STATISTICA BIVARIATA

Relazioni statistiche tra 2 variabili quantitative: la correlazione e la regressione

Come verificare se c'è una relazione tra due variabili?

La metodologia ancora da vedere permette di analizzare come una variabile **quantitativa** sia associata ad un'altra dello **stesso tipo** (metodi di correlazione e regressione).

A partire da un esempio...

Story Name: Age and height
Methods: Scatterplot , Regression , Correlation

Abstract: The height of a child is not stable but increases over time. Since the pattern of growth varies from child to child, one way to understand the general growth pattern is by using the average of several children's heights, as presented in this data set.
The scatterplot of height versus age is almost a straight line, showing a linear growth pattern.

The relationship between height and age provides a simple illustration of linear relationships, correlation, and simple regression.

Descrizione dei dati

Age and height

Description: Mean heights of a group of children in Kalama, an Egyptian village that is the site of a study of nutrition in developing countries.
The data were obtained by measuring the heights of all 161 children in the village each month over several years.

Number of ages (months): 12

Variable Names:

1. **Age: Age in months**
2. **Height: Mean height in centimeters for children at this age**

I dati

Age	Height
18	76.01
19	77.00
20	78.10
21	78.20
22	78.80
23	79.70
24	79.90
25	81.10
26	81.20
27	81.80
28	82.80
29	83.50

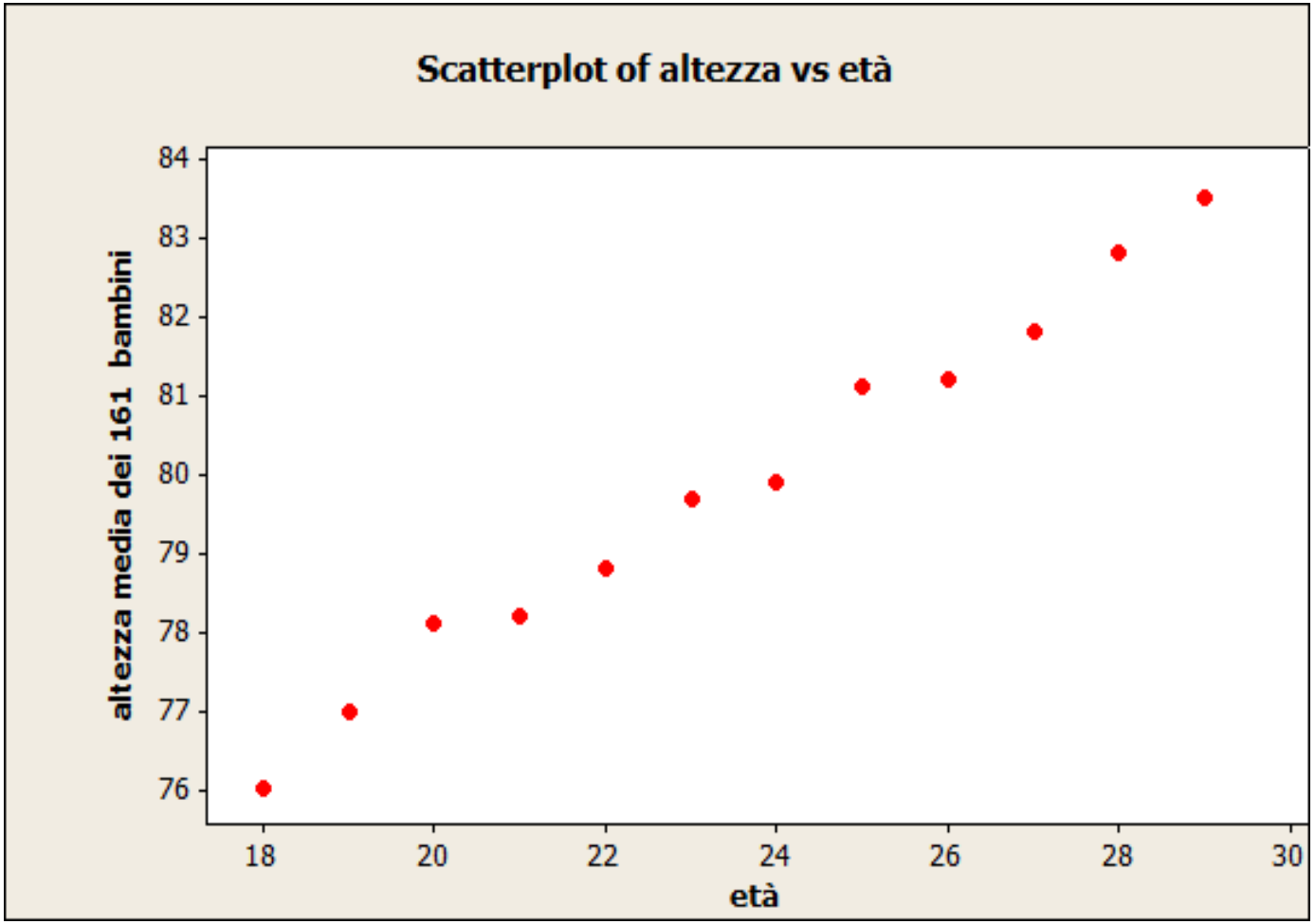
Visualizzazione: scatter plot

Per valutare *qualitativamente* l'esistenza di *associazione (correlazione)* tra due variabili quantitative, si procede ad una prima analisi grafica attraverso la costruzione di un diagramma di dispersione o scatter plot.

Si riportano in ascissa le misure osservate $x_1, x_2,\ldots,x_n$ della variabile X, in ordinata le corrispondenti misure osservate $y_1, y_2,\ldots,y_n$ della variabile Y.

Le singole osservazioni (x_i,y_j) vengono così rappresentate con dei punti su un piano cartesiano.

Lo scatter plot relativo ad età ed altezza



r = 0.994

Continuazione es.

Interpretazione del grafico precedente

- Si cerca l'andamento generale (trend)
- L'andamento generale si descrive attraverso la **forma**, la **direzione** e la **forza** che descrivono il tipo di relazione tra le 2 variabili (nel caso in esame, età ed altezza). Ossia il grafico descrive come, al variare dell'età, varia l'altezza

- La forma: i punti sono disposti attorno a una retta (in questo caso crescente). Associazione lineare
- La direzione: Associazione positiva
- La forza: Associazione forte misurata dall'indice numerico di correlazione che valuta quantitativamente l'entità dell'associazione → **r = 0.994**

Coefficiente di correlazione (Pearson)

La **correlazione** misura la **direzione e la forza** della **relazione lineare** fra due variabili quantitative. La correlazione è solitamente indicata con *r*.

Questo indice misura la tendenza di 2 v. quantitative a co-variare linearmente

Supponiamo di essere in possesso di *n* osservazioni riguardanti le variabili *x* e *y*. Le medie e le deviazioni standard delle due variabili sono $\bar{x}$ e *sx* per i valori *x* e $\bar{y}$ e *sy* per i valori *y*. Il coefficiente di correlazione *r* fra *x* e *y* è dato da

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

$-1 \leq r \leq 1$

Il coefficiente di correlazione

- Perché r non può mai essere più grande di 1?

  La correlazione più forte alla quale si può pensare è quella di una misura con sé stessa. La correlazione di x con x sarebbe

$$r = \sum \frac{(x-\bar{x}) \times (x-\bar{x}) / (n-1)}{s_x^2} = \frac{\mathrm{var}(x)}{s_x^2} = 1$$

L'indice di correlazione dipende dall'unità di misura?

Come si misura l'associazione (correlazione) tra 2 variabili quantitative?

Esempio. Effetto collaterale di un farmaco

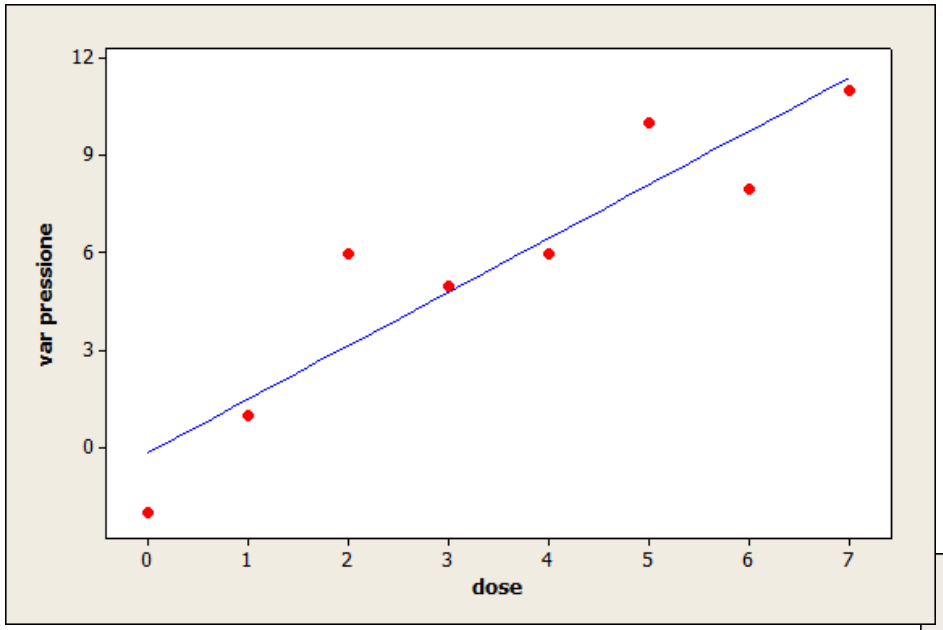
Supponiamo di sospettare che, fra gli effetti indesiderati di un certo farmaco, si annoveri quello di innalzare la pressione arteriosa. Verifichiamo questa ipotesi attraverso un esperimento: somministriamo dosi crescenti del farmaco ad alcuni ratti da esperimento, e misuriamo la variazione della pressione diastolica che si verifica dopo la somministrazione.

In dettaglio, vengono utilizzati 16 ratti, suddivisi in 8 gruppi di 2 animali ciascuno. Il primo gruppo è di controllo e non viene trattato; al secondo gruppo il farmaco viene somministrato in dose di 1 mg/kg, al terzo gruppo in dose di 2 mg/kg e così via. I risultati sono riassunti nella seguente tabella.

Variazione della pressione arteriosa (mm Hg) dopo la somministrazione del farmaco

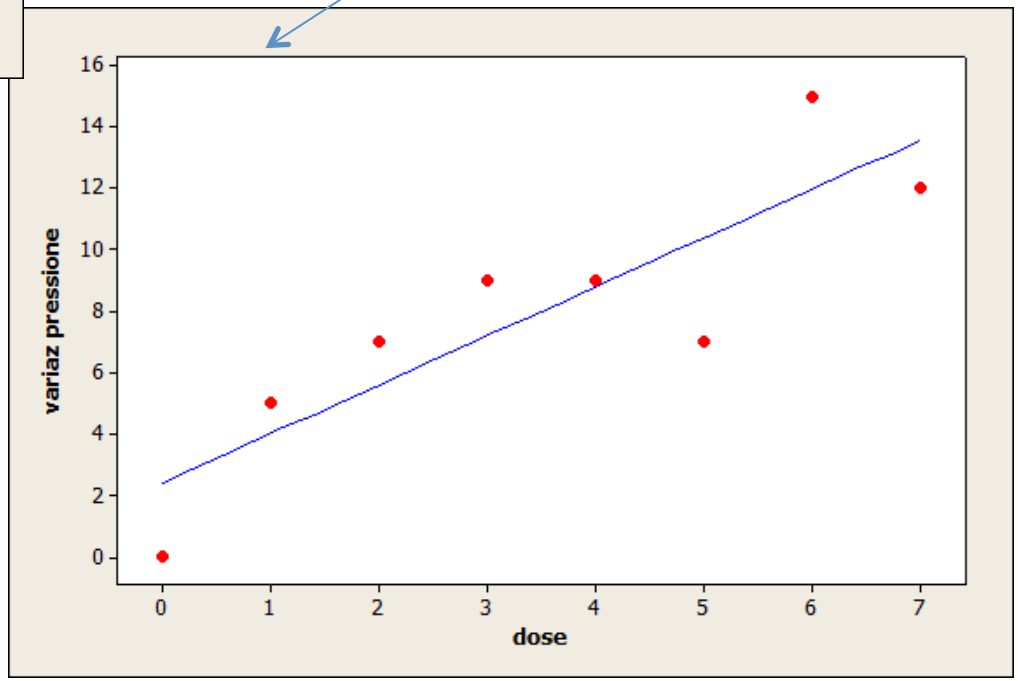
DOSE (mg/kg)	0	1	2	3	4	5	6	7
ratto 1	-2	1	6	5	6	10	8	11
ratto 2	0	5	7	9	9	7	15	12

Esempio: Effetto collaterale farmaco



Ratto 1 r= 0.927

Ratto 2 r= 0.868



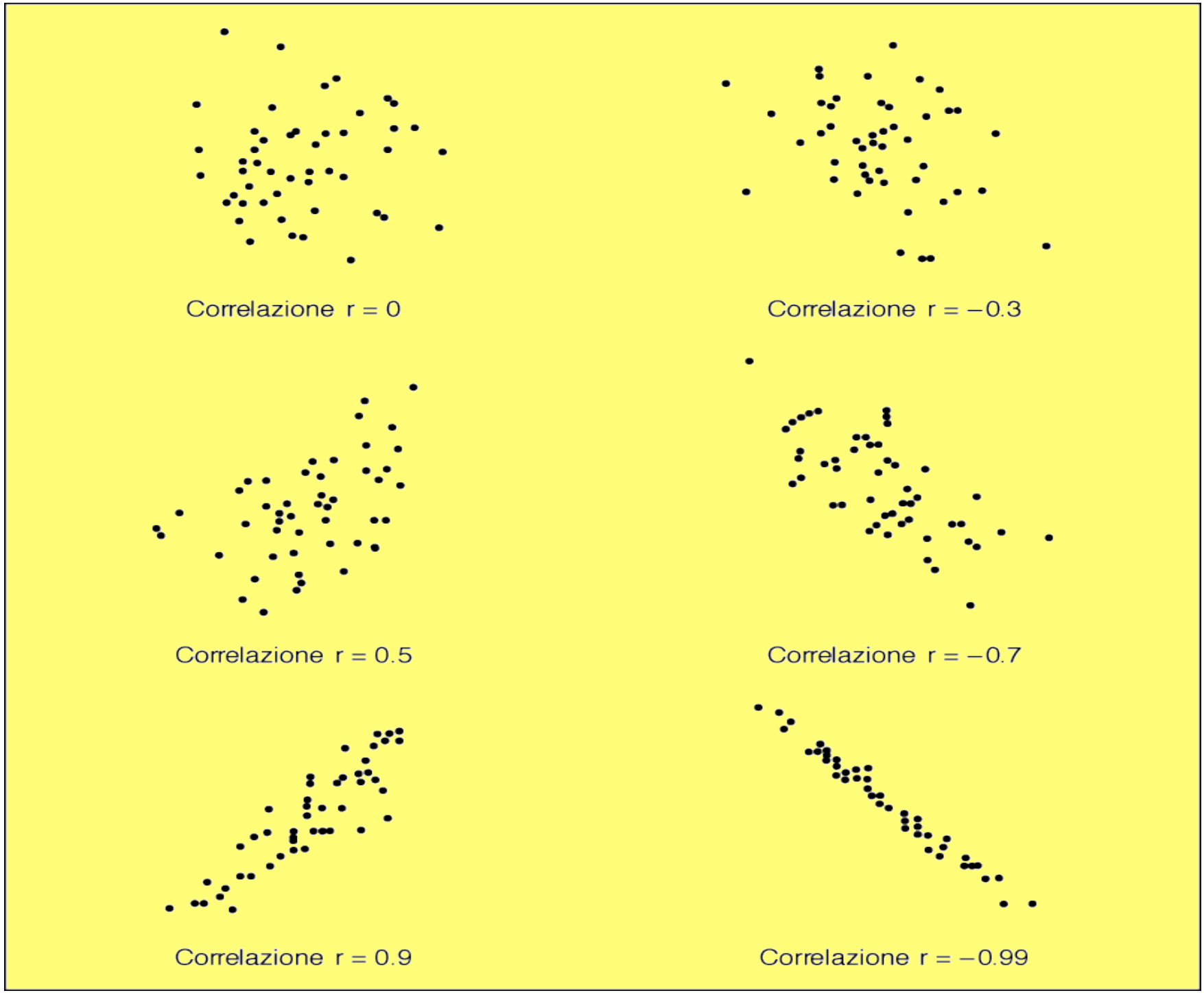
Correlazione tra 2 variabili quantitative

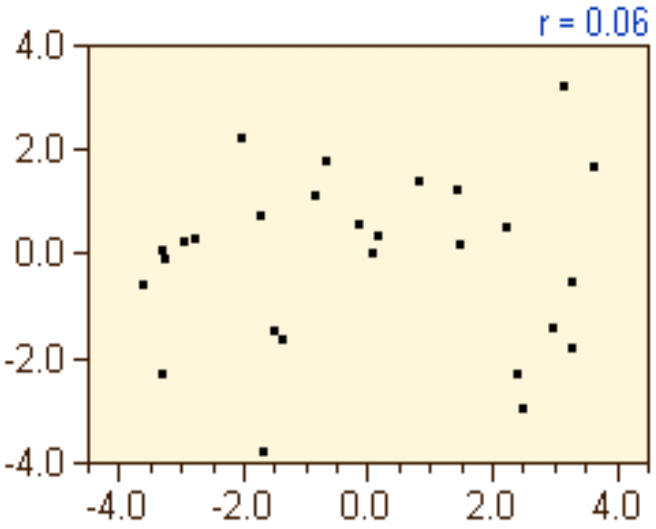
Al crescere della dose la variazione di pressione aumenta

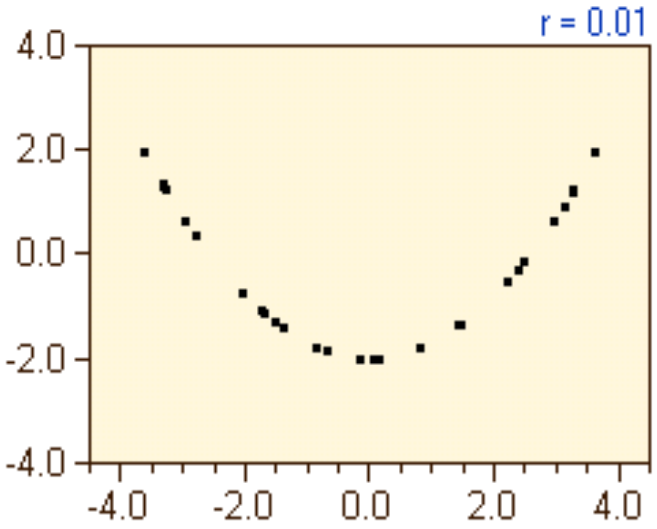
- Associazione **lineare** positiva forte in entrambi i casi (**r = 0.927, r = 0.868**)

L'andamento generale è piuttosto regolare, si può pensare ad un **modello** (matematico, statistico) per descrivere tale andamento.

- Potremmo usare il modello lineare suggerito dal grafico per predire la variazione di pressione al variare della dose del farmaco.

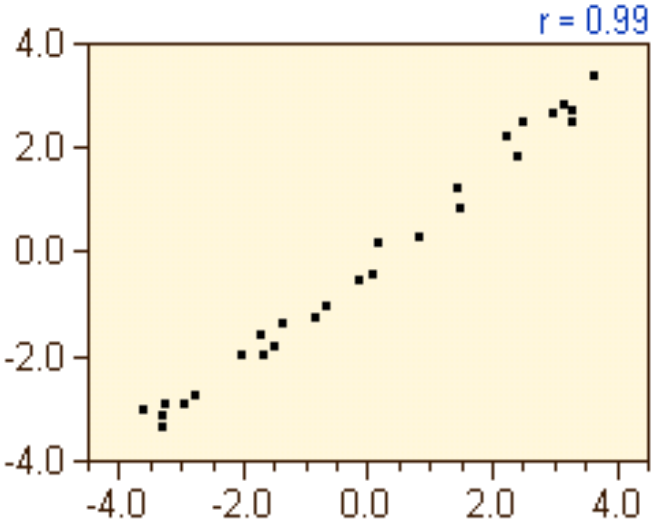

Correlazione r = 0
Correlazione r = −0.3
Correlazione r = 0.5
Correlazione r = −0.7
Correlazione r = 0.9
Correlazione r = −0.99

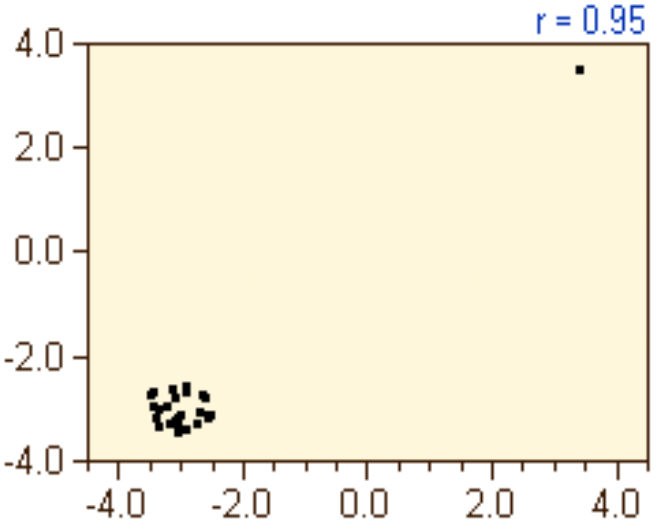




Attenzione!!

Osservate bene I grafici





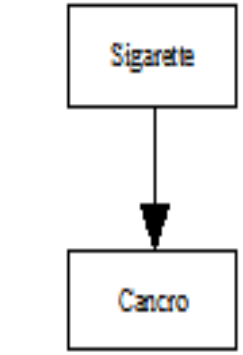
La correlazione

- Una correlazione positiva **non** indica che una variabile influenza l'altra, ma che le 2 variabili evolvono nello stesso senso; la causa di un comune andamento può essere un fatto non preso in considerazione dall'analisi.

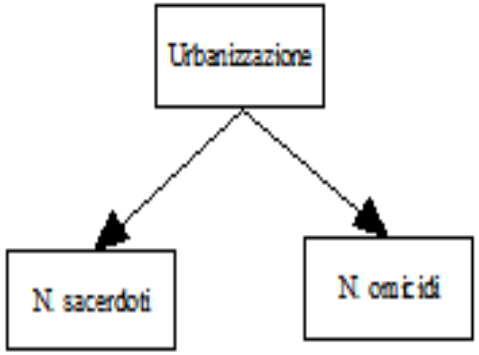
- Ad es., se in un indagine statistica si trova che il n° di figli per famiglia e il consumo di alcool pro capite per famiglia sono correlati positivamente, questo non vuol dire che una variabile influenza direttamente l'altra, ma per esempio si può ipotizzare che la causa comune siano le condizioni economiche e culturali delle famiglie.

