DATA MINING PER IL MARKETING (63 ore)

Marco Riani

mriani@unipr.it

Sito web del corso

http://www.riani.it/DMM

LA REGRESSIONE LINEARE SEMPLICE

LA REGRESSIONE LINEARE

- Esiste una relazione (lineare) tra X e Y?
- In caso affermativo:
- Come varia una variabile (dipendente) in funzione dell'altra (esplicativa)?
- Per convenzione:

Y = variabile dipendente

X = variabile esplicativa

Esempi

- Relazione tra comportamenti di acquisto e caratteristiche dei consumatori
- Relazione tra numero di esami sostenuti nei primi due anni di corso e voto alla maturità
- Relazione tra prezzo di vendita e quantità venduta di un bene

Motivi che spingono ad adottare modelli di regressione lineare

- Semplicità ➔ facilità di interpretazione dei parametri
- $y_i = a + bx_i + e_i$ $\quad\quad i = 1, \ldots, n$

dove:

- $a + bx_i$ ***rappresenta una retta:***
- ***a = ordinata all'origine*** $\rightarrow$ ***intercetta***
- ***b = coeff. angolare*** $\rightarrow$ ***coeff. di regressione***
- e_i ***è un termine di errore (accidentale)***

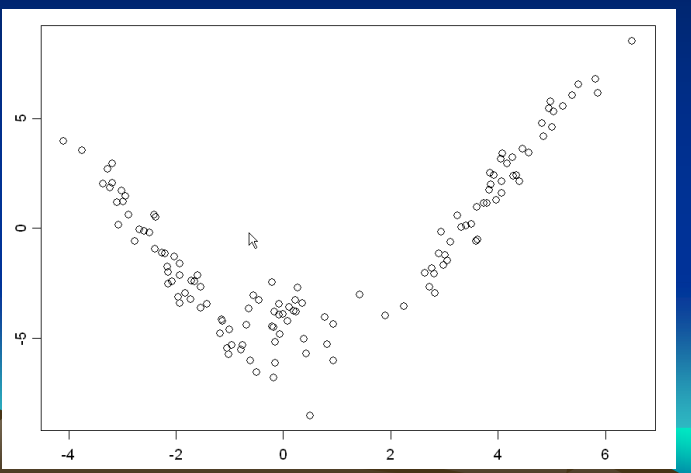
Motivi che spingono ad adottare modelli di regressione lineare

- Effettiva linearità ➔ molte relazioni sono molto vicine alla linearità
- Trasformazioni ➔ la relazione è lineare dopo aver trasformato opportunamente la dipendente e/o l'esplicativa

- Es. $y = a\ b^x$
- $\log y = \log a + (\log b)\ x$
- $y' = a' + b'\ x$

Motivi che spingono ad adottare modelli di regressione lineare

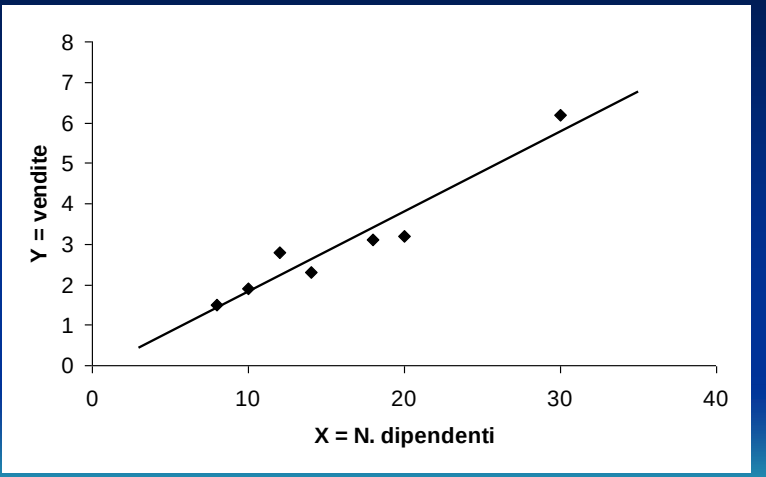
- Limitatezza dell'intervallo



Motivi che spingono ad adottare modelli di regressione lineare

- Ragioni di teoria statistica: lo studio delle funzioni lineari nei parametri ha una trattazione più agevole

