i polinomi necessari.

```
# Srivastava
Y<-c(0.93,0.99,1.11,1.33,1.52,1.60,1.47,1.33)
Tim<-1957:1964
Sriv<-data.frame(cbind(Y,Tim))
Try<-lm(Y~poly(Tim,2),data=Sriv)
summary(Try)
plot(Tim,Y,type="o",pch=19)
Pse<-Tim
Y.new<-data.frame(Trim=Pse)
Y.pred<-predict(Try,newdata=Y.new)
lines(Pse,Y.pred,col="blue")
```



Regressione Broken stick

E' possibile che il modello di regressione lineare debba essere differenziato per gruppi diversi presenti nei dati.

Esempio: Modello con una sola variabile esplicativa in due parti

$$y = \beta_0 + \beta_1 X_1 + e\left[X \le c\right]$$

$$y = \beta_0 + \beta_1 X_1 + e\left[X > c\right]$$

$$y = \beta_0 + \beta_1 X_1 + e\left[X = 0\right]$$

$$y = \beta_0 + \beta_1 X_1 + e\left[X = 1\right]$$

La costante c può rappresentare uno shock ovvero rappresentare la soglia di una variabile binaria.

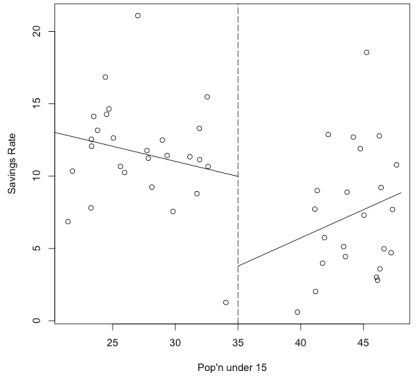
Esempio

La costante è pari a c=35

```
library(faraway)
data(savings)
savings
g1<- lm(sr~pop15, savings, subset=(pop15 < 35))
g2<- lm(sr~pop15, savings, subset=(pop15 > 35))
plot (sr~pop15,savings,xlab="Pop'n under 15",ylab="Savings Rate")
abline (v=35, lty=5)
segments (20, g1$coef[1] +g1$coef[2]*20,35,g1$coef [1] +g1$coef [2] *35)
segments (48, g2$coef[1] +g2$coef[2]*48,35,g2$coef [1] +g2$coef [2] * 35)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.2747     5.2087   3.316  0.00279 **
pop15        -0.2085     0.1895  -1.100  0.28180
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.618 on 25 degrees of freedom
Multiple R-squared: 0.04617,    Adjusted R-squared: 0.008015
F-statistic:  1.21 on 1 and 25 DF,  p-value: 0.2818
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -9.8702    17.5841  -0.561    0.581
pop15         0.3900     0.3964   0.984    0.336

Residual standard error: 4.388 on 21 degrees of freedom
Multiple R-squared: 0.04408,    Adjusted R-squared: -0.001436
F-statistic: 0.9684 on 1 and 21 DF,  p-value: 0.3363
```



Linea spezzata

La stima per parti separati manca di continuità nel punto di giunzione.

Comporta inoltre la stima di più parametri del necessario.

$$X_1 = \begin{cases} X & per \quad X \le c \\ c & per \quad X > c \end{cases}; \quad X_2 = \begin{cases} 0 & per \quad X \le c \\ (X - c) & per \quad X > c \end{cases}$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

Con due rami separati purché $\beta_1 \neq \beta_2$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e\left[X \le c\right] \Rightarrow y = \beta_0 + \beta_1 X + e$$

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e\left[X > c\right] \Rightarrow y = \beta_0 + (\beta_1 - \beta_2)c + \beta_2 X + e$$

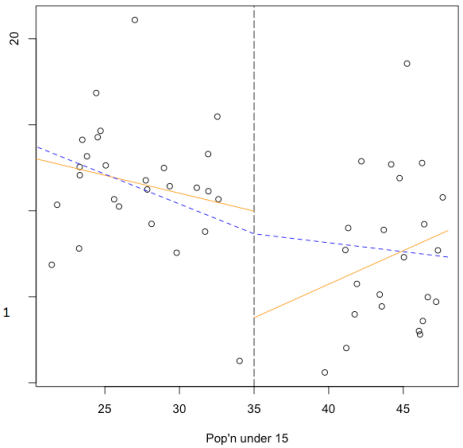
Se X=c+δx, con δx>0 allora $\underset{\delta x \to 0}{Lim}\, y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = y = \beta_0 + \beta_1 X_1$

I coefficienti rappresentano i tassi di aumento per I due diversi rami

Esempio

```
X1<-function(x) ifelse(x <= 35, x, 35)
X2<-function(x) ifelse(x <= 35,0, x-35)
hb<-lm(sr~X1(pop15) + X2(pop15), savings)
x <- seq(20, 48, by=1)
hy <- hb$coef[1]+hb$coef[2]*X1(x)+gb$coef[3]*X2(x)
lines (x, py, lty=2,col="blue")
```

```
Call:
lm(formula = sr ~ X1(pop15) + X2(pop15), data = savings)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7285 -2.6760  0.2065  1.7410 10.9645

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.8092     5.3816   3.867 0.000338 ***
X1(pop15)    -0.3471     0.1930  -1.798 0.078540 .
X2(pop15)    -0.1040     0.1860  -0.559 0.578559
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.053 on 47 degrees of freedom
Multiple R-squared: 0.2152, Adjusted R-squared: 0.1819
F-statistic: 6.446 on 2 and 47 DF,  p-value: 0.003359
```



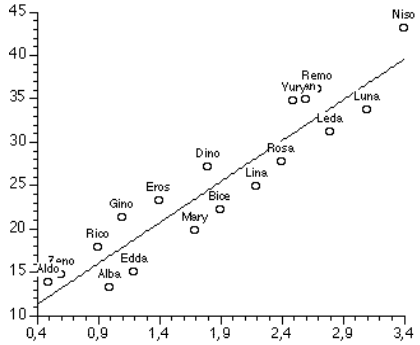
L'uso di variabili qualitative

I modelli di regressione trattano, di solito, con variabili quantitative.
Talvolta però si rende necessario introdurre variabili qualitative o fattori.