Correlazione spuria

- *Un rapporto di correlazione può essere "spuria": esso non implica necessariamente un rapporto di causa e effetto*
- *Esempio: correlazione fra il numero di sacerdoti e il numero di omicidi*

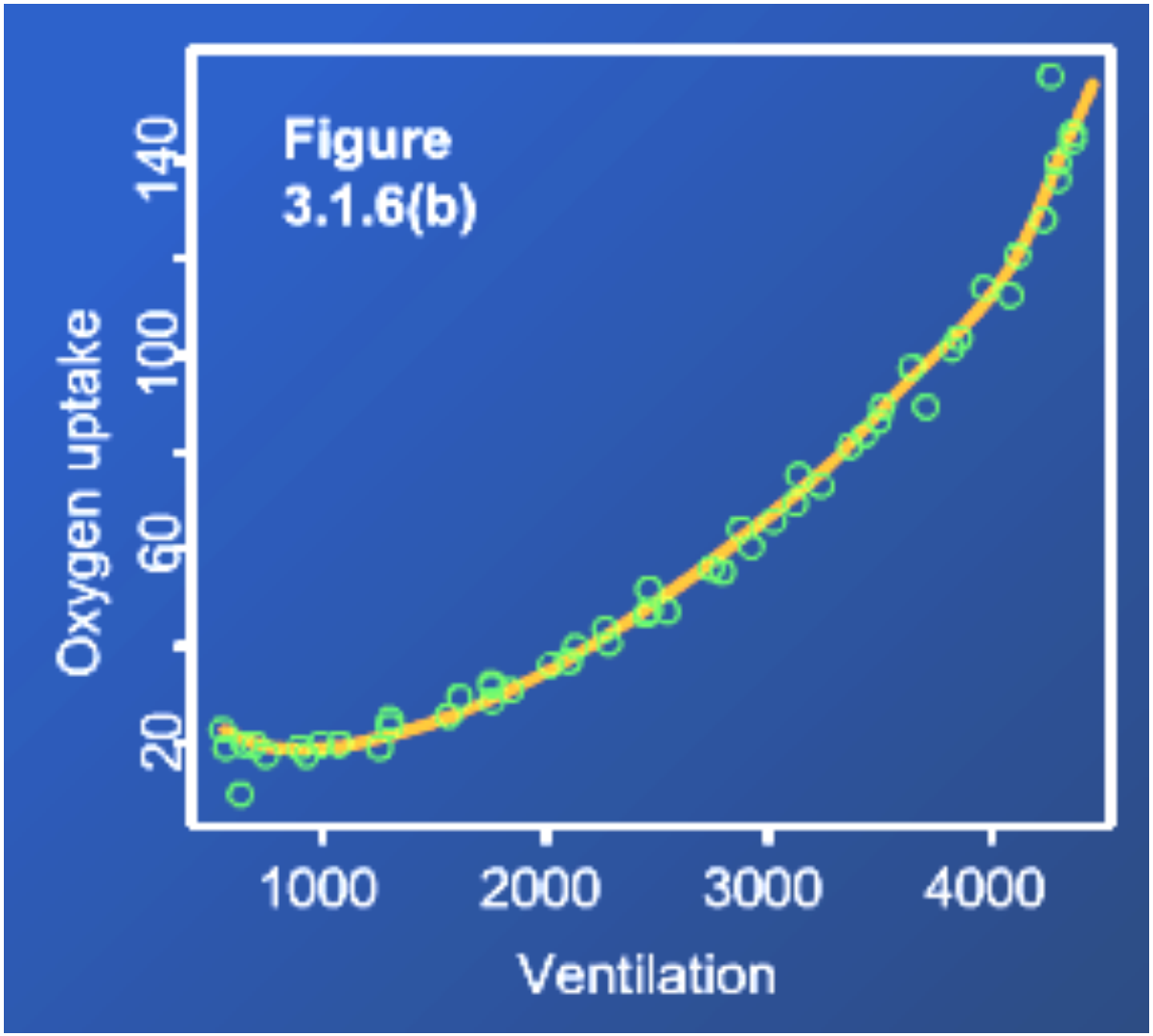


Rapporto causa e effetto



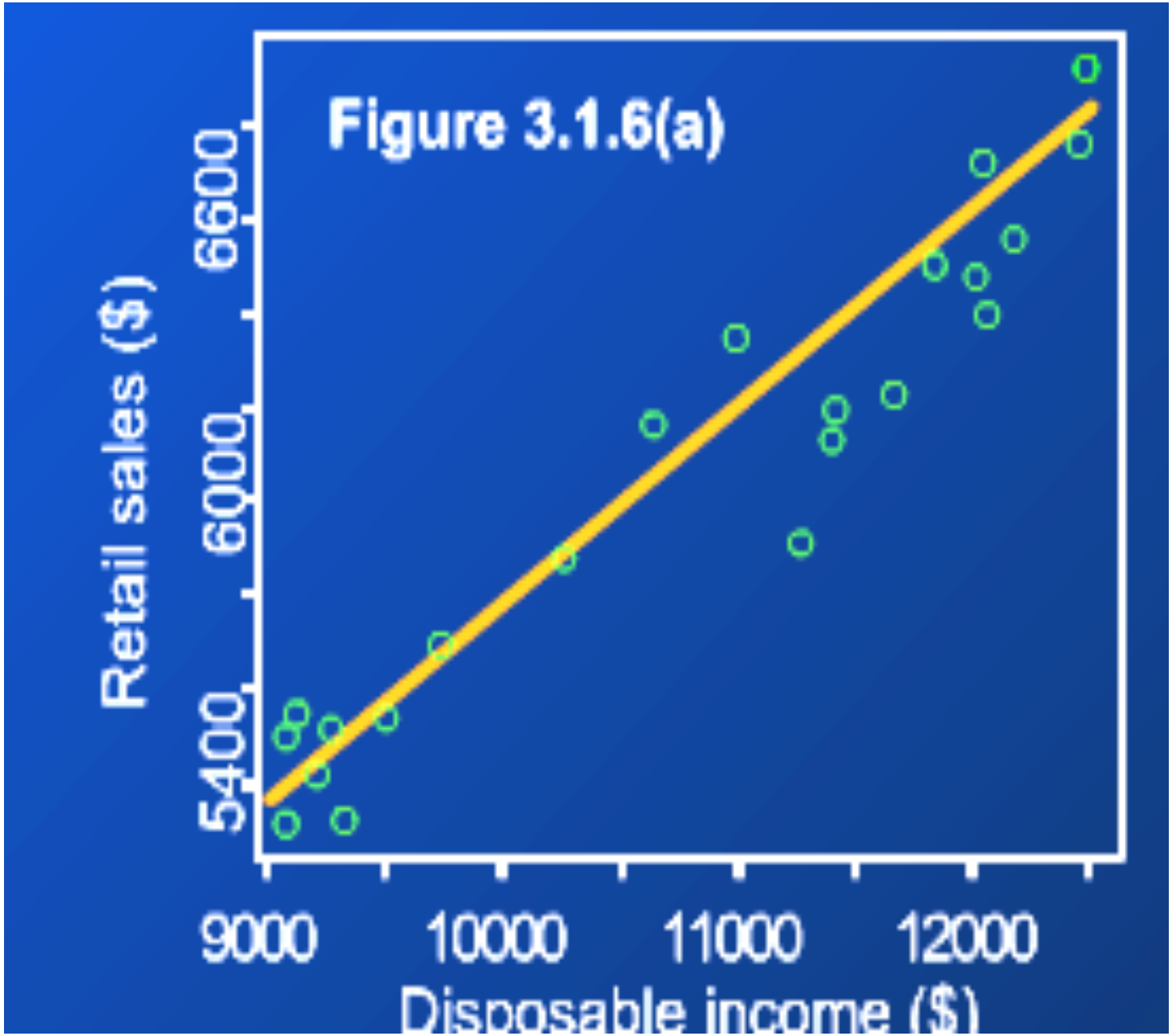
Correlazione spuria

Cosa rivela un diagramma di dispersione?



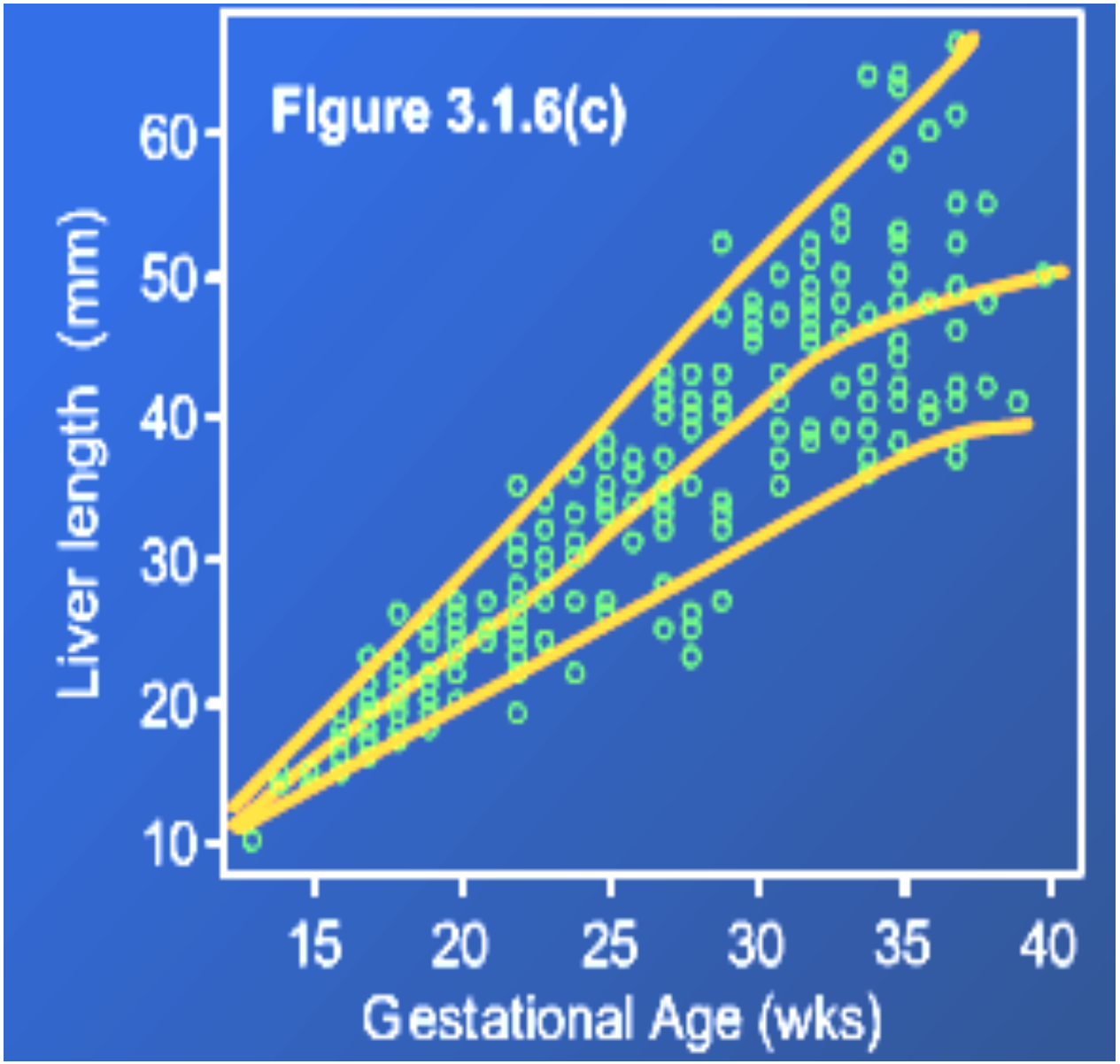
Trend non lineare con poca dispersione dei dati intorno alla curva
Relazione forte

Cosa rivela un diagramma di dispersione?



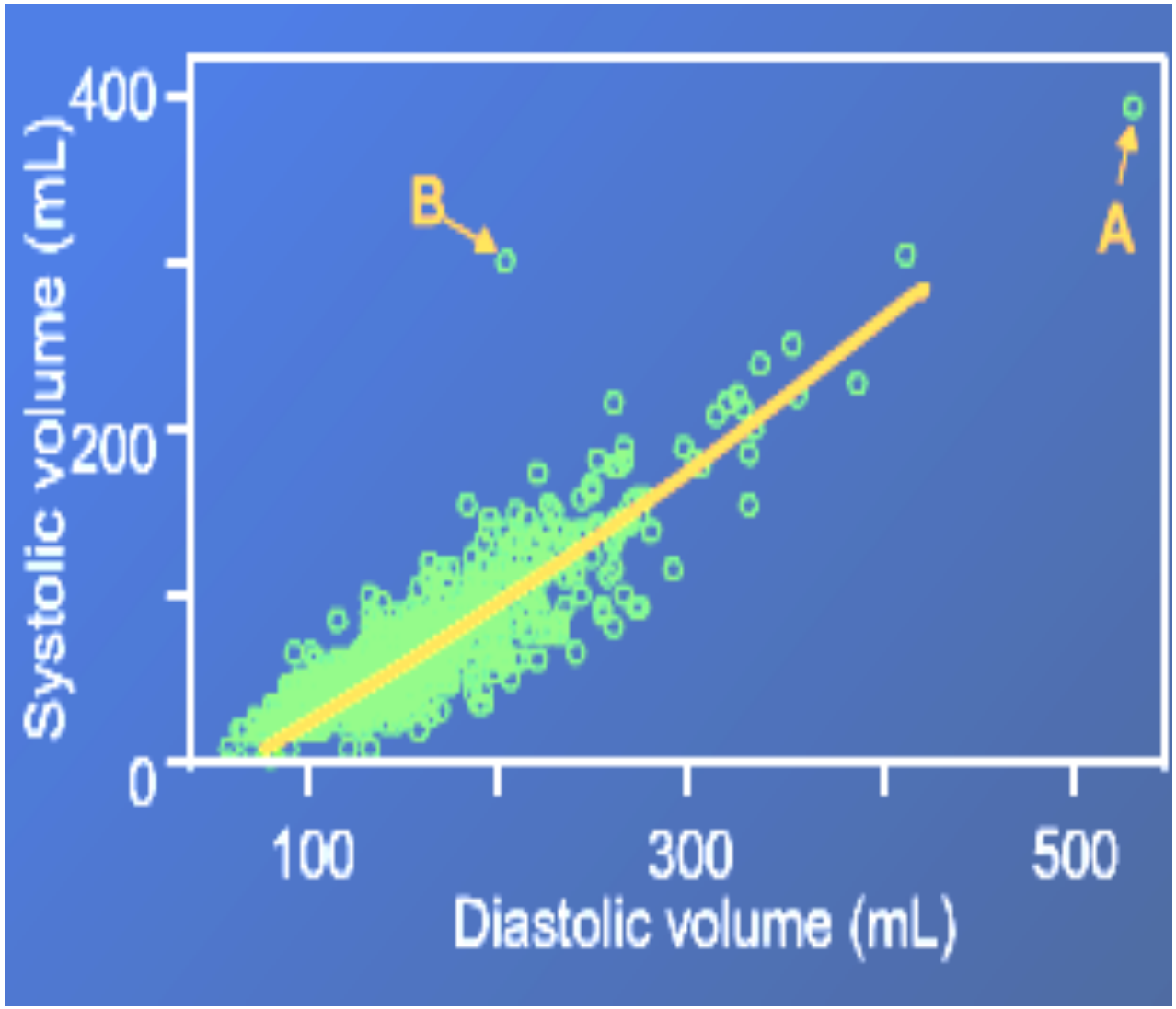
Trend lineare con una dispersione moderata e costante lungo la linea di tendenza

Cosa rivela un diagramma di dispersione?



Trend non lineare con dispersione non costante intorno alla curva
Relazione debole

Cosa rivela un diagramma di dispersione?



A e B sono outlier

Dopo un controllo B si è rivelato un errore mentre A è sembrato un valore possibile

Attenzione all'uso della correlazione

- 1) La relazione tra le 2 variabili deve essere **rettilinea (lineare)** non curvilinea
- 2) Se r=0 (correlazione lineare non significativa) può esistere una relazione non lineare tra le variabili.
- 3) La correlazione **non** è una misura **robusta**.
- 4) Non si può concludere che poiché 2 variabili sono correlate in modo significativo, una è necessariamente **la causa** dell'altra.
- Un fattore nascosto può essere la causa della relazione delle 2 variabili.
- 5) I coefficienti di correlazione che abbiamo visto, negli esempi, **stimano la correlazione nella popolazione** da cui sono estratti i campioni.

Esempio

Si vuole indagare sull'utilità di una ninfa (Plecoptera nymph) come indicatore di fattori ambientali nei corsi d'acqua.

Esiste una correlazione significativa tra durezza dell'acqua e numero di ninfe?

Per un campione di 13 corsi d'acqua si ha:

Durezza dell'acqua ($CaCO_3$ unità): **x**

17 20 22 28 42 55 55 75 80 90 145 145 170

Numero di ninfe: **y**

42 40 30 7 12 10 8 7 3 7 5 2 4

Si può usare il coeff. di correlazione?

Esempio di correlazione positiva

Si vogliono fare inferenze sulla dieta delle foche e uccelli marini a partire dagli otoliti dei pesci mangiati e ritrovati nel cibo rigurgitato o nelle feci.

Si cerca una correlazione tra la lunghezza degli otoliti e la massa del pesce da cui provengono. Un campione casuale di 10 pesci di una particolare specie viene pesato, sezionato e vengono misurati gli otoliti.

lunghezza otoliti (mm): x

6.6 6.9 7.3 7.5 8.2 8.3 9.1 9.2 9.4 10.2

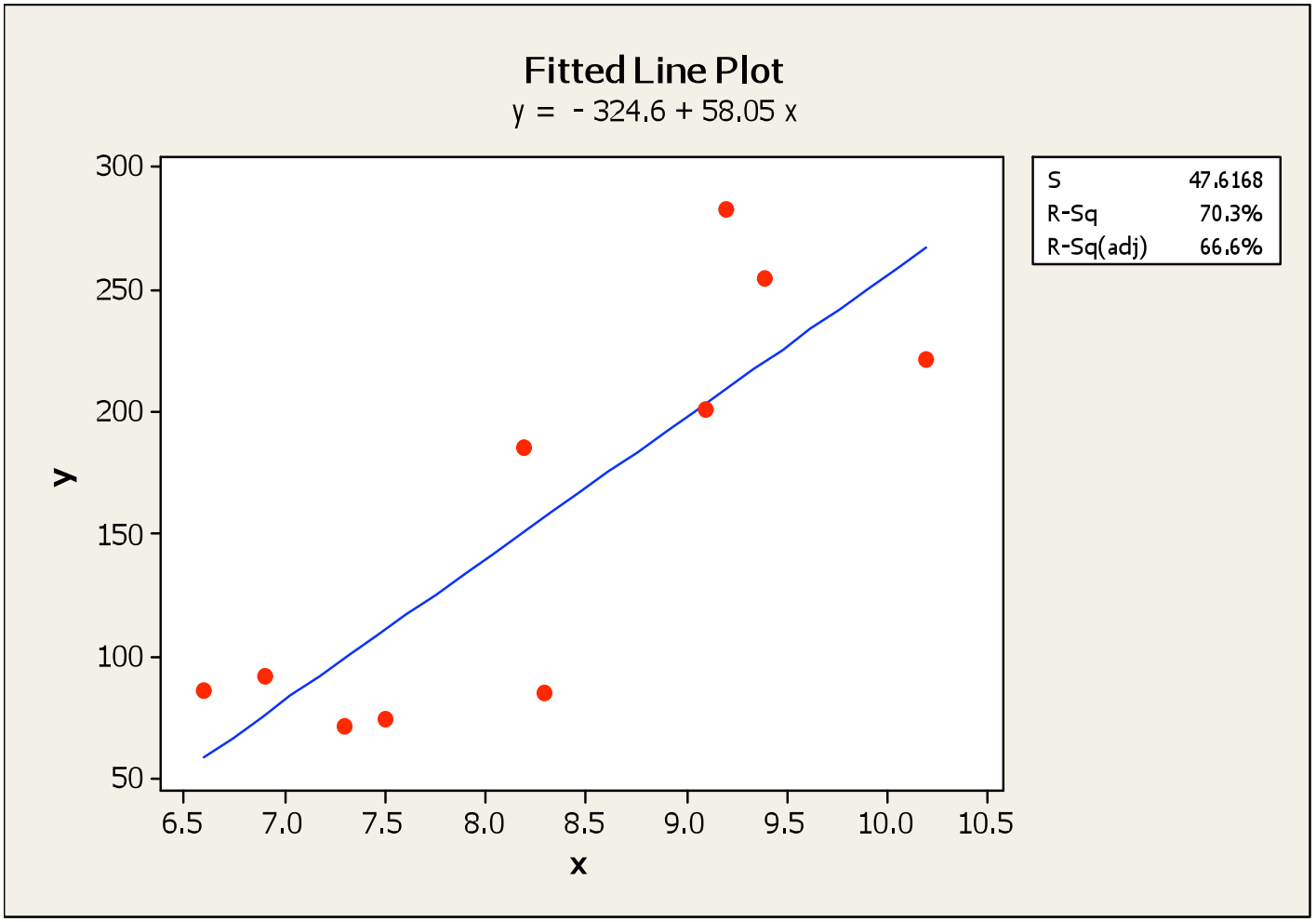
massa del pesce (g): y

86 92 71 74 185 85 201 283 255 222

Non si può stabilire qual è la v. esplicativa e quale di risposta

Esempio (continua)

Coeff. di correlazione di Pearson: 0.83



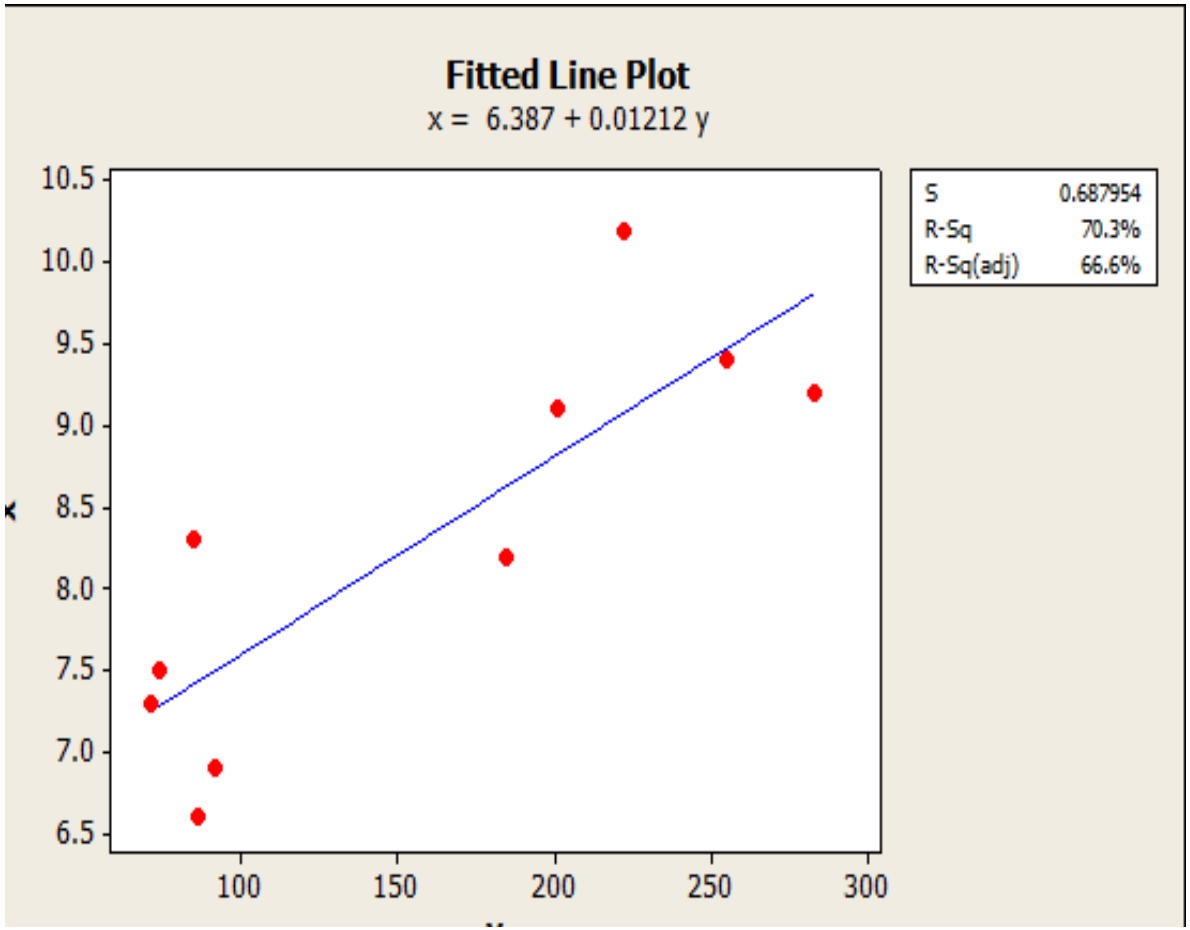
X: lunghezza otoliti

Y: massa del pesce

Esempio (continua)

I valori sugli assi sono scambiati

Coeff. di correlazione di Pearson è uguale ossia: 0.83



X: massa del pesce

Y: lunghezza otoliti

NOTA: malgrado ci sia un unico valore di correlazione tra massa e lunghezza, le rette di regressione sono diverse scambiando la x e la y.

