

MODELLO DI REGRESSIONE LINEARE MULTIPLA

E' la generalizzazione del modello di regressione lineare semplice: per spiegare il fenomeno d'interesse Y vengono introdotte k , con $k > 1$, variabili esplicative.

Tale generalizzazione diventa molto più semplice utilizzando l'algebra delle matrici. Il modello di regressione multipla genera però nuovi problemi:

- 1 scelta delle variabili,
- 2 multicollinearità,
- 3 test multipli.

1. IL PROBLEMA

- Ricerca di un modello matematico in grado di esprimere la relazione esistente tra una variabile di risposta y (quantitativa) e (ad esempio) k variabili esplicative
- Si tratta di una relazione asimmetrica del tipo

$$y = f(x_1 \dots x_k)$$

Nel caso del modello di regr.lineare multipla

abbiamo che: $f(x_1 \dots x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \dots \beta_k x_k$

che geometricamente corrisponde ad un iperpiano a k dimensioni

- Perché si studia tale modello
 - i) facilità con cui può essere interpretato un iperpiano a k dimensioni
 - ii) Facilità di stima dei parametri incogniti β_j ($j = 1 \dots k$)

Anche in tal caso nella realtà studiamo un modello del tipo

$$y = f(x_1 \dots x_k) + \varepsilon$$



Componente
sistematica

componente
casuale

2. IL MODELLO

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots \beta_k x_{ik} + \varepsilon_i$$

In forma matriciale

$$y = X \beta + \varepsilon$$

dove

y : vettore ($n \times 1$) di osservazioni sulla
variabile dipendente

X : matrice ($n \times (k+1)$) di osservazioni su
 k regressori

β : vettore ($(k+1) \times 1$) di parametri incogniti

ε : vettore ($n \times 1$) di disturbi stocastici

Le matrici e i vettori sono così definiti

$$\underset{(n \times 1)}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad \underset{(n \times (k+1))}{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdot & \cdot & \cdot & x_{1k} \\ 1 & x_{21} & x_{22} & \cdot & \cdot & \cdot & x_{2k} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n1} & x_{n2} & \cdot & \cdot & \cdot & x_{nk} \end{pmatrix}$$

$$\underset{((k+1) \times 1)}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_k \end{pmatrix} \quad \underset{(n \times 1)}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{pmatrix}$$

OSSERVAZIONE

La matrice X ha la prima colonna unitaria: per il caso in cui si consideri un modello con intercetta β_0 si introduce una variabile di comodo $X_0 = 1$ per ogni $i=1,2,\dots,n$.

3. LE ASSUNZIONI DEL MODELLO

- 1) Esiste legame lineare tra variabile dipendente e regressori
- 2) Le variabili sono tutte osservabili
- 3) I coefficienti β_i non sono v.c.
- 4) I regressori X sono non stocastici
- 5) Il termine ε non è osservabile
- 6) $E(\varepsilon_i) = 0$
- 7) $Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & \text{per } i \neq j \\ \sigma^2 & \text{per } i = j \end{cases}$

$\Rightarrow$ le ε_i sono omoschedastiche ed incorrelate

$$E[\varepsilon\varepsilon'] = \begin{pmatrix} \sigma^2 & 0 & 0 & . & . & 0 \\ 0 & \sigma^2 & 0 & . & . & 0 \\ . & . & . & . & . & . \\ 0 & 0 & . & . & . & \sigma^2 \end{pmatrix}$$

- 8) X ha rango pieno $\text{rank}(X) = k$
condizione necessaria $n \geq k + 1$
- 9) $\varepsilon \approx N(0, \sigma^2 I)$: ipotesi aggiuntiva da utilizzare nell'analisi inferenziale

OSSERVAZIONE

Anche nel caso multiplo essendo le Y_i funzioni lineari delle v.c. ε_i anch'esse sono variabili casuali e :

$$\begin{aligned} E(Y_i | X_1 = x_{i1}, X_2 = x_{i2}, \dots, X_k = x_{ik}) = \\ = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots \beta_k x_{ik} \end{aligned}$$

Che rappresenta l'equazione di un iperpiano in uno spazio a $k+1$ dimensioni (superficie di regressione). Per le assunzioni fatte sul disturbo discende che:

$$E(Y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots \beta_k x_{ik}$$

e

$$V(Y_i) = V(\varepsilon_i) = \sigma^2$$

4. STIMATORE MINIMI QUADRATI (OLS)

$$y = X\beta + \varepsilon$$

Si cercherà quel vettore $\hat{\beta}$ che minimizza gli scarti al quadrato:

$$\min \sum_{i=1}^n (y_i - X_i \beta)^2$$

dove X_i è la riga i-esima di X

In forma matriciale

$$\varepsilon = (y - X\beta)$$

$$\min \varepsilon' \varepsilon \quad \text{o} \quad \min (y - X\beta)' (y - X\beta)$$

$$\begin{aligned} G(\beta) &= \varepsilon' \varepsilon = (y - X\beta)' (y - X\beta) \\ &= (y' - \beta' X') (y - X\beta) \\ &= y'y - \beta' X' y - \underbrace{y' X \beta}_{= \text{perché scalare}} + \beta' X' X \beta \end{aligned}$$

$$\frac{\partial G}{\partial \beta} = -2X'y + 2X'X\beta = 0 \quad (1)$$

perché

$$\beta' X' y = (\beta_0, \beta_1 \dots \beta_k) \begin{matrix} (1 \times (k+1)) & (k+1 \times n) & (n \times 1) \\ \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ x_{21} & x_{22} & \dots & \dots & x_{2n} \\ x_{31} & x_{32} & \dots & \dots & x_{3n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{k1} & \dots & \dots & \dots & x_{kn} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \end{matrix}$$

è uno scalare

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

dalla (1) si ottiene

$$X' X \beta = X' y$$

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

pre-moltiplicando ambo i membri

$$(X' X)^{-1} (X' X) \beta = (X' X)^{-1} X' y$$

perché **rank (X'X) = rank (X) = k**

$X'X$ è a rango pieno ovvero invertibile

$$\hat{\beta} = (X' X)^{-1} X' y$$

stimatore OLS di β 8

CARATTERISTICHE STIMATORE OLS

Teorema di Gauss-Markov

$\hat{\beta}$ è uno stimatore di tipo BLUE

Best Linear Unbiased Estimator

ovvero ha varianza minima nella classe degli stimatori Lineari e Corretti

1. $\hat{\beta} = (X'X)^{-1} X'y$

La matrice $(X'X)^{-1} X'$ è formata da elementi costanti per cui $\hat{\beta}$ è una trasformazione lineare di y .

2.
$$\begin{aligned}\hat{\beta} &= (X'X)^{-1} X'y = (X'X)^{-1} X'(\underbrace{X\beta + \varepsilon}) \\ &= (X'X)^{-1} X'X\beta + (X'X)^{-1} X'\varepsilon \\ &= \beta + (X'X)^{-1} X'\varepsilon\end{aligned}$$

$$E(\hat{\beta}) = \beta + (X'X)^{-1} X' E(\varepsilon) = \beta$$

È uno stimatore corretto

Inoltre: $(\hat{\beta} - \beta) = (X'X)^{-1} X'\varepsilon$

$$\begin{aligned}
3. \text{Var}(\hat{\beta}) &= E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] \\
&= E\left[(X'X)^{-1} X' \varepsilon \varepsilon' X (X'X)^{-1}\right] \\
&= (X'X)^{-1} X' E[\varepsilon \varepsilon'] X (X'X)^{-1} \\
&= (X'X)^{-1} X' \sigma^2 I X (X'X)^{-1} \\
&= \sigma^2 (X'X)^{-1} X'X (X'X)^{-1} = \sigma^2 (X'X)^{-1}
\end{aligned}$$

Si consideri più in dettaglio $E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right]$:

$$\begin{pmatrix} E(\hat{\beta}_1 - \beta_1)^2 & E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) & \dots & E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_k - \beta_k) \\ E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) & E(\hat{\beta}_2 - \beta_2)^2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ E(\hat{\beta}_k - \beta_k)(\hat{\beta}_1 - \beta_1) & \dots & \dots & E(\hat{\beta}_k - \beta_k)^2 \end{pmatrix}$$