Esempio
Per un gruppo di persone si dispone dei dati relativi al reddito ed alla spesa in abbigliamento annuale (Dati CROSS-SECTION)

Persone	Spesa	Reddito
Gino	1,1	21,3
Alba	1,0	13,3
Rico	0,9	17,8
Edda	1,2	15,1
Remo	2,7	36,1
Rosa	2,4	27,8
Aldo	0,5	13,9
Zeno	0,6	14,8
Lina	2,2	24,9
Eros	1,4	23,3
Bice	1,9	22,2
Dino	1,8	27,1
Ivan	2,6	34,9
Mary	1,7	19,8
Niso	3,4	43,2
Luna	3,1	33,8
Yury	2,5	34,7
Leda	2,8	31,2

E' evidente che ci sono due strutture distinte: uomini e donne



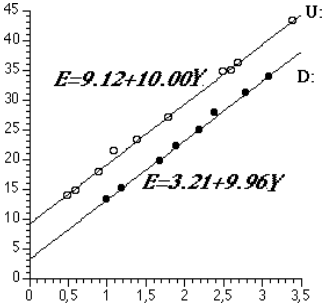
L'uso di variabili qualitative/2

Si potrebbe pensare di stimare i parametri di due relazioni distinte

$$\begin{cases} \text{Uomini:} & E_i^u = b_0 + b_1 Y_i^u + w_i \\ \text{Donne:} & E_i^d = d_0 + d_1 Y_i^d + v_i \end{cases}$$

Qui riteniamo che uomini e donne non solo abbiano un livello minimo di spesa (intercetta) diverso, ma che sia diversa anche la reattività ad un incremento di reddito (coefficiente angolare)

Persone	Spesa	Reddito
Aldo	0,5	13,9
Dino	1,8	27,1
Eros	1,4	23,3
Gino	1,1	21,3
Ivan	2,6	34,9
Niso	3,4	43,2
Remo	2,7	36,1
Rico	0,9	17,8
Yury	2,5	34,7
Zeno	0,6	14,8



Rm06Exem11.csv

Persone	Spesa	Reddito
Alba	1,0	13,3
Bice	1,9	22,2
Edda	1,2	15,1
Leda	2,8	31,2
Lina	2,2	24,9
Luna	3,1	33,8
Mary	1,7	19,8
Rosa	2,4	27,8

Le variabili indicatore o Dummy

La scelta di stimare modelli separati non sempre è obbligatoria. Infatti, nell'esempio i due coefficienti angolari sono praticamente gli stessi.

D'altra parte uno dei due gruppi potrebbe essere così poco numeroso da rendere molto INEFFICIENTE la stima dei parametri.

Per combinare i due sottomodelli (nell'ipotesi che $b_1=d_1$) si introduce una variabile INDICATORE o Variabile DUMMY.

La variabile indicatore è dicòtoma, cioè ha solo due valori: UNO e ZERO.

$$D_{ui} = \begin{cases} 1 & \text{Se la persona è di sesso maschile} \\ 0 & \text{Altrimenti} \end{cases}$$

$$D_{di} = \begin{cases} 1 & \text{Se la persona è di sesso femminile} \\ 0 & \text{Altrimenti} \end{cases}$$

Le variabili dummy/2

$$E_i = b_0 + b_1 D_{ui} + b_2 D_{di} + b_3 Y_i + u_i$$

Lo schema sembra ragionevole, ma ha un grave difetto.

Le prime colonne della matrice dei regressori sarebbero

$$X = \begin{bmatrix} 1 & 0 & 1 & : \\ 1 & 1 & 0 & : \\ 1 & 1 & 0 & : \\ 1 & 0 & 1 & : \\ : & : & : & : \\ 1 & 0 & 1 & : \end{bmatrix}$$

Ad esempio la 2ª colonna si può ottenere dalla 1ª sottraendo la 3ª.

Quindi c'è una colonna linearmente dipendente e non esiste la matrice inversa di

$$(X^t X)$$

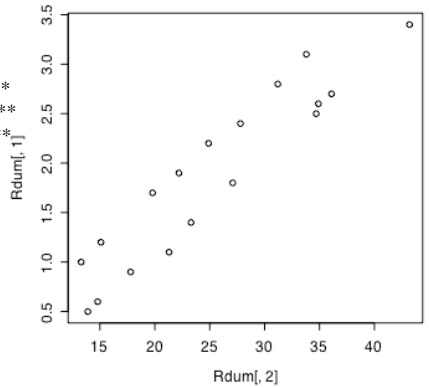
Per superare questo problema è necessario stimare il modello senza b_0

Questo però significa che l'intercetta dipende solo dalle dummies e che non ci sia un livello di base comune.

Esempio_11

```
> Rduml<-read.table(file="Rm06Exem11.csv",sep=",",header=T)
> names(Rdum)
> plot(Rdum[,2],Rdum[,1])
> Ols<-lm(spesa~-1+reddito+du+dd,data=Rdum)
> summary(Ols)
```

```
Estimate Std. Error t value Pr(>ltl)
reddito  0.099553   0.001185  84.042  < 2e-16 ***
du       -0.909051  0.034388 -26.435 5.36e-14 ***
dd       -0.303231  0.031663  -9.577 8.80e-08 ***
Residual standard error: 0.0426 on 15 dof
Multiple R-Squared: 0.9996
Adjusted R-squared: 0.9996
F-statistic: 1.408e+04 on 3 and
 15 DF,  p-value: < 2.2e-16
```



Le variabili politome

Una variabile qualitativa può avere più di due modalità. Ad esempio il pagamento di una transazione può avvenire in vari modi

$$\begin{cases} Contante \\ Assegno \\ Carta\ di\ credito \\ Cambiale \\ Quando\ posso \end{cases}$$

E' da scartare l'idea di utilizzare una pseudo variabile che assuma valori

$$D = \begin{cases} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{cases}$$

Tale codifica rende QUANTITATIVA una variabile QUALITATIVA: pagare in cambiale è quattro volte meglio (o peggio) che pagare in contanti?

E' invece necessario inserire cinque indicatori distinti

$$D_1 = \begin{cases} 1 & Se\ contanti \\ 0 & altrimenti \end{cases}; \quad D_2 = \begin{cases} 1 & Se\ assegno \\ 0 & altrimenti \end{cases}; \quad D_3 = \begin{cases} 1 & Se\ carta\ di\ credito \\ 0 & altrimenti \end{cases};$$

$$D_4 = \begin{cases} 1 & Se\ cambiale \\ 0 & altrimenti \end{cases}; \quad D_5 = \begin{cases} 1 & Se\ quando\ posso \\ 0 & altrimenti \end{cases};$$

Le variabili politome/2

Lo stesso tipo di obiezione vale per le variabili ordinali con categorie in numeri

In un studio sui clienti si potrebbe usare come indipendente la variabile: "Grado di Fedeltà".

$$C_i = \begin{cases} 1 & se\ cliente\ bituale \\ 2 & se\ cliente\ occasionale \\ 3 & se\ non\ cliente \end{cases}$$

Si può usare tale regressore per spiegare l'importo speso "y"

$$y_i = \beta_0 + \beta_1 C_i + u_i \qquad \text{con } C_i = \begin{cases} 1 \\ 2 \\ 3 \end{cases}$$

Avremmo le tre stime

$$\text{Abituale:} \quad y_i = \beta_0 + \beta_1;$$

$$\text{Occasionale:} \quad y_i = \beta_0 + 2\beta_1;$$

$$\text{Non consumatore:} \quad y_i = \beta_0 + 3\beta_1;$$

La differenza tra i primi due livelli è la stessa di quella tra gli ultimi due. Questo non è sensato perché le classi sono arbitrarie

Le variabili politome/3

Anche in questo caso è necessario fare entrare in gioco le variabili indicatore

Modalità	**D1**	**D2**	**D3**
Cliente abituale	1	0	0
Cliente occasionale	0	1	0
Non cliente	0	0	1

Con il vincolo dell'intercetta uguale a zero.

