Regressione semplice in notazione matriciale

Le vendite di dentifricio dipendono 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

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

$$Y = X\beta + u$$

Metodo dei minimi quadrati

Fra i molti possibili criteri per calcolare i parametri incogniti, quello più usato è il metodo dei minimi quadrati

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

$$y = X\beta + u$$

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

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

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

questo è uno scalare ed perciò sempre uguale al suo trasposto

La minimizzazione rispetto a "β" implica la derivazione delo scalare "s' rispetto al vettore "β" ed uguagliando a zero il risultato.

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

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

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

$$\hat{\beta} = \left(X^t X \right)^{-1} X^t y = \frac{1}{29} \begin{bmatrix} 42.25 & -13.5 \\ -13.5 & 5 \end{bmatrix} \begin{bmatrix} 26.2 \\ 79.7 \end{bmatrix} = \begin{bmatrix} 31/29 \\ 44.8/29 \end{bmatrix} = \begin{bmatrix} 1.07 \\ 1.55 \end{bmatrix}$$

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

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*

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

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

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

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

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

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

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

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:

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

Questa colonna non sempre è presente

Le variabili indipendenti X sono anche dette REGRESSORI

Pubblicità e dentifrici/3

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

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

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

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

$$= 152.44 - 151.569 = 0.871$$

Regressione Multipla/2

Il sistema di equazioni per la stima dei parametri rimane lo stesso

$$(X^tX)\beta = X^ty \longrightarrow \hat{\beta} = (X^tX)^{-1}X^ty$$

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

$$X^ty = \begin{bmatrix} \sum y_i \\ \sum y_i x_{i1} \\ \sum y_i x_{i2} \\ \vdots \\ \sum y_i x_{im} \end{bmatrix}$$

Esempio

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

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

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

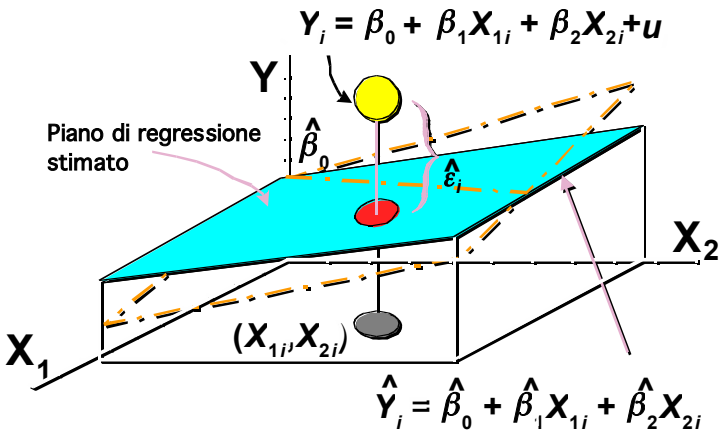
$$(X^t X)^{-1} = \frac{1}{480} \begin{bmatrix} 790 & -80 & -42 \\ -80 & 16 & 0 \\ -42 & 0 & 6 \end{bmatrix}$$

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

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, che rimane incognito.

Somma dei quadrati degli errori

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

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

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

Con le matrici abbiamo:

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

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

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

$$= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}\mathbf{y}$$

Questo è il vettore dei parametri stimati

```
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 di due matrici particolari

- 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)

Misura della bontà di adattamento

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

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

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:

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

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

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:

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

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

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

$Cosdef$ = *Prezzo deflazionato benzina*

$Permed$ = *percorrenza media per auto*

$Numaut$ = *numero auto circolanti*

$Kilit$ = *Km percorsi con un litro*

Pop = *Popolazione presente*

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