Diagramma di dispersione



- **Come variano le vendite in funzione del numero di dipendenti?**

MODELLO DI REGRESSIONE

- $y_i = a + bx_i + e_i$ $\quad i = 1, \ldots, n$

dove:

- $a + bx_i$ *rappresenta una retta:*
- *a = ordinata all'origine* $\rightarrow$ *intercetta*
- *b = coeff. angolare* $\rightarrow$ *coeff. di regressione*
- e_i *è un termine di errore (accidentale)*

RETTA DI REGRESSIONE

$$\hat{y}_i = a + bx_i$$

- $i = 1, \ldots, n$

$\hat{y}_i$ = valore *teorico* (valore *stimato*) di $y_i \rightarrow$ funzione *lineare* di $i = 1, \ldots, n$

Residui

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

Come si calcolano i parametri *a* e *b*?

- **METODO DEI MINIMI QUADRATI**

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

Le incognite sono i parametri della retta: a, b

$$\hat{y}_i = a + bx_i$$

Come si calcolano i parametri *a* e *b*?

METODO DEI MINIMI QUADRATI (p. 224)

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

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \min$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial a} = 0 \qquad \frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b} = 0$$

Sistema di equazioni normali (p. 225)

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} x_i e_i = 0 \end{cases} \qquad \begin{cases} \sum_{i=1}^{n} (y_i - a - bx_i) = 0 \\ \sum_{i=1}^{n} (y_i - a - bx_i) x_i = 0 \end{cases}$$

2 equazioni e 2 incognite (*a* e *b*)

Dalla prima equazione (p. 225)

$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

$$na = \sum_{i=1}^{n}(y_i - bx_i)$$

$$a = \bar{y} - b\bar{x}$$

Sostituendo il valore trovato di a nella seconda equazione $a = \bar{y} - b\bar{x}$

$$\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0$$

$$\sum_{i=1}^{n}[y_i - (\bar{y} - b\bar{x}) - bx_i]x_i = 0$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Espressioni alternative per *a* e *b* *(eq. 8.4, 8.5, p. 224)*

$$a = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

ESEMPIO (7 supermercati) r_{xy}=0,96

	N. dipendenti (X)	Fatturato in milioni di € (Y)
A	10	1,9
B	18	3,1
C	20	3,2
D	8	1,5
E	30	6,2
F	12	2,8
G	14	2,3

Calcolo di a e b

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
A	10	1,9	100	3,61	19
B	18	3,1	324	9,61	55,8
C	20	3,2	400	10,24	64
D	8	1,5	…	…	…
E	30	6,2	…	…	…
F	12	2,8	…	…	…
G	14	2,3	…	…	…
Tot.	112	21	2128	77,28	402,6

$$a = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$a = \frac{21 \cdot 2.128 - 112 \cdot 402,6}{7 \cdot 2.128 - 112^2} = -\frac{403,2}{2.352} = -0,17$$

Calcolo di a e b

	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
A	10	1,9	100	3,61	19
B	18	3,1	324	9,61	55,8
C	20	3,2	400	10,24	64
D	8	1,5	…	…	…
E	30	6,2	…	…	…
F	12	2,8	…	…	…
G	14	2,3	…	…	…
Tot.	112	21	2128	77,28	402,6

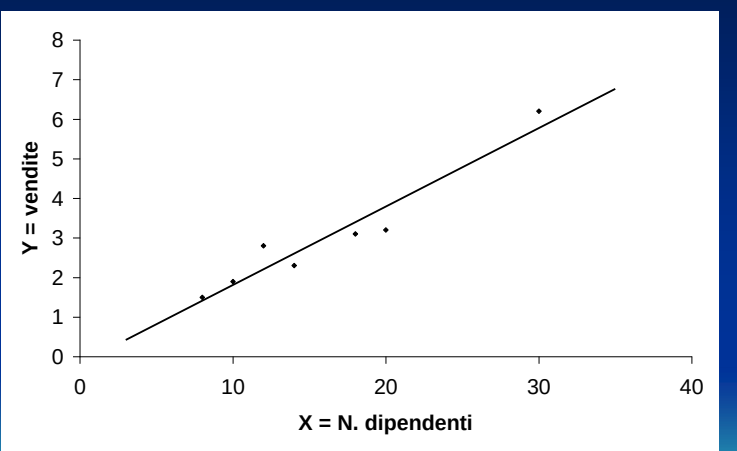
$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{7 \cdot 402,6 - 112 \cdot 21}{7 \cdot 2.128 - 112^2} = \frac{466,2}{2.352} = 0,198$$