L'effetto differenziale tra "abituale' e "occasionale" sarà $(\beta_1 - \beta_2)$

e quello tra "occasionale" e "non cliente" da $(\beta_2 - \beta_3)$

Questa stima evita l'imposizione della scala arbitraria conseguente all'uso di una codifica per livelli

Discretizzazione

Talvolta le variabili quantitative non possono entrare nel modello perché poco precise o poco attendibili.

Ad esempio in un modello che leghi le spese alimentari alle spese non alimentari, al livello dei prezzi e al reddito:

$$A_i = \beta_0 + \beta_1 N_i + \beta_2 P_i + \beta_3 R_i + u_i$$

il reddito potrebbe essere ritenuto così "infedele" che entra solo per livelli

$$R_i = \begin{cases} 1 \text{ se le entrate sono inferiori a 12Mln} \\ 2 \text{ se le entrate ricadono in [12Mln-36 Mln]} \\ 3 \text{ se le entrate sono superiori a 36Mln} \end{cases}$$

Questo regressore non può essere usato perchè risponde con lo stesso incremento in "A" a variazioni molto diverse in "R"

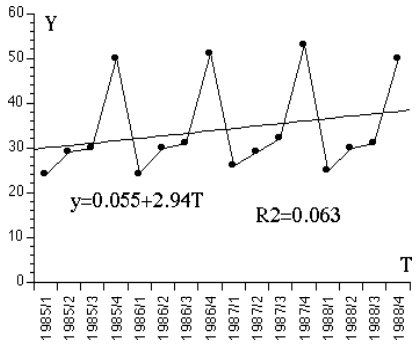
Non resta perciò che definire tre variabili dummy

$$D_{i1} = \begin{cases} 1 \text{ se } R_i < 12 \\ 0 \text{ altrimenti} \end{cases} \quad D_{i2} = \begin{cases} 1 \text{ se } 12 \le R_i \le 36 \\ 0 \text{ altrimenti} \end{cases} \quad D_{i3} = \begin{cases} 1 \text{ se } R_i > 36 \\ 0 \text{ altrimenti} \end{cases}$$

La stagionalità e le dummies

Supponiamo di voler esaminare i dati trimestrali della vendita di gioielli (dati TIME SERIES) in una certa regione

Periodo	Vendite
1985.1	24
1985.2	29
1985.3	30
1985.4	50
1986.1	24
1986.2	30
1986.3	31
1986.4	51
1987.1	26
1987.2	29
1987.3	32
1987.4	53
1988.1	25
1988.2	30
1988.3	31
1988.4	50



Il modello di regressione lineare delle vendite sul tempo che non tiene conto dell'incremento di vendite del 4^trimestre (periodo natalizio) è un modello insoddisfacente.

Esempio

In un campione di 60 consumatori è stata rtilevata la spesa mensile in gasolio, percorrenza media, regione di residenza (3 livelli) e classi di età (3 livelli)

Simate i parametri scomponendo le variabili politome

```
Call:
lm(formula = Spesa ~ -1 + Percorrenza + DumReg1 + DumReg2 + DumAge1 +
    DumAge2 + DumAge3)

Residuals:
    Min      1Q  Median      3Q     Max
-52.146 -11.906   0.662  14.394  31.473

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Percorrenza  4.628e-01  2.055e-03  225.16   <2e-16 ***
DumReg1     -1.950e+02  5.104e+00  -38.21   <2e-16 ***
DumReg2     -1.078e+02  4.720e+00  -22.84   <2e-16 ***
DumAge1      2.455e+03  9.544e+00  257.25   <2e-16 ***
DumAge2      2.459e+03  9.346e+00  263.13   <2e-16 ***
DumAge3      2.454e+03  9.392e+00  261.26   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.67 on 84 degrees of freedom
Multiple R-squared:     1,  Adjusted R-squared:     1
F-statistic: 8.031e+05 on 6 and 84 DF,  p-value: < 2.2e-16
```

Esempio

Case	VENDITE	PERIODO	DS1	DS2	DS3	DS4
1	24.000	1.000	1.000	0.000	0.000	0.000
2	29.000	2.000	0.000	1.000	0.000	0.000
3	30.000	3.000	0.000	0.000	1.000	0.000
4	50.000	4.000	0.000	0.000	0.000	1.000
5	24.000	5.000	1.000	0.000	0.000	0.000
6	30.000	6.000	0.000	1.000	0.000	0.000
7	31.000	7.000	0.000	0.000	1.000	0.000
8	51.000	8.000	0.000	0.000	0.000	1.000
9	26.000	9.000	1.000	0.000	0.000	0.000
10	29.000	10.000	0.000	1.000	0.000	0.000
11	32.000	11.000	0.000	0.000	1.000	0.000
12	53.000	12.000	0.000	0.000	0.000	1.000
13	25.000	13.000	1.000	0.000	0.000	0.000
14	30.000	14.000	0.000	1.000	0.000	0.000
15	31.000	15.000	0.000	0.000	1.000	0.000
16	50.000	16.000	0.000	0.000	0.000	1.000

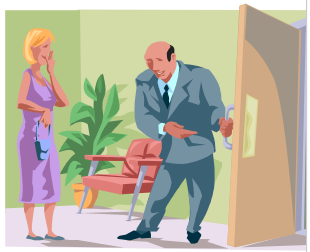
Dependent Var =	VENDITE	N of Cases =	16
Multiple R =	.999761313	F =	4606.9
R-sqr =	.999522683	p =	0.000000
Adjusted R-sqr =	.999305721	df =	5, 11
Std. error of est. =	.936021562		
Intercept =	forced to 0		

STATISTICA REGRESSION STATS — REGRESSION WEIGHTS

variable	BETA	St. Err. of BETA	B	St. Err. of B	t(11)	p-level
PERIODO	.0221162	.0142429	.08125	.0523252	1.55279	.1487534
DS1	.3403532	.0083648	24.18125	.5943000	40.68863	.0000000
DS2	.4060662	.0088378	28.85000	.6279023	45.94663	.0000000
DS3	.4260353	.0093449	30.26875	.6639323	45.59011	.0000000
DS4	.7063934	.0098809	50.18750	.7020162	71.49052	.0000000

Esempio

Spese di rappresentanza in relazione ai costi generali, alle venditi ed alla dimensione del bacino clienti: locale, regionale, nazionale, europea



```
Call:
lm(formula = Rappres. ~ -1 + Costi + Vendite + C1 + C2 + C3 +
    C4)

Residuals:
     Min       1Q   Median       3Q      Max
-0.34598 -0.05448 -0.01579  0.09190  0.33939

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
Costi     0.6709     0.2939   2.283 0.032951 *
Vendite   0.2650     0.1197   2.215 0.037970 *
C1        0.7143     0.4213   1.695 0.104801
C2        1.1929     0.4194   2.844 0.009719 **
C3        1.6808     0.4150   4.050 0.000577 ***
C4        2.2645     0.3860   5.867    8e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1584 on 21 degrees of freedom
Multiple R-squared: 0.9998, Adjusted R-squared: 0.9997
F-statistic: 1.795e+04 on 6 and 21 DF,  p-value: < 2.2e-16
```

a) Stimare i parametri del modello

b) Calcolare l'R²

Dummies sui coefficienti

In precedenza l'effetto delle dummies si è avuto solo sull'intercetta per ogni status. Nulla esclude che si possa avere anche sugli altri coefficienti

Per un lotto di auto usate sono stati rilevati il prezzo di vendita, l'anno di immatricolazione ed il chilometraggio; inoltre si è acquisito il tipo dialimentazione.