```
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
```

Campioni e popolazione

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

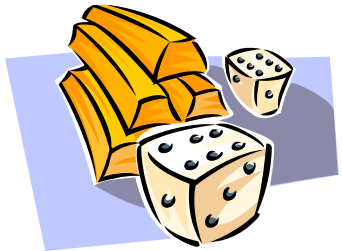
Ogni campione può dare una sola stima del modello (fermo restando l'ampiezza campionaria)

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

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

Poiché i campioni variano, variano anche le stime. Cosa possiamo dire sulle stime che non abbiamo?

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



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"

$$y_i = \sum_{j=0}^{m-1} \beta_j x_i^j + u_i$$

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** $\overline{R}^2 \leq 0$

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

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

Le costanti c possono essere nulle, ma non tutte insieme

Stimatori dei parametri β soddisfacenti e agevoli da trattare si ottengono con una funzione lineare dei dati osservati nella dipendente y

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

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

Soluzione

Due dei requisiti richiesti ad uno stimatore sono:

- **Essere corretto (non distorto)**

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

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

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

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

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

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 ***Dipendono da c***

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

B.L.U.E.

Restano perciò da determinare i pesi della combinazione. Sostituendo θ abbiamo

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

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

BLUE/2

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

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

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

Ripetendo le operazioni per ogni parametro si arriva a

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

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

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 ha forma

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

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

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

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/2

La devianza dei residui SSE è espressa da una forma quadratica

$$SSE = y^t (I - H) y$$

Dove H è la matrice cappello simmetrica e idempotente.

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

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

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

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

Stima della varianza per gli OLS

Consideriamo la forma quadratica

$$x^t A x = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

Ipotizziamo che E(x)=*μ* e Var(x)=V. Poiché $E\left(xx^t\right) = V + \mu\mu^t$

avremo

$$E\left(x^t A x\right) = E\left[Tr\left(x^t A x\right)\right] = Tr\left[E\left(A x^t x\right)\right] = Tr\left[AE\left(x^t x\right)\right]$$

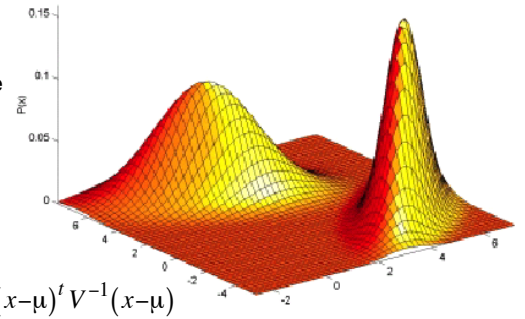
E dunque

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

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

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

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:

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

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

- **Gli errori del modello hanno distribuzione gaussiana multivariata**

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

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 dei minimi quadrati del parametro e la sua deviazione standard

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

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.

Conseguenze

- **La gaussianità degli errori si estende alle osservazioni sulla y**

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

- **Anche gli stimatori dei parametri hanno distribuzione gaussiana**

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

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

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 (perché il risultato è dovuto al caso)

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

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

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

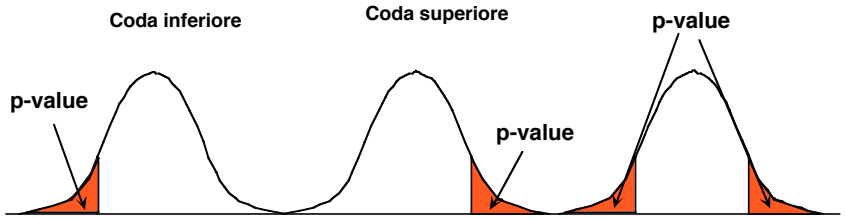
Il modello senza intercetta è migliorativo una volta su tre. Non è consigliabile rifiutare H_0.

Precisazioni

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

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

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



p-Value/2

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

Nel caso della gaussiana si ha:

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

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

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)

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

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

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

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

=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

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

Esempio

```
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
```

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

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%

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.

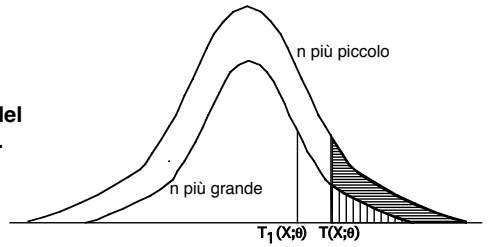
Questa relazione è simile a quella che lega la t-Student del coefficiente angolare al test-F nel modello di regressione lineare semplice.

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 significativi, ma 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
```

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

La leva dell'osservazione

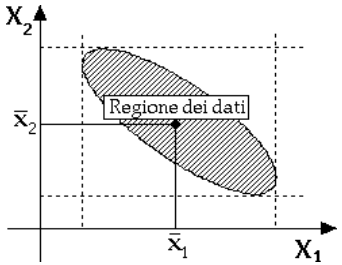
Un'indicazione della variabilità dei regressori si ha dalla diagonale della matrice cappello

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

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

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

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

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.

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

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 (ipotetica) 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

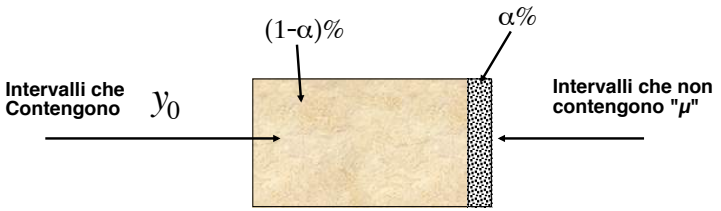
$$\hat{y}_0 - t_{1-\frac{\alpha}{2},n-m-1}\hat{\sigma}\sqrt{x_0^t\left(X^tX\right)^{-1}x_0} < y_0 < \hat{y}_0 + t_{1-\frac{\alpha}{2},n-m-1}\hat{\sigma}\sqrt{x_0^t\left(X^tX\right)^{-1}x_0}$$

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

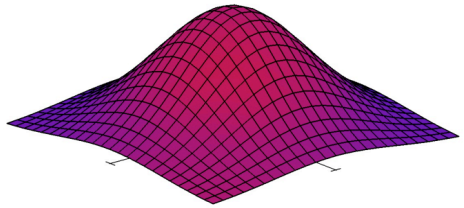
$$Confidenza\ \ al\ \ 99\% : 20.25 - 9.92\sqrt{50.025} < y_0 < 20.25 + 9.92\sqrt{50.025}$$
$$Previsione\ \ al\ \ 99\% : 20.25 - 9.92\sqrt{55.725} < y_0 < 20.25 + 9.92\sqrt{57.725}$$