Pertanto la varianza $E(\hat{\beta}_j - \beta_j)^2$ di ogni parametro $\hat{\beta}_j$ si desume prendendo il corrispondente valore sulla diagonale principale della $(X'X)^{-1}$, moltiplicato per σ^2 :

$$\text{Var}(\hat{\beta}_j) = [(X'X)^{-1}]_{jj} \sigma^2$$

Definiamo uno stimatore alternativo lineare e corretto

$$\beta^* = \hat{\beta} + C' y$$

dove C è una matrice ($n \times k$)

$$\beta^* = (X'X)^{-1} X' y + C' y$$

$$E(\beta^*) = (X'X)^{-1} X' (X\beta + \varepsilon) + C'X\beta + C'\varepsilon = \beta + C'X\beta \Rightarrow C'X = 0$$

$$V[\beta^*] = E[(\beta^* - \beta)(\beta^* - \beta)'] = E\left[\left\{(X'X)^{-1} X' + C'\right\} \varepsilon \varepsilon' \left\{(X'X)^{-1} X' + C'\right\}'\right]$$

$$= \sigma^2 \left\{ (X'X)^{-1} X'X (X'X)^{-1} + C'X (X'X)^{-1} + (X'X)^{-1} X'C + C'C \right\}$$

ma $C'X = 0 = X'C$

$$= \sigma^2 (X'X)^{-1} + \sigma^2 (C'C)$$

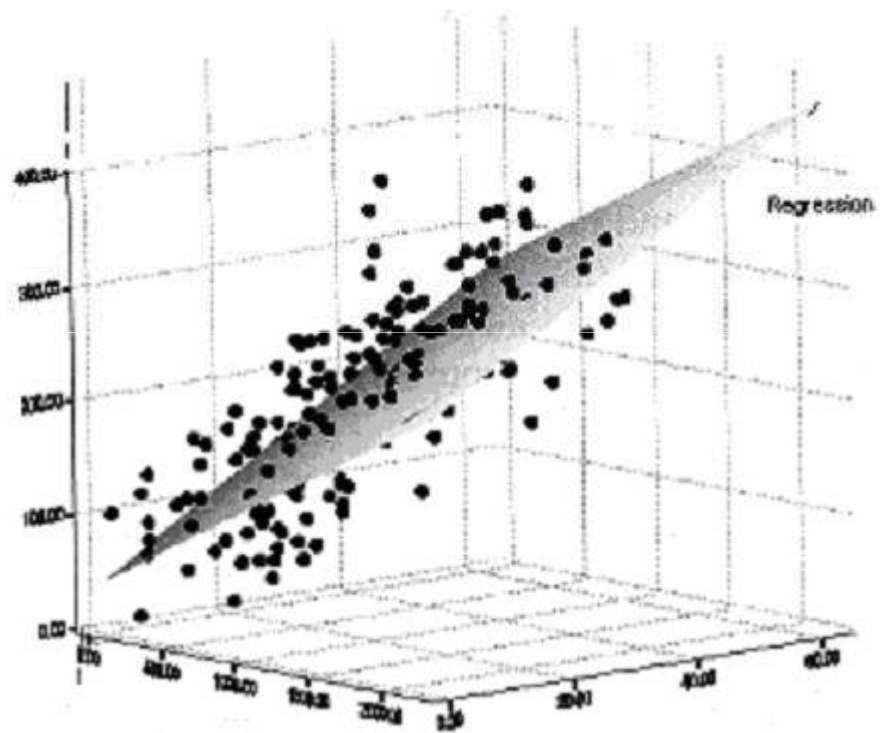
$$= \text{Var}(\hat{\beta}) + \sigma^2 (C'C) \geq \text{Var}(\hat{\beta})$$

Pertanto la $\text{Var}(\hat{\beta})$ è la minima nella classe degli stimatori lineari e corretti, e risulta provato il teorema di Gauss-Markov

Geometricamente l'equazione (stimata con gli OLS)

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \dots + \hat{\beta}_k x_{ik}, \quad i=1,2,\dots,n$$

definisce l'iperpiano nello spazio a $k + 1$ dimensioni, che tra gli infiniti piani, rende minima la somma dei quadrati delle lunghezze dei segmenti congiungenti i punti osservati al piano stesso (v. figura).



Stima del parametro σ^2

Consideriamo innanzitutto l'identità

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

che rappresenta la scomposizione della devianza totale in devianza residua e devianza spiegata (o della regressione), ovvero

$$\text{SQT} = \text{SQE} + \text{SQR}.$$

La stima della varianza delle v.c. errori è data da

$$s^2 = \frac{\tilde{e}\hat{e}}{n-p-1} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-p-1}$$

Che è uno stimatore corretto

ESEMPIO

$$G_i = \beta_0 + \beta_1 P_{gi} + \beta_2 y_i + \beta_3 P_{qi} + \varepsilon_i$$

$$i : 1960 \dots 1986, \quad n = 27$$

$G_i =$ consumo di benzina in \$

$P_{gi} =$ indice dei prezzi benzina

$Y_i =$ reddito pro-capite in \$

$P_{qi} =$ indice dei prezzi auto nuove

Vettore y	x_0	x_2	x_3	x_4
121.01034	1	0.9250000	6036.0000	1.0450000
130.20306	1	0.9140000	6113.0000	1.0450000
136.62968	1	0.9190000	6271.0000	1.0410000
134.39852	1	0.9180000	6378.0000	1.0350000
150.34150	1	0.9140000	6727.0000	1.0320000
171.88391	1	0.9490000	7027.0000	1.0090000
175.44395	1	0.9700000	7280.0000	0.9910000
172.03874	1	1.0000000	7513.0000	1.0000000
198.65222	1	1.0470000	7891.0000	1.0440000
208.37573	1	1.0560000	8134.0000	1.0760000
214.38531	1	1.0630000	8322.0000	1.1200000
228.52113	1	1.0760000	8562.0000	1.1100000
237.37202	1	1.1810000	9042.0000	1.1110000
234.34193	1	1.5990000	8867.0000	1.1750000
222.32567	1	1.7080000	8944.0000	1.2760000
228.16247	1	1.7790000	9175.0000	1.3570000
242.33362	1	1.8820000	9381.0000	1.4290000
248.32557	1	1.9630000	9735.0000	1.5380000
240.93266	1	2.6560000	9829.0000	1.6600000
229.58893	1	3.6910000	9722.0000	1.7930000
227.13648	1	4.1090000	9769.0000	1.9020000
210.44373	1	3.8940000	9725.0000	1.9760000
236.85998	1	3.7640000	9930.0000	2.0260000
255.36365	1	3.7070000	10421.000	2.0850000
243.75057	1	3.7380000	10563.000	2.1520000
277.31965	1	2.9210000	10780.000	2.2400000

Matrice $X'X$;

27.000000	51.357000	229865.00	37.296000
51.357000	133.15081	473127.10	83.319118
229865.00	473127.10	2.0120502e+09	331319.22
37.296000	83.319118	331319.22	56.280428

Matrice inv ($X'X$);

2.6605735	0.51586178	-0.00029970528	-0.76246362
0.51586178	0.30384762	-6.4047001e-07	-0.78790617
-0.00029970528	-6.4047001e-07	6.6199636e-08	-0.00019015563
-0.76246362	-0.78790617	-0.00019015563	2.8089108

Stime $\beta = \text{inv}(X'X) * X'y$;

-89.761482
-12.588147
0.039938109
-14.443884

RICAPITOLANDO

$$\hat{\beta} = (X'X)^{-1} X' y$$

$$E[\hat{\beta}] = \beta$$

$$V[\hat{\beta}] = E\left(\left(\hat{\beta} - \beta\right) \left(\hat{\beta} - \beta\right)'\right) = (X'X)^{-1} \sigma^2$$

$$s^2 = \hat{\sigma}^2 = \frac{\sum e_i^2}{n-k-1}$$

$$E[\hat{\sigma}^2] = \sigma^2$$

Fino ad ora nessuna ipotesi è stata posta per la distribuzione degli errori nel problema della stima.

Aggiungiamo :

$$\varepsilon_i \approx N(0, \sigma^2)$$

$$\varepsilon \approx N(0, \sigma^2 I)$$

TEST PER LA VERIFICA DI IPOTESI SU SINGOLI PARAMETRI

Dal teorema di GAUSS-MARKOV :

$$\hat{\beta} \approx N\left(\beta, \sigma^2 (X'X)^{-1}\right)$$

Vogliamo testare

$$H_0 : \beta_i = 0, \quad H_1 : \beta_i \neq 0$$

Ovvero vogliamo verificare se il regressore X_i è effettivamente sulla variabile dipendente Y .