La formulazione usuale è

$$P_i = \beta_0 + \beta_1 K_i + \beta_2 I_i + u_i$$

Prezzo	Km	Imm	Alim
7.8	55	81	B
11.4	97	86	D
12.6	68	85	B
8.1	77	83	B
12.9	65	87	D
13.1	52	84	B
15.3	93	84	D
17.6	48	88	B
8.3	72	84	B
11.3	84	86	D
14.4	90	87	D

A questa formulazione vogliamo aggiungere sia una distinzione sull'intercetta che nei coefficienti

$$P_i = \beta_1 B_i + \beta_2 D_i + \beta_3 K_i + \beta_4 I_i + \beta_5 B_i K_i + \beta_6 D_i K_i + \beta_7 B_i I_i + \beta_8 D_i I_i + u_i$$

Dummies sui coefficienti /2

La nuova formulazione equivale ai due modelli

$$Benzina: \; P_i = \beta_1 + (\beta_3 + \beta_5) K_i + (\beta_4 + \beta_7) I_i + u_i$$

$$Diesel: \; P_i = \beta_2 + (\beta_3 + \beta_6) K_i + (\beta_4 + \beta_8) I_i + u_i$$

il vantaggio, rispetto alla stima di due modelli separati, è la presenza degli stessi errori "u" in entrambe le relazioni.

I termini

$$B_i K_i, D_i K_i, B_i I_i, D_i I_i$$

sono dei particolari REGRESSORI noti come TERMINI DI INTERAZIONE e segnalano che l'effetto sulla dipendente ad esempio del K cambia secondo il tipo di B/D (alimentazione)

L'idea di tali regressori può essere estesa al prodotto di due regressori qualsiasi

Dummies sui coefficienti/3

La stima dei parametri del modello con le interazioni non presenta novità.

P_i	S_i	int	B_iK_i	D_iK_i	B_iK_i	D_iK_i
7.8	1	1	55	0	81	81
11.4	-1	1	0	97	0	86
12.6	1	1	68	0	85	0
8.1	1	1	77	0	83	0
12.9	-1	1	0	65	0	87
13.1	1	1	52	0	84	0
15.3	-1	1	0	93	0	84
17.6	1	1	48	0	88	0
8.3	1	1	72	0	84	0
11.3	-1	1	0	84	0	86
14.4	-1	1	0	90	0	87

In questo caso si è esclusa l'esistenza di un effetto "Klm" o "Imm" che prescinda dal tipo di alimentazione ed i regressori K e I non compaiono da soli.

Questo assicura che il rango della matrice dei regressori, dopo l'accorpamento delle due dummies, sia di rango pieno

$$P_i = -25.16 - 38.82 B_i + 38.82 D_i - 0.169 B_i K_i + 0.014 D_i K_i + 1.021 B_i I_i - 0.021 D_i I_i$$

$$R^2 = 0.836$$

I due modelli separati sono

$$Benzina: \; P_i = -63.98 - 0.169 K_i + 1.021 I_i$$

$$Diesel: \; P_i = -13.66 + 0.014 K_i - 0.021 I_i$$

Esempio

Si studiano i tempi di adozione di una nuova polizza da parte delle compagnie di assicurazione.

Y: Variabile dipendente: Mesi precedenti l'adozione
X1: Variabile esogena: Fatturato (in milioni di euro)
X2: Tipologia società: Mutua oppure Spa o Srl

$$y_i = \beta_0 + \beta_1 M_i + \beta_2 A_i + \beta_3 M_i F_i + \beta_4 A_i F_i + u_i$$

$$M_i = \begin{cases} 1 \ se \ è \ una \ Muta \\ 0 \ altrimenti \end{cases}, \quad A_i = \begin{cases} 1 \ se \ è \ una \ Spa \ o \ srl \\ 0 \ altrimenti \end{cases}$$

Soc	Y	F	Tipo
1	17	151	Mu
2	26	92	Mu
3	21	175	Mu
4	30	31	Mu
5	14	246	Spa
6	30	124	Spa
7	13	305	Spa
8	20	166	Spa