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

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

Altro 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
```

Linearità del modello di regressione

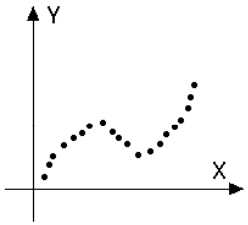
La linearità del modello di regressione dipende solo da come vi compaiono i parametri.

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

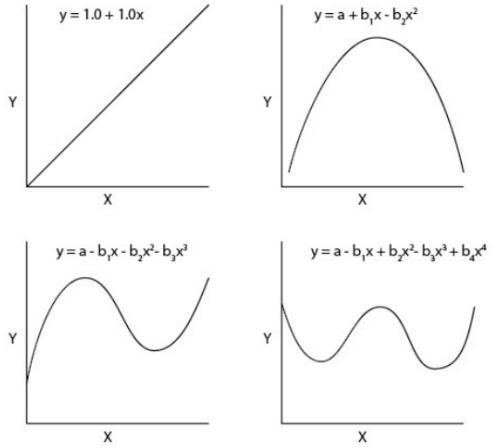
$$y_i = b_0 + b_1x_i + b_2x_i^2 + b_3x_i^3 + u_i$$
$$= b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + u_i$$

Anche in questo caso il modello è lineare

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.



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

Ci sono però delle difficoltà

Un grado elevato comporta problemi di OVERFLOW e UNDERFLOW nella 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 capacità di rappresentazione del computer.

Un grado elevato comporta problemi di condizionamento nella matrice dei coefficienti

Le potenze elevate ravvicinate hanno andamenti simili, almeno in alcuni intervall, e questo determina problemi di dipendenza lineare (collinearità).

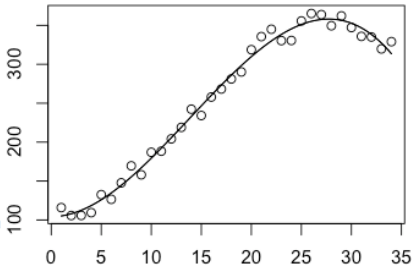
Esempio

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?



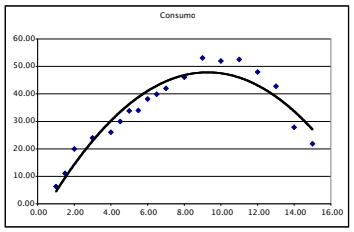

```
Call:
lm(formula = NIPI ~ Trim + I(Trim^2) + I(Trim^3), data

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 può centrare la variabile su cui poi si calcolano le potenze.

Esempio: una curva di domanda

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

Lo scatterplot suggerisce una 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
```

Cosa succede se invece utilizziamo una polinomiale di grado superiore?

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

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 scelti 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 semplici potenze della variabile esplicativa x, ma polinomi separati in x, vincolati ad essere ortogonali.

I vantaggi sono che i parametri di ogni polinomio in ogni potenza si calcolano 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

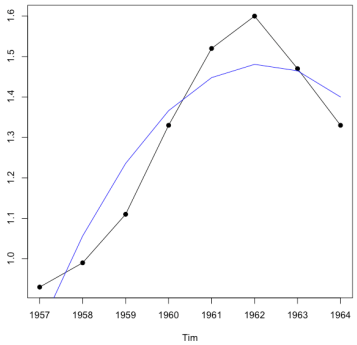
L'ortogonalità dei polinomi implica che

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

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

Variabile dicotomizzata

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

$$y = \beta_0 + \beta_1 X_1 + e\left[X = 0\right]$$
$$y = \beta_0 + \beta_1 X_1 + e\left[X = 1\right]$$

Variabile binaria

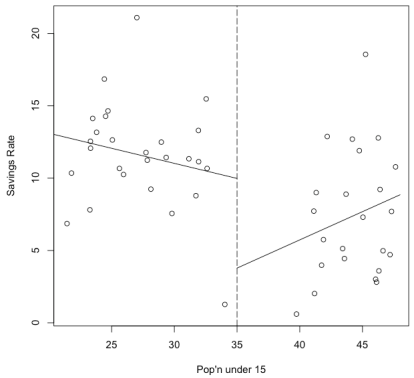
La costante c può rappresentare uno shock ovvero rappresentare la soglia di una variabile binaria.

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 con giunzione

La stima per parti separati manca di continuità nel punto di giunzione. Comporta inoltre la stima di più parametri del necessario.

C'è un'alternativa

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

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

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

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

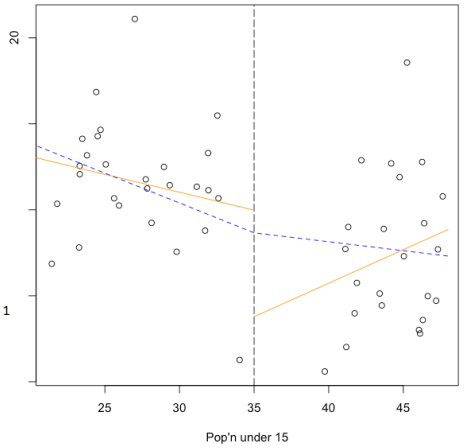
Se X=c+δx, con δx>0 allora $\underset{\delta x \to 0}{Lim}\, y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 = y = \beta_0 + \beta_1 X_1$

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
```