Nel caso (improbabile) che sia nota σ^2 la statistica test è:

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\sigma^2 [(X'X)^{-1}]_{ii}}}$$

Sotto $H_0 : \beta_i = 0$ si distribuisce come una normale standardizzata

Se il valore cade all'esterno dell'intervallo di confidenza, per esempio al 95%, della $N(0,1)$ rifiutiamo H_0 ed il parametro β_i sarà “significativamente” diverso da zero; altrimenti non rifiutiamo H_0 e concludiamo che il parametro β_i non sarà “significativo”

In generale per un sistema di ipotesi $H_0 : \beta_i = c$ contro $H_0 : \beta_i \neq c$ rifiuto, al livello $100 \alpha \%$ di significatività, quando

$$\left| \frac{\hat{\beta}_i - c}{\sqrt{\sigma^2 [(X'X)^{-1}]_{ii}}} \right| > z_{\alpha/2}$$

QUANDO σ^2 NON E' NOTA

- Utilizziamo la sua stima $\hat{\sigma}^2 = s^2$

$$a_{ii} = \left[(X'X)^{-1} \right]_{ii} \quad \hat{\sigma}^2 = \frac{e'e}{(n-k-1)}$$

- In questo caso la statistica test è

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma} \sqrt{a_{ii}}} \approx t_{n-k-1}$$

dove $a_{ii} = \left[(X'X)^{-1} \right]_{ii}$

è l'elemento generico di posto ii nella diagonale della $(X'X)$

Le ipotesi su β_i possono essere verificate sostituendo i valori nella **statistica test** e controllando poi che la statistica superi o meno i valori della regione critica della distribuzione t_{n-k-1} .

Quindi per verificare la significatività di β_i procederò nel seguente modo:

$$H_0 : \beta_i = 0 \text{ contro } H_1 : \beta_i \neq 0$$

Statistica test:

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma} \sqrt{a_{ii}}} = \frac{\hat{\beta}_i - \beta_i}{\text{s.e.}(\hat{\beta}_i)}$$

Che sotto H_0 si distribuisce come una $t(n-k)$.
Pertanto fissato α se il valore della statistica test cade all'esterno dell'intervallo di confidenza

$$\hat{\beta}_i - t_{\frac{\alpha}{2}} \text{s.e.}(\hat{\beta}_i), \hat{\beta}_i + t_{\frac{\alpha}{2}} \text{s.e.}(\hat{\beta}_i)$$

Rifiuto H_0 di non significatività del parametro, altrimenti non rifiuto H_0 e concludo che il parametro non è significativo.

5. ADATTAMENTO DEL MODELLO

Come nel caso del modello di regressione semplice, il coefficiente di determinazione rappresenta la proporzione di variabilità totale spiegata dal modello, ovvero una misura dell'adattabilità del modello ai dati osservati.

La formula per esprimere il coefficiente è analoga a quella della regressione semplice, solo che in questo caso per variabilità spiegata dal modello si intende la variabilità spiegata dall'insieme dei regressori

$$R^2 = \frac{SQR}{SQT} = \frac{SQT - SQE}{SQT} = 1 - \frac{SQE}{SQT}$$

Alternativamente si può scrivere:

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

$$0 \leq R^2 \leq 1$$

Il coefficiente di determinazione è un indicatore del legame lineare tra Y e i regressori.

- Ha però un difetto:
può aumentare anche se viene aggiunto un regressore anche se non “spiega” y .

$$R^2 = 1 - \frac{SQE}{SQT} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2}$$

- Se dividiamo le devianze per i gradi di libertà andiamo a pesare il contributo a R^2 di ogni regressore

$$\hat{R}^2 = 1 - \frac{\sum e_i^2 / (n - k - 1)}{\sum (Y_i - \bar{Y})^2 / (n - 1)}$$

$$\hat{R}^2 = 1 - \frac{(n - 1)}{(n - k - 1)} (1 - R^2) = 1 - \frac{SQE / (n - k - 1)}{SQT / (n - 1)}$$

Possiamo anche nel caso multiplo costruire la
TABELLA ANOVA

Fonte di variazione	Devianza	g.d.l	Medie dei Quadrati (Stime var.)
Regressione	SQR	k	$SQR/(k)$
Residuo	SQE	n-k-1	$SQE/(n-k-1)$
Totale	SQT	n-1	$SQT/(n-1)$

Nota: direttamente dalla tabella ANOVA si può costruire il coefficiente di determinazione.

Inoltre la tavola ANOVA ed il Test F consentono di condurre test globale su tutti i parametri del modello ma anche per verificare ipotesi riguardanti insiemi di più coefficienti, permettendo così di selezionare il modello di regressione.

Per valutare la significatività del modello si ricorre al test F: tramite la ANOVA per un modello di regressione lineare multipla, consiste in un test globale su tutti i parametri del modello (eccetto β_0) e in particolare nel confronto tra la devianza del modello saturo (o non vincolato) $Y = X\beta + \varepsilon$ e quella del modello vincolato (o ridotto) $Y = \beta_0 1_n + \varepsilon$. Le ipotesi saranno

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

H_1 : almeno uno dei $\beta_i \neq 0$

Si costruisce la statistica test F

$$F = \frac{SQR/(k)}{SQE/(n-k-1)}$$

Che sotto l'ipotesi nulla ha distribuzione $F_{(k, n-k-1)}$

Fissato il livello di significatività α si individua il valore critico della distribuzione

• Se $F > F_{(\alpha; k, n-k-1)}$ si rifiuta H_0

ovvero si accetta la significatività congiunta di tutte le variabili esplicative.

Esempio

$$Y = X\beta + u$$

$$\begin{array}{l} n = 12 \\ k = 2 \end{array} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Facendo riferimento ai valori

$$\bar{Y} = 9 \quad \bar{X}_2 = 2 \quad \bar{X}_3 = 1$$

$$\sum x_2^2 = 10 \quad \sum x_3^2 = 15 \quad \sum y^2 = 200$$

$$\sum x_2 y = 12 \quad \sum x_3 y = 9 \quad \sum x_2 x_3 = -11$$

Determinare il vettore di stime OLS

Se consideriamo il modello in forma di scarti
dalle medie

$$\begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (X'X)^{-1} X'y$$

Dove

$$X = \begin{bmatrix} x_{21} & x_{31} \\ x_{22} & x_{32} \\ \cdot & \cdot \\ \cdot & \cdot \\ x_{2n} & x_{3n} \end{bmatrix}$$

$$x_{2i} = X_{2i} - \bar{X}_2$$

$$x_{3i} = X_{3i} - \bar{X}_3$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3$$

$$(X'X) = \begin{bmatrix} \sum X_{2i}^2 & \sum X_{2i} X_{3i} \\ \sum X_{2i} X_{3i} & \sum X_{3i}^2 \end{bmatrix}$$

$$(X'X)^{-1} = \frac{1}{|X'X|} \begin{pmatrix} (-1)^2 \sum X_{3i}^2 & (-1)^3 \sum X_{2i} X_{3i} \\ (-1)^3 \sum X_{2i} X_{3i} & (-1)^4 \sum X_{2i}^2 \end{pmatrix}$$

$$= \frac{1}{\sum X_{2i}^2 \sum X_{3i}^2 - (\sum X_{2i} X_{3i})^2} \begin{pmatrix} \sum X_{3i}^2 & -\sum X_{2i} X_{3i} \\ -\sum X_{2i} X_{3i} & \sum X_{2i}^2 \end{pmatrix}$$

$$X' y = \begin{pmatrix} \sum X_{2i} Y_i \\ \sum X_{3i} Y_i \end{pmatrix}$$

da cui

$$\begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \frac{1}{\sum X_2^2 \sum X_3^2 - (\sum X_2 X_3)^2} \times \\ \times \begin{pmatrix} \sum X_3^2 \sum X_2 Y - \sum X_2 X_3 \sum X_3 Y \\ \sum X_2^2 \sum X_3 Y - \sum X_2 X_3 \sum X_2 Y \end{pmatrix}$$

$$\hat{\beta}_2 = \frac{15 \times 12 - (-11) \times 9}{10 \times 15 - 121} = \frac{180 + 99}{29} \cong 9.62$$

$$\hat{\beta}_3 = \frac{10 \times 9 - (-11) \times 12}{10 \times 15 - 121} = \frac{90 + 132}{29} \cong 7.65$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}_2 - \hat{\beta}_3 \bar{X}_3 = 9 - 2 \times 9.62 - 7.65 = -17.89$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} -17.89 \\ 9.62 \\ 7.65 \end{pmatrix}$$