L'intercetta deve essere omessa se sono presenti entrambe le dummy

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
M  32.40283    3.86426   8.385  0.00111 **
A  37.86763    5.43810   6.963  0.00224 **
MF -0.07931    0.03083  -2.573  0.06180 .
AF -0.08855    0.02454  -3.609  0.02258 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.44 on 4 degrees of freedom
Multiple R-squared: 0.9881, Adjusted R-squared: 0.9762
F-statistic:  82.9 on 4 and 4 DF,  p-value: 0.0004228
```

Considerazioni sui vincoli

Nascono dalla teoria o da un'esigenza di flessibilità perchè si possono prefissare uno o più parametri a dati livelli

$$\beta_1 = 1, \quad \beta_3 = 2 \Rightarrow \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta 2 \\ \beta_1 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Le informazioni sui vincoli o altre relazioni esterne al data set sono incluse nella relazione matriciale Rβ=r

dove la matrice R ha (m+1) colonne e h righe, ma il rango di R deve essere

inferiore a (m+1) altrimenti non ci sarebbe problema di stima. Infatti, si avrebbe subito

$$\left(R^t R\right)\beta = R^t r \Rightarrow \beta = \left(R^t R\right)^{-1} R^t r$$

Vincoli di uguaglianza sui paramteri

Diversi problemi si risolvono usando modelli di regressione con parametri soggetti a vincoli LINEARI DI UGUAGLIANZA

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

$$soggetto \ a \quad \begin{cases} \beta_0 + \beta_1 + \beta_2 + \beta_3 = 1 \\ \beta_1 = 2\beta_2 \end{cases}$$

I vincoli lineari possono essere espressi con le matrici

$$y = X\beta$$
$$Soggetto \ a \ \ R\beta = r$$

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & -2 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta 2 \\ \beta_1 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

I vincoli di disuguaglianza del tipo Rβ≤r richiedono metodi ad hoc. Non sempre sono compabili con i dati: se lo sono, i vincoli sono superflui e se non lo sono o i dati o il modello o entrambi sono sbagliati

Impostazione

L'accorpamento delle osservazioni e delle informazioni a priori avviene con la tecnica delle matrici a blocchi

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X \\ R \end{bmatrix} \beta + \begin{bmatrix} u \\ 0 \end{bmatrix}$$

Poichè R ha rango h possiamo costruire una sottomatrice quadrata R_1 di ordine h dotata di inversa e definire la suddivisione seguente

$$\begin{bmatrix} X \\ R \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \\ R_1 & R_2 \end{bmatrix}$$

Ne consegue che

$$\begin{bmatrix} y \\ r \end{bmatrix} = \begin{bmatrix} X_1 & X_2 \\ R_1 & R_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u \\ 0 \end{bmatrix} \Rightarrow \begin{matrix} y = X_1\beta_1 + X_2\beta_2 + u \\ r = R_1\beta_1 + R_2\beta_2 \end{matrix}$$

Impostazione /2

Analizziamo la 2ª relazione

$$r = R_1\beta_1 + R_2\beta_2$$

Risolvendola rispetto a β_1 (cioè sottraendo il 2° addendo e moltiplicando per l'inversa di R_1 (che esiste per costruzione) si ha

$$\beta_1 = R_1^{-1}\left(r - R_2\beta_2\right)$$

Inserendo questo risultato nella prima si perviene a

$$y = X_1R_1^{-1}\left(r - R_2\beta_2\right) + X_2\beta_2 + u \Rightarrow y = X_1R_1^{-1}r - R_1^{-1}R_2\beta_2 + X_2\beta_2 + u$$

$$y - X_1R_1^{-1}r = \left(X_2 - R_1^{-1}R\right)\beta_2 + u \Rightarrow z = W\beta_2 + u$$

$$z = y - X_1R_1^{-1}r$$

$$W = \left(X_2 - X_1R_1^{-1}R_2\right)$$

In pratica, i vincoli ci hanno permesso di eliminare alcuni parametri incogniti.

Esempio

La superficie in mq degli appartamenti può essere legata

- **al reddito del nucleo familiare,**
- **al numero di componenti ;**
- **agli anni di scolarità aggiuntivi all'obbligo dei percettori di reddito**

$$y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_3x_{i3} + u_i$$

$$soggetto\ \ a\ \begin{cases}\beta_0 + \beta_2 = 0\\ \beta_1 = \dfrac{\beta_2}{2}\end{cases}$$

Unità abit.	Superf. Mq	Reddito Mgl euro	compon. numero	Scolar. anni
1	90	28	2	10
2	110	46	4	14
3	120	50	5	15
4	150	80	3	21
5	80	30	2	12
6	180	100	6	18
7	100	40	5	9

Calcolo delle stime

La stima dei minimi quadrati nel modello con le nuove variabili è

$$\hat{\beta}_2 = (W^tW)^{-1}W^tz$$

A questo punto non rimane che completare la stima dei parametri accantonati

$$\hat{\beta}_1 = R_1^{-1}(r - R_2\hat{\beta}_2)$$

Controlliamo che le stime verifichino i vincoli

$$R\hat{\beta} = \begin{bmatrix}R_1 & R_2\end{bmatrix}\begin{bmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{bmatrix} = R_1\hat{\beta}_1 + R_2\hat{\beta}_2 =$$

$$= RR_1^{-1}(r - R_2\hat{\beta}_2) + R_2\hat{\beta}_2 = r - R_2\hat{\beta}_2 + R_2\hat{\beta}_2 = r$$

Esempio (continua)

Occorre per prima cosa predefinire tanti parametri per quanti vincoli. Quali parametri è indifferente (almeno statisticamente)

$$y = \begin{bmatrix}90\\110\\120\\150\\80\\180\\100\end{bmatrix};X = \begin{bmatrix}1 & 28 & 2 & 10\\1 & 46 & 4 & 14\\1 & 50 & 5 & 15\\1 & 80 & 3 & 21\\1 & 30 & 2 & 12\\1 & 100 & 6 & 18\\1 & 40 & 5 & 9\end{bmatrix};R = \left[\begin{array}{cc|cc}1 & 0 & 1 & 0\\0 & 1 & -\frac{1}{2} & 0\end{array}\right];r = \begin{bmatrix}0\\0\end{bmatrix} \qquad \beta = \begin{bmatrix}\beta_0\\\beta_1\\\beta_2\\\beta_3\end{bmatrix}$$

La matrice identità è molto comoda in questo caso

$$R_1 = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix};R_1^{-1} = \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix};X_1 = \begin{bmatrix}1 & 28\\1 & 46\\1 & 50\\1 & 80\\1 & 30\\1 & 100\\1 & 40\end{bmatrix};X_2 = \begin{bmatrix}2 & 10\\4 & 14\\5 & 15\\3 & 21\\2 & 12\\6 & 18\\5 & 9\end{bmatrix}$$

Esempio/Continua_2

Definizione della matrice dei regressori e del vettore della dipendente nel modello RESIDUALE

$$z = y - X_1 R_1^{-1} r = y \qquad \text{(dato che } r = 0\text{)}$$

$$W = X_2 - X_1 R_1^{-1} R_2 = X_2 - X_1 R_2$$

$$W = \begin{bmatrix} 2 & 10 \\ 4 & 14 \\ 5 & 15 \\ 3 & 21 \\ 2 & 12 \\ 6 & 18 \\ 5 & 9 \end{bmatrix} - \begin{bmatrix} 1 & 28 \\ 1 & 46 \\ 1 & 50 \\ 1 & 80 \\ 1 & 30 \\ 1 & 100 \\ 1 & 40 \end{bmatrix} * \begin{bmatrix} 1 & 0 \\ -\frac{1}{2} & 0 \end{bmatrix} = \begin{bmatrix} 2 & 10 \\ 4 & 14 \\ 5 & 15 \\ 3 & 21 \\ 2 & 12 \\ 6 & 18 \\ 5 & 9 \end{bmatrix} - \begin{bmatrix} -13 & 0 \\ -22 & 0 \\ -24 & 0 \\ -39 & 0 \\ -14 & 0 \\ -49 & 0 \\ -19 & 0 \end{bmatrix} = \begin{bmatrix} 15 & 10 \\ 26 & 14 \\ 29 & 15 \\ 42 & 21 \\ 16 & 12 \\ 55 & 18 \\ 24 & 9 \end{bmatrix}$$

A questo punto abbiamo un modello con soli due parametri

Esempio continua

$$W^t W = \begin{bmatrix} 7363 & 3229 \\ 3229 & 1511 \end{bmatrix}; \quad \left(W^t W\right)^{-1} = \frac{1}{699052} \begin{bmatrix} 1511 & -3229 \\ -3229 & 7363 \end{bmatrix}$$

$$W^t z = \begin{bmatrix} 27570 \\ 12490 \end{bmatrix}; \quad \hat{\beta}_2 = \begin{bmatrix} 1.90 \\ 4.21 \end{bmatrix}$$

$$\hat{\beta}_1 = R_1^{-1}\left(r - R_2 \hat{\beta}_2\right) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ -0.5 & 0 \end{bmatrix}\begin{bmatrix} 1.90 \\ 4.21 \end{bmatrix}\right) = \begin{bmatrix} -1.9 \\ 0.95 \end{bmatrix}$$

In definitiva $\hat{\beta} = \begin{bmatrix} 1.9 \\ 4.21 \\ -1.9 \\ 0.95 \end{bmatrix}$ **che verifica i vincoli Imposti nella stima.**

A questo punto si possono avviare le verifiche di adattamento sul modello ridotto.

Esercizio

Per i dati seguenti

Y= risparmio

Anno	Risparmio	Reddito	Rendimento	Prezzi
1	80	210	9	100
2	91	241	10	102
3	87	234	11	104
4	105	278	8	105
5	115	325	12	101
6	113	296	7	98
7	111	290	9	95
8	109	289	4	92
9	110	284	5	96
10	112	292	7	103

$$R = \begin{bmatrix} R_1 & R_2 \end{bmatrix} = \left[\begin{array}{cc|cc} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}\right]; \quad r = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_0 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Si consideri il seguente modello macroeconomico

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + u_i$$

$$soggetto\ a \begin{cases} \beta_0 - \beta_2 = 0 \\ \beta_1 + \beta_3 = 1 \end{cases}$$

a) Stimare i parametri;
b) Verificare l'adattamento