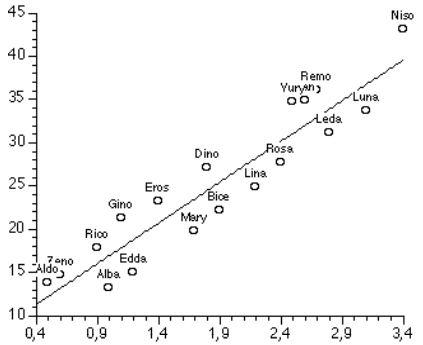
L'uso di variabili qualitative

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



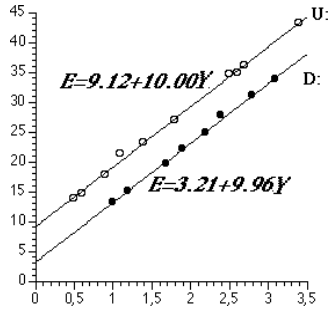
L'uso di variabili qualitative/2

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 binarie 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 binaria 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

$$(X^t X)$$

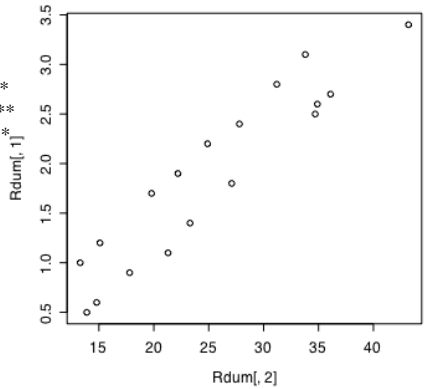
Per superare questo problema è necessario stimare il modello senza b_0

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

Esempio

```
> 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

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

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

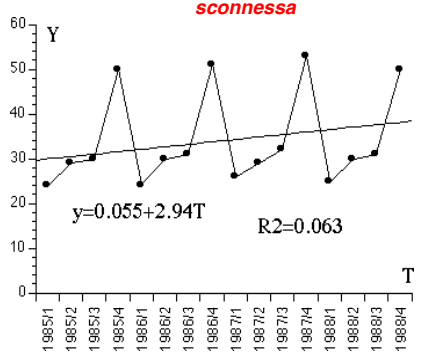
```
Call:
lm(formula = Spesa ~ -1 + Percorrenza + DumReg1 + DumReg2 + DumAge1 +
    DumAge2 + DumAge3)