Continua ESERCIZIO

$$H_0 : \beta_2 = \beta_3 = 0 \quad (F_{0.01, 2, 9} = 8.02)$$

è anche $F = \frac{R^2/(k)}{(1-R^2)/(n-k-1)} = \frac{SQR/(k)}{SQE/(n-k-1)}$ Ricordiamo:
n = 12
k = 2 con

inoltre poichè risulta $R^2 = \frac{\hat{\beta}' X' y}{y' y}$ intercetta $\Rightarrow$
2 var. esplicative

in forma di scarti

$$(\hat{\beta}_2 \quad \hat{\beta}_3) \times \left(\begin{array}{ccc|c} x_{21} & \cdot & \cdot & x_{2n} \\ x_{31} & \cdot & \cdot & x_{3n} \end{array} \right) \times \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ y_n \end{pmatrix} =$$

$$(\hat{\beta}_2 \quad \hat{\beta}_3) \times \begin{pmatrix} \sum X_2 y \\ \sum X_3 y \end{pmatrix} = \hat{\beta}_2 \sum X_2 y + \hat{\beta}_3 \sum X_3 y$$

$$R^2 = \frac{9.62 \times 12 + 7.65 \times 9}{200} = \frac{184.29}{200} \cong 0.92$$

$$F = \frac{0.92/2}{(1-0.92)/9} = 11.5 \times \frac{9}{2} \cong 51.75 \rightarrow \text{valore empirico di } F$$

Si rifiuta H_0 : F empirico = 51.75 $> F_{0.01, 2, 9} = 8.02$

Se avessimo voluto testare

$$H_0 : \beta_2 = 0$$

Ovvero la significatività di X_2

$$t = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\hat{\sigma}^2 a_{22}}} \quad t_{n-k} \quad (o \quad F(1, n-k)) \quad (t_{99.9} = 2.82)$$

$$a_{22} = \frac{\sum X_3^2}{\sum X_2^2 \sum X_3^2 - (\sum X_2 X_3)^2} = \frac{15}{150 - 121} = \frac{15}{29} \cong 0.51$$

$$\hat{\sigma}^2 = \frac{e'e}{n-k} = \frac{TSS - ESS}{9} = \frac{200 - 184.29}{9} \cong 1.74$$

$$t = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 a_{22}}} = \frac{9.62}{\sqrt{1.74 \times 0.51}} = \frac{9.62}{1.50} \cong 6.4 \rightarrow \text{valore empirico di } t$$

Anche adesso rifiutiamo $H_0 \Rightarrow$ il regressore X_2 è significativo

Il test F per selezionare il modello di regressione: Confronto tra modelli annidati

Il test F viene anche impiegato per verificare l'ipotesi che due o più parametri siano congiuntamente pari a zero.

In particolare si utilizza un test di tipo ANOVA per confrontare un modello vincolato (ridotto) ed **annidato** con in modello non vincolato (saturo):

Si supponga in generale di voler verificare l'ipotesi:

$$H_0 : \beta_{h+1} = \beta_{h+2} = \dots = \beta_k = 0$$

per qualche $h \leq k$ contro l'alternativa

H_1 : almeno una eguaglianza in H_0 non è vera.

L'ip. nulla afferma che almeno $k-h$ variabili non sono utili a spiegare la relazione lineare con la variabile dipendente, ossia sotto l'ipotesi nulla vale il modello ridotto (ed annidato):

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_h x_{ih} + \varepsilon_i$$

in contrapposizione al modello completo (saturo)

con tutte le variabili esplicative:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_h x_{ih} + \beta_{h+1} x_{ih+1} \dots + \beta_k x_{ik} + \varepsilon_i$$

L'adattamento ai dati osservati del modello completo è comunque migliore del modello ridotto.

Considerando l'ANOVA ed indicando con SQE la somma dei quadrati degli errori del modello completo e con SQE_0 la somma dei quadrati degli errori del modello ridotto si ha che: $SQE \leq SQE_0$

➤ Se la differenza $SQE_0 - SQE$ è grande H_0 dovrebbe essere rifiutata poiché le ultime $k-h$ variabili aumentano considerevolmente l'adattamento del modello;

➤ Se la differenza $SQE_0 - SQE$ è piccola H_0 non dovrebbe essere rifiutata poiché le ultime $k-h$ variabili non aumentano considerevolmente l'adattamento del modello

Per stabilire se tale differenza è significativamente grande o piccola si utilizza il seguente test F

$$F = \frac{SQE_0 - SQE/(k-h)}{SQE/(n-k-1)} \approx F_{(k-h, n-k-1)} \quad \text{sotto } H_0$$

Al denominatore troviamo la stima di σ^2 ottenuta dal modello completo: quanto più il valore di F è grande tanto più H_0 sarà rifiutata a favore della ipotesi alternativa.

Stabilito α la regione di rifiuto è data da:

$$\mathbf{R.C. : } F \geq F_{\alpha}$$

Quando $F < F_{\alpha}$ l'ipotesi nulla non può essere rifiutata e quindi le $k-h$ variabili esplicative possono essere escluse dal modello.

PROBLEMI DI PREVISIONE

Si vuole prevedere il valore di Y_{n+1} per un insieme di valori X non osservati come:

$$C = \{1 \quad X_{2,n+1} \quad X_{3,n+1} \quad \dots \quad X_{k,n+1}\}$$

E' possibile fare una previsione puntuale o stimare un intervallo di previsioni.

$$\begin{aligned} Y_{n+1} &= \beta_1 + \beta_2 X_{2,n+1} + \dots + \beta_k X_{k,n+1} + \varepsilon_{n+1} \\ &= \underset{1 \times k}{C'} \underset{k \times 1}{\beta} + \underset{1 \times 1}{\varepsilon_{n+1}} \end{aligned}$$

$$E[Y_{n+1}] = C' \beta$$

Utilizzando le proprietà BLUE di $\hat{\beta}$ avremo il

PREVISORE PUNTUALE

$$\hat{Y}_{n+1} = C' \hat{\beta} \quad \text{sarà BLUFF}$$

Best Linear Unbiased Forecasting Function

Per ottenere un intervallo di previsione
è necessario individuare la distribuzione di

$$E[C'\hat{\beta}] = C'\beta$$

$$\begin{aligned} \text{Var}[C'\hat{\beta}] &= E\left[(C'\hat{\beta} - C'\beta)(C'\hat{\beta} - C'\beta)'\right] \\ &= E\left[C'(\hat{\beta} - \beta)(\hat{\beta} - \beta)'C\right] = C'(X'X)^{-1}\sigma^2C \end{aligned}$$

$$C'\hat{\beta} \approx N(C'\beta, \sigma^2 C'(X'X)^{-1}C)$$

$$\frac{\frac{C'\hat{\beta} - C'\beta}{\cancel{\sigma} \sqrt{C'(X'X)^{-1}C}}}{\sqrt{\frac{e'e}{\cancel{\sigma}^2} / (n-k)}} \approx t_{n-k}$$

Quindi una stima intervallare con un livello
fiduciario del $100(1-\alpha)\%$:

$$C'\hat{\beta} \pm t_{\alpha/2} \sigma \sqrt{C'(X'X)^{-1}C}$$

$$C'\hat{\beta} - t_{\alpha/2} \sigma \sqrt{\phantom{C'(X'X)^{-1}C}} < C\beta < C'\hat{\beta} + t_{\alpha/2} \sigma \sqrt{\phantom{C'(X'X)^{-1}C}}$$

APPLICAZIONE

$$Y = \beta_1 + \beta_2 X + \varepsilon$$

Voglio prevedere Y dato $X = X_0$. Per calcolare l'intervallo devo determinare

$$(X'X) = \begin{bmatrix} n & \sum X \\ \sum X & \sum X^2 \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & X_0 \end{bmatrix}$$

$$C'(X'X)^{-1}C = \frac{\sum X^2 - 2X_0 \sum X + n X_0^2}{n \sum X^2 - (\sum X)^2}$$