Quale modello per i dati osservati negli esempi?

- E' possibile descrivere la relazione tra queste coppie di dati viste nei diversi esempi facendo uso di un modello statistico?
- Se lo scatterplot evidenzia che i punti sono disposti attorno a una **retta crescente** o **decrescente**, si parla di **correlazione lineare**.
- In tal caso si può tracciare una **retta di regressione** (interpolante) a partire dai dati.

- Perciò si può pensare a un **modello lineare**.
- Si potrà usare per fare previsioni sulla variazione della y al variare di x …….

Quale modello per questi dati?

- Si può pensare a un **modello lineare negli esempi considerati**.

- Si potrà usare, nell'es. sul farmaco, per predire la variazione di pressione per dosi di farmaco differenti da quelle considerate nell'es.
- Si potrà usare, nell'es. dei pesci, per predire la massa del pesce a partire dalla misura degli otoliti

Regressione lineare

- **Retta di regressione. Modello statistico che descrive la relazione lineare tra due variabili quantitative**
- Una **retta di regressione:**
- descrive come cambia (linearmente) una variabile di **risposta *y*** quando cambia la **variabile esplicativa *x*,**
- spesso viene usata per prevedere nuovi valori di *y da nuovi* valori di *x*,
- determina quanta parte della variabilità (incertezza) di y può essere spiegata dalla relazione lineare con x, e quanta di questa variabilità resta non spiegata.

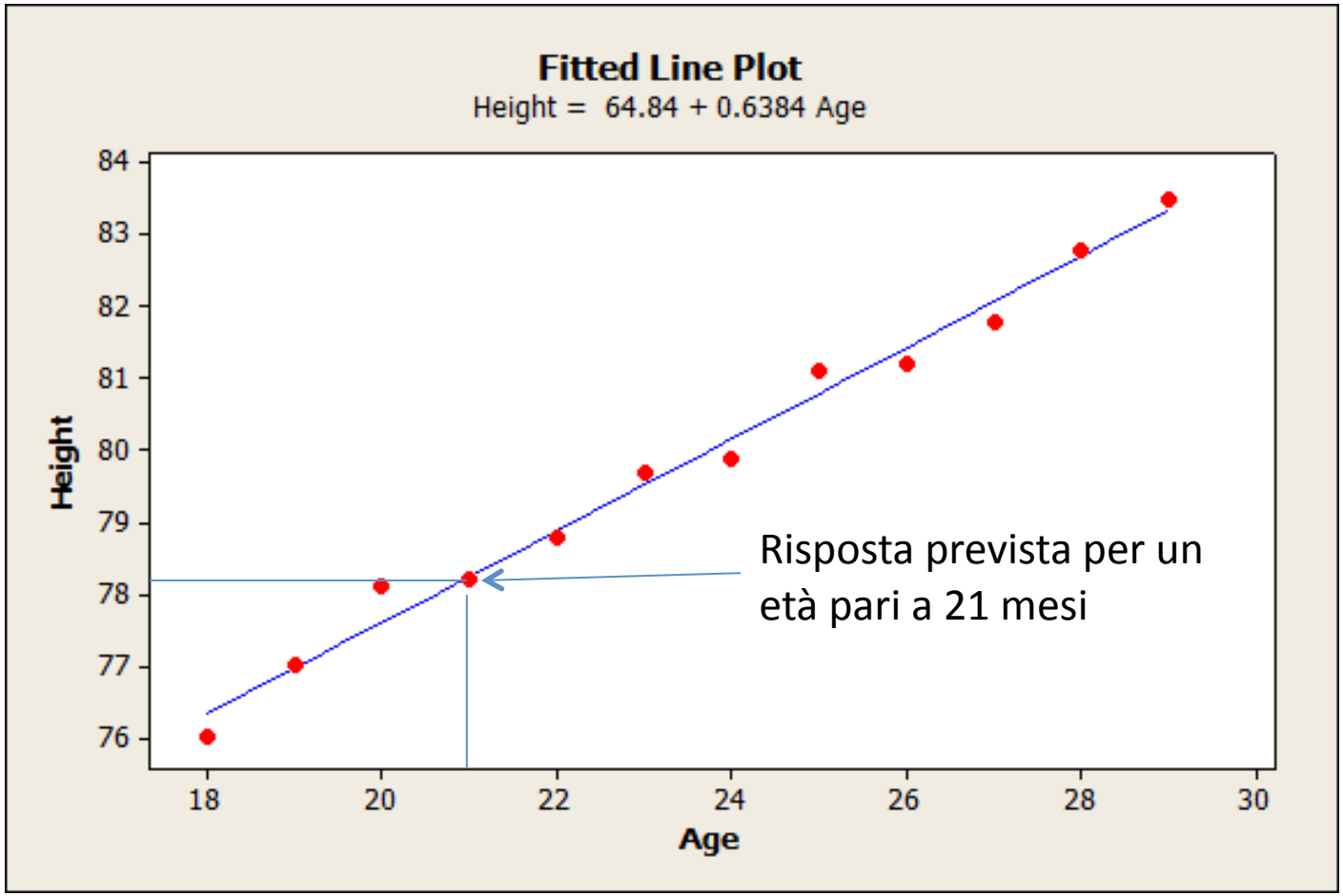
Es precedente: età e altezza

Variabile esplicativa

Age	Height
18	76,01
19	77,00
20	78,10
21	78,20
22	78,80
23	79,70
24	79,90
25	81,10
26	81,20
27	81,80
28	82,80
29	83,50

Variabile di risposta

La retta interpolante
Es: età, altezza



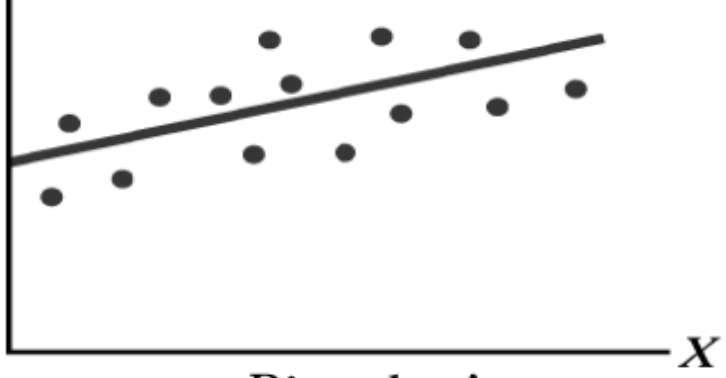
r = 0.994

Significato della retta

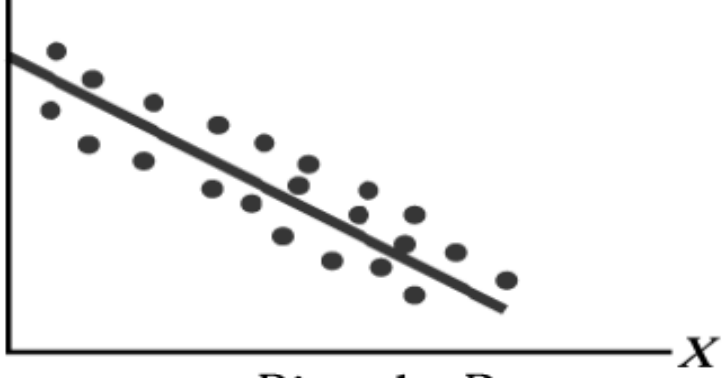
- La retta che interpola i punti sperimentali rappresenta il **modello dei dati** e permette di prevedere, data l'età di un bambino, quale dovrebbe essere in media la sua altezza.

- Se osserviamo altezze molto lontane dalla retta, abbiamo a che fare con dei casi anomali.

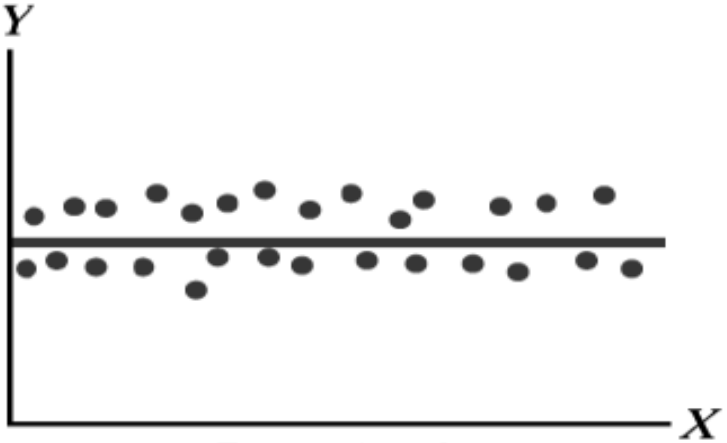
- Si noti che nel contesto della regressione occorre scegliere quale variabile è esplicativa.



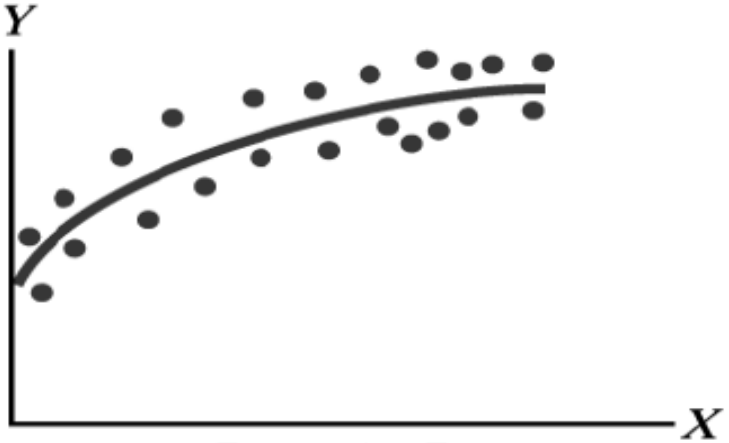
Riquadro A
Esempio di relazione lineare diretta



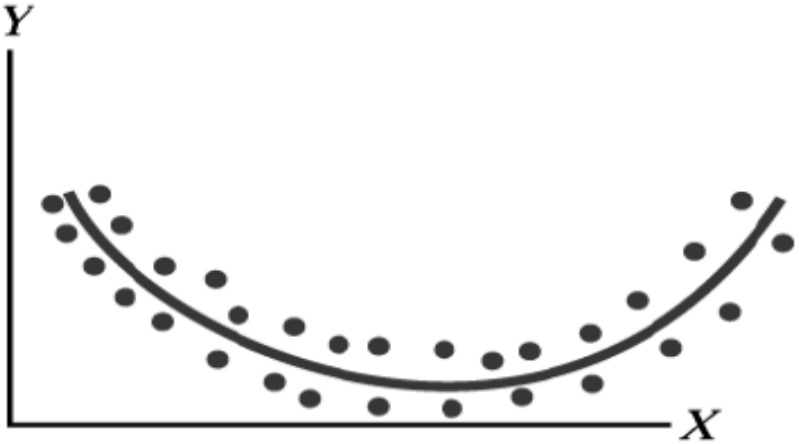
Riquadro B
Esempio di relazione lineare inversa



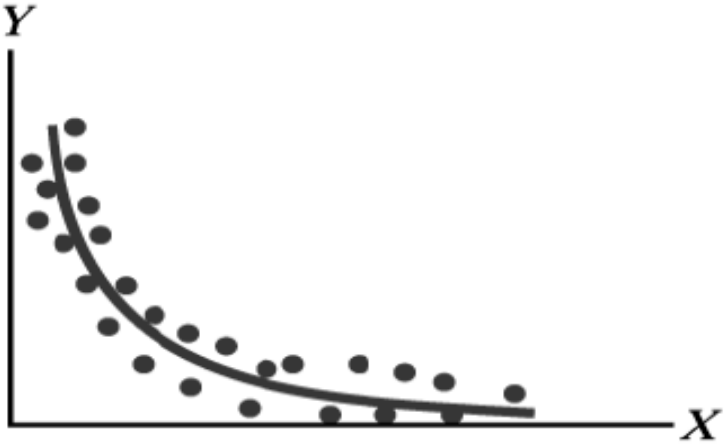
Riquadro C
Nessuna relazione tra X e Y



Riquadro D
Esempio di relazione polinomiale diretta



Riquadro E
Esempio di relazione curvilinea a U



Riquadro F
Esempio di relazione polinomiale inversa

La retta dei minimi quadrati

- Poiché la relazione espressa dal modello di regressione lineare semplice consiste nell'equazione di una retta, si tratterà di trovare la retta che **meglio approssima** i punti osservati.
- Secondo il metodo dei minimi quadrati si sceglie la "migliore" retta di regressione minimizzando la somma dei quadrati delle **distanze verticali** tra i punti osservati e la retta stessa. Tali distanze sono dette

residui o errori di previsione.

I residui sono anche chiamati errori stimati.

Esempio dei topi

Esempio. Effetto collaterale di un farmaco

Supponiamo di sospettare che, fra gli effetti indesiderati di un certo farmaco, si annoveri quello di innalzare la pressione arteriosa. Verifichiamo questa ipotesi attraverso un esperimento:

somministriamo dosi crescenti del farmaco ad alcuni ratti da esperimento, e misuriamo la variazione della pressione diastolica che si verifica dopo la somministrazione.

In dettaglio, vengono utilizzati 16 ratti, suddivisi in 8 gruppi di 2 animali ciascuno. Il primo gruppo è di controllo e non viene trattato; al secondo gruppo il farmaco viene somministrato in dose di 1 mg/kg, al terzo gruppo in dose di 2 mg/kg e così via. I risultati sono riassunti nella seguente tabella.

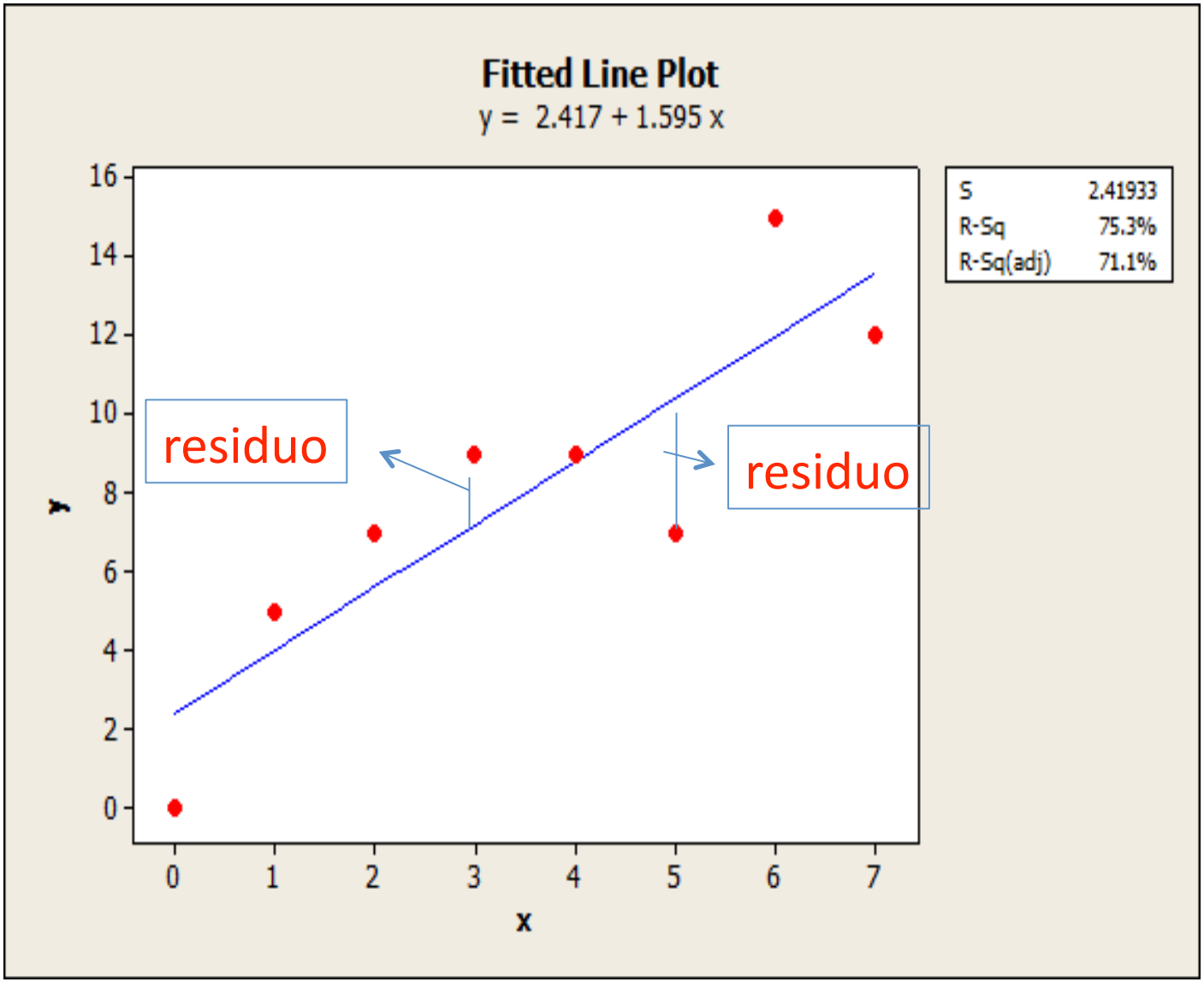
Variazione della pressione arteriosa (mm Hg) dopo la somministrazione del farmaco

DOSE (mg/kg)	0	1	2	3	4	5	6	7
ratto 1	-2	1	6	5	6	10	8	11
ratto 2	0	5	7	9	9	7	15	12

Solo 2° gruppo di

Topi trattati

x	y
0	0
1	5
2	7
3	9
4	9
5	7
6	15
7	12



Qual è la risposta osservata? Qual è la risposta prevista?

Residui

Il **residuo** è la differenza fra un valore osservato della variabile di risposta e il valore previsto dalla retta di regressione. Vale a dire

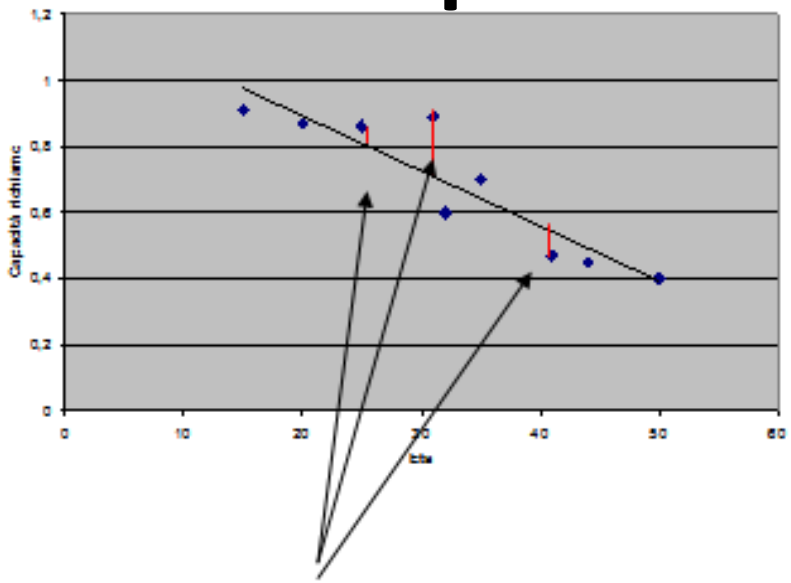
residuo = y osservato-y previsto

$= y - \hat{y}$

La somma dei residui dei minimi quadrati è pari a zero e perciò la loro media è sempre zero

Scarti quadratici (o **Residui al quadrato**)

- ***Alcune distanze sono "positive"***
- ***Alcune sono negative***
- ***L'ottimizzazione minimizza la somma degli scarti quadratici*** **(o residui al quadrato)**



Distanza verticali

$$\sum (Y - \hat{Y})^2$$

La somma degli scarti o residui è pari a zero

Retta di regressione dei minimi quadrati: formule

Supponiamo di avere dei dati su una variabile esplicativa *x* e su una variabile di risposta *y* per *n* unità. In base ai dati, ricaviamo le medie $\bar{x}$ e $\bar{y}$ e le deviazioni standard *sx* e *sy* delle due variabili e la loro correlazione *r*. La retta di regressione dei minimi quadrati è la linea

$$\hat{y} = a + b\,x$$

con **coefficiente angolare**

$$b = r\,\frac{s_y}{s_x}$$

e **intercetta**

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

La retta di regressione dei m. q.

- Il coefficiente angolare misura quanto cambia $\hat{y}$ quando x aumenta di 1 e dipende dalle unità di misura di x e y. COME???????

- Ossia, il coefficiente angolare misura quanto cambia $\hat{y}$ per un cambiamento unitario di X.

L'intercetta è il valore di $\hat{y}$ quando x=0.

Regressione lineare

- Si noti che il termine "lineare" si riferisce alla combinazione dei parametri, non alla forma della relazione. Nessun parametro appare all'esponente o è moltiplicato o diviso per un altro parametro.
- $E(y) = \alpha + \beta x^2$
- $E(y) = \beta_0 + \beta_1 x + \beta_2 x^2$

  questi sono esempi di modelli statistici lineari.
- E(y) è funzione lineare dei parametri incogniti

  α e β, ma non necessariamente una funzione lineare di x.

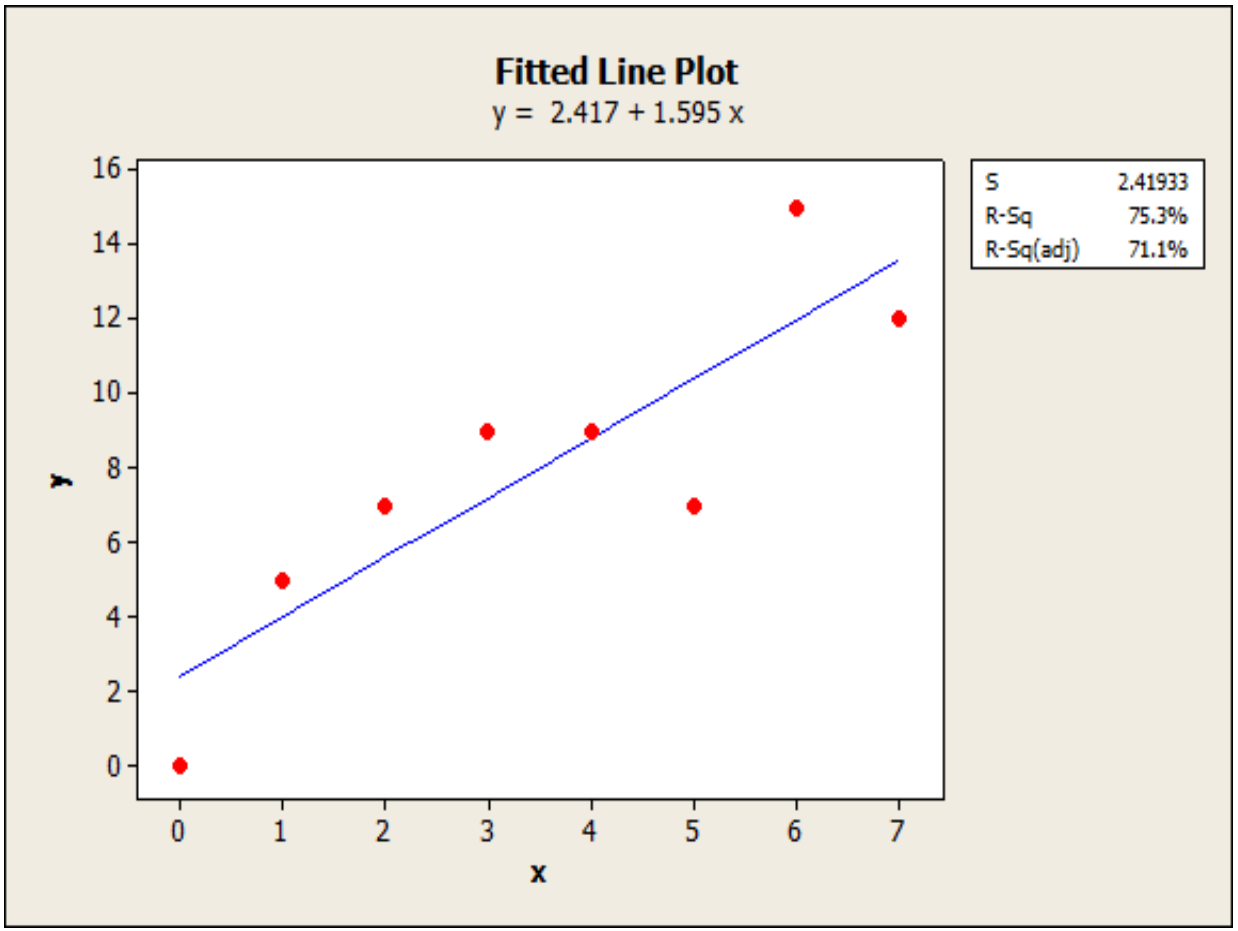
Regressione: continuazione esempio (solo ratto 2)

$\hat{y} = 2.417 + 1.595x$

Variazione pressione = 2.417+1.595
dose farmaco

r= 0.868

x	y
0	0
1	5
2	7
3	9
4	9
5	7
6	15
7	12



La retta di regressione: esperimento sui ratti (**solo ratto 2**)

- La retta di regressione dei minimi quadrati è data da:

$$\hat{y} = 2.417 + 1.595x$$

- Il coefficiente angolare misura quanto cambia $\hat{y}$ quando x aumenta di 1 e dipende dalle unità di misura di x e y.
- Nell'esempio, b = 1.595 ci dice che per ogni dose di 1mg/kg di farmaco in più è possibile prevedere circa 1.595mmHg di variazione di pressione
- L'intercetta è il valore di $\hat{y}$ quando x=0.
- Nell'esempio, x=0 equivale a dose pari a 0, con valore

$\hat{y}$=2.417 (poco significato)

- Si possono fare previsioni con la retta…….