Interpretazione dei parametri ESEMPIO (7 supermercati)

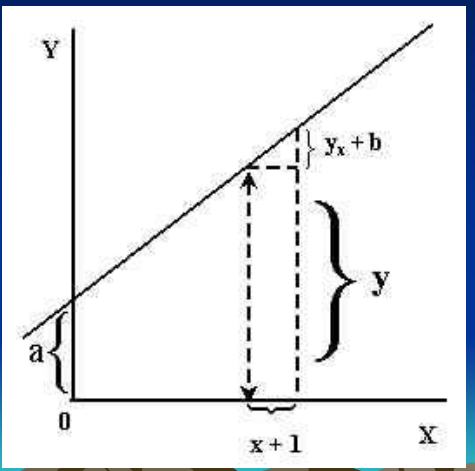
- ***a*** **= –0,17 → fatturato teorico quando N. di dipendenti = 0**
- ***b*** **= 0,198 → incremento medio nel fatturato quando il numero di dipendenti aumenta di 1 unità**

Scatter con retta di regressione



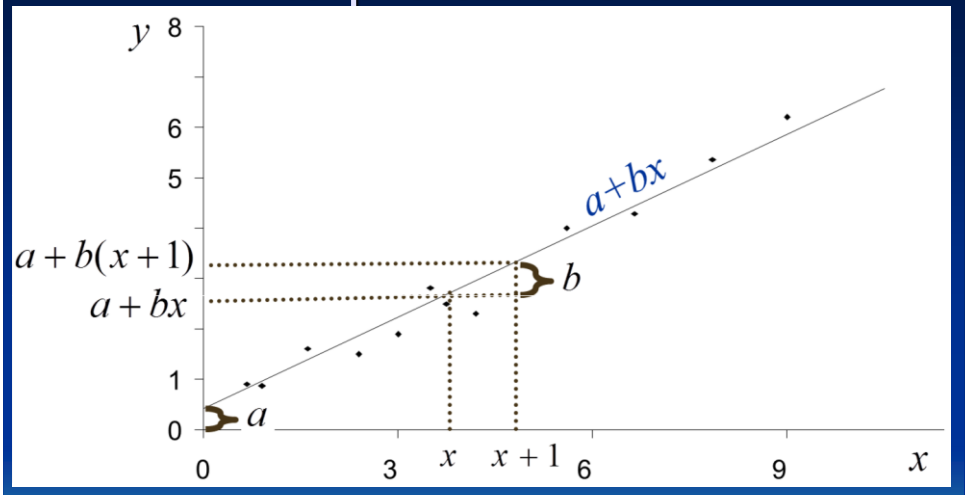
- **Come variano le vendite in funzione del numero di dipendenti?**

Interpretazione di b



- b= indica l'entità della variazione teorica della variabile dipendente in corrispondenza di un incremento unitario della variabile esplicativa

Interpretazione di b



- b= indica l'entità della variazione teorica della variabile dipendente in corrispondenza di un incremento unitario della variabile esplicativa

BONTA' DI ADATTAMENTO

- Occorre analizzare i residui $e_i = (y_i - \hat{y}_i)$

DEVIANZA RESIDUA

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

- **L'adattamento è buono quando *DEV*(*E*) è "piccola"**
- **Problemi:**
- ***DEV*(*E*) cresce all'aumentare del *numero di osservazioni* (*n*)**
- ***DEV*(*E*) dipende dall'*unità di misura* e dall'*ordine di grandezza* di *Y***

In qualsiasi modello di regressione con o senza intercetta è valida la relazione che segue

$$\sum_{i=1}^{n} y_i^{\,2} = \sum_{i=1}^{n} \hat{y}_i^{\,2} + \sum_{i=1}^{n} e_i^{\,2}$$