Infatti

$$\begin{aligned} & \begin{bmatrix} 1 & X_0 \end{bmatrix} \frac{1}{n \sum X^2 - (\sum X)^2} \begin{bmatrix} \sum X^2 & -\sum X \\ -\sum X & n \end{bmatrix} \begin{pmatrix} 1 \\ X_0 \end{pmatrix} = \\ & = \frac{1}{n \sum X^2 - (\sum X)^2} [\sum X^2 - X_0 \sum X, X_0 n - \sum X] \begin{pmatrix} 1 \\ X_0 \end{pmatrix} = \\ & = \frac{(\sum X^2 - X_0 \sum X) + (X_0^2 n - X_0 \sum X)}{n \sum X^2 - (\sum X)^2} = \end{aligned}$$

$$= \frac{\sum X^2 - 2X_0 \sum X + nX_0^2}{n \sum X^2 - (\sum X)^2} = \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum X^2}$$

L'intervallo fiduciario sarà

$$C'\beta \pm t_{\varepsilon/2} \sigma^2 \sqrt{C'(X'X)^{-1}C} \Rightarrow$$

$$\Rightarrow (\hat{\beta}_1 - \hat{\beta}_2 X_0) \pm t_{\varepsilon/2} \sigma^2 \sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum X^2}}$$

A parità di dati osservati l'intervallo sarà tanto più largo quanto più X_0 è distante dalla media di X

CENNI SULLE VARIABILI DUMMY (Variabili di comodo)

Fino ad ora abbiamo assunto che nella equazione generale $Y = X\beta + \varepsilon$ le variabili X siano variabili quantitative. E' possibile introdurre variabili cosiddette "di comodo" che riescano a rappresentare diversi fattori :

- EFFETTI TEMPORALI
- EFFETTI SPAZIALI
- VARIABILI QUALITATIVE

E' il caso in cui un modello economico possa subire mutamenti strutturali .

ESEMPIO

FUNZIONE DI CONSUMO

$$C = \alpha_1 + \beta Y + \varepsilon$$

Tempo di guerra

$$C = \alpha_2 + \beta Y + \varepsilon$$

Tempo di pace

Ipotizzando che la propensione marginale al consumo $\frac{\partial C}{\partial Y} = \beta$ rimanga invariata in entrambi i periodi

invece di considerare i due modelli separatamente (stime meno precise) questi possono essere uniti in una sola relazione

$$C = \alpha_1 X_1 + \alpha_2 X_2 + \beta Y + \varepsilon$$

Dove X_1 e X_2 sono variabili dummy :

$$X_1 = \begin{cases} 1 & \text{anni di guerra} \\ 0 & \text{anni di pace} \end{cases}$$

$$X_2 = \begin{cases} 0 & \text{anni di guerra} \\ 1 & \text{anni di pace} \end{cases}$$

La matrice β dei coefficienti sarà $\beta = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{pmatrix}$

e la matrice dei dati

$$X = (X_1 \quad X_2 \quad Y) = \begin{bmatrix} 0 & 1 & Y_1 \\ 0 & 1 & Y_2 \\ \cdot & \cdot & Y_3 \\ 0 & 1 & \cdot \\ 1 & 0 & \cdot \\ 1 & 0 & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 0 & \cdot \\ 0 & 1 & \cdot \\ \cdot & \cdot & \cdot \\ 0 & 1 & Y_n \end{bmatrix}$$

ATTENZIONE

Quando si utilizzano le variabili dummy è necessario fare attenzione a come viene costruito il modello, per non rendere la matrice $(X'X)$ singolare.

Infatti se nel modello precedente lasciavamo una intercetta :

$$C = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \beta Y + \varepsilon$$

$$X = \begin{bmatrix} 1 & 0 & 1 & Y_1 \\ 1 & 0 & 1 & Y_2 \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 1 & \cdot \\ 1 & 1 & 0 & \cdot \\ 1 & 1 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 1 & 0 & \cdot \\ 1 & 0 & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ 1 & 0 & 1 & Y_n \end{bmatrix}$$

$$1 \times X_0 - 1 \times X_1 - 1 \times X_2 + 0 \times Y = 0$$

Abbiamo che le 4 colonne di X sono linearmente dipendenti $rank(X) = rank(X'X) = 3 \neq k$
 $(X'X)$ non è invertibile

Volendo utilizzare una regressione con intercetta si utilizzerà così solo una dummy :

$$C = \gamma_1 + \gamma_2 X_2 + \beta Y + \varepsilon$$

$$X_2 = \begin{cases} 0 & \text{anni di guerra} \\ 1 & \text{anni di pace} \end{cases}$$

β = PMC in entrambi i periodi

$\alpha_1 = \gamma_1$ = intercetta anni di guerra

$\alpha_2 = \gamma_1 + \gamma_2$ = intercetta anni di pace

$\alpha_1 - \alpha_2 = \gamma_2$ = differenza tra l'intercetta del periodo guerra e pace

- ~~Cambiamento di coefficiente angolare~~

$$C = \alpha + \beta_1 Y + (\beta_2 - \beta_1) X_2 Y + \varepsilon$$

$$X_2 = \begin{cases} 0 & \text{anni di guerra} & C = \alpha + \beta_1 Y + \varepsilon \\ 1 & \text{anni di pace} & C = \alpha + \beta_2 Y + \varepsilon \end{cases}$$

$\beta_2 - \beta_1$ = differenza propensione marginale_{4f} al consumo nei due periodi

ASPETTI CONNESSI ALLA SPECIFICAZIONE DEL MODELLO

In ogni studio econometrico, la scelta del modello è la prima fase del lavoro. Gli aspetti fondamentali sono:

- a) La scelta della forma funzionale
- b) La scelta dei regressori
- c) La verifica sulle assunzioni del modello.

a. La scelta della forma funzionale

Abbiamo parlato di modelli di regressione lineari, intendendo lineari nei parametri, ovvero anche di quei modelli che possono essere resi lineari tramite una opportuna trasformazione delle variabili. Ad esempio si consideri la funzione di produzione Cobb-Douglas (Y produzione, L lavoro, K capitale:

$$Y = \alpha L^{\beta} K^{\gamma}$$

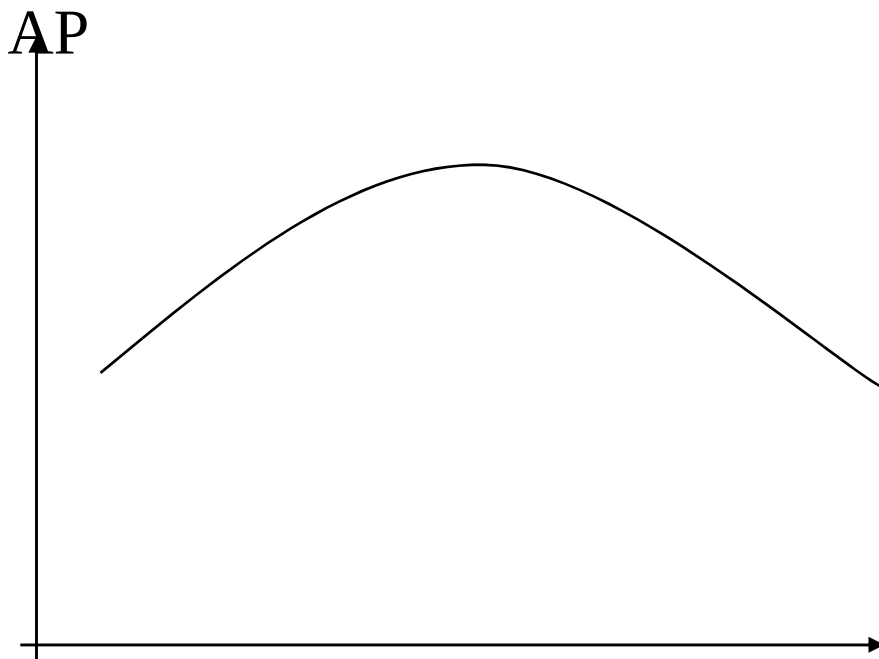
Potrebbe sembrare non lineare, tuttavia dopo aver applicato la trasformazione logaritmica otteniamo:

$$\ln(Y) = \ln(\alpha) + \beta \ln(L) + \gamma \ln(K)$$

Il modello così trasformato è lineare nei parametri e può essere facilmente trattato ed interpretato.

Esistono forme di modelli che risultano lineari nei parametri, ma sui quali fare attenzione soprattutto in fase di interpretazione.

