DATA MINING PER IL MARKETING (63 ore)

Marco Riani

mriani@unipr.it

Sito web del corso

http://www.riani.it/DMM

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

Casi particolari

R=(0 , …, 0, 1, 0. …0) e r=0

$$H_0 : \beta_i = 0.$$

$$R\hat{\beta} - r = \hat{\beta}_i$$

$$R(X'X)^{-1}R'$$

individua unicamente l'i-esimo elemento S^{ii} sulla diagonale principale della matrice $(X'X)^{-1}$. Il rapporto F diventa

$$F = \frac{\hat{\beta}_i^2}{S^{ii}s^2} \sim F(1, n-k)$$

Relazione con il test t per testare $\beta_i=0$

- L'equazione

$$F = \frac{\hat{\beta}_i^2}{S^{ii} s^2} \sim F(1, n-k)$$

- non è altro che il quadrato del test t

Relazione con il test t

$$T^2(n-k) = \left(\frac{N(0,1)}{\sqrt{\frac{\chi^2(n-k)}{n-k}}} \right)^2 = \frac{\chi^2(1)/1}{\chi^2(n-k)/(n-k)} = F(1, n-k).$$

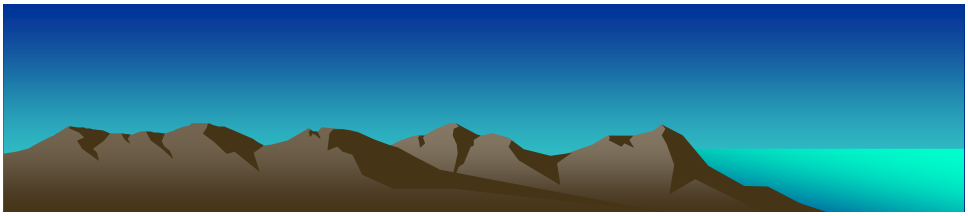
Set di variabili esplicative non rilevanti

Si può dimostrare, infine, che qualora l'ipotesi nulla sia della forma

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

ossia che gli ultimi q coefficienti siano pari a 0 nell'universo, l'equazione che definisce il test può essere scritta come segue:

$$F = \frac{(e_r' e_r - e'e)/q}{e'e/(n-k)} \sim F(q, n-k)$$



Procedura

1. regredire y sulle variabili esplicative $X_1, \ldots, X_{k-q}$ che non rientrano nel sottoinsieme da testare e calcolare la devianza residua $e_r' e_r$
2. effettuare la regressione completa e calcolare la devianza residua $e'e$. La differenza $e_r' e_r - e'e$ misura la diminuzione nella devianza residua dovuta all'inclusione dell'insieme $X_{k-q+1}, X_{k-q+2}, \ldots X_k$ nella regressione;
3. La quantità $(e_r' e_r - e'e)/r$ viene confrontata con la quantità $e'e/(n-k)$. Se il valore del test F che ne deriva supera un prefissato valore critico si rifiuta l'ipotesi che le variabili $X_{k-q+1}, X_{k-q+2}, \ldots, X_k$ non abbiano influenza sulla y.

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

ossia

$$H_0 : R^2 = 0$$

contro una alternativa

$H_1 : R^2 > 0,$ cioè *almeno* uno dei coefficienti significativamente diverso da 0.

$$F = \frac{(e_r' e_r - e'e)/q}{e'e/(n-k)} \sim F(q, n-k)$$

- In questo esempio cos'è $e'_r e_r$? cos'è $e'e$?

$$F = \frac{(e_r' e_r - e'e)/q}{e'e/(n-k)} \sim F(q, n-k)$$

- $e'_r e_r$= Devianza totale
- $e'e$ = Devianza residua

$$F_c = \frac{\text{DEV.Regr}/\,(k-1)}{\text{DEV.Res}/\,(n-k)} = \frac{R^2/\,(k-1)}{(1-R^2)\,/\,(n-k)}$$

Output della funzione REGR.LIN

	A	B	C	D	E	F	G
1	b_k	b_{k-1}	…	b_2	b_1	b_0	a
2	s_k	s_{k-1}	…	s_2	s_1	s_0	s_a
3	R2	se_y					
4	F	d_f					
5	ss_{reg}	ss_{resid}					
6							

- Test F

Output del componente aggiuntivo analisi dati

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	5841.06918	2920.53	107.86051	2.14126E-08
Residuo	12	324.923484	27.0769		
Totale	14	6165.99266			

$$F_c = \frac{\text{DEV.Regr}/\left(k-1\right)}{\text{DEV.Res}/\left(n-k\right)} = \frac{R^2/\left(k-1\right)}{\left(1-R^2\right)/\left(n-k\right)}$$