```
rm(list = ls())
Ris<-read.table(file="Risparmio.csv",sep=";",header=T,row.names=1)
n<-nrow(Ris)
X1<-cbind(rep(1,n),Ris[,2])X2<-Ris[,3:4];y<-Ris[,1]
R1<-diag(c(1,1));R1m1<-diag(c(1,1));R2<-diag(c(1,1));r<-c(0,1)
z<-y-X1%*%R1m1%*%r;z
W<-X2-X1%*%R1m1%*%R2;W<-as.data.frame(W);W
Ols<-lm(z~-1+.,data=W);summary(Ols)
B2<-as.matrix(Ols$coef);B2
B1<-R1m1%*%(r-R2%*%B2);B1
R<-cbind(R1,R2);Beta<-rbind(B1,B2)
rc<-R%*%Beta;rc
```

Vincoli lineari come ipotesi sui parametri

La relazione $R\beta = r$ **Può essere considerata come una ipotesi da verificare sui parametri**

$$\begin{cases} H_0 : R\beta = r \\ H_1 : R\beta \neq r \end{cases}$$

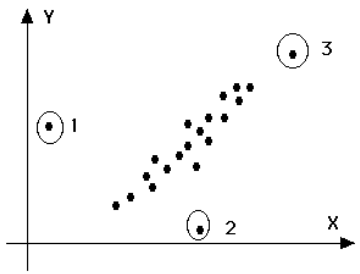
Ad esempio, nell'esercizio precedente si ha H_0 come: $H_0 : \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_0 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$

```
> library(car)
> Ols<-lm(z~.,data=Ris)
> linearHypothesis(Ols, c("1*(Intercept)+1*Rendimento=0","1*Reddito+1*Prezzi=1"))
Linear hypothesis test

Hypothesis:
(Intercept) + Rendimento = 0
Reddito + Prezzi = 1

Model 1: restricted model
Model 2: z ~ Risparmio + Reddito + Rendimento + Prezzi

  Res.Df    RSS Df Sum of Sq          F    Pr(>F)
1      7 428.18
2      5   0.00  2    428.18 2.9255e+28 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In questo caso l'ipotesi dovrebbe essere respinta ovvero i vincoli sono effettivamente stringenti

Valori remoti

Alcune osservazioni possono risultare così "remote" dalle altre da avere una influenza eccessiva sul modello e determinare un pessimo FITTING



I punti "1" e "2" sembrano in netto contrasto con la configurazione complessiva dei dati. Il "3", pur essendo anomalo, sembra collocarsi nel trend del fenomeno

Da notare che il punto "1" è anomalo rispetto alla "X", il "2" rispetto alla "Y" ed il "3" rispetto ad entrambe

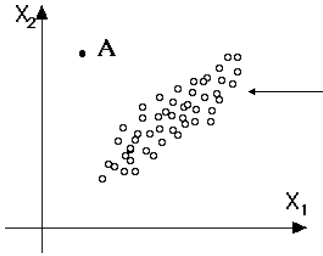
L'effetto della "1" e della "2" potrebbe non essere eccessivo: il valore della Y è coerente con l'andamento generale. La "2" invece infuisce molto (in modo negativo) dato che il suo "Y" è molto scentrato.

ATP2

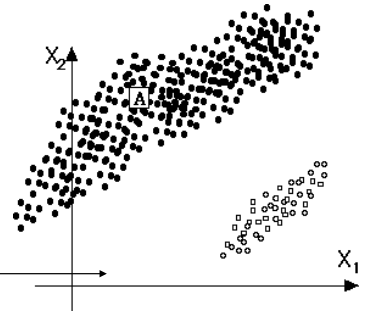
Valori remoti/2

Nelle applicazioni a dati reali qualche rilevazione scaturisce da circostanze inusuali: catastrofi naturali, problemi internazionali, cambiamenti politici, scioperi o serrate, etc.

C'è poi il rischio che certi dati siano sbagliati per mero errore materiale



In questi casi è necessario accertarsi che i valori remoti o anomali ci siano, ma su questa strada c'è incertezza



Non c'è alcuna garanzia che il punto "A" sia ANOMALO e gli altri NORMALI.

Un ampliamento delle rilevazioni potrebbe dar luogo ad uno scatter diverso

Diagnostiche per i valori remoti

Più importante è valutare l'influenza dei valori remoti sul modello.

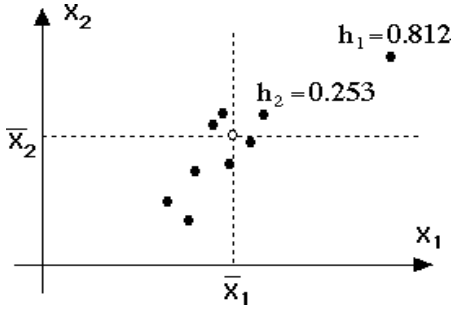
Nel caso della regressione lineare semplice è sufficiente lo studio dello scatterplot. Se i dati sono multidimensionali è necessario ricorrere a speciali formule.

Nel prosieguo studieremo tre diagnostiche:

- Un indice che esprima la posizione della i-esima osservazione rispetto alle altre
- Un indice che esprima l'effetto di eliminare l' i-esima osservazione sui valori teorici
- Un indice che esprima l'effetto di eliminare l i-esima osservazione sulla stima dei parametri

Esistono anche misure basate sull'effetto di cancellazione di più di una osservazione, ma non saranno considerate nel nostro corso

Uso della matrice cappello



Una leva prossima ad uno indica che l'osservazione è molto discosta dal "nucleo" dei dati ed un valore prossimo a zero significa che si colloca in prossimità del punto medio

Se la leva dell'i-esimo dato è grande essa contribuisce fortemente a determinare il valore teorico della risposta.

Poichè le teoriche sono una combinazione lineare delle osservate

$$\hat{y} = Hy$$

Maggiore è h_i maggiore sarà il peso di X_i sul valore stimato. Al limite, se fosse $h_i=1$ allora

$$\hat{y}_i = y_i$$

quindi il modello sarebbe VINCOLATO a stimare esattamente y_i col rischio di viziare l'adattamento delle altre osservazioni

Un valore di soglia per la leva

Quant'è che il valore della leva è tanto grande da preoccupare per il fitting del modello?