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
```

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

Binarizzazione di una politoma sconnessa



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

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

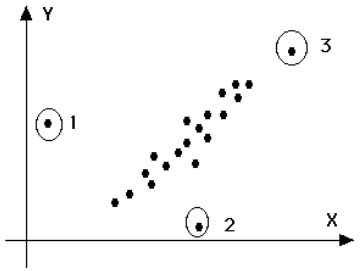
Dependent Var = VENDITE N of Cases = 16
Multiple R = .999761313 F = 4606.9
R-sqr = .999522683 p = 0.000000
Adjusted R-sqr = .999305721 df = 5, 11
Std. error of est. = .936021562
Intercept = forced to 0

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

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

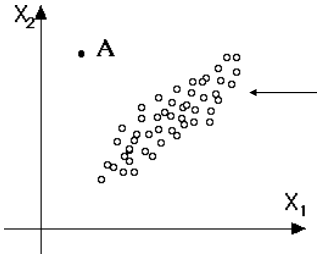
Da notare che il punto "1" è anomalo rispetto alla "X", il "2" rispetto alla "Y" ed il "3" rispetto ad entrambe

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.

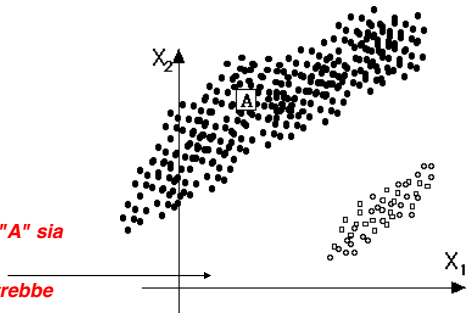
Valori remoti/2

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.

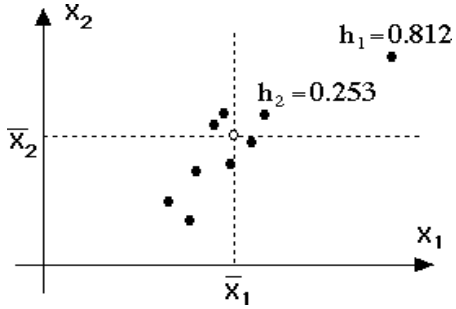
Nel caso della regressione lineare semplice è sufficiente lo studio dello scatterplot. Se i dati sono multidimensionali è necessario ricorrere a speciali formule.

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 verosimilmente eccessiva se $h_i \geq 2\left(\frac{m+1}{n}\right)$

Le indicazioni ottenute con la leva prescindono dai valori osservati sulla dipendente, ma quantificano la "forza di trascinamento" 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

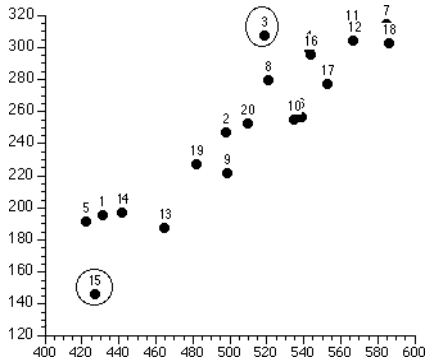
Esempio

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

Residui SD (studentized deleted)

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

A questo dine sono utili i residui ottenuti dopo aver cancellato l'i-esima osservazione.

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 degli SD è effettuato con quantità già ottenute dal modello ordinario di regressione

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

Maggiore è d_i più influente è l'osservazione i-esima 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).

Distanza di Cook

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

Una sintesi di queste variazioni è la DISTANZA DI COOK

$$c_i = \frac{\hat{e}_i^2}{m\hat{\sigma}}\sqrt{\frac{h_i}{(1-h_i)^2}}$$

dipende da due fattori: residuo (quindi dall'adattamento) e leva (dalla collocazione rispetto agli altri punti).

maggiore è il residuo oppure è la leva 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**

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

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

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

Che fare in caso di anomalia?

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

Se A_i è considerato anomalo si può…

- Escluderlo dal data set con guadagno sul fitting del modello. Si compromette però l'indipendenza tra le osservazioni perché ora sono condizionate dall'esclusione
  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.
- Considerare il dato anomalo come se fosse mancante ed applicare un metodo di imputazione. Ma quale?
- Utilizzare un criterio alternativo ai minimi quadrati che sia meno sensibile ai valori remoti.

Rango della matrice dei regressori

Come è noto, la matrice di centramento è simmetrica, idempotente ed ha rango pari a (n-1).

Poiché si suppone che il numero dei casi sia molto maggiore del numero dei regressori (cioè il numero di colonne di X) allora si può ritenere che

$$ran(\mathbf{W}) \le ran(\mathbf{X})$$

Il rango di X è dato dal numero di colonne (regressori) linearmente indipendenti

La X ha RANGO PIENO se ran(X)=m+1.

Se invece una o più colonne possono essere espresse come combinazione lineare di altre colonne, allora la matrice dei regressori è SINGOLARE e la usuale matrice inversa non può essere determinata.

In generale esistono sempre relazioni tra i regressori che comportano un certo grado di dipendenza lineare. Questo fenomeno è detto COLLINEARITA'

Problemi nella matrice dei regressori

L'operatività del modello di regressione è ancorata alla inversa della matrice dei prodotti incrociati

$$\left(\mathbf{X}^t\mathbf{X}\right)$$

Che è la matrice delle varianze-covarianze quando i regressori sono centrati

$$\left[(\mathbf{CX})^t(\mathbf{CX})\right] = (\mathbf{XCCX}) = (\mathbf{XCX}) = \mathbf{W}$$

Il rango di W non può essere maggiore di quello delle matrici componenti