- **Questa relazione sfrutta la terza proprietà delle stime dei minimi quadrati (vincolo della derivata parziale rispetto a b posta uguale a 0)**

$$\sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0$$

Dimostrazione

$$y_i = a + bx_i + e_i$$

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

$$y_i^2 = (\hat{y}_i + e_i)^2$$

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

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

L'ultimo termine è zero dato che $\sum_{i=1}^{n} x_i e_i = 0$ $\sum_{i=1}^{n} e_i = 0$

Esempio supermercati (continua)

$y_i = -0{,}17 + 0{,}198x_i$

	x_i	y_i	Valori teorici	Residui	$x_i \times residuo_i$	y_i^2	(Valori teorici)^2	residui^2
A	**10**	**1,9**	1,81	0,09	0,89	3.61	3.279	0.008
B	**18**	**3,1**	3,40	-0,30	-5,34	9.61	11.536	0.088
C	**20**	**3,2**	3,79	-0,59	-11,86	10.24	14.386	0.351
D	**8**	**1,5**	1,41	0,09	0,69	2.25	2.000	0.007
E	**30**	**6,2**	5,78	0,43	12,75	38.44	33.351	0.181
F	**12**	**2,8**	2,21	0,59	7,11	7.84	4.871	0.351
G	**14**	**2,3**	2,60	-0,30	-4,25	5.29	6.779	0.092
Tot.	**112**	**21**	**21**	**0**	**0**	77.28	76.201	1.079

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

77.28=76.201+1.079

$$\sum_{i=1}^{n} y_i^{\,2} = \sum_{i=1}^{n} \hat{y}_i^{\,2} + \sum_{i=1}^{n} e_i^{\,2}$$

Indice di bontà di adattamento nei modelli di regressione senza intercetta

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

Varia nell'intervallo [0 1]

BONTA' DI ADATTAMENTO

- **Retta di regressione:** $\hat{y}_i = a + bx_i$

DEVIANZA TOTALE

$$DEV(Y) = \sum_{i=1}^{n} (y_i - M_y)^2$$

DEVIANZA DI REGRESSIONE

$$DEV(\hat{Y}) = \sum_{i=1}^{n} (\hat{y}_i - M_y)^2$$

DEVIANZA RESIDUA

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

Scomposizione della devianza di *Y* *(modelli di regressione con intercetta)*

$$DEV(Y) = DEV(\hat{Y}) + DEV(E)$$

- **Questa relazione sfrutta le Proprietà 1 e 3 delle stime dei minimi quadrati**

- **Proprietà 1**

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \Rightarrow \sum_{i=1}^{n} e_i = 0$$

- **Proprietà 3**

$$\sum_{i=1}^{n} x_i (y_i - \hat{y}_i) = 0$$

Dimostrazione

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

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

$$= DEV(\hat{Y}) + DEV(E) + 2\sum_{i=1}^{n} (\hat{y}_i - \bar{y}) e_i$$

$$= DEV(\hat{Y}) + DEV(E) + 2\sum_{i=1}^{n} \hat{y}_i e_i - 2\bar{y}\sum_{i=1}^{n} e_i$$

Indice di determinazione lineare (R^2)

$$\delta = \frac{DEV(\hat{Y})}{DEV(Y)} = 1 - \frac{DEV(E)}{DEV(Y)}$$

- $\delta = 1$ se $\sum (y_i - \hat{y}_i)^2 = 0$
- $\delta = 0$ se $\sum (\hat{y}_i - M_y)^2 = 0$

Esempio 7 supermercati (continua)

$\hat{y}_1 = -0{,}17 + 0{,}198 * 10$ Calcolo di R^2 (δ)

	x_i	y_i	$\hat{y}_i$	e_i^2	$(\hat{y}_i - M_y)^2$
A	10	1,9	1,81	0.008	1,416
B	18	3,1	3,394	0.088	0,155
C	20	3,2	3,79	0.351	0,624
D	8	1,5	1,414	0.007	…
E	30	6,2	5,77	0.181	…
F	12	2,8	2,206	0.351	…
G	14	2,3	2,602	0.092	…
Tot.	112	21	21	1,079	13,201