Sessione al computer: verificare

$$F_c = \frac{\text{DEV.Regr}/\,(k-1)}{\text{DEV.Res}/\,(n-k)} = \frac{R^2/\,(k-1)}{(1-R^2)\,/\,(n-k)}$$

4. Porre

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

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

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

Verifica della bontà di adattamento del modello

Analisi dei residui

Diverse tipologie di residui

Residui standardizzati $e_i/s \qquad i = 1, \ldots, n.$

Residui studentizzati $r_i = \dfrac{e_i}{s\sqrt{1-h_{ii}}} \qquad i = 1, \ldots, n.$

Residui studentizzati di cancellazione $r_i^* = \dfrac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$

Come si trova s(i)

$$(n-k-1)s^2_{(i)} = (n-k)s^2 - e_i^2/(1-h_{ii})$$

Quali sono le osservazioni più importanti nella stima di beta cappello?

- Punto di partenza.
- Un intervallo di confidenza al livello (1-gamma) per il vettore beta è dato da:

$$(\beta - \hat{\beta})' X' X (\beta - \hat{\beta}) \leq k s^2 f_{k,n-k,\gamma},$$

Quali sono sono più importanti nella stima di beta cappello?

- La distanza di Cook

$$D_i = (\hat{\beta}_{(i)} - \hat{\beta})' X' X (\hat{\beta}_{(i)} - \hat{\beta}) / (k s^2). \qquad (4.53)$$

dove $\hat{\beta}_{(i)}$ è la stima dei minimi quadrati di β omettendo l'osservazione i-esima (eq.

Con semplici passaggi

$$D_i = \frac{e_i^2 h_{ii}}{k s^2 (1 - h_{ii})^2} = \frac{h_{ii}}{1 - h_{ii}} \frac{r_i^2}{k}$$

- Distanza di Cook modificata (Atkinson, 1985)

$$\begin{aligned} C_i &= \left\{ \frac{n-k}{k} \right\}^{1/2} \left\{ \frac{h_{ii}}{(1-h_{ii})^2} \frac{e_i^2}{s_{(i)}^2} \right\}^{1/2} \\ &= \left\{ \frac{n-k}{k} \frac{h_{ii}}{1-h_{ii}} \right\}^{1/2} |r_i^*|. \end{aligned}$$

Intervallo di confidenza del valore y0 associato ad uno specifico insieme di valori delle variabili esplicative

$$y_0 = \beta' x_0 + \varepsilon_0 .$$

Es. investimenti PIL e trend

Anni	investimenti (y)	P.I.L. (X_1)	Trend (X_2)
1982	209.952	1060.859	1
1983	207.825	1073.783	2
1984	214.923	1101.366	3
1985	215.985	1132.313	4
1986	220.371	1164.465	5
1987	230.058	1200.523	6
1988	245.872	1246.966	7
1989	256.720	1282.905	8
1990	266.044	1310.659	9
1991	268.273	1325.582	10
1992	263.361	1333.072	11
1993	229.628	1317.668	12
1994	230.785	1346.267	13
1995	246.659	1385.830	14
1996	249.619	1395.408	15
Fonte: ISTAT			

$$x_0' = (1, 1405, 16)$$

$$y_0 = \beta' x_0 + \varepsilon_0.$$

Strategia

- Passiamo attraverso e_0 e poi esplicitiamo y_0

$$e_0 = y_0 - \widehat{y}_0 = \left(\beta - \hat{\beta}\right)' x_0 + \varepsilon_0.$$

$$\frac{e_0 - E(e_0)}{\sqrt{var(e_0)}} \sim N(0,1)$$

Troviamo E(e_0) e var(e_0)

$$e_0 = y_0 - \widehat{y}_0 =$$

$$E(e_0) = E\left(\left(\beta - \hat{\beta}\right)' x_0 + \varepsilon_0\right) = E(\beta' x_0) - E\left(\hat{\beta}' x_0\right) + E(\varepsilon_0) = 0.$$

Var(e_0)

$$\begin{aligned} var(e_0) &= var(y_0 - x_0'\hat{\beta}) \\ &= var(y_0) + var(x_0'\hat{\beta}) \\ &= var(y_0) + x_0' var(\hat{\beta}) x_0 \\ &= var(y_0) + x_0'\sigma^2(X'X)^{-1}x_0 \end{aligned}$$

$$var(e_0) = \sigma^2(1 + x_0'(X'X)^{-1}x_0)$$

$$\widehat{var}(e_0) = s^2(1 + x_0'(X'X)^{-1}x_0)$$

Ob. trovare intervallo di conf. per y_0

$$\frac{e_0 - E(e_0)}{\sqrt{s^2(1 + x_0'(X'X)^{-1}x_0)}} = \frac{e_0 - E(e_0)}{\sqrt{\widehat{Var}(e_0)}} \sim t(n-k)$$

$$\Pr\left(-t_\gamma \leq \frac{e_0 - E(e_0)}{\sqrt{\widehat{var}(e_0)}} \leq t_\gamma\right) = 1 - \gamma$$