La retta di regressione dei m. q.

- Si possono fare **previsioni** con la retta.
- Uno dei motivi più importanti per adattare un modello di regressione ai dati è la possibilità di prevedere il valore della variabile di risposta in corrispondenza di un particolare valore della variabile esplicativa.

Regressione: previsioni (**solo 2° ratto**)

$$\hat{y} = 2.417 + 1.595x$$

Variazione pressione = 2.417+1.595 dose farmaco

Previsioni:

Quale sarà l'incremento di pressione somministrando 5.5mg del farmaco?

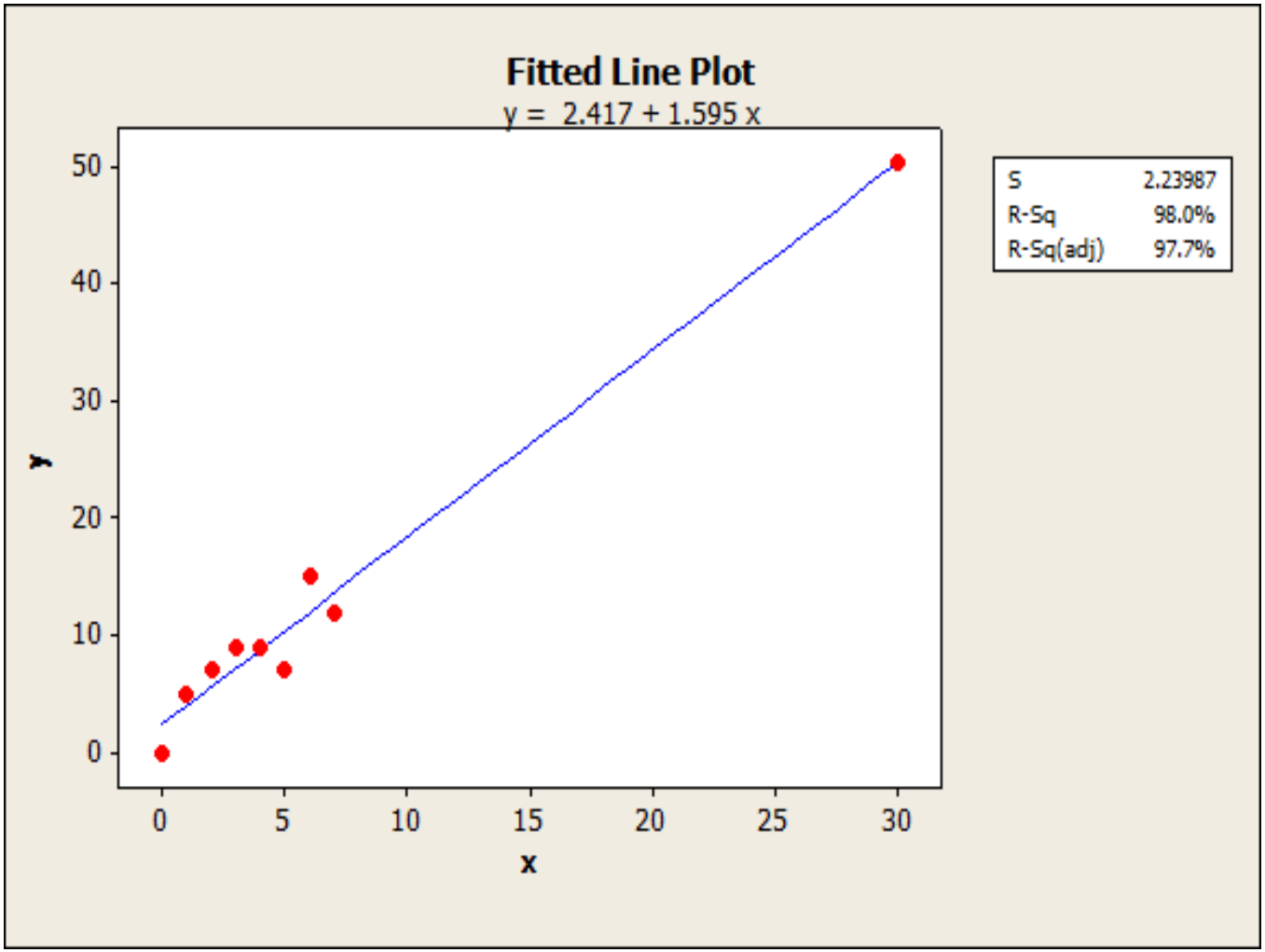
$\hat{y}$= 2.417+ (1.595*5.5) = 11.19 mm Hg

E somministrando **30**mg? 50.27 mm Hg →

→ (**outlier estremo)**

Regressione: Esempio (**solo ratto 2**)

$\hat{y}$ = 2.417 + 1.595x30→ $\hat{y}$ = 50.27



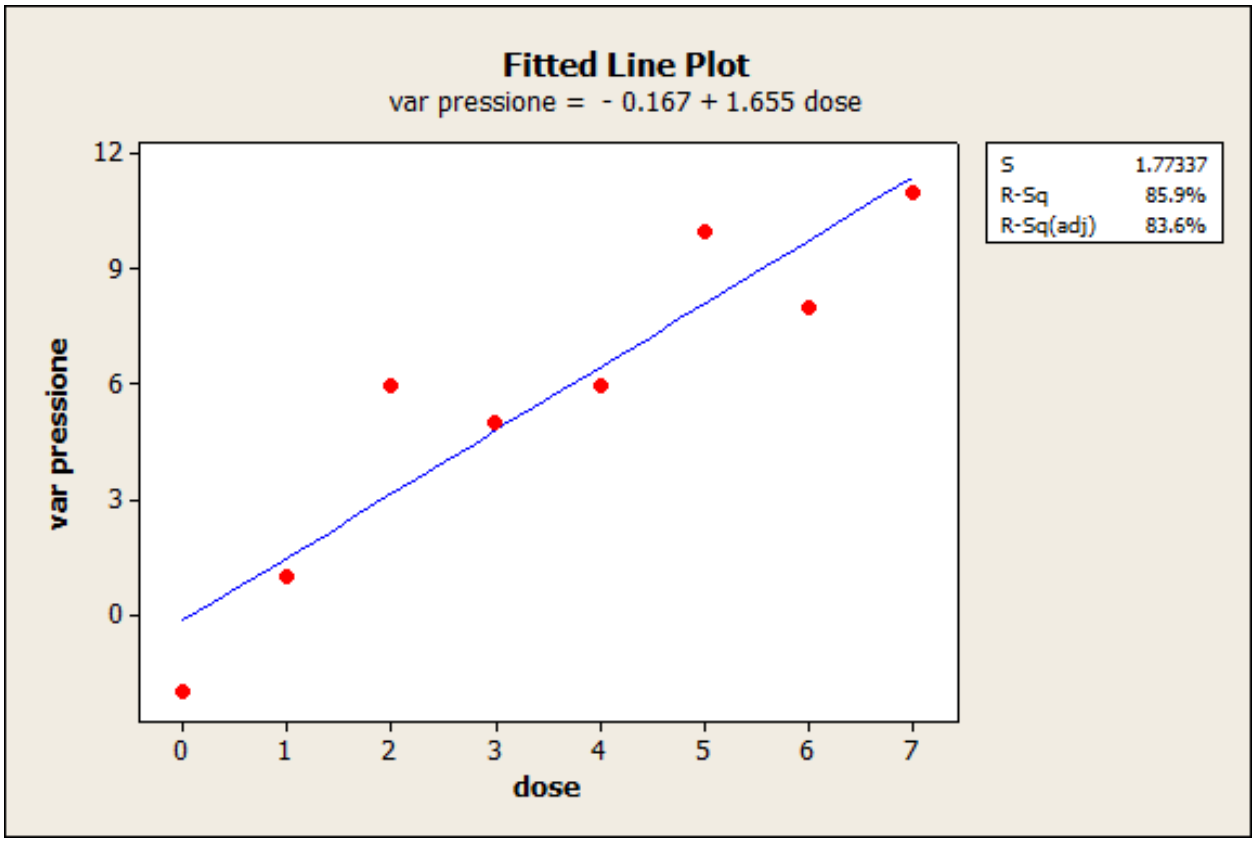
r = 0.99

Attenzione a Interpretare correttamente r osservando il grafico

Regressione: Esempio (Ratti)

$$\hat{y} = -0.17 + 1.65x$$

Variazione pressione = -0.17+1.65 dose farmaco



Solo
I° ratto

r= 0.927

La retta di regressione dei m. q.: esperimento sui topi

$$\hat{y} = -0.17 + 1.65x \text{ (I ratto)}$$

- Esiste una correlazione positiva fra la dose del farmaco e l'innalzamento della pressione diastolica **(I ratto)**

- la correlazione è abbastanza forte, infatti
  r = 0.927 e b = 1.65mm/Hg/mg/kg

- in assenza del farmaco la variazione di pressione è prossima a 0

- Si tratta di un esperimento → il farmaco può provocare un aumento della pressione diastolica nel ratto

Regressione: previsioni (**solo 1° ratto**)

$$\hat{y} = -0.17 + 1.65x$$

Quale sarà l'incremento di pressione somministrando 5.5mg del farmaco?

$\hat{y}$ = -0.17+ (1.65*5.5) = 8.93

E somministrando **30**mg? → 49.47mm Hg (**outlier estremo)**

Es precedente: età e altezza

TAB. 1

Variabile esplicativa

Age	Height
18	76,01
19	77,00
20	78,10
21	78,20
22	78,80
23	79,70
24	79,90
25	81,10
26	81,20
27	81,80
28	82,80
29	83,50

Variabile di risposta

Come si valuta la bontà del modello? r^2

- Per valutare la bontà del modello si può considerare r^2
- r^2 è la frazione di variabilità dei valori di y spiegata dalla retta di regressione dei **minimi quadrati** di y su x.

Nell'esempio delle età, $r^2 = 0.988$, ossia il **98.8%** della variabilità delle altezze nella tabella 1 è spiegata tramite la relazione lineare con le età.

Osservando il grafico di dispersione, si vede che le altezze y variano da 76 a 83.5cm . Gran parte di questa variabilità di y è spiegata dal fatto che l'età varia da 18 a 29 mesi.

Nel contesto della **regressione dei m. q.** ci sono 2 fonti di variabilità per la risposta y.

Come si valuta la bontà del modello? r^2

- Nel contesto della regressione è possibile dividere la **variabilità (varianza) di y osservato** in 2 parti (A e B) .
- A) le altezze variano con le età. Quando l'età x cresce da 18 a 29 mesi si trascina dietro l'altezza y lungo la retta di regressione. La retta spiega questa parte di variabilità delle altezze
- B) le altezze osservate non giacciono esattamente sulla retta, ma sono sparse sopra e sotto. La retta non può spiegare questa parte di variabilità delle altezze y.
- **Usiamo r^2 per misurare la parte A della variabilità come frazione della variabilità totale delle altezze.**
- Tutto ciò può essere spiegato numericamente ma non lo faremo.

Come si valuta la bontà del modello: il coefficiente di determinazione

$$r^2 = \frac{\text{variabilità delle } \hat{y} \text{ trascinate dalle } x \text{ lungo la retta}}{\text{variabilità totale dei valori osservati } y} =$$

$$= \text{coefficiente di determinazione}$$

Ma la **parte B della variabilità** al di sopra o al di sotto della retta **non può essere spiegata dalla relazione lineare tra x e y**.

Nell'esempio: **$r^2 = 0.988$**.

1% circa della variabilità totale è la variabilità interna delle altezze che **non** è spiegabile dalla relazione lineare.

Nell'esempio dei topi ($r = 0.927$, $r^2 = 0.859$).

Più dell' 85% della variabilità della pressione è spiegata dal modello lineare.

Il 15% è la variabilità fra le variazioni di pressione (**non** spiegata dal modello)

Come si valuta la bontà del modello?

- Ma esiste anche una certa variabilità al di sopra o al di sotto della retta che **non può essere spiegata dalla relazione lineare tra x e y**.
- Nell'esempio: $r = 0.994$ e $r^2 = 0.99$. 1% della variabilità totale è la variabilità interna delle altezze che **non** è spiegabile dalla relazione lineare.

- Nell'esempio dei topi($r = 0.927$, $r^2 = 0.859$).
- Più dell' 85% della variabilità della pressione è spiegata dal modello lineare.
- il 15% è la variabilità fra le variazioni di pressione
- r^2 (che varia tra 0 e 1) **non** è il fattore più importante per valutare la bontà del modello. Si possono osservare valori di r^2 vicini a 1 in regressioni in cui l'analisi dei residui mostra l'inadeguatezza del modello.

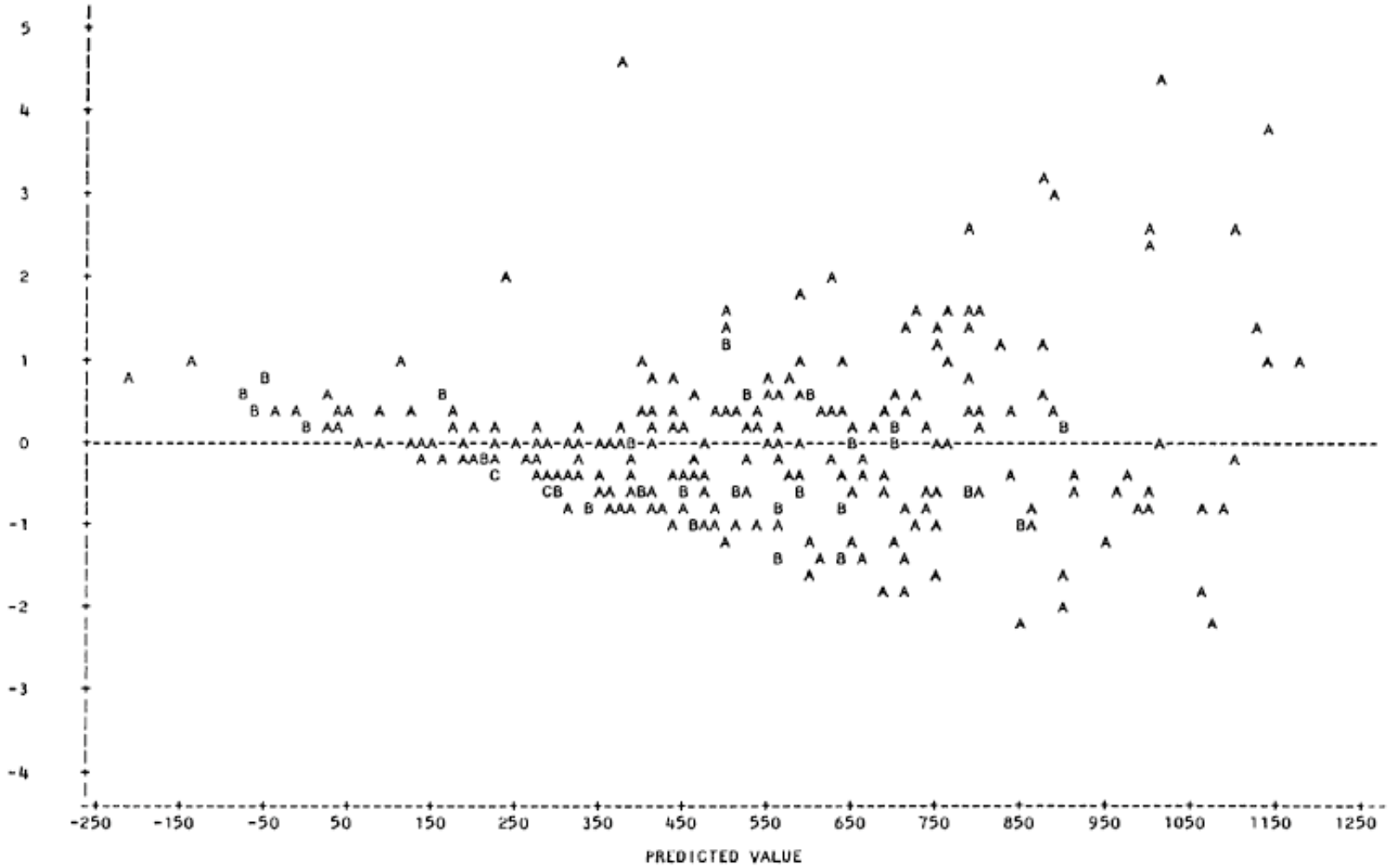
Valutazione della bontà del modello

- **r^2** (che varia tra 0 e 1) **non** è il fattore più importante per valutare la bontà del modello. Si possono osservare valori di r^2 vicini a 1 in regressioni in cui l'analisi dei residui mostra l'inadeguatezza del modello.
- Per controllare l'adeguatezza del modello, ossia per valutare l'adattamento ai dati della retta di regressione si può osservare il grafico dei residui.
- Tale grafico è un diagramma a dispersione con i residui sull'asse delle y e la variabile esplicativa x oppure i valori $\hat{y}$ sull'asse delle x.
- Amplifica ogni scostamento rispetto alla linearità

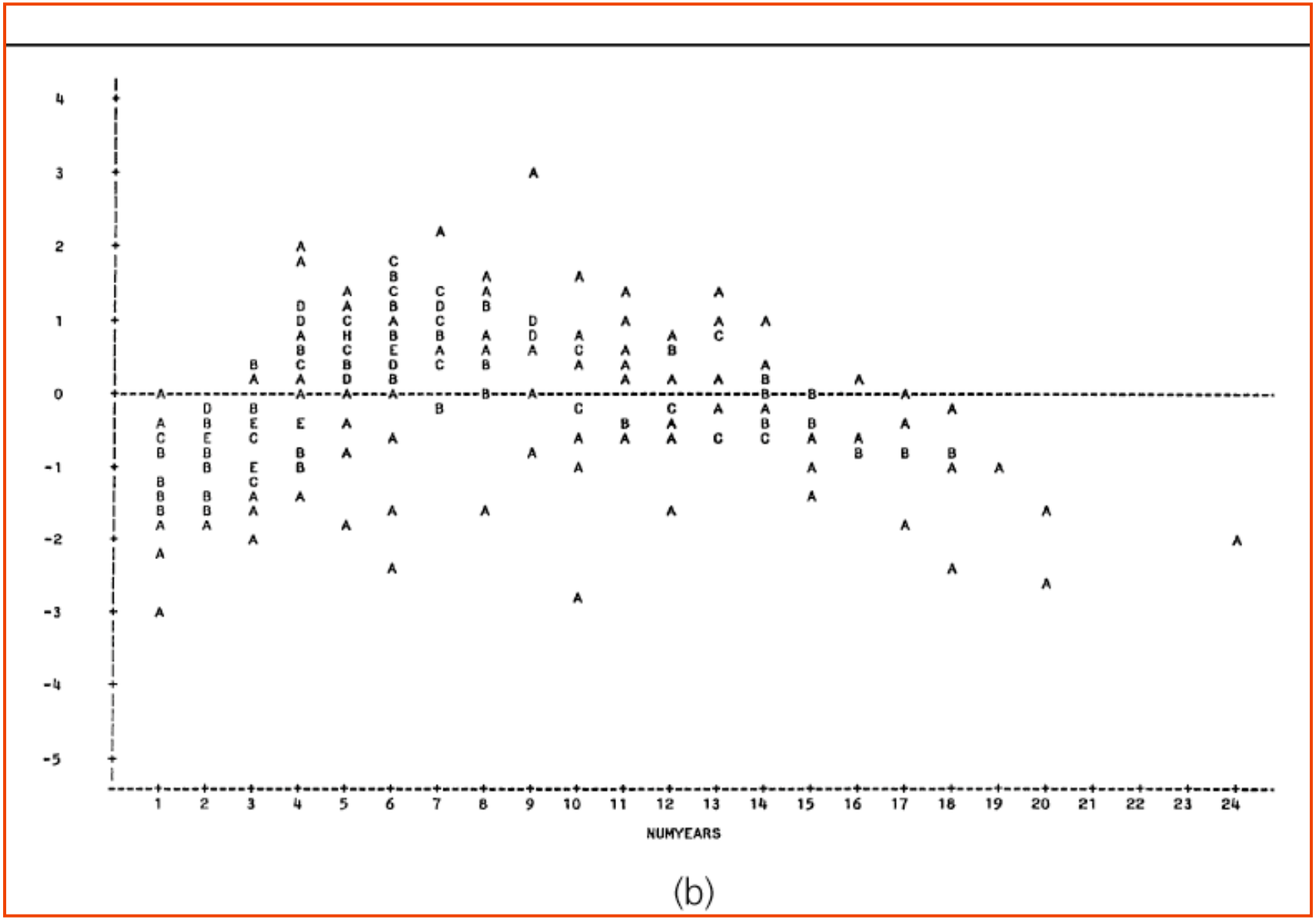
I residui

- Perché ci sia un ***buon adattamento del modello*** ai dati vogliamo che:
- i residui non individuino alcun andamento ulteriormente interpolabile con termini di ordine superiore.
- il segno dei residui sia "casuale", sia cioè, in qualche modo, ripartito equamente tra + e-, per escludere errori sistematici
  - rappresentando su di un grafico i punti (x_i, r_i) se la relazione tra **X** e **Y** è lineare dovremmo osservare dei punti che oscillano casualmente sopra e sotto lo 0.

La variabilità dei residui non è costante, ma aumenta all'aumentare di x.



(a)



(b)

I residui si dispongono lungo una curva: probabilmente la dipendenza della y dalla x non è lineare.