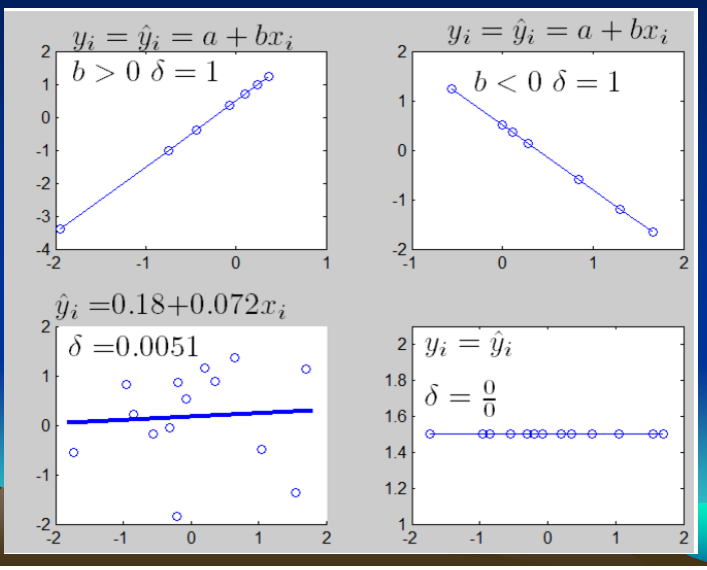
- $DEV(Y) = 7 \cdot (1{,}428)^2 = 14{,}28$
  $M_y = 3$

$Dev_{TOT} = Dev_{REGR} + Dev_{RES}$

$14{,}28 = 13{,}201 + 1{,}079$

$$\delta = \frac{13{,}201}{14{,}28} = 1 - \frac{1{,}079}{14{,}28} = 0{,}924$$

Figura 8.4 situazioni estreme per l'indice di determinazione lineare



Relazione tra indice di determinazione ***δ e coefficiente di correlazione lineare r_{xy} (p. 235)***

$$\delta = \frac{DEV(\hat{Y})}{DEV(Y)} = 1 - \frac{DEV(E)}{DEV(Y)}$$

$$r_{xy} = \frac{COV(X,Y)}{\sqrt{VAR(X)\,VAR(Y)}}$$

$$\delta = r_{xy}^2$$

Nell'esempio precedente

$$\delta = \frac{13{,}201}{14{,}28} = 1 - \frac{1{,}079}{14{,}28} = 0{,}924$$

$$\delta = (0{,}9615)^2 = 0{,}924$$

Relazione tra δ e r_{xy}

(p. 235)

$$\delta = \frac{DEV(\hat{Y})}{DEV(Y)} =$$

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

$$= \frac{\sum_{i=1}^{n}(a + bx_i - (a + b\bar{x}))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$= \frac{b^2 \sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = b^2 \frac{\mathrm{var}(X)}{\mathrm{var}(Y)}$$

$$= \frac{\mathrm{cov}(X,Y)^2}{\mathrm{var}(X)^2} \frac{\mathrm{var}(X)}{\mathrm{var}(Y)} =$$

$$= \frac{\mathrm{cov}(X,Y)^2}{\mathrm{var}(X)\,\mathrm{var}(Y)} = r_{xy}^2$$

Passi riassuntivi del modello di regressione

- Stima dei parametri (interpretazione)
- Valutazione bontà di adattamento
- Analisi dei residui
- Previsioni (estrapolazioni)