In media, il valore di h_i è pari a $\bar{h} = \frac{\sum_{i=1}^{n} h_i}{n} = \frac{m+1}{n}$

sarà considerato "eccessiva" una leva SUPERIORE al doppio della media

h_i è da considerarsi eccessivo se $h_i \geq 2\left(\frac{m+1}{n}\right)$ **Purché n<2(m+1)**

Le indicazioni ottenute con la leva prescindono dai valori osservati sulla dipendente, ma quantificano la "forza" che eserciterà la osservata y_i sulla stimata $\hat{y}_i$

Maggiore è la leva h_i, maggiore sarà l'influenza del punto i-esimo sulla regressione

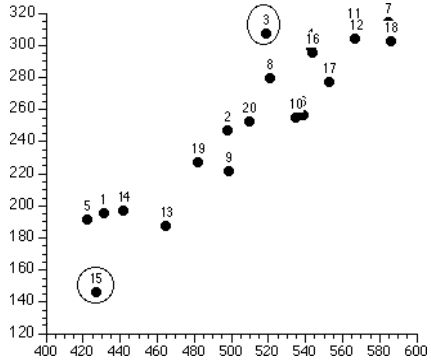
Esempio

Ecco alcuni dati regionali: due regressori e la leva. Il valore di soglia è

Regione	X_{i1}	X_{i2}	h_i
1	19.5	43.1	0.201
2	24.7	49.8	0.059
3	30.7	51.9	0.372
4	29.8	54.3	0.111
5	19.1	42.2	0.248
6	25.6	53.9	0.129
7	31.4	58.5	0.156
8	27.9	52.1	0.096
9	22.1	49.9	0.115
10	25.5	53.5	0.110
11	31.1	56.6	0.120
12	30.4	56.7	0.109
13	18.7	46.5	0.178
14	19.7	44.2	0.148
15	14.6	42.7	0.333
16	29.5	54.4	0.059
17	27.7	55.3	0.106
18	30.2	58.6	0.197
19	22.7	48.2	0.067
20	25.2	51.0	0.050

$$2\left(\frac{m+1}{n}\right) \Rightarrow 2\left(\frac{2+1}{20}\right) = 0.3$$

che evidenzia come anomale le rilevazioni "3" e "15".



Entrambe le osservazioni hanno leva molto alta rispetto alla terza leva più grande

Diagnostica dell'adattamento

Una volta individuate le osservazioni con leva molto alta occorre stabilire se il loro effetto sull'adattamento inficia realmente la validità del modello.

Una prima idea può ottenersi dall'esame dei residui $\hat{e}_i = y_i - \hat{y}_i$

In particolare dai residui standardizzati

$$e_i^* = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_i}} \qquad i = 1, 2, \ldots, n; \qquad \hat{\sigma} = \sqrt{\frac{SSE}{n-(m+1)}} \qquad SSE = \sum_{i=1}^{n} e_i^2$$

In realtà, la formula descrive i residui *studentized* cioè rapportati non tanto al loro scarto quadratico medio $\hat{\sigma}$, ma a questo corretto per il fattore $\sqrt{(1-h_i)}$ che li rende maggiormente comparabil.

Residui SD

I residui s*tudentized* non sono usati in pratica. Sono più informativi i residui ottenuti dopo aver cancellato l'i-esima osservazione (deleted).

In questo modo il valore teorico $\hat{y}_i$ non può essere influenzato da forzature verso il valore osservato y_i in quanto la i-esima osservazione è esclusa

Il calcolo dei residui SD (*Studentized Deleted*) può essere effettuato con le quantità già ottenute dal classico modello di regressione

$$d_i^* = e_i \sqrt{\frac{n-(m+1)}{\hat{SSE}*(1-h_i)-\hat{e}_i^2}}$$

Maggiore è d_i più influente è l'osservazione per determinare y_i

Tali valori andrebbero confrontati con i quantili della t-Student con n-(m+1) gradi di libertà

In linea di massima se $d_i^* > 1.6$

si può ritenere che l'effetto della i-esima osservazione sia eccessivo ovvero che sia un valore anomalo

Esempio

Regione	h_i	e_i	d^*_i
1	0.201	-1.68	-0.75
2	0.059	3.64	1.58
3	0.372	-3.17	-1.70
4	0.111	-3.16	-1.39
5	0.248	0.00	0.00
6	0.129	-0.36	-0.15
7	0.156	0.72	0.31
8	0.096	4.02	1.82
9	0.115	2.66	1.15
10	0.110	-2.48	-1.07
11	0.120	0.34	0.14
12	0.109	2.23	0.95
13	0.178	-3.95	-1.88
14	0.148	3.45	1.57
15	0.333	0.57	0.27
16	0.059	0.64	0.26
17	0.106	-0.85	-0.35
18	0.197	-0.78	-0.34
19	0.067	-2.86	-1.21
20	0.050	1.04	0.42

Se il residuo SD è grande, il dato corrispondente potrebbe essere anomalo.

Lo studio dei residui SD evidenzia le osservazioni 3, 8, 13 come "anomale".

In realtà anche i residui semplici avrebbero dato la stessa indicazione, ma segnalando anche come remote altre osservazioni che invece risultano normali.

Da notare che solo per l'osservazione 3 coincidono le indicazioni della leva e degli SD

N.B. più grande è l'ampiezza del campione, maggiore sarà il numero di osservazioni che potrebbe apparire anomalo (senza esserlo).

Diagnostica sui parametri

Abbiamo già visto che, senza ripetere i calcoli, è possibile misurare l'effetto sui β stimati della esclusione della osservazione i-esima

$$\hat{\beta}_n = \hat{\beta}_o - \frac{\hat{e}_i}{1-h_i} W_o^{-1} x_i^t$$

Una sintesi di queste variazioni è la DISTANZA DI COOK

$$c_i = \frac{(\hat{\beta}_o - \hat{\beta}_n)^t W_o^{-1} (\hat{\beta}_o - \hat{\beta}_n)}{m\,\hat{\sigma}} = \frac{\left(\hat{y}_o - \hat{y}_n\right)^t \left(\hat{y}_o - \hat{y}_n\right)}{m\,\hat{\sigma}}$$

che mostra l'equivalenza tra variazione nei parametri e variazione nei valori stimati dovuta alla cancellazione della i-esima osservazione

Il calcolo della formula è basato sulle quantità già usate per le altre diagnostiche $c_i = \frac{\hat{e}_i^2}{m\,\hat{\sigma}}\left[\frac{h_i}{(1-h_i)}\right]^2$

Significato della distanza di Cook

La formula dei c_i evidenzia la dipendenza da due fattori: dal residuo (quindi dall'adattamento) e dalla leva (dalla collocazione rispetto agli altri punti).

maggiore è il residuo $\hat{e}_i$ oppure la leva h_i più grande sarà la distanza. Ne consegue che una osservazione può essere INFLUENTE perché

- **è associato ad un residuo elevato, ma con leva moderata**
- **è associato ad un residuo piccolo, ma con leva elevata**
- **è associato ad un residuo ed una leva entrambi elevati**

E' per questo che la distanza di Cook si affianca bene alle altre misure e le completa

Ancora sulla distanza di Cook

E' difficile stabilire un valore di soglia per le c_i. A questo fine si usano i quantili della F-Fisher (m,n-m), che sono solo parzialmente appropriati.