Regressione lineare e R^2

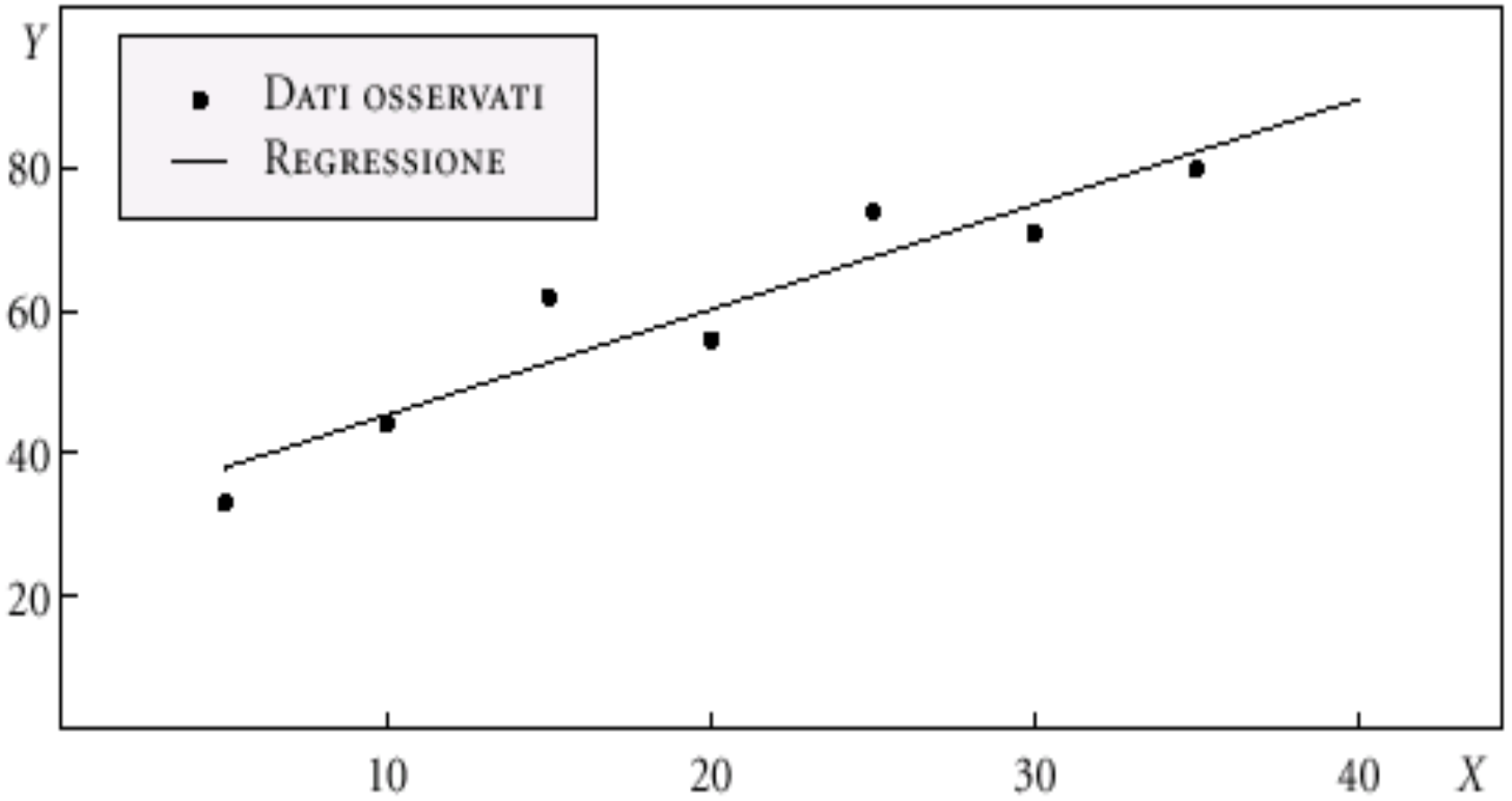
Un possibile ed efficace utilizzo di R^2 è nel confronto di diversi modelli di regressione. Consideriamo il seguente esempio legato ad un problema agrario.

Si vuole misurare la crescita delle radici del mais in funzione del contenuto di saccarosio nel nutrimento fornito.

I seguenti dati si riferiscono alla crescita in millimetri della radice del mais (Y), coltivata in vitro per 10 giorni, in relazione al contenuto di saccarosio (X), misurato in g/l (grammi per litro), nel terreno di coltura:

$$\boldsymbol{x} = \{5, 10, 15, 20, 25, 30, 35\}; \qquad \boldsymbol{y} = \{33, 44, 62, 56, 74, 71, 80\}.$$

Regressione lineare R^2=0.889



Regressione lineare R^2=0.92

R^2 ha un valore superiore rispetto al modello precedente

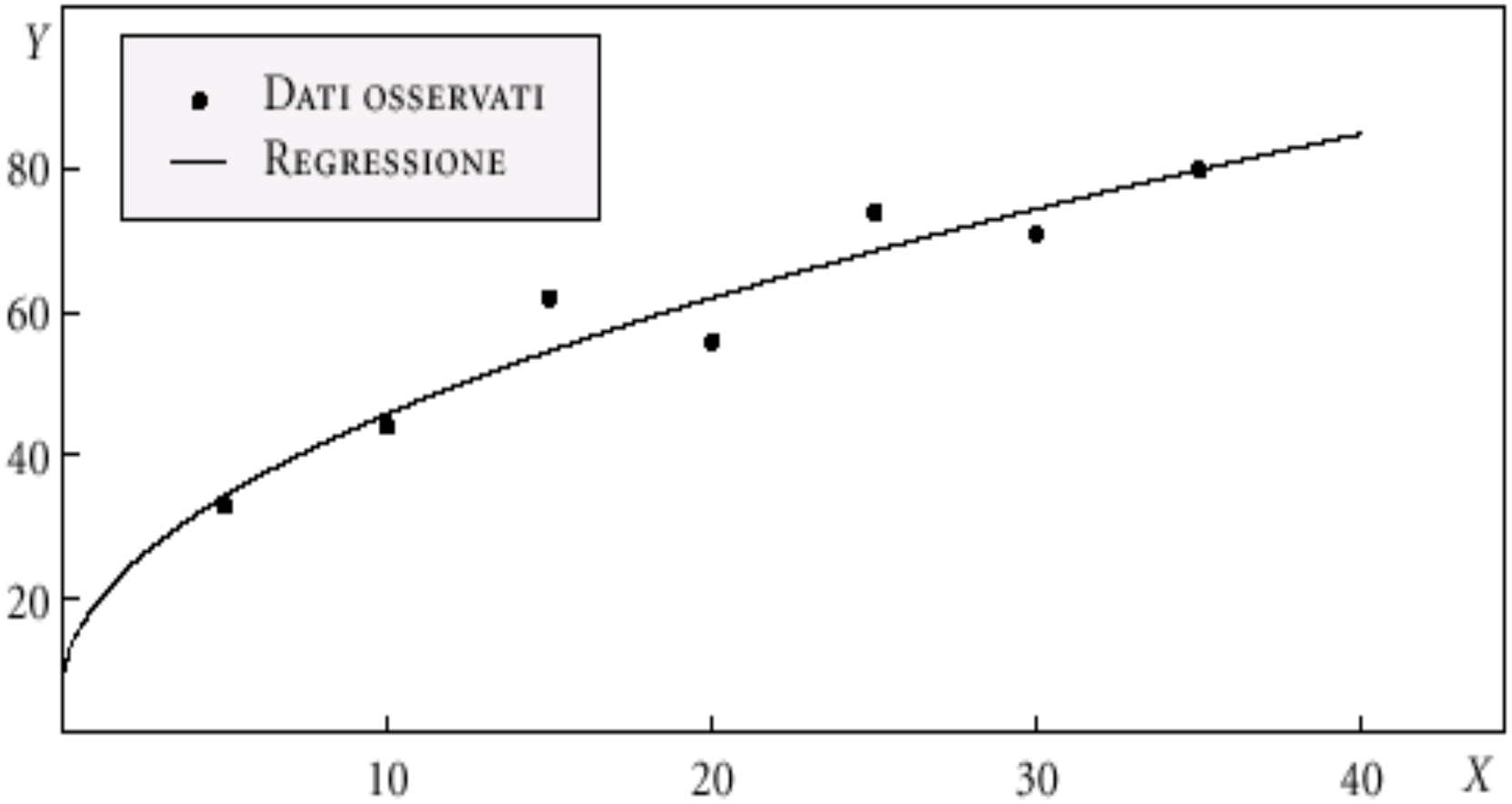
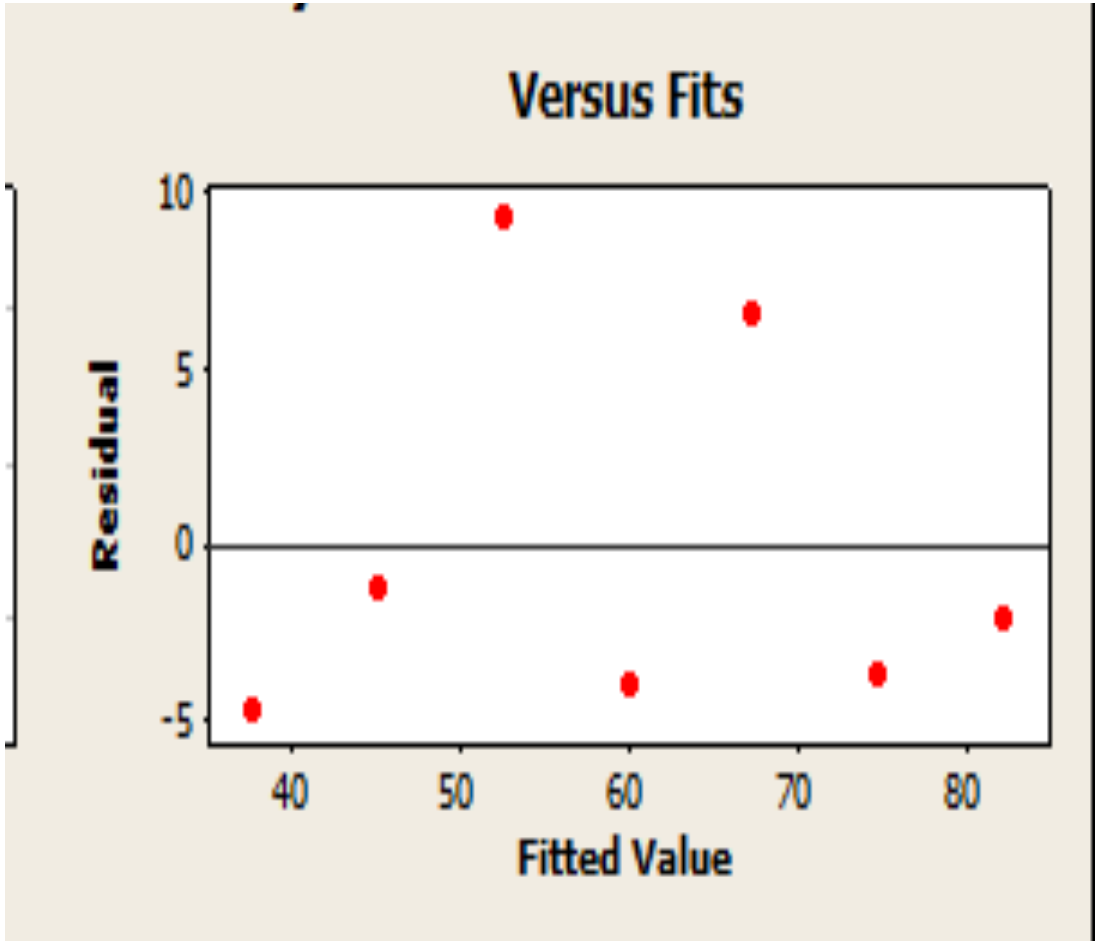


Grafico dei residui dell'es. precedente

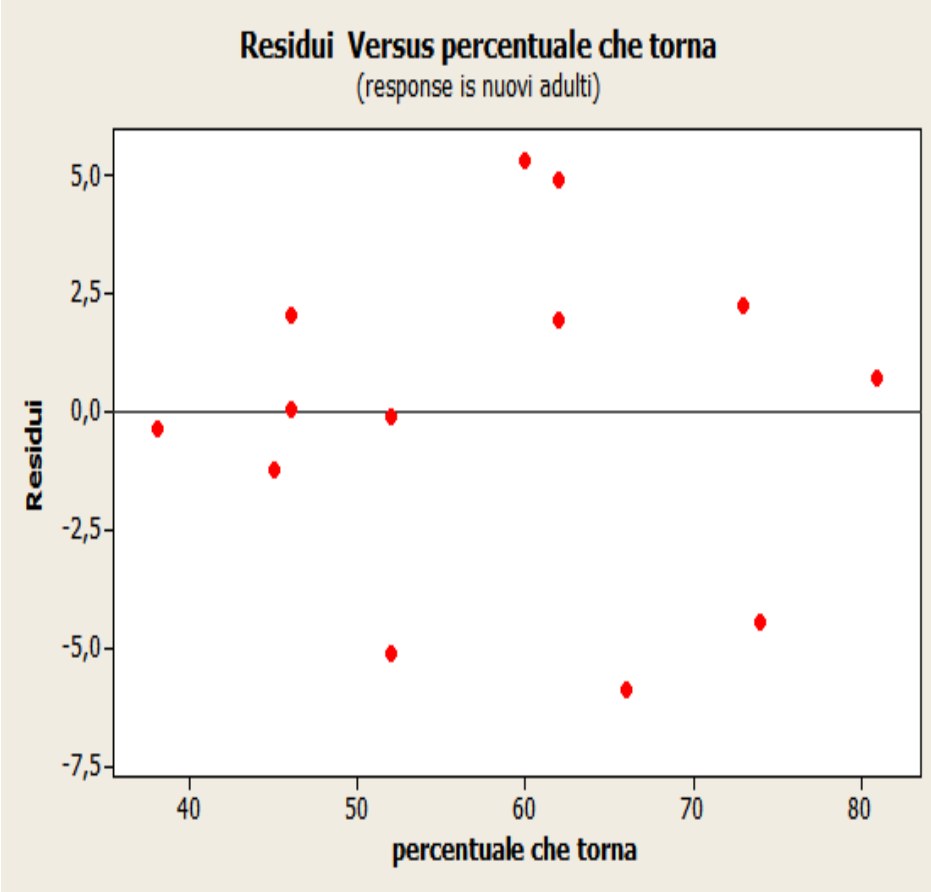


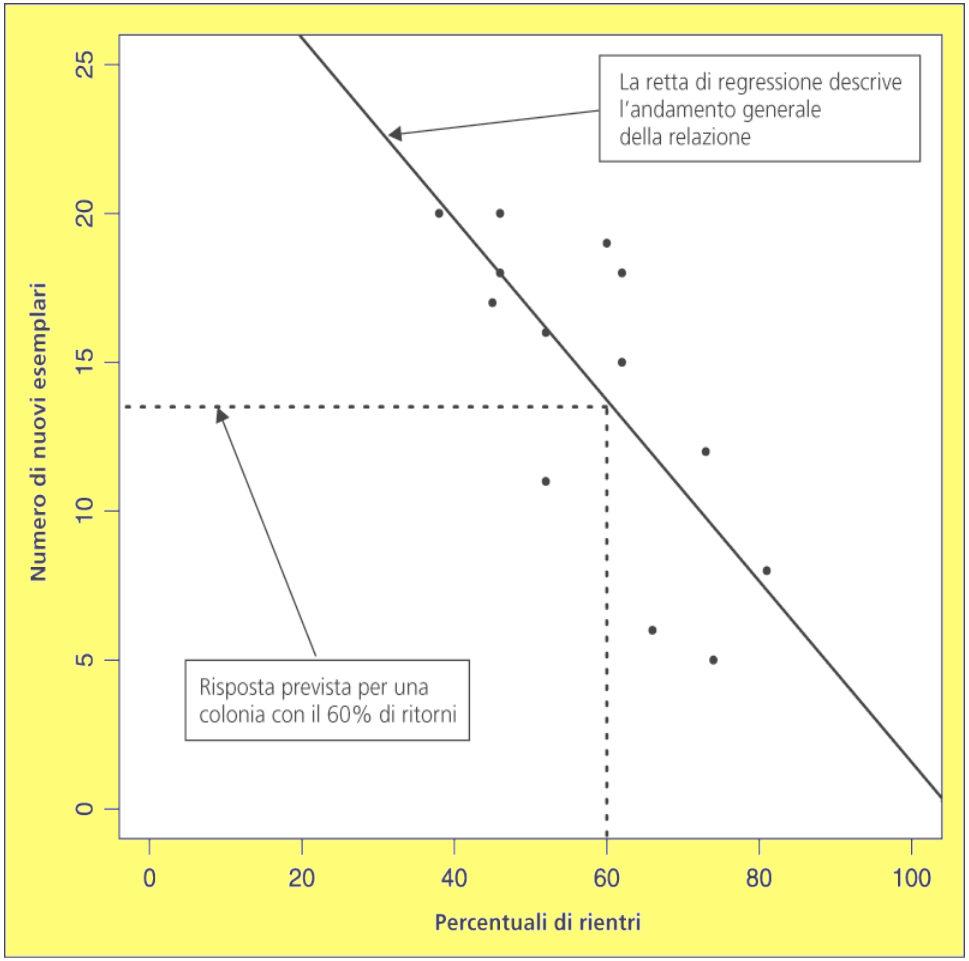
Interpretazione del fenomeno

- L'andamento dei residui suggerisce che ci sia una crescita maggiore vicino a zero e poi un effetto tipo saturazione dell'assorbimento del nutrimento.
- Anche R^2 è coerente con questo effetto e suggerisce una funzione più verticale di una retta vicino a zero e poi meno inclinata.

Grafico dei residui

Se il grafico dei residui non mostra alcuna configurazione particolare (ossia presenta una **configurazione casuale**) l'equazione di regressione è un buon modello per rappresentare l'associazione tra le 2 variabili





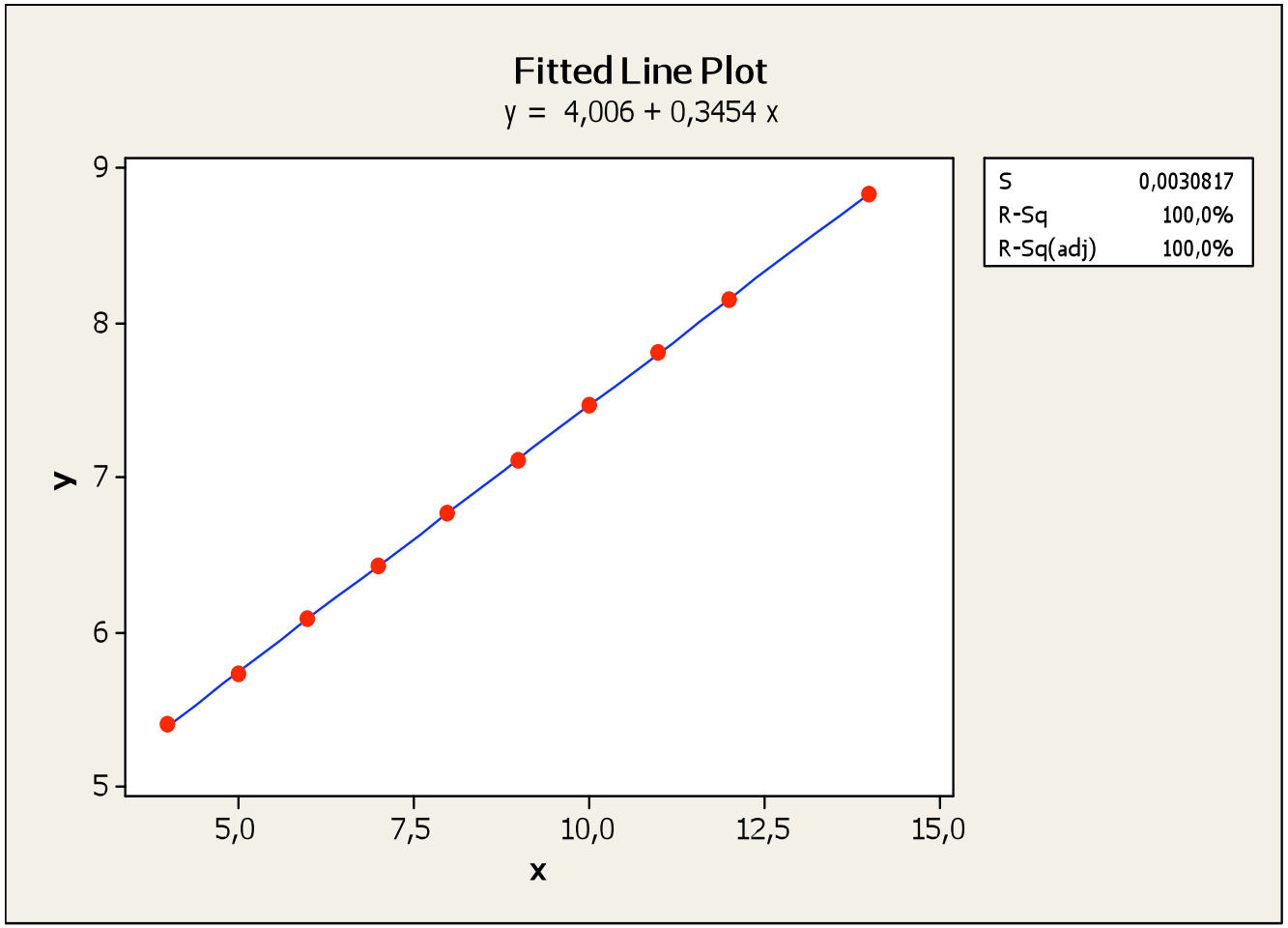
Regressione:analisi dei residui e coefficiente di determinazione

r = 1

r^2 = 100%

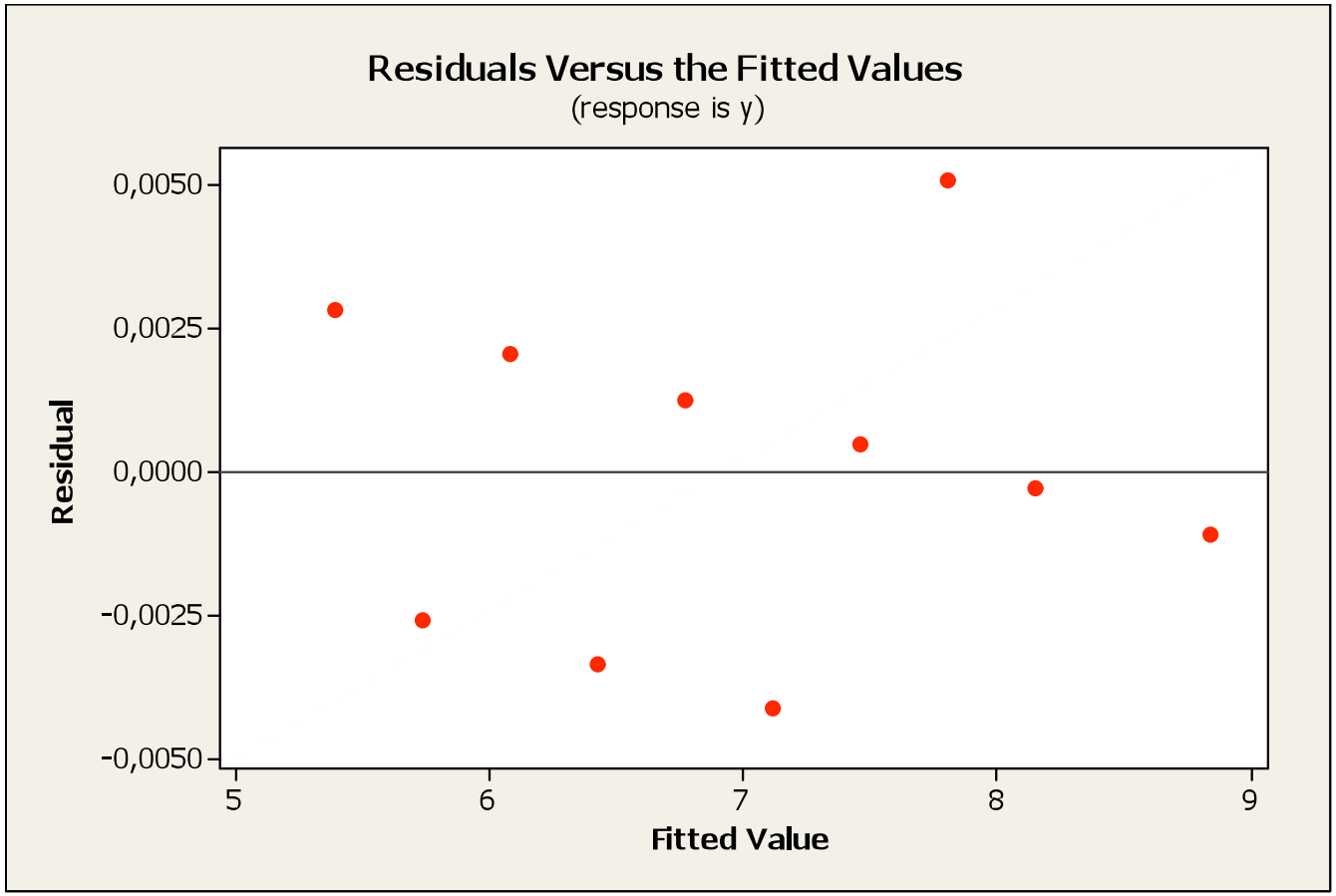
Retta di regressione per i dati osservati

x	y
10	7,46
8	6,77
9	7,11
11	7,81
14	8,84
6	6,08
4	5,39
12	8,15
7	6,42
5	5,73



Regressione:analisi dei residui e coefficiente di determinazione

Il grafico dei residui per l'es precedente mostra una **configurazione casuale**



L'equazione di regressione è un buon modello per rappresentare l'associazione tra le 2 variabili