Visualizzazione grafica dei residui

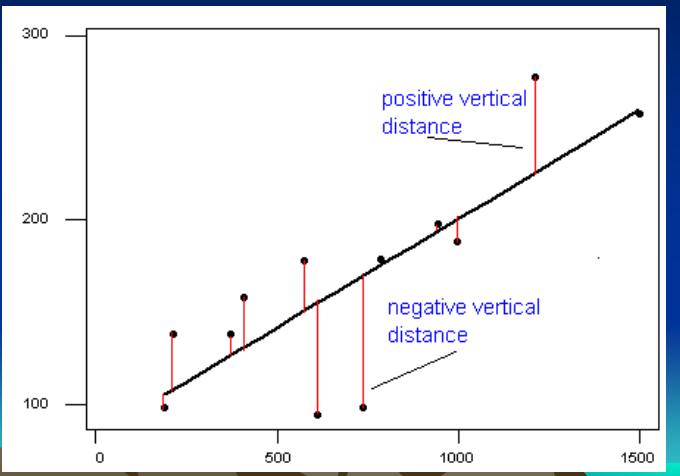
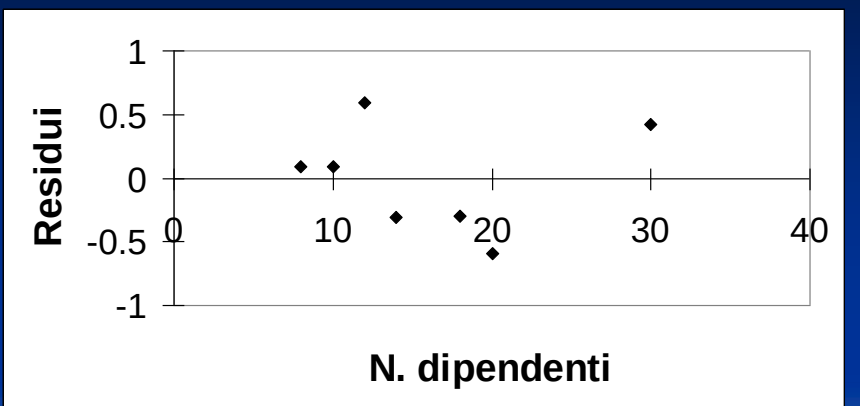


Grafico dei residui

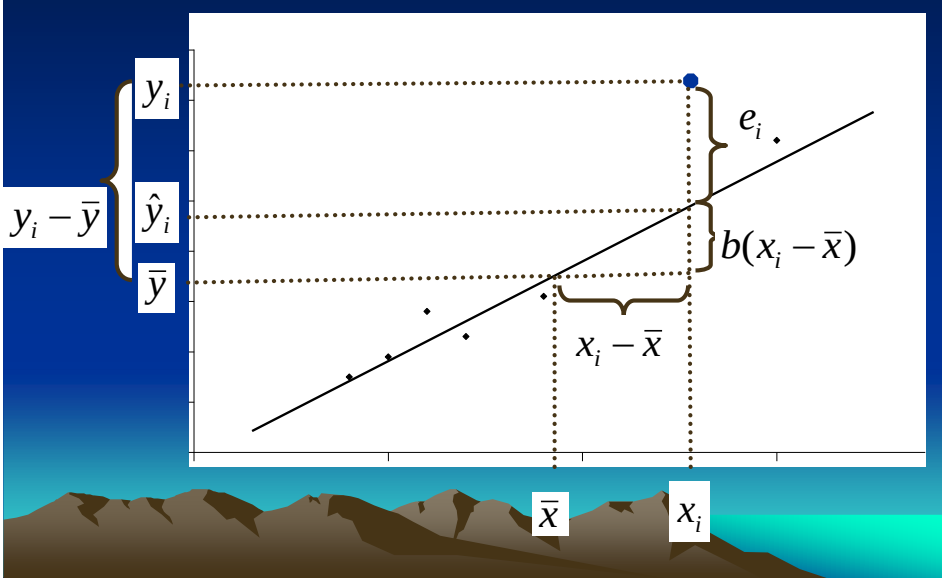


- **Modello soddisfacente: distribuzione casuale dei residui → componente erratica**

ESTRAPOLAZIONE

- **Si tenta di valutare in maniera attendibile il valore che assumerà la variabile dipendente in corrispondenza di un valore noto della variabile esplicativa.**
- ***CONDIZIONI***
  - **Validità della retta di regressione (δ prossimo ad 1)**
  - **valore noto della variabile esplicativa non lontano dai valori utilizzati nel calcolo della retta**

(Vendite, nr. di dip.) scomposizione di y_i



Stimare i parametri della retta di regressione, trovare i valori stimati, verificare i vincoli del sistema di equazioni normali e la bontà di adattamento con Excel

Funzione regr.lin

- Ordine in cui vengono restituite le statistiche aggiuntive di regressione dalla funzione di Excel 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							

Componente aggiuntivo analisi dei dati