Ob. trovare intervallo di conf. per y_0

$$\Pr\left(-t_\gamma \leq \frac{e_0 - E(e_0)}{\sqrt{\widehat{var}(e_0)}} \leq t_\gamma\right) = 1 - \gamma$$

$$\Pr\left(-t_\gamma \sqrt{\widehat{var}(e_0)} \leq e_0 \leq t_\gamma \sqrt{\widehat{var}(e_0)}\right) = 1 - \gamma$$

$$\Pr\left(-t_\gamma \sqrt{\widehat{var}(e_0)} \leq y_0 - \hat{y}_0 \leq t_\gamma \sqrt{\widehat{var}(e_0)}\right) = 1 - \gamma$$

$$\Pr\left(\hat{y}_0 - t_\gamma \sqrt{\widehat{var}(e_0)} \leq y_0 \leq \hat{y}_0 + t_\gamma \sqrt{\widehat{var}(e_0)}\right) = 1 - \gamma$$

Es. investimenti PIL e trend

Anni	investimenti (y)	P.I.L. (X_1)	Trend (X_2)
1982	209.952	1060.859	1
1983	207.825	1073.783	2
1984	214.923	1101.366	3
1985	215.985	1132.313	4
1986	220.371	1164.465	5
1987	230.058	1200.523	6
1988	245.872	1246.966	7
1989	256.720	1282.905	8
1990	266.044	1310.659	9
1991	268.273	1325.582	10
1992	263.361	1333.072	11
1993	229.628	1317.668	12
1994	230.785	1346.267	13
1995	246.659	1385.830	14
1996	249.619	1395.408	15
Fonte: ISTAT			

$$x_0' = (1, 1405, 16)$$

Es. investimenti PIL e trend

$$\hat{y}_0 = x_0'\hat{\beta} = \begin{pmatrix} 1 & 1405 & 16 \end{pmatrix} \begin{pmatrix} -441.272 \\ 0.625 \\ -12.522 \end{pmatrix} = 236.818$$

$$\widehat{var}(e_0) = s^2(1 + x_0'(X'X)^{-1}x_0)$$

$$x_0'(X'X)^{-1}x_0 = \begin{pmatrix} 1 & 1405 & 16 \end{pmatrix} \begin{pmatrix} 136.422 & -0.13 & 3.227 \\ -0.1302 & 0.0001 & -0.003 \\ 3.22733 & -0.003 & 0.081 \end{pmatrix} \begin{pmatrix} 1 \\ 1405 \\ 16 \end{pmatrix} = 0.4963$$

$$\widehat{var}(e_0) = 5.20355^2\sqrt{1 + 0.4963} = 40.515$$

Intervallo di confidenza per y_0

$$\Pr\left(\hat{y}_0 - t_\gamma \sqrt{\widehat{var}(e_0)} \leq y_0 \leq \hat{y}_0 + t_\gamma \sqrt{\widehat{var}(e_0)}\right) = 1 - \gamma$$

$$\Pr\left(236.818 - 3.0545 \times 40.515 \leq \quad y_0 \quad \leq 236.818 + 3.0545 \times 40.515\right) = 0.99$$

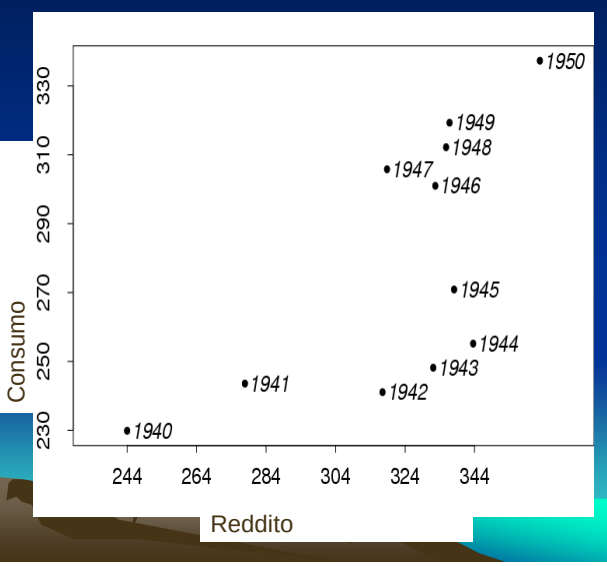
$$\Pr\left(217.375 \leq \quad y_0 \quad \leq 256.260\right) = 0.99$$