Ci si può però basare sul valore di equilibrio della leva

$$h_i = \left(\frac{m+1}{2}\right)$$

nonché su di un valore standard per $\left|\frac{e_i}{m\,\hat{\sigma}}\right| \cong 2$

Molto empiricamente, consideriamo elevata la distanza di Cook se

$$c_i > \frac{4}{n-(m+1)}$$

Esempio

Regione	h_i	e_i	c_i
1	0.201	-1.68	0.046
2	0.059	3.64	0.045
3	0.372	-3.17	0.488
4	0.111	-3.16	0.072
5	0.248	0.00	0.000
6	0.129	-0.36	0.001
7	0.156	0.72	0.006
8	0.096	4.02	0.098
9	0.115	2.66	0.054
10	0.110	-2.48	0.044
11	0.120	0.34	0.001
12	0.109	2.23	0.035
13	0.178	-3.95	0.212
14	0.148	3.45	0.125
15	0.333	0.57	0.013
16	0.059	0.64	0.001
17	0.106	-0.85	0.005
18	0.197	-0.78	0.010
19	0.067	-2.86	0.032
20	0.050	1.04	0.003

La distanza di Cook conferma come dati anomali quelli relativi alla 3^osservazione.

La 13^ è sospetta perché la sua c_i è vicina al valore di soglia: 0.24

Ma si tratta di reali anomalie?

La differenza nella stima dell'i-esimo valore della dipendente è

$$\hat{y}_{io} - \hat{y}_{in} = \frac{h_i \hat{e}_i}{1 - h_i}$$

$$\text{in particolare } \hat{y}_{3o} - \hat{y}_{3n} = \frac{h_3 e_3}{1 - h_3} = \frac{0.372 * (-3.17)}{1 - 0.372} = 1.88$$

che rispetto al valore osservato: 30.7 costituisce appena il 6.3%. Nonostante le indicazioni, la 3^non è un valore anomalo

Esempio

Nel modello del ciclo vitale formulato da Modigliani il tasso di risparmio (risparmi privati/reddito disponibile) è spiegato con la relazione

$$T = \beta_0 + \beta_1 POP15 + \beta_2 POP75 + \beta_3 RDM + \beta_1 POP15 + \beta_4 TCR + u$$

Dove

- **POP15= % di pop.res. con età inferiore a 15 anni**
- **POP75= % di pop.res. con età superiore a 75 anni**
- **RDM= % Reddito disponibile procapite**
- **TCR= % Tasso di crescita di RDM**

Indice	Paese	TR	POP15	POP75	RDM	TCR
1	Australia	11.43	29.35	2.87	2329.68	2.87
2	Austria	12.07	23.32	4.41	1507.99	3.93
3	Belgium	13.17	23.80	4.43	2108.47	3.82
4	Bolivia	5.75	41.89	1.67	189.13	0.22
5	Brazil	12.88	42.19	0.83	728.47	4.56
6	Canada	8.79	31.72	2.85	2982.88	2.43
7	Chile	0.60	39.74	1.34	662.86	2.67
8	Taiwan	11.90	44.75	0.67	289.52	6.51

I dati sono misurati come medie decennali

Esempio sul ciclo vitale/2

Il DATASET relativo all'esercizio prevede rilevazioni per 50 paesi. La stima dei parametri è la seguente

$$TR_i = 28.56 - 0.46 POP15_i - 1.69 POP75_i - 0.00034 RDM_i + 0.41 TCR_i$$

$$R^2 = 0.33; \quad \hat{\sigma} = 3.802$$

Questa fase dei calcolo non presenta particolari novità.

Per il calcolo delle diagnostiche è necessario calcolare residui e leve. In realtà, i packages moderni provvedono le misure necessarie. In mancanza si può procedere con il foglio elettronico.

Paese	y_i	h_i	$\beta_0+\beta_1 x_{i1}+\dots$	e_i	d^*_i	c_i
Australia	11.43	0.0677	10.56	0.866	0.236	0.000
Austria	12.07	0.1203	11.45	0.619	0.174	0.000
Belgium	13.17	0.0874	10.95	2.221	0.614	0.003
Bolivia	5.75	0.0894	6.45	-0.697	-0.192	0.000
Brazil	12.88	0.0695	9.32	3.555	0.980	0.005
Canada	8.79	0.1584	9.10	-0.315	-0.090	0.000
Chile	0.6	0.0372	8.84	-8.240	-2.339	0.007
Taiwan	11.9	0.0779	9.36	2.538	0.699	0.003

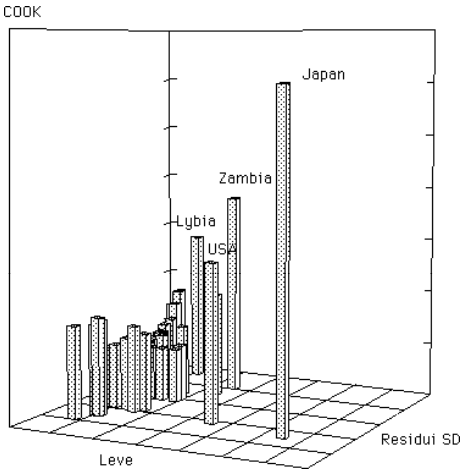
I valori di soglia sono

$$h_i: 2 * \frac{5}{50} = 0.2; \quad d_i^*: 1.6$$

$$c_i: \frac{4}{45} = 0.089;$$

C'è un solo sospetto: il Cile, ma non ci sono tutti i riscontri

Esempio sul ciclo vitale/3



Il modo migliore di analizzare le tre diagnostiche è di collocarle in un grafico 3D.

Il grafico evidenzia i maggiori *outliers* ovvero i paesi che sembrano più discostarsi dal complesso dei dati

Che fare?

L'osservazione $A=(y_i, x_{i1},x_{i2},\ldots,X_{im})$ è giudicata anomala se sembra NON seguire la struttura del modello laddove la stragrande maggioranza degli altri dati vi si adatta bene

Se A è considerato anomalo si può…

Escluderlo dal data set con guadagno sul fitting del modello. Attenzione! Per alcuni fenomeni non è serio eliminare dei dati (pensate ad esempio alle osservazioni sulle massime dei fiumi, delle piogge, delle eruzioni vulcaniche, etc.

Farlo intervenire con un peso ridotto in modo da attenuarne l'impatto. Attenzione! Si aggiunge un problema: la scelta dei pesi.

Utilizzare un criterio alternativo ai minimi quadrati che sia meno sensibile ai valori remoti.

Brownlee's stack-loss data

The data consist of n=21 daily observations measured in a plant for the oxidation of ammonia to nitric acid.

a) Stimare il modello di regressione lineare multipla.

b) Rappresentare graficamente le diagnostiche sui valori remoti.

c) Verificarne numericamente la presenza.

File Stackloss.csv

Stack Loss	Air Flow	Water Temper.	Acid Concentr.
42	80	27	89
37	80	27	88
37	75	25	90
28	62	24	87
18	62	22	87
18	62	23	87
19	62	24	93
20	62	24	93
15	58	23	87
14	58	18	80
14	58	18	89
13	58	17	88
11	58	18	82
12	58	19	93
8	50	18	89
7	50	18	86
8	50	19	72
8	50	19	79
9	50	20	80
15	56	20	82
15	70	20	91