Modelli polinomiali: consideriamo un esempio. In microeconomia si studiano funzioni di produzione, se consideriamo la relazione tra prodotto medio ottenuto da aziende produttrici di materiale elettrico (AP: average product) e l'input (I) necessario alla produzione



È evidente che la relazione non è costante e quindi non può essere rappresentata da un modello “lineare nelle variabili”. La relazione può essere espressa da un polinomio:

$$AP = \alpha + \beta I + \gamma I^2 + \varepsilon$$

Questa forma funzionale ha una forma non lineare ma risulta ancora un modello di regressione lineare essendo lineare nei parametri. Tali parametri si stimano con OLS e gli stimatori hanno tutte le “buone” proprietà; ma attenzione all’interpretazione!

I parametri che si stimano non sono di per se’ le pendenze, le quali invece sono date da

$$\frac{dE (AP)}{dI} = \beta + 2 \gamma I$$

E pertanto cambia per ogni valori di I con i parametri β e γ .

Modelli con interazioni: quando in un modello si inserisce il prodotto tra due variabili esplicative (interazione) l'effetto che si ottiene è quello di alterare la relazione di ognuna di esse con la variabile dipendente del modello. Per capire l'effetto consideriamo un esempio: studiamo l'effetto di reddito (Y) ed età (ETA') sul consumo di pizza C, supponiamo di avere i dati su un campione di individui con età superiore a 18 anni. Il modello senza interazione:

$$C = \alpha + \beta \text{ETA}' + \gamma Y + \varepsilon$$

$dE(C)/d\text{ETA}' = \beta$ per qualsiasi livello di reddito la spesa attesa per pizza varia di β (solitamente $\beta < 0$) per un incremento di un anno di età.

$dE(C)/dY = \gamma$ per qualsiasi età la spesa attesa per pizza varia di γ per un incremento di un euro di reddito ($\gamma > 0$).

In realtà sembrerebbe più ragionevole pensare che da una certa età in poi, con il crescere dell'età, la propensione marginale a spendere in pizza diminuisca. Siamo cioè nel caso in cui l'effetto di una variabile è modificato da un'altra. Per tenere conto di ciò il modello che dobbiamo specificare è il seguente:

$$C = \alpha + \beta \text{ETA} + \gamma Y + \lambda (\text{ETA} * Y) + \varepsilon$$

Gli effetti di Y e ETA' sono:

$dE(C)/d\text{ETA} = \beta + \lambda Y$ al crescere dell'età ci si aspetta che la spesa per pizza si riduca, inoltre siccome presumibilmente $\lambda < 0$, maggiore è il reddito, maggiore è la riduzione della spesa per pizza.

$dE(C)/dY = \gamma + \lambda \text{ETA}$ la propensione marginale a spendere in pizza dipende da ETA', quindi la propensione diminuisce sempre più al crescere dell'età.

b. La scelta dei regressori

Nella scelta delle variabili esplicative di un modello di regressione, si cerca di seguire la teoria relativa al fenomeno trattato, la logica e l'esperienza. Tuttavia può accadere che nella scelta si siano omesse importanti variabili o inserite variabili irrilevanti, vediamo quali problemi si incontrano in questi casi.

- Variabili rilevanti omesse: è come introdurre restrizioni (parametro=0) non vere sul modello. La stima OLS dei restanti parametri del modello risulta generalmente distorta, inoltre gli standard error di tali parametri sono sottostimati. Il caso in cui gli stimatori OLS non sono distorti si ha quando le variabili omesse sono incorrelate con le variabili inserite.

Per realizzare che alcune variabili rilevanti del modello sono state omesse si deve proprio fare attenzione a segni o valori dei coefficienti inaspettati. Si potrebbe pensare che per ovviare a questo problema il ricercatore dovrebbe inserire nel modello tutte le variabili che ha a disposizione; in questo modo tuttavia si potrebbe complicare il modello eccessivamente ed inoltre introdurre variabili irrilevanti.

■ Variabili irrilevanti inserite: gli stimatori OLS che si ottengono sono corretti, tuttavia la varianza degli stimatori dei parametri relativi alle variabili “buone” risulta maggiore di quella che avremmo ottenuto specificando il modello correttamente. Il motivo di questa sovrastima è legato al fatto che il Teorema di Gauss Markov dice che lo stimatore b.l.u.e. è lo stimatore OLS relativo ad un modello correttamente specificato.

LE VIOLAZIONI DELLE IPOTESI DEL MODELLO

- a) Multicollinearità
- b) Eteroschedasticità
- c) Autocorrelazione dei residui

a. MULTICOLLINEARITA'

Quando due o più variabili esplicative di un modello di regressione lineare si muovono sistematicamente “insieme” esiste un problema di *multicollinearità*. Le conseguenze di una tale situazione in un modello di tipo econometrico possono essere riassunte così:

- Se esiste una relazione lineare esatta tra le variabili esplicative (due o più) si parla di *esatta multicollinearità* → non si possono determinare le stime OLS dei parametri.
- Se la dipendenza lineare tra le variabili è quasi perfetta, ma non perfetta (coefficiente di correlazione prossimo a 1), siamo nel caso di *quasi multicollinearità* → le stime OLS si determinano ma sono molto instabili a causa degli elevati standard error, si determinano intervalli di confidenza molto larghi.

Cosa fare?

Nel caso di esatta multicollinearità si può fare una sostituzione di variabile.

Esempio:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$X_2 = \gamma X_1 + \theta$$

$$\begin{aligned} Y &= \alpha + \beta_1 X_1 + \beta_2 \theta + \beta_2 \gamma X_1 + \varepsilon \\ &= \lambda_1 + \lambda_2 X_1 + \varepsilon \end{aligned}$$

$$\lambda_1 = \alpha + \beta_2 \theta$$

$$\lambda_2 = \beta_1 + \gamma \beta_2$$

Nel caso in cui due o più regressori siano **quasi-collineari, si incontrano i problemi maggiori:**

- Varianze campionarie molto alte
- Covarianze sovrastimate
- Forte instabilità dei coefficienti stimati per piccole variazioni dei dati.

Per comprendere il perché di questi effetti si consideri il modello di regressione a tre variabili:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1} X' y$$

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$
$$= \frac{\sigma^2}{\sum X_1^2 \sum X_2^2 - (\sum X_1 X_2)^2} \begin{pmatrix} \sum X_2^2 & -\sum X_1 X_2 \\ -\sum X_1 X_2 & \sum X_1^2 \end{pmatrix}$$

$$\begin{aligned}
V(\hat{\beta}_1) &= \frac{\sigma^2 \sum X_2^2}{\sum X_1^2 \sum X_2^2 - (\sum X_1 X_2)^2} = \\
&= \frac{\sigma^2 \sum X_2^2}{\sum X_1^2 \sum X_2^2 \frac{(\sum X_1^2 \sum X_2^2 - (\sum X_1 X_2)^2)}{\sum X_1^2 \sum X_2^2}} \\
&= \frac{\sigma^2}{\sum X_1^2 (1 - r_{12}^2)}
\end{aligned}$$

$$V(\hat{\beta}_2) = \frac{\sigma^2}{\sum X_2^2 (1 - r_{12}^2)}$$

Si vede che valori molto alti di r_{12}^2 rendono le stime OLS molto imprecise.

Inoltre, nell'esempio che segue vediamo che piccole variazioni nella matrice dei dati possono provocare grandi variazioni nella stima dei parametri.

ESEMPIO:
instabilità delle stime

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$\begin{array}{ll} \text{Dati : } \sum X_{1i}^2 = 200 & \sum X_{1i} X_{2i} = 150 \\ \sum X_{2i}^2 = 113 & \sum X_{1i} Y_i = 350 \\ & \sum X_{2i} Y_i = 263 \end{array}$$

$$\hat{\beta}_1 = \frac{\sum X_2^2 \sum X_1 Y - \sum X_1 X_2 \sum X_2 Y}{\sum X_1^2 \sum X_2^2 - (\sum X_1 X_2)^2} =$$

$$\frac{113 \times 350 - 150 \times 263}{200 \times 113 - 150^2} = \frac{39550 - 39450}{22600 - 22500} = \frac{100}{100} = 1$$

$$\hat{\beta}_2 = \frac{52600 - 52500}{22600 - 22500} = \frac{100}{100} = 1$$

$$r_{X_2 X_3}^2 = \frac{(\sum X_2 X_3)^2}{\sum X_2^2 \sum X_3^2} = \frac{150^2}{200 \times 113} = 0.995$$

Se togliamo solo una osservazione cosicchè:

$$\sum X_1^2 = 199$$

$$\sum X_2 X_1 = 149$$

$$\sum X_2^2 = 112$$

$$\sum X_1 Y = 327.5$$

$$\sum X_2 Y = 261.5$$

$$\hat{\beta}_1 = \frac{112 \times 347.5 - 149 \times 261.5}{199 \times 112 - 149^2} = \frac{-43.5}{87} = -\frac{1}{2}$$

$$\hat{\beta}_2 = \frac{199 \times 261.5 - 149 \times 347.5}{199 \times 112 - 149^2} = \frac{261}{87} = 3$$

Le stime si modificano di molto

Come identificare un problema di multicollinearità?