Sessione al computer: calcolare l'intervallo di confidenza per y_0

Regressione con variabili categoriche

Es. consumo e reddito

Anni	Reddito	Consumo
1940	244	229.9
1941	277.9	243.6
1942	317.5	241.1
1943	332.1	248.2
1944	343.6	255.2
1945	338.1	270.9
1946	332.7	301.0
1947	318.8	305.8
1948	335.8	312.2
1949	336.8	319.3
1950	362.8	337.3



Aggiunta di una variabile dummy

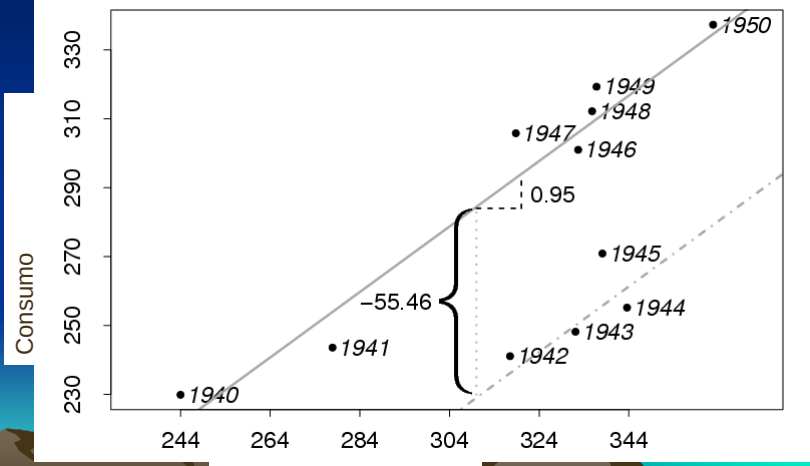
$$X = \begin{pmatrix} \text{Intercetta} & \text{Reddito} & \text{Variabile dummy} \\ 1 & 244 & 0 \\ 1 & 277.9 & 0 \\ 1 & 317.5 & 1 \\ 1 & 332.1 & 1 \\ 1 & 343.6 & 1 \\ 1 & 338.1 & 1 \\ 1 & 332.7 & 0 \\ 1 & 318.8 & 0 \\ 1 & 335.8 & 0 \\ 1 & 336.8 & 0 \\ 1 & 362.8 & 0 \end{pmatrix}$$

Risultati del modello di regr. linere multiplo

	Coefficienti	*Errore standard*	*Stat t*	*Livello di significatività*
Intercetta	-10.0649	28.44336	-0.35386	0.732591
Reddito	0.959595	0.089481	10.72398	5.03E-06
Dummy	-55.4624	5.902399	-9.39659	1.35E-05

Tabella 1.4: Stime dei coefficienti (standard errors, statistiche t e livelli di significatività (p-values)) calcolati sui dati della tabella 1.3, dopo aver aggiunto la variabile dummy

Rappresentazione grafica dell'effetto della variabilie dummy = diminuire la stima teorica dell'ammontare dei consumi di un ammontare pari a -55.46



Confronto (con e senza dummy)

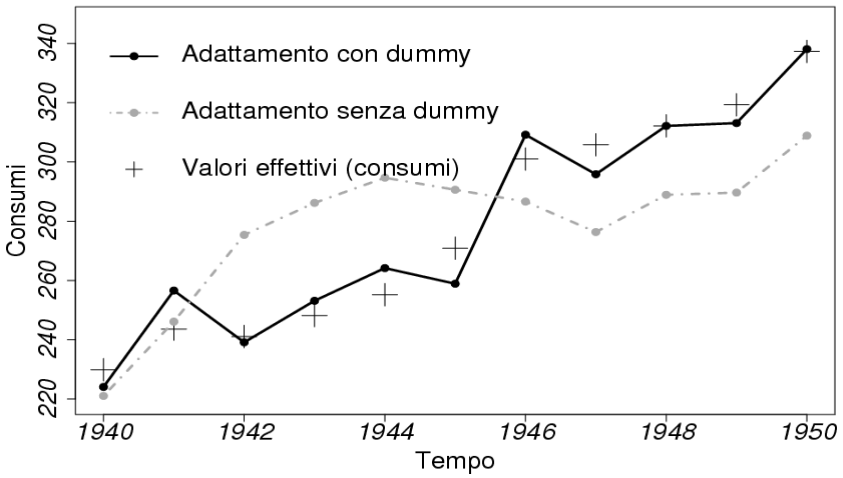


Figura 1.5: Confronto tra valori osservati (+), valori adattati con un modello di regressione che include la variabile dummy (nero) e senza variabile dummy (grigio) al variare del tempo