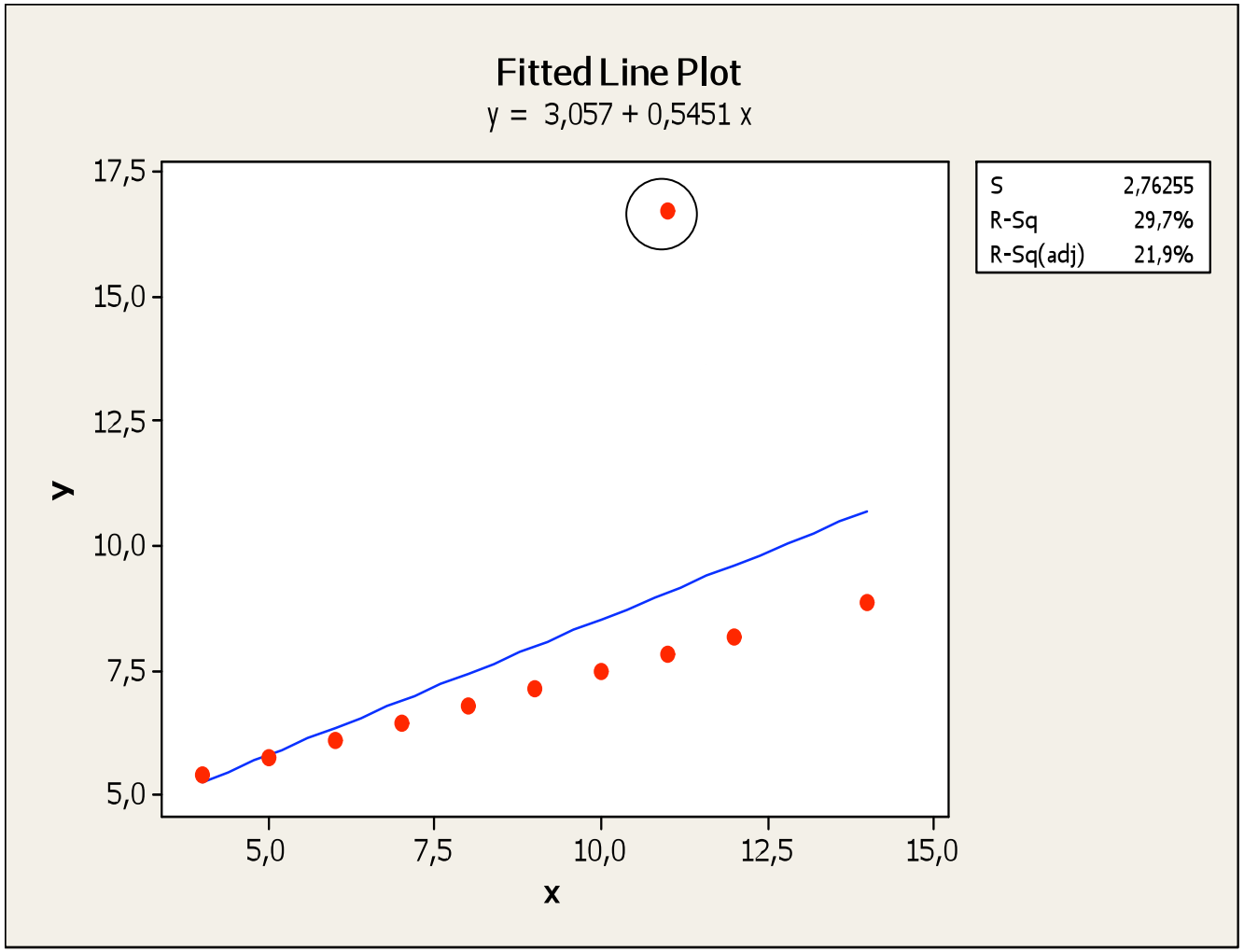
Regressione:analisi dei residui e coefficiente di determinazione

Al campione precedente è stata **aggiunta un'osservazione (outlier)**. Il modello lineare non è un buon modello (r^2 = 29.7%)

r = 0.545

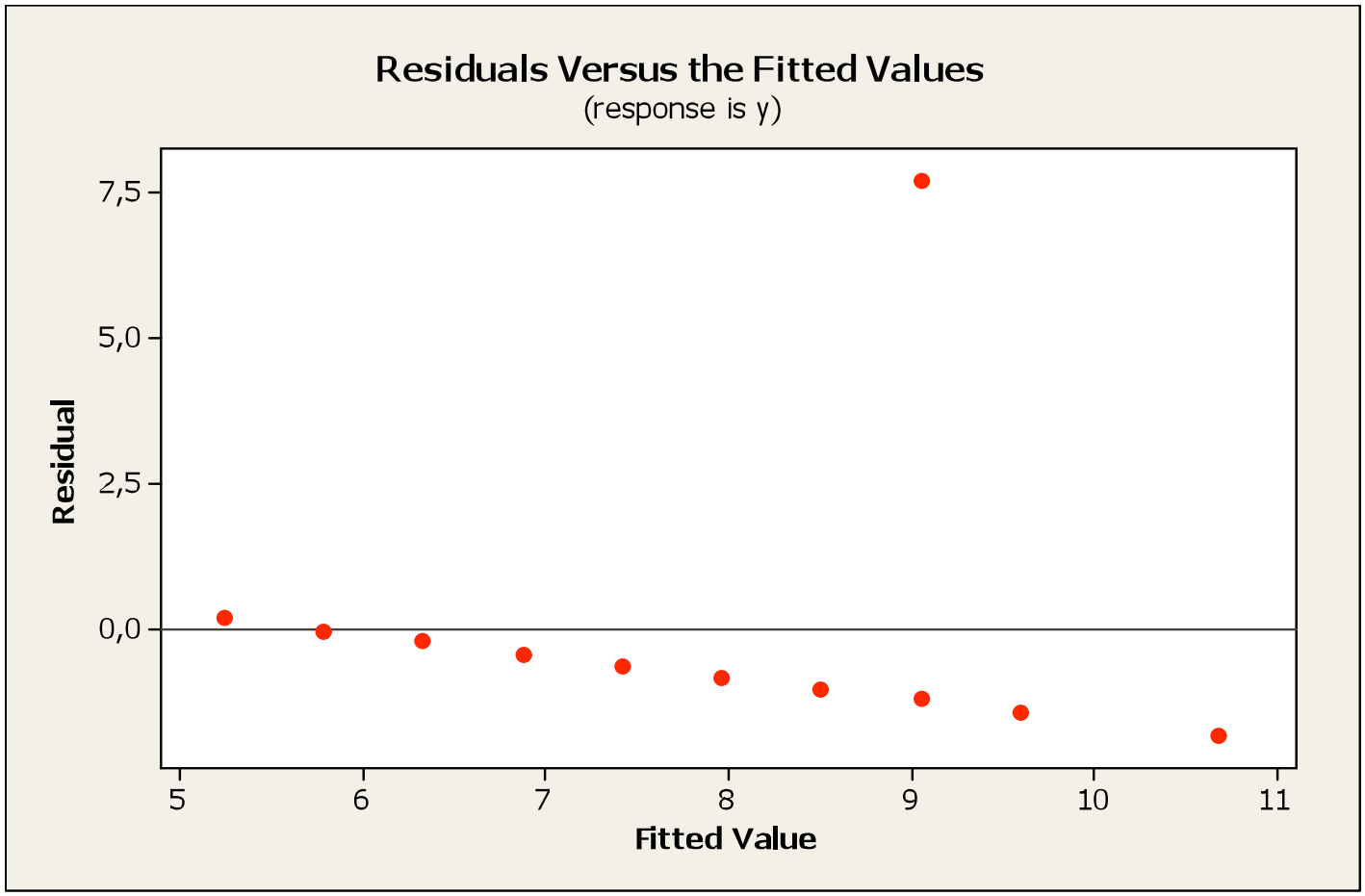
r^2 = 29.7%

x	y
10	7,46
8	6,77
11	16,74
9	7,11
11	7,81
14	8,84
6	6,08
4	5,39
12	8,15
7	6,42
5	5,73



Regressione:analisi dei residui e coefficiente di determinazione

Grafico dei residui (es. precedente, campione con outlier) **non** mostra una configurazione casuale. Il modello lineare **non** è un buon modello

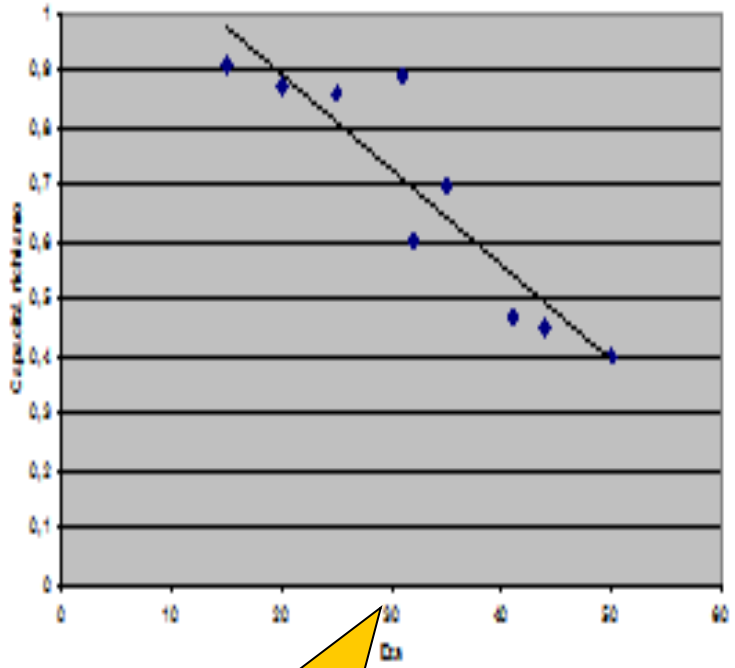


Osservazioni influenti

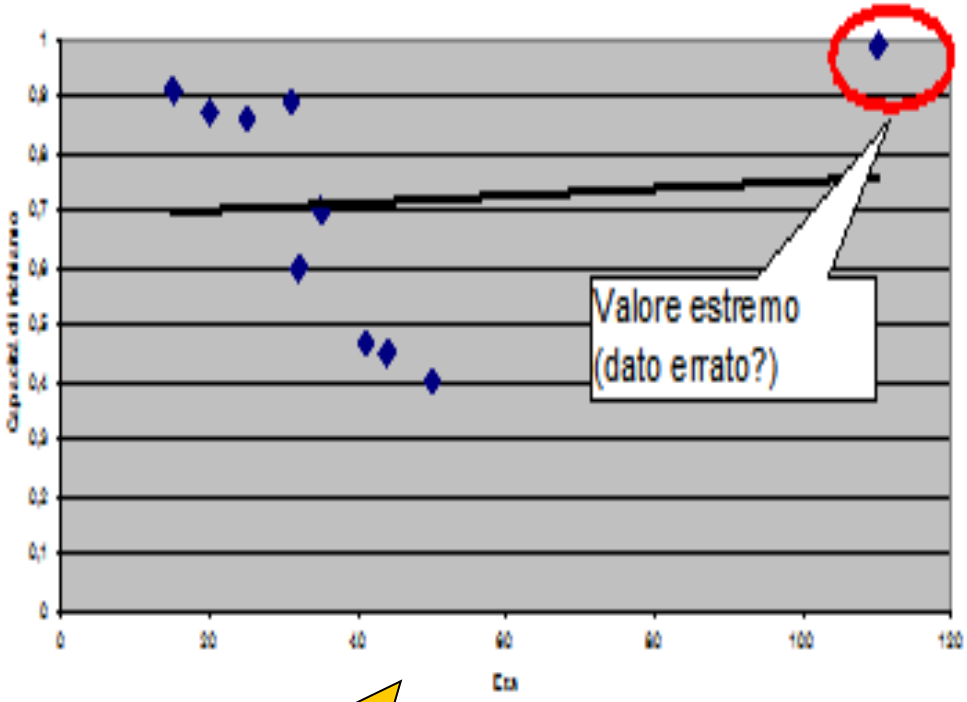
- **Outlier e osservazioni influenti nella regressione**
- Un **outlier** è un'osservazione che non segue il modello generale assunto dalla maggior parte delle osservazioni. I punti che, guardando un diagramma di dispersione, possiamo **considerare** **outlier in direzione di *y***, hanno residui elevati.

- Un'osservazione è **influente** se, eliminandola, cambierebbe profondamente il risultato. I punti che, in un diagramma di dispersione, possiamo considerare **outlier in direzione della *x*** sono spesso punti influenti nella determinazione della retta di regressione dei minimi quadrati.

Osservazioni influenti



retta di regressione con un buon adattamento ai dati



ai dati originali è stata aggiunta una osservazione influente

Attenzione all'estrapolazione

- **Estrapolazione**
- L'**estrapolazione** è l'utilizzo della retta di regressione per prevedere al di fuori dell'intervallo di valori, della variabile esplicativa *x,* utilizzati per ottenere la linea. Previsioni di questo tipo sono spesso piuttosto imprecise.

Attenzione alla variabile nascosta

Variabile nascosta

Una **variabile nascosta** è una variabile che ha un effetto importante sulla relazione fra le variabili analizzate, ma che non è stata presa in considerazione nello studio.

La correlazione e la regressione possono essere ingannevoli se si ignorano variabili nascoste significative.

Le associazioni non implicano rapporti di causa ed effetto

- **Le associazioni non implicano causa/effetto**
- Un'associazione fra una variabile esplicativa *x* e una di risposta *y*, anche se molto forte, non comporta necessariamente che *y* venga modificata ad opera di *x*.
- Esempio: esiste un'alta correlazione positiva tra il numero di televisori per persona x e la vita media y per diverse nazioni del mondo: nelle nazioni con molti televisori si vive più a lungo.
- Correlazione spuria. Non c'è rapporto causa/effetto.

Le associazioni non implicano rapporti di causa ed effetto

Subito dopo la II guerra mondiale in Inghilterra si è osservato:

aumento delle nidificazioni di cicogne (x)

aumento delle nascite (y)

Si è visto che esisteva una correlazione positiva tra x e y.

Ma si può considerare la nidificazione delle cicogne come possibile " causa" delle nascite?

Certamente No. Esiste una comune causa (fine della guerra e dei bombardamenti) che influisce positivamente su entrambe le variabili.

Ma non esiste nessun rapporto diretto tra le 2 variabili.

Regressione alla media

- Il termine regressione fu usato per la prima volta da F. Galton che studiava le leggi sull'ereditarietà.
- Galton si accorse che i figli di padri eccezionalmente alti o bassi tendono ad essere più nella media dei loro genitori. Ossia vi è quella che lui chiamò "una regressione verso la media della popolazione".

Regressione alla media

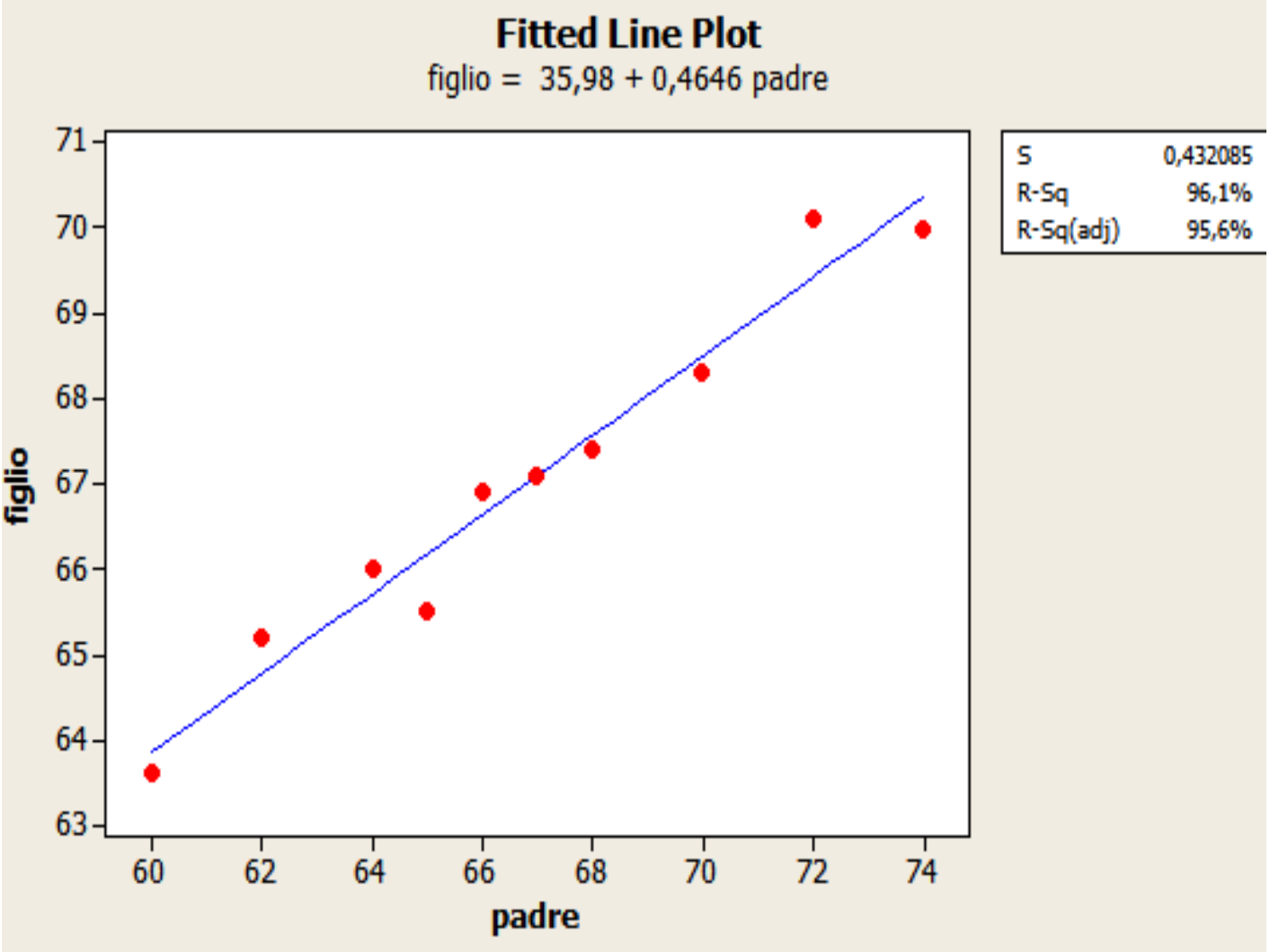
- Una giustificazione biologica moderna del fenomeno della regressione verso la media dovrebbe basarsi sul fatto che ogni figlio ottiene una selezione casuale di metà dei geni di ciascuno dei genitori. Si potrebbe dire, in modo impreciso, che il figlio di un padre molto alto o molto basso avrà generalmente meno geni “della statura” del padre.

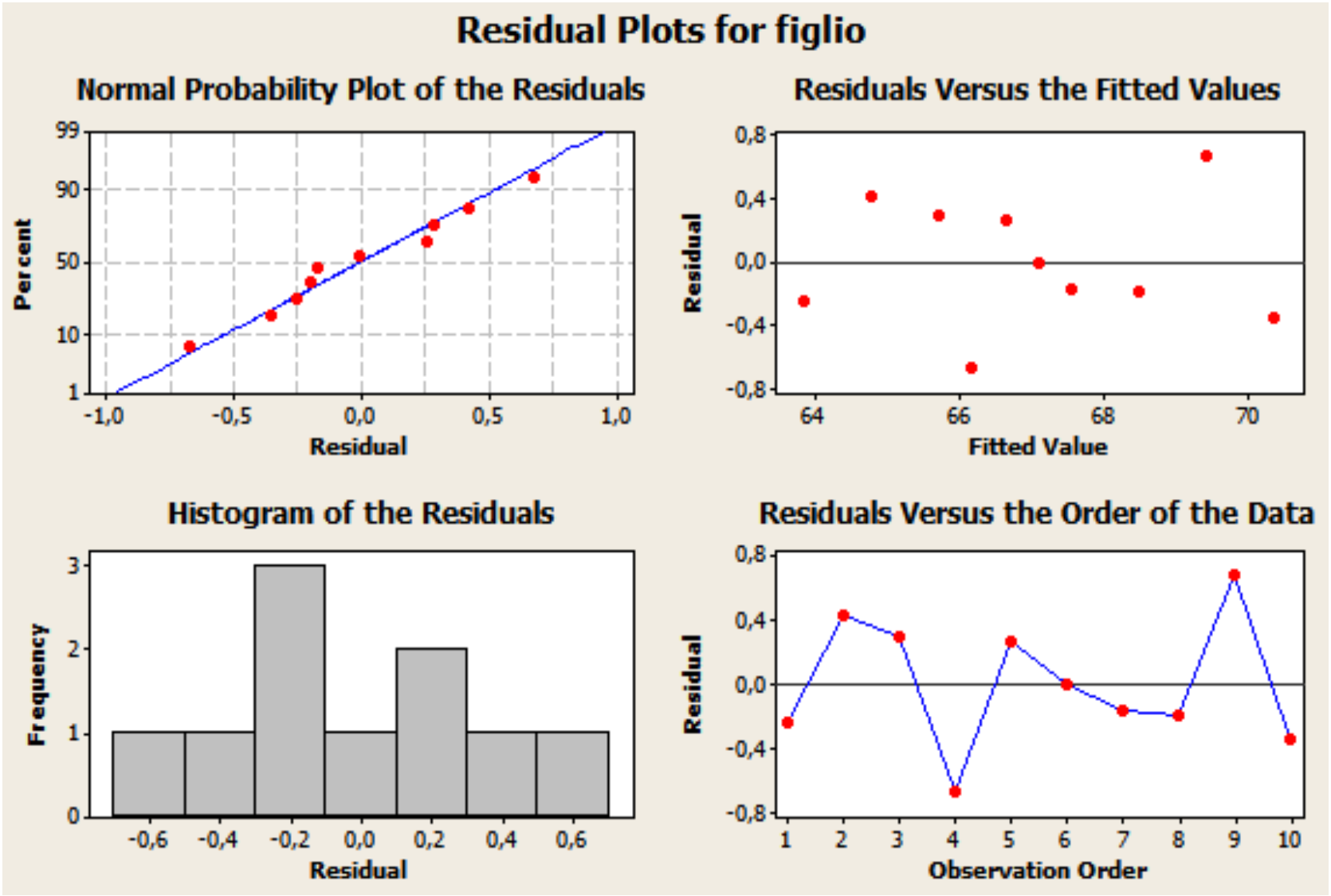
Esempio

- Per mostrare la regressione verso la media dei caratteri ereditari, K. Pearson ha confrontato le stature (in pollici) di 10 figli maschi scelti a caso con quelle dei loro padri.

padre	60	62	64	65	66	67	68	70	72	74
figlio	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

Correlazione r = 0,98; r^2 = 96.1




Residual Plots for figlio
Normal Probability Plot of the Residuals
Percent
99
90
50
10
1
-1,0
-0,5
0,0
0,5
1,0
Residual
Residuals Versus the Fitted Values
Residual
0,8
0,4
0,0
-0,4
-0,8
64
66
68
70
Fitted Value
Histogram of the Residuals
Frequency
3
2
1
0
-0,6
-0,4
-0,2
0,0
0,2
0,4
0,6
Residual
Residuals Versus the Order of the Data
Residual
0,8
0,4
0,0
-0,4
-0,8
1
2
3
4
5
6
7
8
9
10
Observation Order