La via più intuitiva è quella di osservare la matrice di correlazione delle variabili, se identifichiamo coefficienti di correlazione prossimi a 0.9 (in valore assoluto) abbiamo ragione di credere che il problema della quasi multicollinearità sia presente.

Tuttavia con il suddetto metodo si identificano problemi per coppie di variabili, resta il dubbio su cosa fare se sono più di due le variabili a creare multicollinearità. Una strategia è quella di fare “regressioni ausiliarie” tra una variabile “sospetta” e le altre esplicative; se il coefficiente di determinazione che si ottiene è prossimo a 1 sicuramente il coefficiente di regressione della variabile sospetta –nella regressione originale- risente del problema della multicollinearità.



LA MULTICOLLINEARITA'

Si verifica quando il rango della matrice X non è massimo e si traduce nella presenza di un'elevata correlazione tra le variabili esplicative. Le variabili collineari non forniscono informazioni aggiuntive e risulta difficile individuare l'effetto che ciascuna di esse ha sulla variabile risposta. Una misura della multicollinearità è data dall'indice VIF (Variance Inflationary Factor). In particolare, per la j -esima variabile si ha

$$VIF_j = \frac{1}{1 - R_j^2},$$

dove R_j^2 è il coefficiente di determinazione che caratterizza il modello in cui la variabile dipendente è X_j e tutte le altre variabili esplicative sono incluse nel modello.

ETEROSCHEDASTICITA'

Avevamo ipotizzato che:

$$E[\varepsilon'\varepsilon] = \sigma^2 I$$

tale assunzione è in molte situazioni non valida.

In effetti, se noi consideriamo come variabile dipendente di un modello la spesa per alimenti Y e come variabile indipendente il reddito X , è poco plausibile assumere omoschedasticità perché al crescere del reddito ci sono molti più fattori di soggettività nella scelta degli alimenti e quindi nella relativa spesa. Il modo più semplice per valutare la validità dell'ipotesi di omoschedasticità è considerare i residui OLS del modello stimato e tracciare un diagramma cartesiano in cui in corrispondenza di ogni valore di X si riporta il corrispondente residuo stimato. Se i residui risultano casualmente dispersi attorno allo zero, si può supporre che l'ipotesi di omoschedasticità sia plausibile, se essi hanno un andamento sistematico a ventaglio o quadratico o sinusoidale la nostra ipotesi risulta presumibilmente non vera. Nell'esempio i residui saranno disposti a ventaglio, dato che al crescere del reddito essi cresceranno.

Ma quali sono le conseguenze dell'eteroschedasticità negli stimatori OLS dei parametri?

Innanzitutto è opportuno comprendere quale diventa la nuova formulazione dell'ipotesi sul termine stocastico:

$$E[\varepsilon] = 0$$

$$E[\varepsilon'\varepsilon] = \sigma^2\Omega, \quad \text{con} \quad E[\varepsilon_i^2] = \sigma_i^2$$

Le stime OLS dei parametri sono:

$$\hat{\beta} = (X'X)^{-1} X' y$$

$$y = X\beta + \varepsilon$$

$$E[\hat{\beta}] = \beta + (X'X)^{-1} X' E(\varepsilon) = \beta$$

Quindi STIMATORI OLS ancora lineari e corretti, tuttavia si perde l'efficienza, infatti

$$\begin{aligned} V(\hat{\beta}) &= (X'X)^{-1} X' E(\varepsilon'\varepsilon) X (X'X)^{-1} \\ &= (X'X)^{-1} X' \Omega (X'X)^{-1} \sigma^2 \end{aligned}$$

Ne consegue che gli intervalli di confidenza e i risultati della verifica di ipotesi possono essere fuorvianti.

Per individuare la presenza di eteroschedasticità la via più intuitiva è quella di fare un'analisi dei residui, tuttavia essa può essere complessa se le variabili esplicative sono molte.

Ci sono tuttavia alcuni test che si basano in generale sempre sui residui.

Esempio: GOLDFELD – QUANDT TEST

- Si ordinano le osservazioni secondo la variabile X_j che si ipotizza sia la causa dell'eteroschedasticità
- Si divide il campione in tre parti di numerosità n_1 n_2 n_3 .
- Dopo la stima OLS nei tre sottocampioni si calcola

$$F = \frac{e_1' e_1}{e_3' e_3} \approx F_{n_1-k, n_2-k}$$

Sotto H_0 : omoschedasticità : (il valore di F è piccolo)

$$F_{\text{empirico}} > F_{\text{teorico}} \Rightarrow \text{Rifiuto } H_0$$

I RIMEDI

1. σ_i $i = 1, \dots, n$ siano valori noti.

si applicano i **MINIMI QUADRATI PESATI (WLS)**

ovvero si applica OLS al modello trasformato

$$y_i^* = \frac{y_i}{\sigma_i} \quad ; \quad x_{ij}^* = \frac{x_{ij}}{\sigma_i} \quad ; \quad \varepsilon_i^* = \frac{\varepsilon_i}{\sigma_i}$$

Ovvero
$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \dots + \beta_k x_{ik}^* + \varepsilon_i^*$$

Dove
$$Var(\varepsilon_i^*) = Var\left(\frac{\varepsilon_i}{\sigma_i}\right) = \frac{1}{\sigma_i^2} Var(\varepsilon_i) = \frac{\sigma_i^2}{\sigma_i^2} = 1$$

Nella pratica σ_i non sono noti quindi il metodo non è applicabile in pratica

2. In caso di relazione tra la componente stocastica e uno dei regressori, ad esempio

$$y_i = \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i$$

$$Var \varepsilon_i = C x_{i2}^2$$

Trasformiamo il modello

$$y_i^* = \frac{y_i}{x_{i2}} \quad ; \quad x_{ij}^* = \frac{x_{ij}}{x_{i2}} \quad ; \quad \varepsilon_i^* = \frac{\varepsilon_i}{x_{i2}}$$

$$\Rightarrow \frac{y_i}{x_{i2}} = \beta_1 \frac{1}{x_{i2}} + \beta_2 + \dots + \beta_k \frac{x_{ik}}{x_{i2}} + \frac{\varepsilon_i}{x_{i2}}$$

$$Var(\varepsilon_i^*) = Var\left(\frac{\varepsilon_i}{x_{i2}}\right) = \frac{1}{x_{i2}^2} Var(\varepsilon_i) = C$$

applico OLS e ottengo stimatori B.L.U.E. per i parametri di interesse.

3. Si stima il modello originale ottenendo stimatori lineari e corretti, per il calcolo degli s.e. dei parametri si ricorre allo stimatore di White che tutti i software prevedono.

ESERCIZIO

La stima di un modello lineare sulla base dei valori del Reddito e del Consumo di 30 famiglie americane fornisce i seguenti valori :

$$\hat{C} = 1480 + 0.788y \quad R^2 = 0.97$$

(3.29) (29.37)

La stima dello stesso modello sulle prime 12 e sulle ultime 12 osservazioni fornisce i seguenti valori:

$$\hat{C} = 846.7 + 0.837y \quad R^2 = 0.91$$

(0.74) (9.91)

$$SEQ = 1069000$$

$$\hat{C} = 2306.7 + 0.747y \quad R^2 = 0.71$$

(0.79) (5.00)

$$SEQ = 3344000$$

Verificare l'ipotesi H_0 di omoschedasticità

$$F = \frac{3344000}{1069000} = 3.12 \quad F_{10,10} = 1.83$$

Rifiuto H_0 : c'è eteroschedasticità

c. AUTOCORRELAZIONE DEI RESIDUI

Nelle analisi di dati *cross-sectional* le osservazioni sono generalmente individui o famiglie o aziende che costituiscono un campione casuale di una popolazione. Il fatto che il campione sia casuale, generalmente implica l'incorrelazione dei termini casuali.

$$E[\varepsilon'\varepsilon] = \sigma^2 I$$

Quando si hanno invece serie storiche o comunque osservazioni che seguono un ordine temporale tale ipotesi si altera ed i termini di errore risultano generalmente tra loro correlati.