$$ran(\mathbf{W}) \le Min\left\{ran(\mathbf{CX}),\ ran\left(\mathbf{X}^t\mathbf{C}\right)\right\} = ran(\mathbf{CX})$$

Ma anche la matrice centrata nasce da un prodotto di matrici, per cui si ha

$$ran(\mathbf{W}) \le Min\left\{ran(\mathbf{C}),\ ran(\mathbf{X})\right\}$$

Correlazione e collinearità

La perfetta correlazione tra due regressori genera singolarità. Poiché

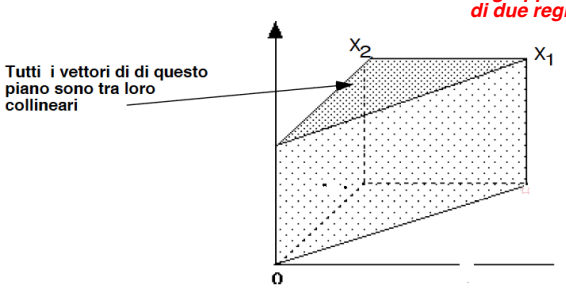
$$\left|r_{ij}\right| = 1 \Rightarrow x_i = a + bx_j \quad \text{ovvero} \quad x_j = c + dx_i$$

abbiamo

$$\lambda_1 x_i + \lambda_2 x_j = 0 \quad \text{con} \quad \lambda_1 \ \text{o} \ \lambda_2 \neq 0$$

per cui una colonna è LID

La collinearità tuttavia è un fenomeno di gruppo che in genere riguarda più di due regressori



Esempio

L'analisi della matrice di correlazione può segnalare le singolarità o le quasi singolarità (correlazioni superiori a 0.9).

Correlation Analysis

Pearson Corr Coeff /Prob>|r| under $H_0: \rho = 0$

	Y	X1	X2
Y	1.00000	0.90932	0.93117
	0.0	0.0120	0.0069
X1	0.90932	1.00000	0.74118
	0.0120	0.0	0.0918
X2	0.93117	0.74118	1.00000
	0.0069	0.0918	0.0

Diagonale unitaria

r_{Y1} r_{Y2} r_{12}

Se una delle entrate è superiore a 0.98 le stime dei parametri negli OLS sono da considerarsui poco affidabili

Correlazione e collinearità/2

L'assenza di correlazioni elevate non esclude la collinearità. Essa nasce da

$$\sum_{k \in K} \lambda_k x_k \cong 0 \quad \text{dove } K = \{\text{insieme di interi tra 1 e m}\}$$

cioè una relazioe di uguaglianza approssimata

può quindi succedere che r_{12}, r_{13}, r_{23} **siano piccoli, ma la regressione**

$$x_1 = \alpha_0 + \alpha_1 x_2 + \alpha_2 x_3$$

possa dare un R^2 **pari ad uno**

Natura del problema

Consideriamo i seguenti dati

Dato	y	x_1	x_2
1	23	2	6
2	83	8	9
3	63	6	8
4	103	10	10

Due persone, separatamente, stimano modello

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

Stima di A: $y_i = -87 + x_{i1} + 18x_{i2}$

Stima di B: $y_i = -7 + 9x_{i1} + 2x_{i2}$

Entrambi i modelli si adattano perfettamente alle osservate.

Come è possibile?

Dato	y_i	Stima-A	Stima-B
1	23	23	23
2	83	83	83
3	63	63	63
4	103	103	103

Natura del problema/2

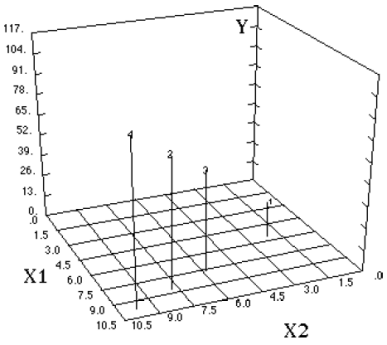
In realtà ci sono infiniti modelli, tutti diversi e tutti perfetti.

La causa è che tra i regressori c'è una relazione lineare esatta

$x_{i2} = 5 + 0.5x_{i1}$ (è ovvio che $r_{12} = 1$)

i punti si allineano ed una delle dimensioni è superflua.

D'altra parte gli scarti tra teoriche ed osservate sono tutti nulli.



Conclusioni provvisorie:

La collinearità tra i regressori è compatibile con valori elevati ;

A causa di essa esistono modelli diversi che danno un buon adattamento;

I parametri non possono essere interpretati come effetto di un regressore TENUTI COSTANTI GLI ALTRI perché variando un regressore, varia anche l'altro.

Cause della collinearità

La collinearità è una sorta di quasi-singolarità della matrice dei regressori.

Quali ne sono le cause?

- **Le tecniche di rilevazione dei dati: ad esempio l'intervallo limitato di alcuni regressori oppure presenza di errori di misurazione simili su regressori diversi.**
- **Correlazione spuria o latente: i regressori pur essendo, in principio, poco legati, risentono di fenomeni esterni che agiscono su entrambi.**
- **Non coerenza dei dati di un regressore con la specificazione del modello (ad esempio quando si usa una polinomiale di grado più elevato del necessario).**
- **Il modello è applicato ad un numero ridotto di casi.**

Effetti della collinearità

La collinearità danneggia la stima dei parametri e la loro precisione.

Ad esempio, nel modello con due regressori si ha

$$\beta_1^* = \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}; \quad s(\beta_1^*) = \frac{\hat{\sigma}}{\sqrt{1 - r_{12}^2}}$$

$$\beta_2^* = \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}; \quad s(\beta_2^*) = \frac{\hat{\sigma}}{\sqrt{1 - r_{12}^2}}$$

al tendere di $|r_{12}| \to 1$ i valori dei parametri e la loro Dev.Std. diventano infinitamente grandi per cui la stima è poco attendibile.

Da notare che, essendo

$$t_1 = \frac{r_{1y} - r_{12}r_{2y}}{\hat{\sigma}\sqrt{1 - r_{12}^2}}; \quad t_2 = \frac{r_{2y} - r_{12}r_{1y}}{\hat{\sigma}\sqrt{1 - r_{12}^2}}$$

La significatività dei parametri tenderebbe all'infinito anche se i regressori fossero irrilevanti per spiegare la dipendente

Regressori non correlati