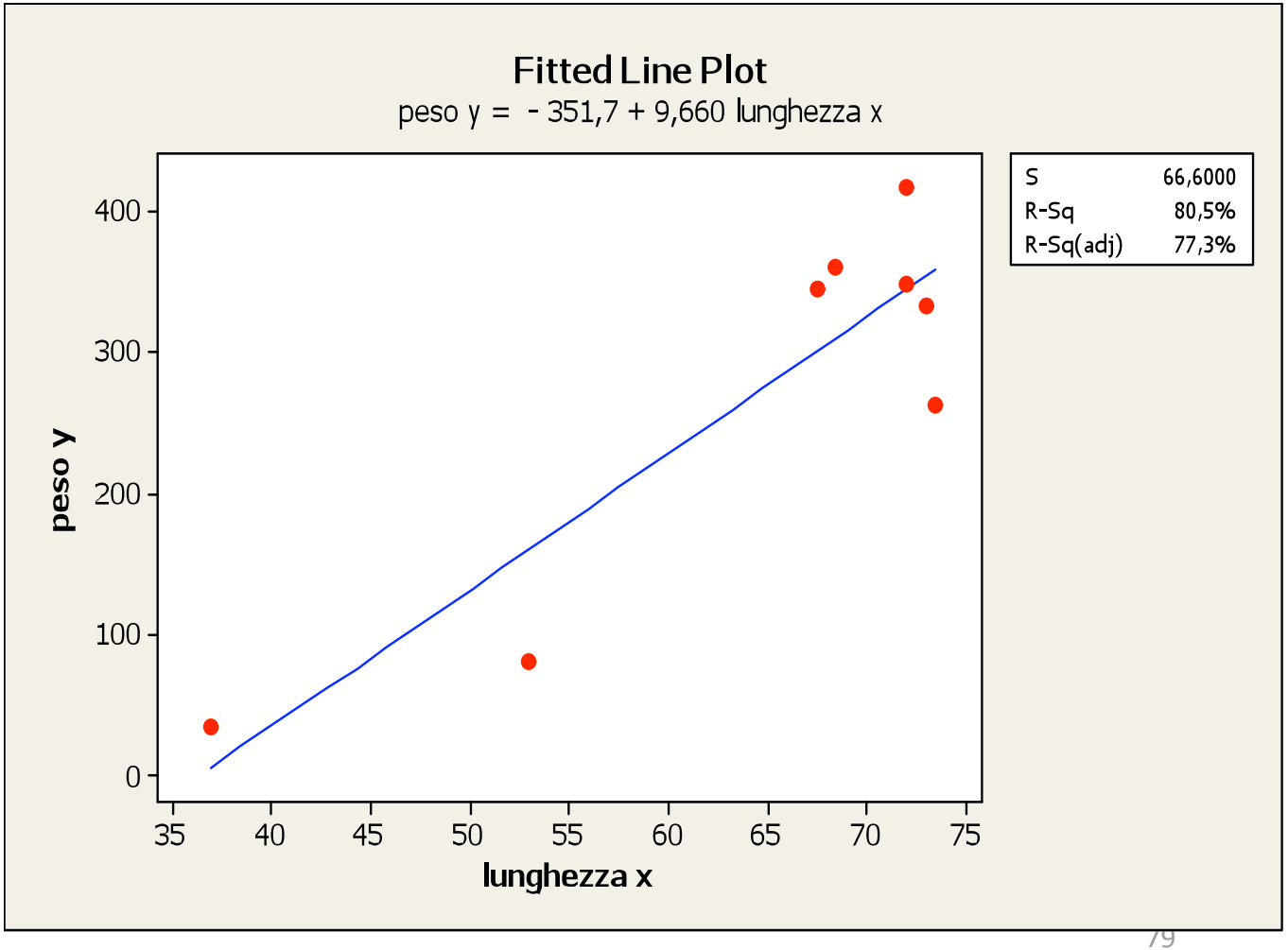
Regressione: Lunghezza e peso di un campione di 8 orsi maschi

la retta di regressione interpola bene i dati

$r = 0.897$

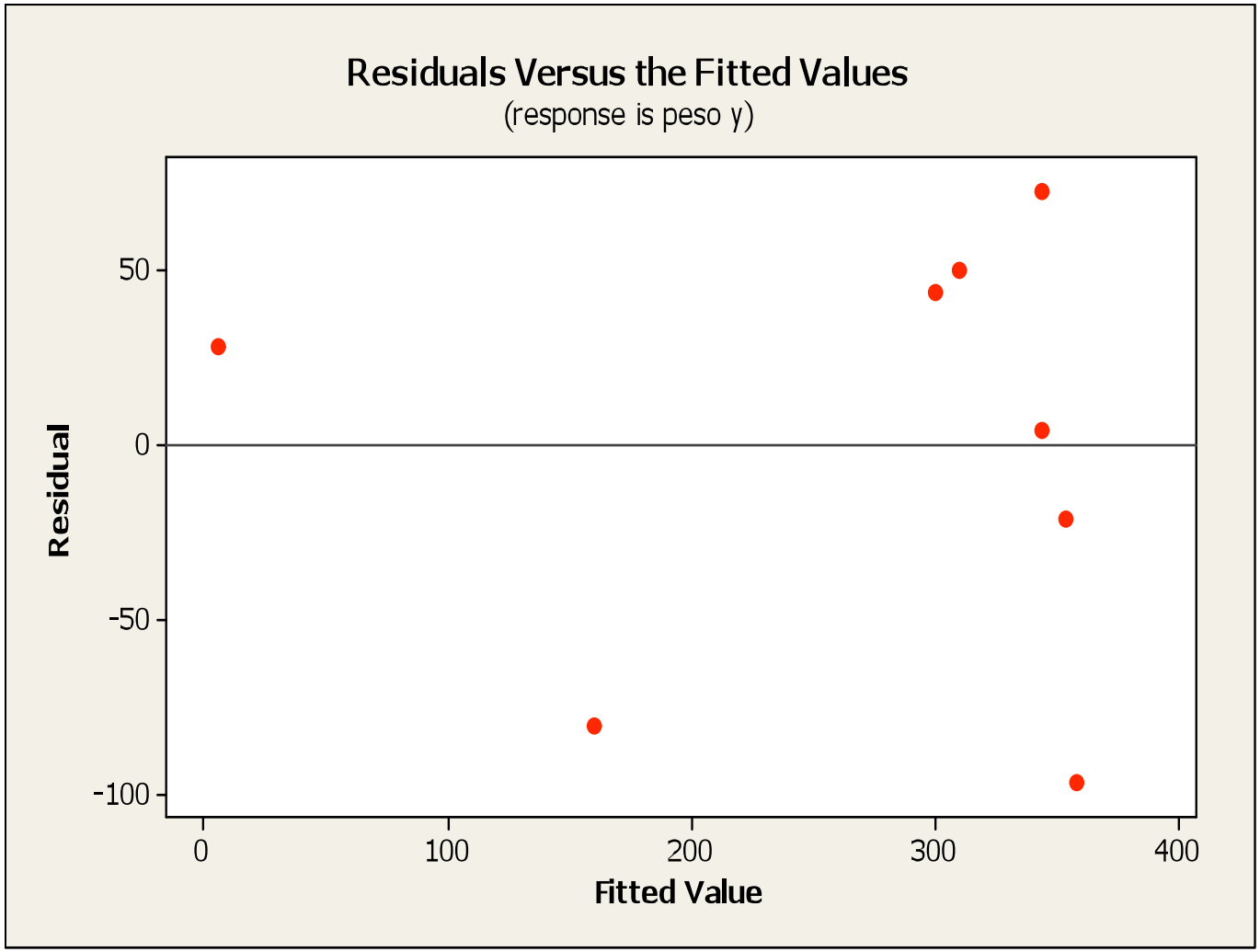
$r^2 = 80.5\%$

lunghezza (pollici)	peso (libbre)
53,0	80
67,5	344
72,0	416
72,0	348
73,5	262
68,5	360
73,0	332
37,0	34



Regressione: continuazione esempio

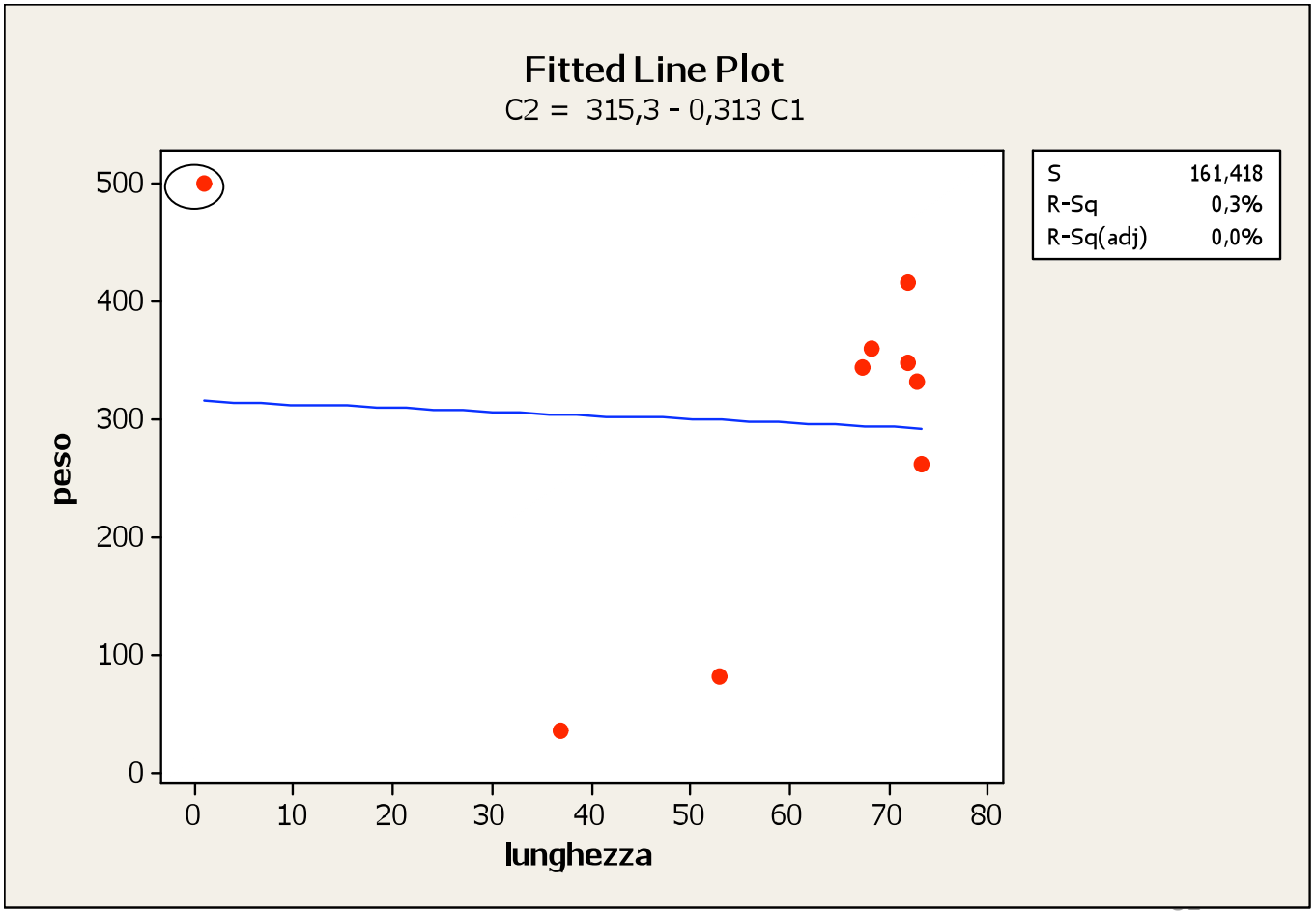
Il grafico dei residui presenta una configurazione casuale



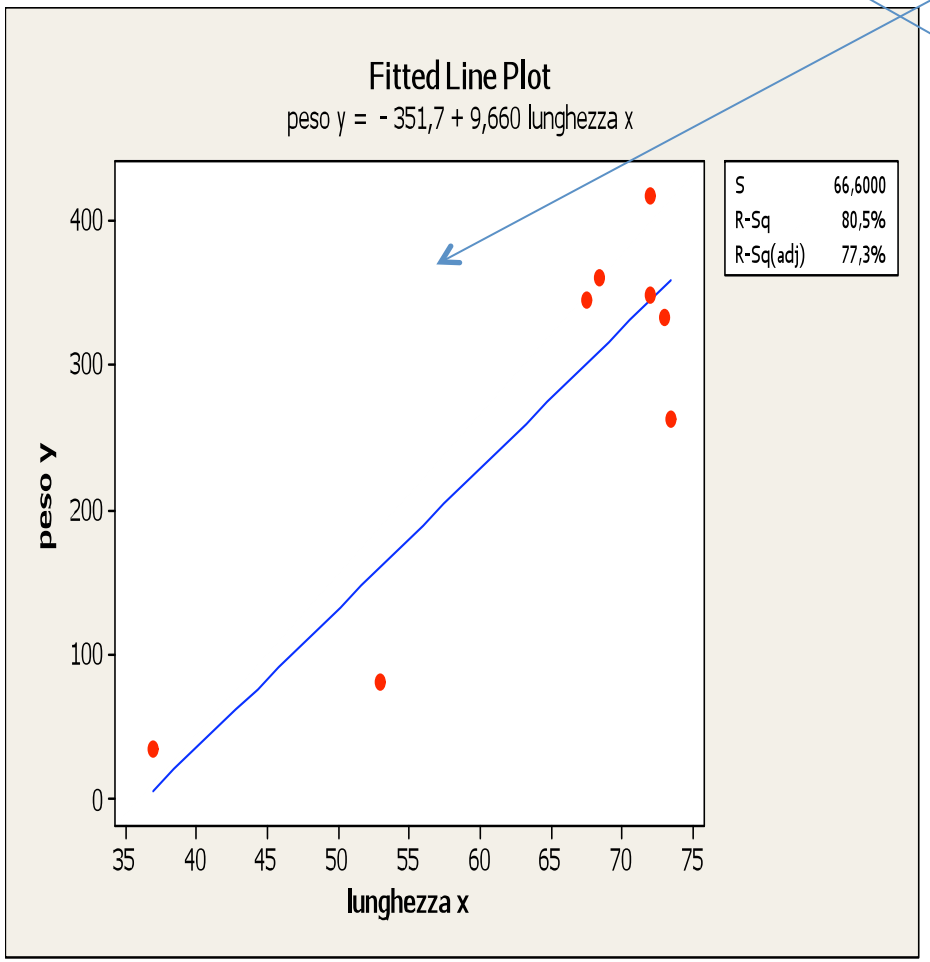
Regressione: continuazione esempio

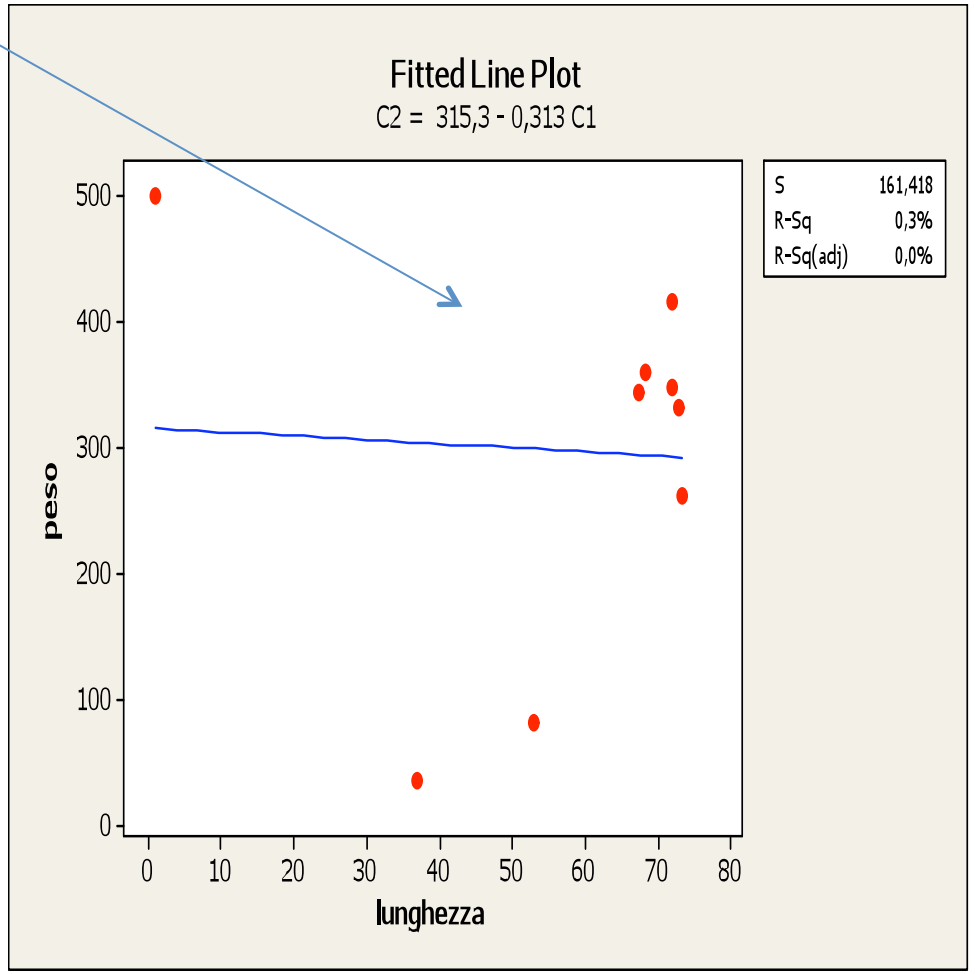
Se aggiungessimo al campione un nuovo orso lungo 1 pollice e di peso pari a 500 libbre!!! avremmo un punto influente. Infatti si avrebbe un effetto notevole sulla retta di regressione (cambia l'equazione).

x	y
53,0	80
67,5	344
72,0	416
72,0	348
73,5	262
68,5	360
73,0	332
37,0	34
1,0	**500**



Le due rette a confronto (senza osservazione influente: $r^2 = 80,5\%$ e con osservazione influente: $r^2 = 0,3\%$)

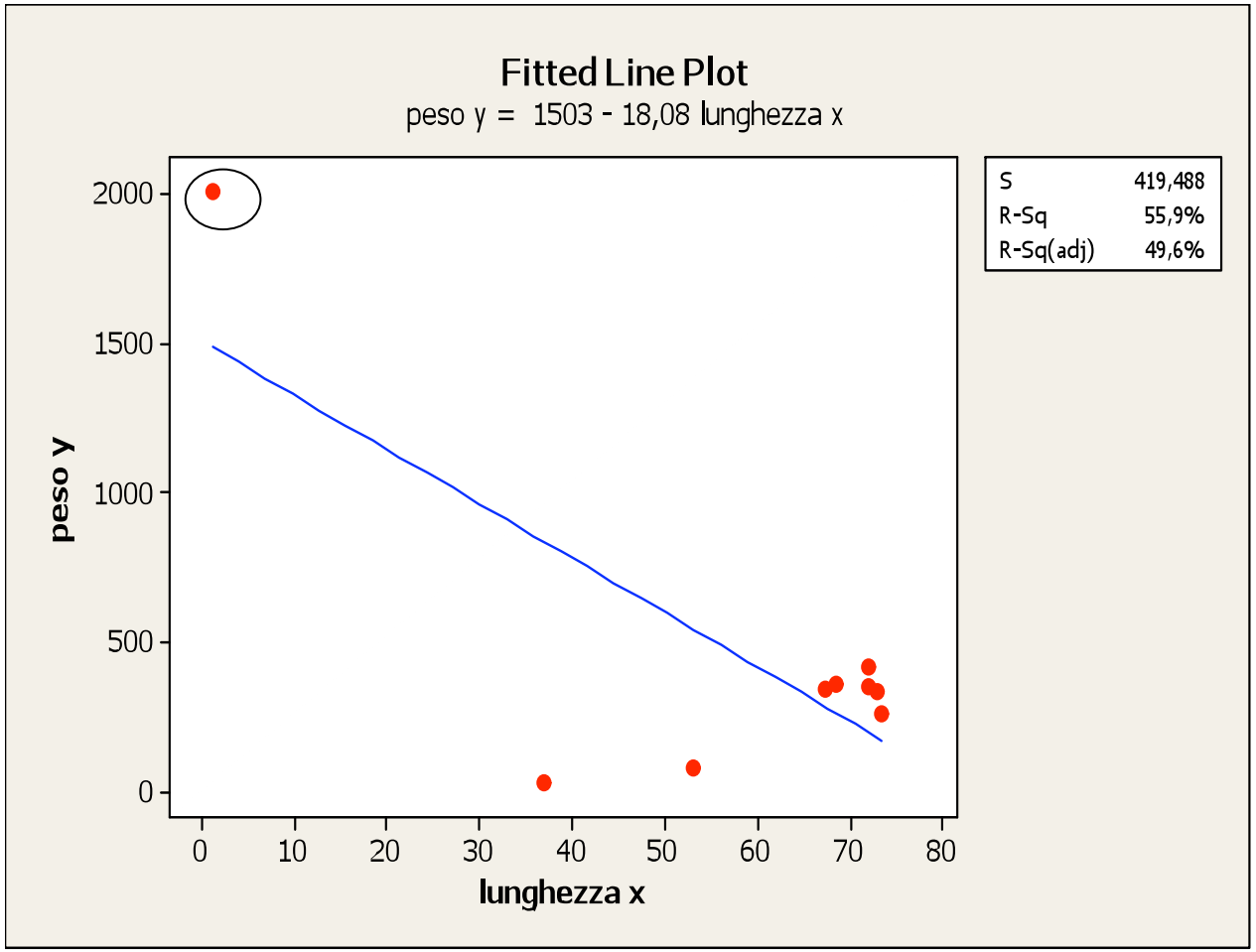




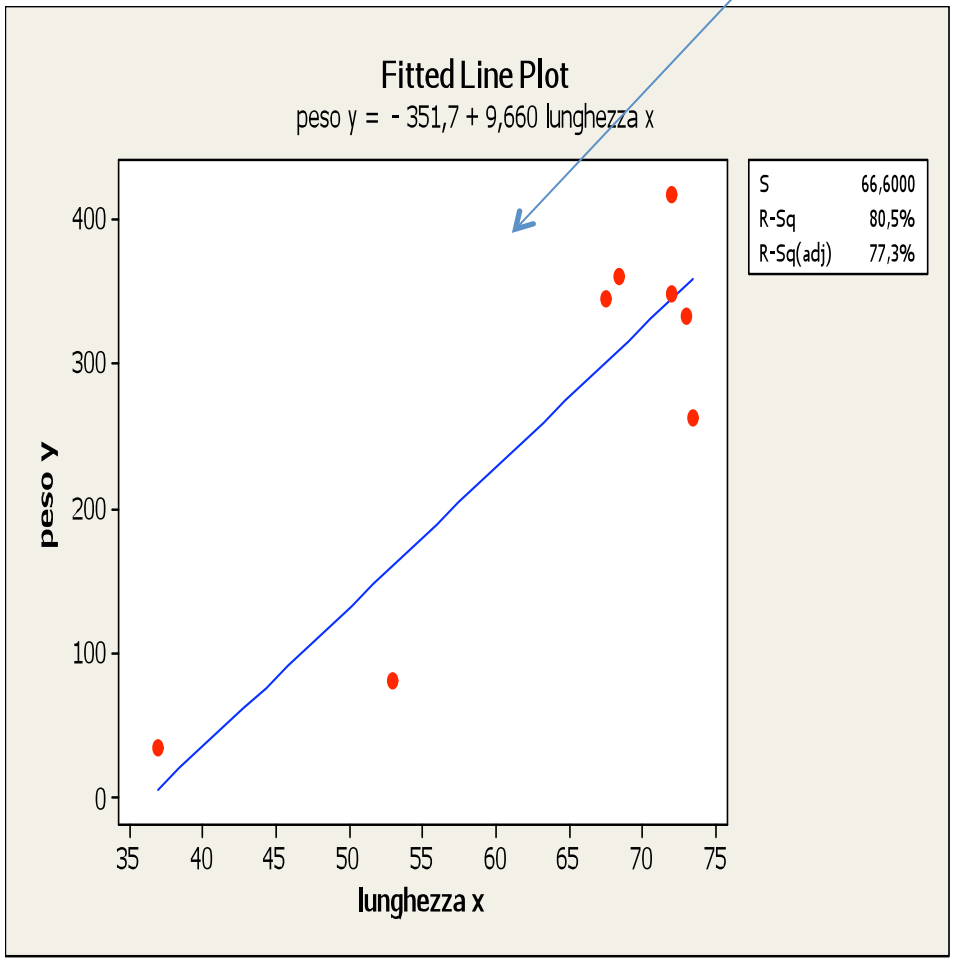
Regressione: continuazione esempio

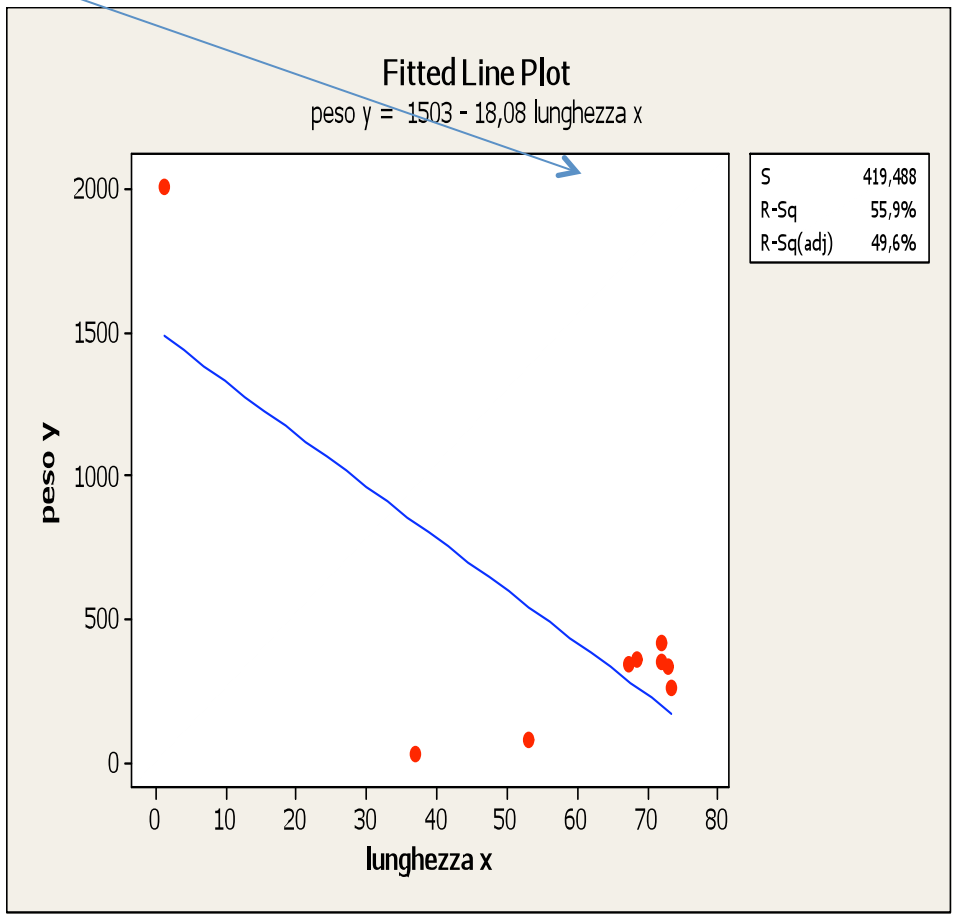
Se aggiungessimo al campione un nuovo orso lungo 1 pollice e di peso pari a 2000 libbre!!! avremmo di nuovo un punto influente (e anche outlier nella direzione della y). Infatti si avrebbe un effetto drammatico sulla retta di regressione.