Per illustrare il problema consideriamo una semplice relazione a due variabili

$$\begin{aligned} y_t &= \alpha + \beta X_t + \varepsilon_t \\ \varepsilon_t &= \rho \varepsilon_{t-1} + u_t \end{aligned}$$

Le ipotesi aggiuntive su tale modello, detto *modello autoregressivo del primo ordine* AR(1) sono:

$$|\rho| < 1$$

$$E[u_t] = 0$$

$$E[u_t \quad u_{t-s}] = \begin{cases} \sigma_u^2 & s = 0 \\ 0 & s \neq 0 \end{cases}$$

Quindi:

$$\begin{aligned} \varepsilon_t &= \rho \varepsilon_{t-1} + u_t \\ &= \rho(\rho \varepsilon_{t-2} + u_{t-1}) + u_t \\ &= u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \dots = \\ &= \sum_{r=0}^{\infty} \rho^r u_{t-r} \end{aligned}$$

$$E(\varepsilon_t) = \sum_{r=0}^{\infty} \rho^r E(u_{t-r}) = 0$$

$$\begin{aligned} E(\varepsilon_t^2) &= E(u_t^2) + \rho^2 E(u_{t-1}^2) + \rho^4 E(u_{t-2}^2) + \dots \\ &\quad + 2\rho E[u_t u_{t-1}] + 2\rho^2 E[u_t u_{t-2}] + \dots \\ &\quad + 2\rho E[u_{t-1} u_{t-2}] + \dots \\ &= \sigma_u^2 (1 + \rho^2 + \rho^4 + \dots) \\ &= \frac{\sigma_u^2}{1 - \rho^2} = \boxed{\sigma_u^2} \end{aligned}$$

$$\begin{aligned}
E[\varepsilon_t \varepsilon_{t-1}] &= E\left[(u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \dots) \times (u_{t-1} + \rho u_{t-2} + \dots)\right] = \\
&= \rho \sigma_u^2 + \rho^3 \sigma_u^2 + \rho^5 \sigma_u^2 + \dots = \\
&= \rho \sigma_u^2 (1 + \rho^2 + \rho^4 + \dots) = \\
&= \rho \frac{\sigma_u^2}{1 - \rho^2} = \rho \sigma_\varepsilon^2
\end{aligned}$$

$$\begin{aligned}
E[\varepsilon_t \varepsilon_{t-2}] &= E\left[\begin{aligned} &(u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3}) \times \\ &\times (u_{t-2} + \rho u_{t-3} + \rho^2 u_{t-4} + \dots) \end{aligned}\right] = \\
&= \rho^2 \sigma_u^2 + \rho^4 \sigma_u^2 + \rho^6 \sigma_u^2 + \dots = \\
&= \rho^2 \frac{\sigma_u^2}{1 - \rho^2} = \rho^2 \sigma_\varepsilon^2
\end{aligned}$$

$$E[\varepsilon_t \varepsilon_{t-s}] = \rho^s \sigma_\varepsilon^2$$

$$E[\varepsilon \varepsilon'] = V = \sigma_u^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdot & \rho^{n-1} \\ \rho & 1 & \rho & \cdot & \rho^{n-2} \\ \rho^2 & \rho & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \rho \\ \rho^{n-1} & \rho^{n-2} & \cdot & \rho & 1 \end{bmatrix}$$

CONSEGUENZE per OLS

1. Stime OLS di β lineari e corrette
2. Varianze di $\hat{\beta}$ molto grandi $\rightarrow$ **stimatori inefficienti**
3. Sottostima delle varianze dell'errore
4. Conseguente non validità dei test t ed F

Infatti si può dimostrare che

$$E(e'e) \cong \sigma_{\varepsilon}^2 \left[\varepsilon - \frac{1+\rho^2}{1-\rho^2} \right]$$

Solo se $\rho^2 = 0$

$$E\left[\frac{e'e}{n-1}\right] = E[\hat{\sigma}^2] = \sigma_{\varepsilon}^2$$

Con $N=20$; $\rho = 0.5$:

$$E\left[\frac{e'e}{n-1}\right] = \frac{18.3}{19} \sigma_{\varepsilon}^2 \quad \text{sottostima 4\%}$$

Con $N=20$; $\rho = 0.8$

$$E\left[\frac{e'e}{n-1}\right] = \frac{15.4}{19} \sigma_{\varepsilon}^2 \quad \text{sottostima 19\%}$$

E' MOLTO USATO IL TEST DI DURBIN - WATSON

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

$$\hat{e} = y - X\hat{\beta}$$



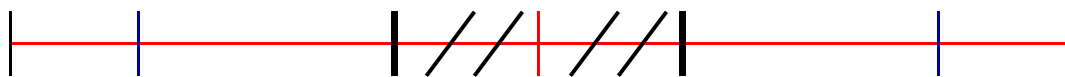
residui nella
stima OLS

$$d = \frac{\left[\sum_{t=2}^n e_t^2 + \sum_{t=2}^n e_{t-1}^2 \right] - 2 \sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}$$

per n grande

$$d \cong 2 - 2 \frac{\sum e_t e_{t-1}}{\sum e_t^2} = 2 \left[1 - \frac{\sum e_t e_{t-1}}{\sum e_t^2} \right] = 2(1 - r)$$

$$0 \leq d \leq 4$$



0

d_L

d_H

2

$4-d_H$

$4-d_L$

4

autocorr.(+) ?

No autocorr.

?

Autocorr.(-)

Il limite tra la zona di accettazione e quella di rifiuto è funzione della matrice X .

- Il valore della statistica di Durbin-Watson è sempre compreso tra 0 e 4.
- Un valore di 2 indica che non appare presente alcuna autocorrelazione. Valori piccoli di d indicano che i residui successivi sono, in media, vicini in valore l'uno all'altro, o correlati positivamente. Valori grandi di d indicano che i residui successivi sono, in media, molto differenti in valore l'uno dall'altro, o correlati negativamente.
- La distribuzione teorica della statistica di Durbin-Watson non è nota; tuttavia gli stessi Durbin e Watson hanno tabulato, con un esercizio di simulazione condotto col metodo Montecarlo, i valori critici della statistica.

Per verificare la presenza di **autocorrelazione positiva** al livello di significatività α , la statistica test d viene confrontata con dei valori critici inferiori e superiori ($d_{L,\alpha}$ and $d_{U,\alpha}$):

- Se $d < d_{L,\alpha}$ si ha una prova statistica di autocorrelazione positiva degli errori.
- Se $d > d_{U,\alpha}$, si ha una prova statistica di **non** autocorrelazione positiva degli errori.
- Se $d_{L,\alpha} < d < d_{U,\alpha}$ il test non è conclusivo.

Per verificare la presenza di **autocorrelazione negativa** al livello di significatività α , la statistica test d viene confrontata con dei valori critici inferiori e superiori ($d_{L,\alpha}$ and $d_{U,\alpha}$):

- Se $(4 - d) < d_{L,\alpha}$ si ha una prova statistica di autocorrelazione negativa degli errori.
- Se $(4 - d) > d_{U,\alpha}$, si ha una prova statistica di **non** autocorrelazione negativa degli errori.
- Se $d_{L,\alpha} < (4 - d) < d_{U,\alpha}$ il test non è conclusivo.

I valori critici $d_{L,\alpha}$ e $d_{U,\alpha}$ variano secondo il livello di significatività (α), secondo il numero di osservazioni e il numero di parametri di regressione e vengono generalmente ottenuti da apposite tavole.

METODI RISOLUTIVI

GLS – Generalized Least Squares

Se ho una stima di ρ

$$\hat{\rho} = \frac{\sum e_t e_{t-1}}{\sum e_t^2} \Rightarrow \Omega = \begin{bmatrix} 1 & \hat{\rho} & \cdot & \hat{\rho}^{n-1} \\ \hat{\rho} & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 \end{bmatrix}$$

Riesco a trovare la matrice $T : T' T = \Omega^{-1}$

e trasformo il modello in $Ty = TX\beta + Tu$

$$\text{Var}(Tu) = \sigma^2 I \Rightarrow$$

stima OLS