La produttività di un gruppo di operai è collegata all'ampiezza del gruppo ed all'entità del bonus di produzione

Ciclo i	Numero X_{i1}	Bonus X_{i2}	Produt y_i
1	4	2	42
2	4	2	39
3	4	3	48
4	4	3	51
5	6	2	49
6	6	2	53
7	6	3	61
8	6	3	60

$$R = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; \quad r_{1y} = 0.7419 \quad r_{2y} = 0.6384$$

I due regressori sono ortogonali e la matice di correlazione coincide col la matrice identità

$$\textit{due variabili esplicative}: \quad y_i = 0.375 + 5.375x_{i1} + 9.250x_{i2}$$

$$\textit{variabile esplicativa } 1: \quad y_i = 23.5 + 5.375x_{i1}$$

$$\textit{variabile esplicativa } 2: \quad y_i = 27.5 + 9.250x_{i2}$$

Se i regressori sono incorrelati il loro effetto è lo stesso sia che siano tutti presenti che se presenti separatamente

Regressori correlati

La quota di spesa che le famiglie dedicano all'abbigliamento è legata a quella del vitto e della casa. Ecco i dati medi per regione:

Regione	A_i	C_i	V_i
Piemonte	11.9	19.5	43.1
Val d'Aosta	22.8	24.7	49.8
Lombardia	18.7	30.7	51.9
Trentino	20.1	29.8	54.3
Veneto	12.9	19.1	42.2
Friuli	21.7	25.6	53.9
Liguria	27.1	31.4	58.5
Emilia-R.	25.4	27.9	52.1
Toscana	21.3	22.1	49.9
Umbria	19.3	25.5	53.5
Marche	25.4	31.1	56.6
Lazio	27.2	30.4	56.7
Abruzzi	11.7	18.7	46.5
Molise	17.8	19.7	44.2
Campania	12.8	14.6	42.7
Puglia	23.9	29.5	54.4
Basilicata	22.6	27.7	55.3
Calabria	25.4	30.2	58.6
Sicilia	14.8	22.7	48.2
Sardegna	21.1	25.2	51.0

La spesa per vestiario dipende anche dal reddito, dall'età, dal risparmio. Queste variabili non sono considerate. E' una procedura corretta?

I due regressori: vitto e casa sono correlati

$$r_{cv} = 0.92$$

Quale effetto si ha sulla stima del modello?

Regressori correlati/2

Valutiamo la stima dei modelli con i due regressori e separatamente

$$\text{Regressione di y su } x_1 \text{ e } x_2 : A_i = -19.174 + 0.222C_i + 0.659V_i;\quad \hat{\sigma}=6.47;\quad R^2=0.778$$

$$\text{Regressione di y su } x_1 : A_i = -1.496 + 0.857C_i;\quad \hat{\sigma}=7.95;\quad R^2=0.711$$

$$\text{Regressione di y su } x_2 : A_i = -23.634 + 0.857V_i;\quad \hat{\sigma}=6.30;\quad R^2=0.771$$

Da notare che, a causa della intercorrelazione tra i regressori

- **Il coefficiente di C_i cambia se si impiegano entrambi i regressori. Il suo effetto è perciò legato alla presenza o meno di V_i**
- **Lo stesso succede per il coefficiente di V_i**

In presenza di collinearità non si può riscontrare un aumento della varianza spiegata sicuramente attribuibile ad uno specifico regressore.

i coefficienti non riflettono il legame GLOBALE tra il regressore e la Y, ma quello PARZIALE tenuto conto di quanti e quali altri regressori sono inclusi nel modello

Esercizio

Studiamo le seguenti variabili

$y = Ln(\text{indice prezzi al consumo})$
$x_1 = Ln(\text{indice a prezzi al correnti})$
$x_2 = Ln(\text{PIL a prezzi costanti})$
$x_3 = x_1 - x_2 = Ln(\text{deflattore PIL})$

Anno	IPC	Pil-Cor	Pil-Cos	Defl.
1950	4.37	6.28	5.66	-0.62
1960	4.52	6.60	6.23	-0.37
1970	4.73	6.98	6.89	-0.09
1980	5.54	7.30	7.87	0.57
1982	5.58	7.30	8.03	0.73
1983	5.60	7.33	8.10	0.77
1984	5.54	7.43	8.24	0.81

coinvolgiamole nei modelli

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2};\quad y_i = \alpha_1 x_{i1} + \alpha_2 x_{i3}$$

nessuno dei regressori ha influenza sulla dipendente (livello di inflazione). L'unica sarebbe la terza, che però nel primo è assente.

$$y = \underset{(2.997)}{0.273}x_1 + \underset{(4.972)}{0.442}x_2;\quad R^2=0.9994;\ \hat{\sigma}=0.127$$

$$y = \underset{(92.356)}{0.715}x_1 + \underset{(4.972)}{0.442}x_3;\quad R^2=0.9994;\ \hat{\sigma}=0.127$$

t-Student

a) perchè aumenta il t-student della X_1?

b) Perchè cambia solo la precisione della X_1?

c) Cosa c'è di diverso nei due modelli?

Misura del condizionamento/2

Un'altra misura è il *condition number*

$$C(X) = \frac{\lambda_{max}\left[(X'X)\right]}{\lambda_{min}\left[(X'X)\right]}$$

Un valore prossimo ad uno indica una matrice di regressori con equivariabilità verso ogni direzione. Valori elevati indicano collinearità

L'ordine di grandezza di C(X) esprime il numero di cifre che si degradano nel calcolo della matrice inversa.

Se X è misurata con 7 cifre significative e C(X)=10'000 allora la stima dei parametri sarà accurata fino alla terza cifra decimale (7-4=3).

Valori di C(X)<1'000 sono piuttosto tranquilli. Preoccupano valori C(X)>100'000

Relazioni tra i regressori

Per individuare le quasi-singolarità dovremmo esplorare tutti i submodelli definibili tra i regressori. Il loro numero è

$$m * \frac{(m-1)^2}{2}$$

per cui se m=2 si studia solo un submodello; ma se m=6 occorre studiarne 75 e questo non sempre è possibile.

In genere ci si limita a quelli che pongono ciascun regressore in relazione con tutti gli altri badando in particolare a valore di R^2 in

$$X_i = \sum_{j \neq i} \beta_j X_j$$

Non è però necessario effettuare materialmente i calcoli. Gli R^2 parziali sono infatti ottenibili dai calcoli già effettuati per la stima ordinaria

Variance inflation factors

La Dev.Std delle stime nel modello con regressori standardardizzati è

$$s\left(\beta_i^*\right) = \hat{\sigma}\sqrt{v_{ii}^*}$$

dove v_{ii}^* è l'elemento sulla diagonale dell'inversa della matrice di correlazione

Questi elementi sono noti come VIF *(Variance Inflation Factors)* ed indicano quanto la variabilità del parametro dipenda dai legami tra tutti i regressori

Si dimostra infatti che $v_{ii}^* = \dfrac{1}{\left(1 - R_i^2\right)}$

dove R_i^2 è il coefficiente di determinazione multipla del modello in cui x_i è come dipnendente rispetto agli altri regressori.

Dato che R_i^2 varia tra zero ed uno, i VIF hanno un campo di variazione che parte da uno e va all'infinito

Significato dei VIF

Se il regressore i-esimo non è legato linearmente agli altri si avrà $R_i^2 = 0$ ed il suo VIF sarà pari ad uno.

Man mano che R_i^2 aumenta anche il VIF cresce e tende ad infinito al tendere di R_i^2 ad uno

Ricordiamo che $$Var\left(\hat{\beta}_j\right) = VIF_j \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}$$

Il secondo fattore è la varianza della stima del parametro nella regressione semplice usando come regressore solo x_i

Il VIF dell'i-esimo regressore esprime l'incremento di variabilità generatosi nel modello aggiungendo dei regressori diversi dalla x_i

Altre misure di collinearità

Il maggiore tra gli "m" VIF (uno per ogni regressore) è un indicatore di collinearità.

Empiricamente si ritiene che

$$\max_{1 \le i \le m}\left\{VIF_i\right\} \ge 10$$

sia una soglia che indichi livelli pericolosi di collinearità nella X.

Un'altro indicatore, ma che aggiunge poco al precedente, è la media dei VIF

$$\overline{VIF}_i = \frac{\sum_{i=1}^{m} VIF_i}{m}$$

Esempio

Nella tabella ci sono dei risultati relativi alla stima dei dati regionali già visti

Regressore	Parametri β_i^*	Determin. R_i^2	Tolleranza $1 - R_i^2$	VIF $\dfrac{1}{1 - R_i^2}$
x1	4.264	0.9986	0.0014	708.84
x2	-1.562	0.9904	0.0096	104.61
x3	-2.983	0.9982	0.0012	564.34

Tutti i VIF sono in questo caso molto elevati. I dati sono poco coerenti con il modello, ovvero la specificazione di quest'ultimo non è corretta.

La tolleranza è talvolta adoperata per indicare se includere o meno il regressore (dovrebbe essere superiore a 0.01).

Da notare che $$\text{Tolleranza}_i = \frac{1}{VIF_i}$$

Rimedi per la collinearità

Per i modelli polinomiali

Tali modelli comportano spesso gradi elevati dei polinomi e quindi collinearità dei regressori: questa aumenta con il crescere del grado e con il ridursi della variabilità relativa delle X.

in questi casi Il rimedio più efficace è centrare i regressori.

Lo stesso vale nel modello con le dummies: centrare elimina la colonna di "uno" associata all'intercetta che è una delle fonti della collinearità

Con i dati sotto controllo

Si possono aggiungere nuove rilevazioni che spezzino la struttura di collinearità.

Si possono stimare i parametri di una parte dei regressori utilizzando un dataset diverso (combinando dati cross-section con dati time-series).

Si possono eliminare alcune variabili.

Esercizio

Si è condotto uno studio su alcuni centri di assistenza (anziani, disabili, etc.) privati al fine da stabilire il contributo che il servizio sanitario deve mediamente pagare per paziente

Le variabili rilevate sono

- y = contributo richiesto
- x_1 = % di letti con assistenza 24/24
- x_2 = numero di posti letto
- x_3 = unità di personale infermieristico

y_i	x_{i1}	x_{i2}	x_{i3}
56.00	67	252	48
59.75	14	213	32
64.00	23	223	32
61.00	8	168	21
67.00	18	165	25
63.00	50	200	37
63.75	48	252	42
64.00	45	195	21
55.00	57	169	18
70.00	13	150	18

Ecco alcuni risultati

Param.	Valore	Precis.	R^2
β_1	-0.133	0.080	0.274
β_2	0.004	0.096	0.824
β_3	-0.027	0.326	0.819

a) Calcolare i t-student

b) calcolare gli indici di presenza

c) Calcolare i VIF