x	y
53,0	80
67,5	344
72,0	416
72,0	348
73,5	262
68,5	360
73,0	332
37,0	34
1,0	**2000**



Le due rette a confronto (senza osservazione influente: $r^2 = 80,5\%$ e con osservazione influente: $r^2 = 55,9\%$)

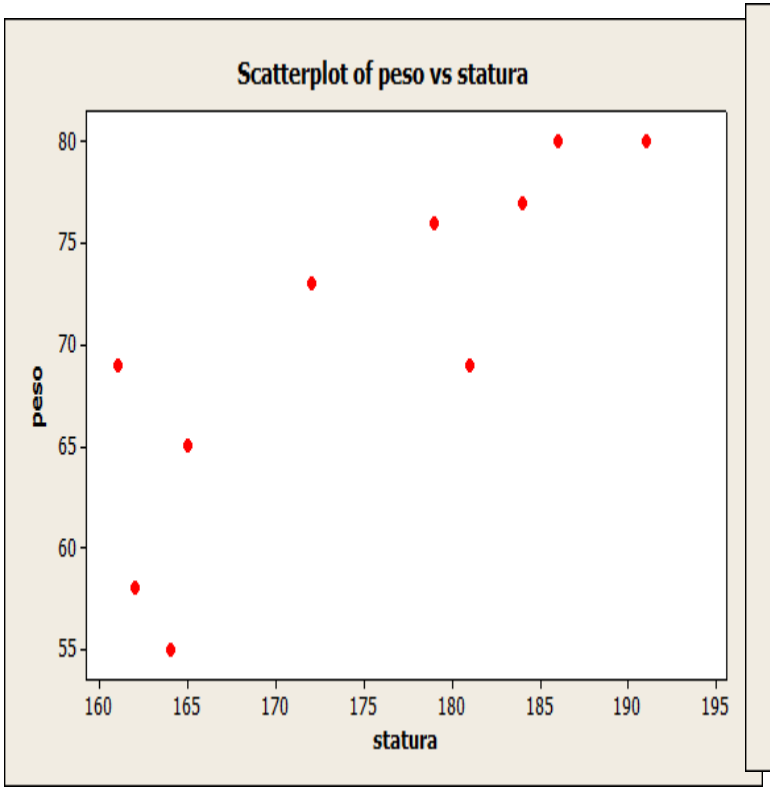


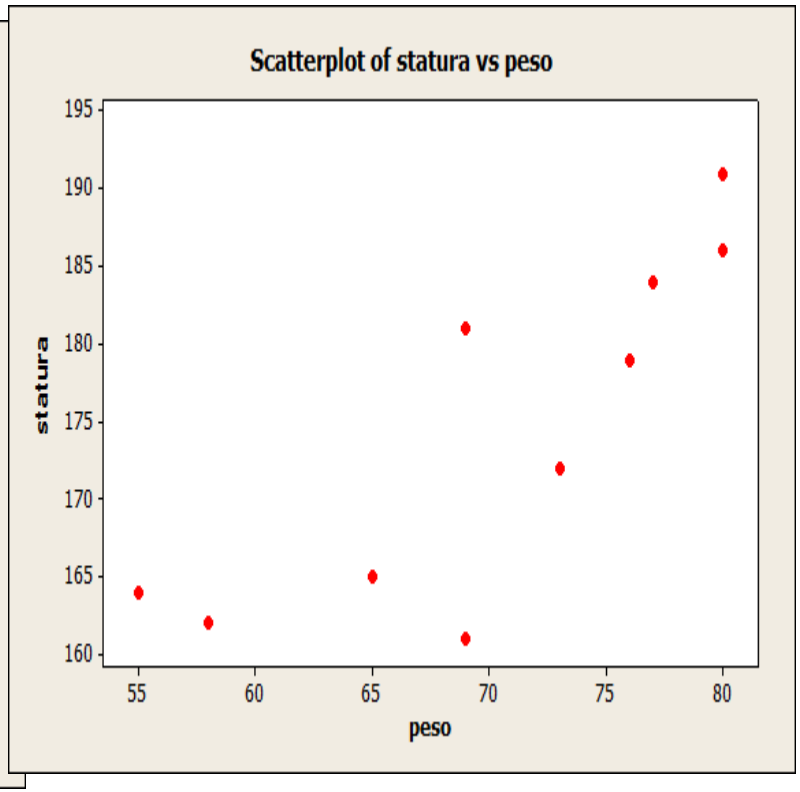


La retta di regressione dei minimi quadrati

Vogliamo determinare la retta di regressione lineare di due variabili
X = statura e Y = peso di 10 individui

statura	peso
161	69
162	58
164	55
165	65
172	73
179	76
181	69
184	77
186	80
191	80

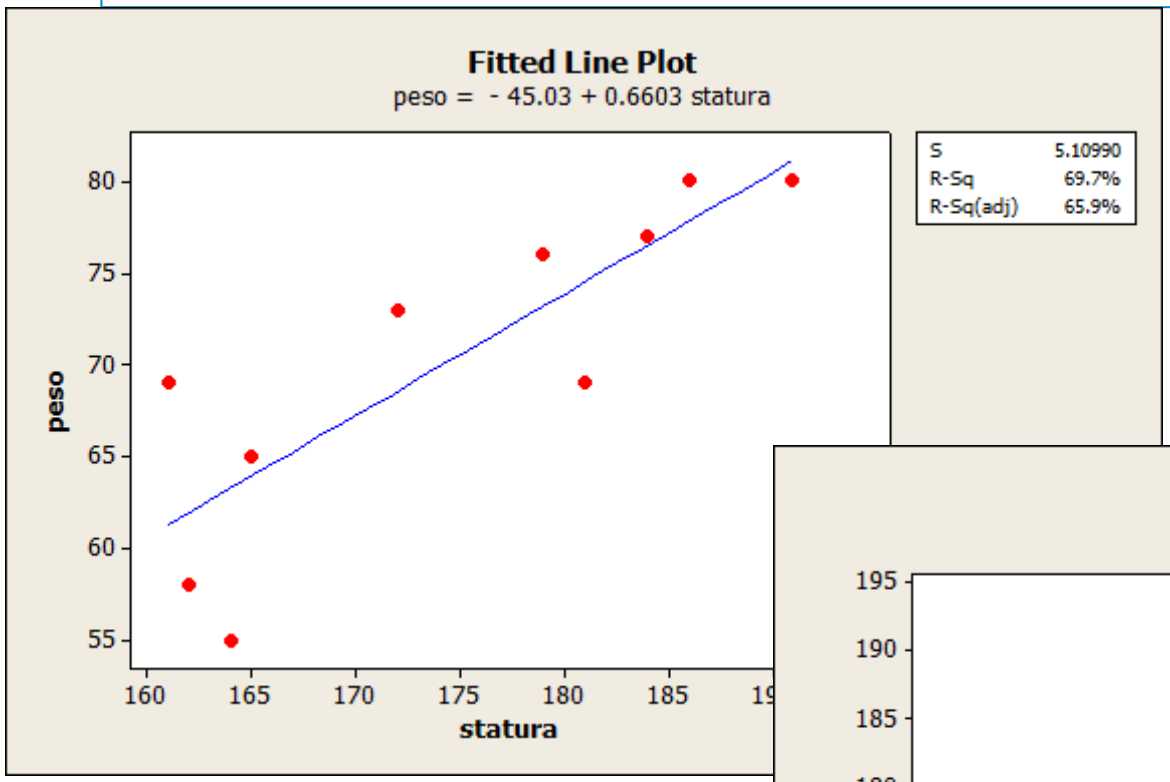




r = ?

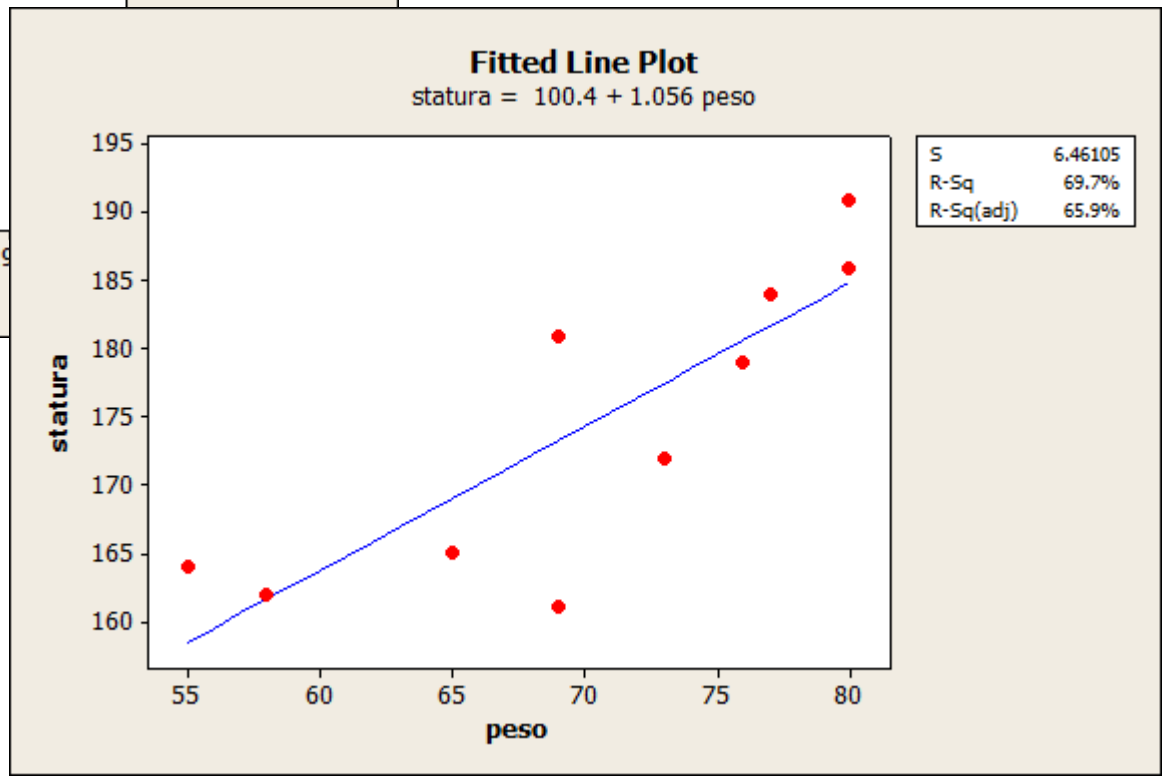
r = ?

La retta di regressione

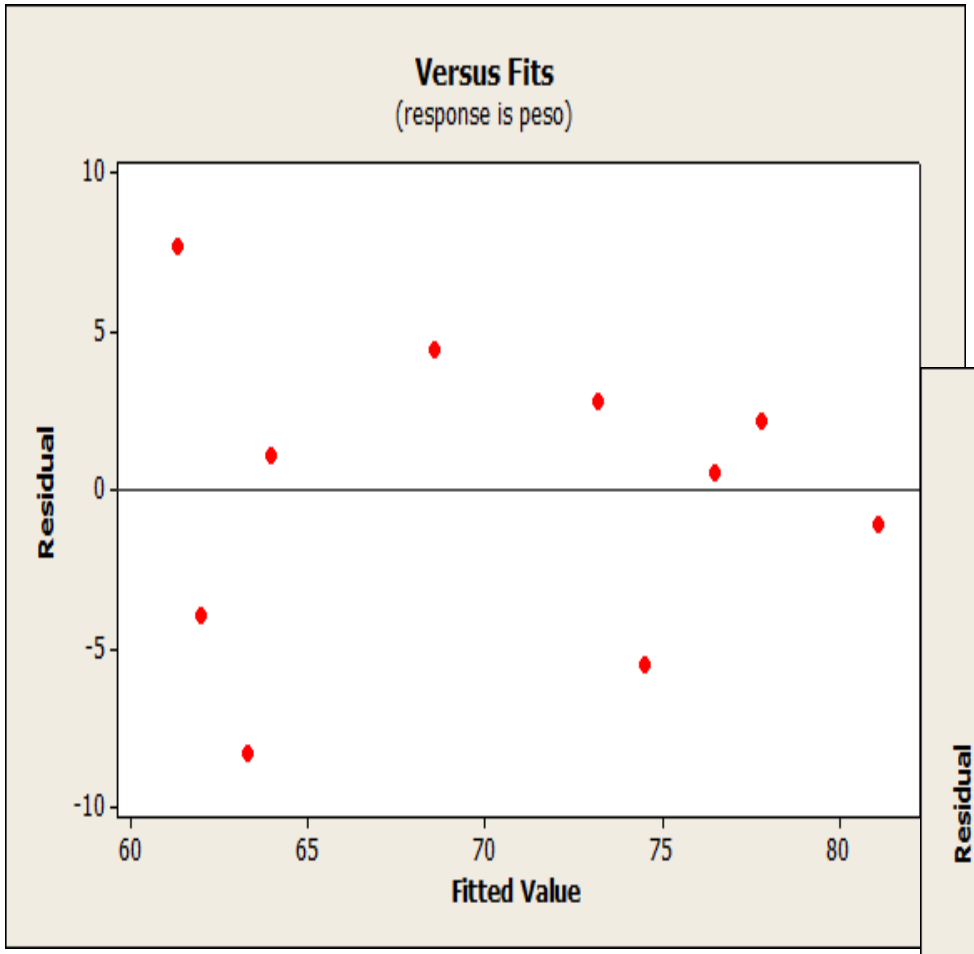


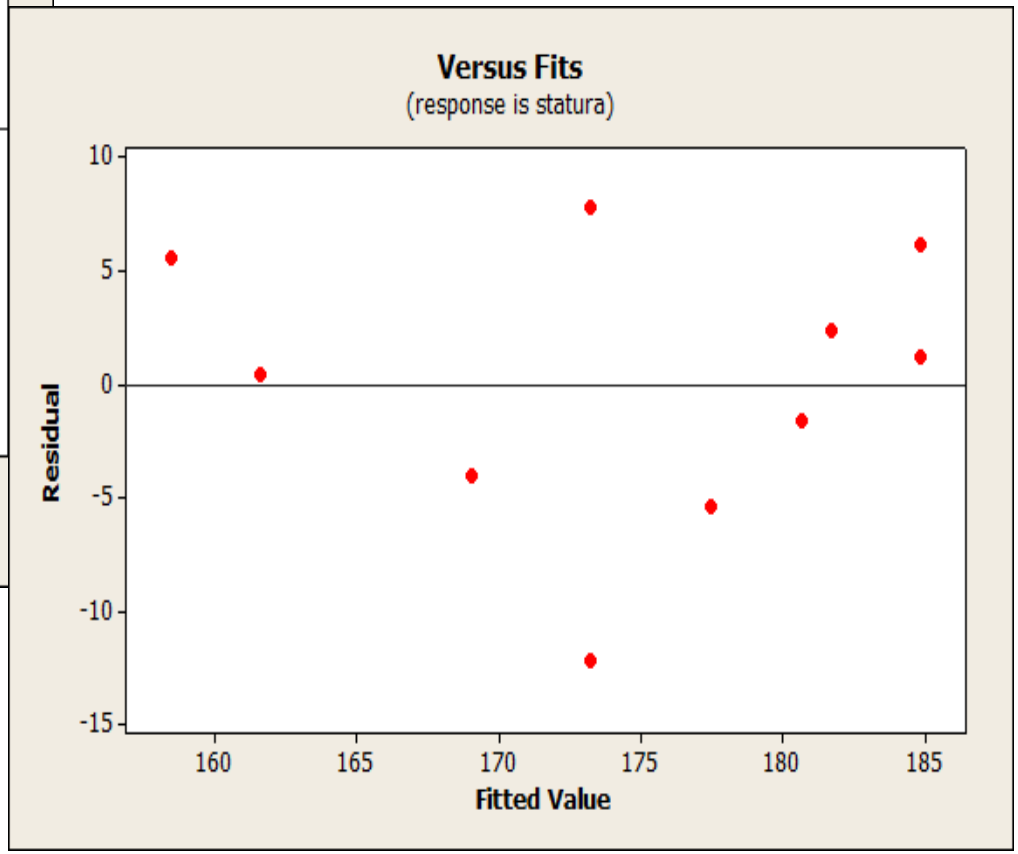
r = 0.835

Qui si può scegliere come v. esplicativa sia la statura sia il peso



Il grafico dei residui (peso, statura)





Esercizio

Nella tabella che segue sono riportate le dosi (in mg) di un farmaco antipertensivo e le relative diminuzioni di pressione (mm/Hg) misurate in 8 pazienti dopo 20 giorni di cura. **Esiste una buona relazione lineare tra le 2 variabili?**

L'equazione della retta di regressione è data da:

Diminuzione (y) = 41.5 - 0.500 dose (x)

Dim.(y)	dose (x)
40	28
22	21
35	15
30	18
32	24
45	11
24	21
24	22

Verificate se il valore del residuo corrispondente all'osservazione x = 28 e y = 40 è pari a 12.50

12. Il valore del coefficiente di determinazione è: $r^2 = 10.4\%$

Retta di regressione relativa all'es. precedente

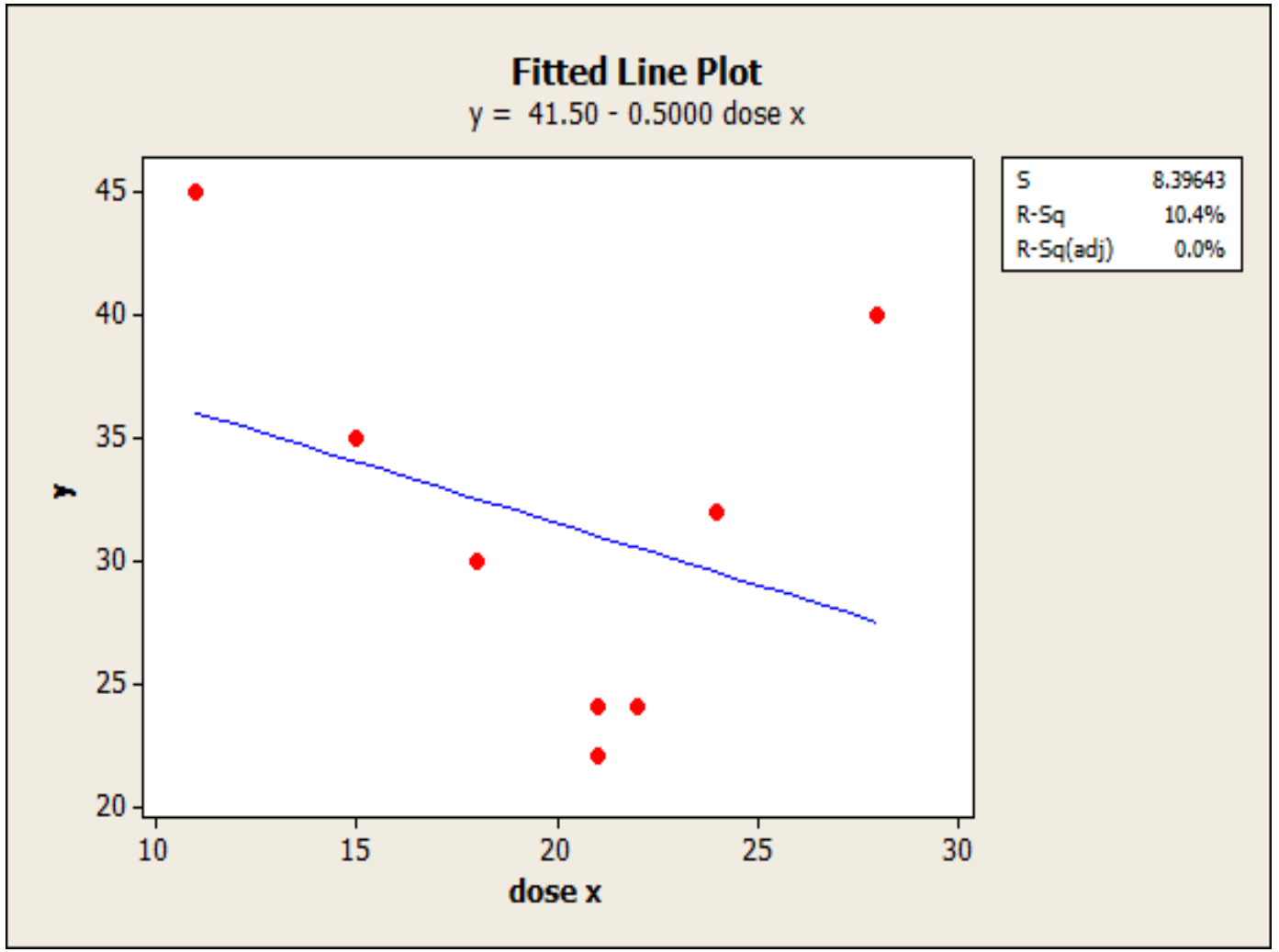
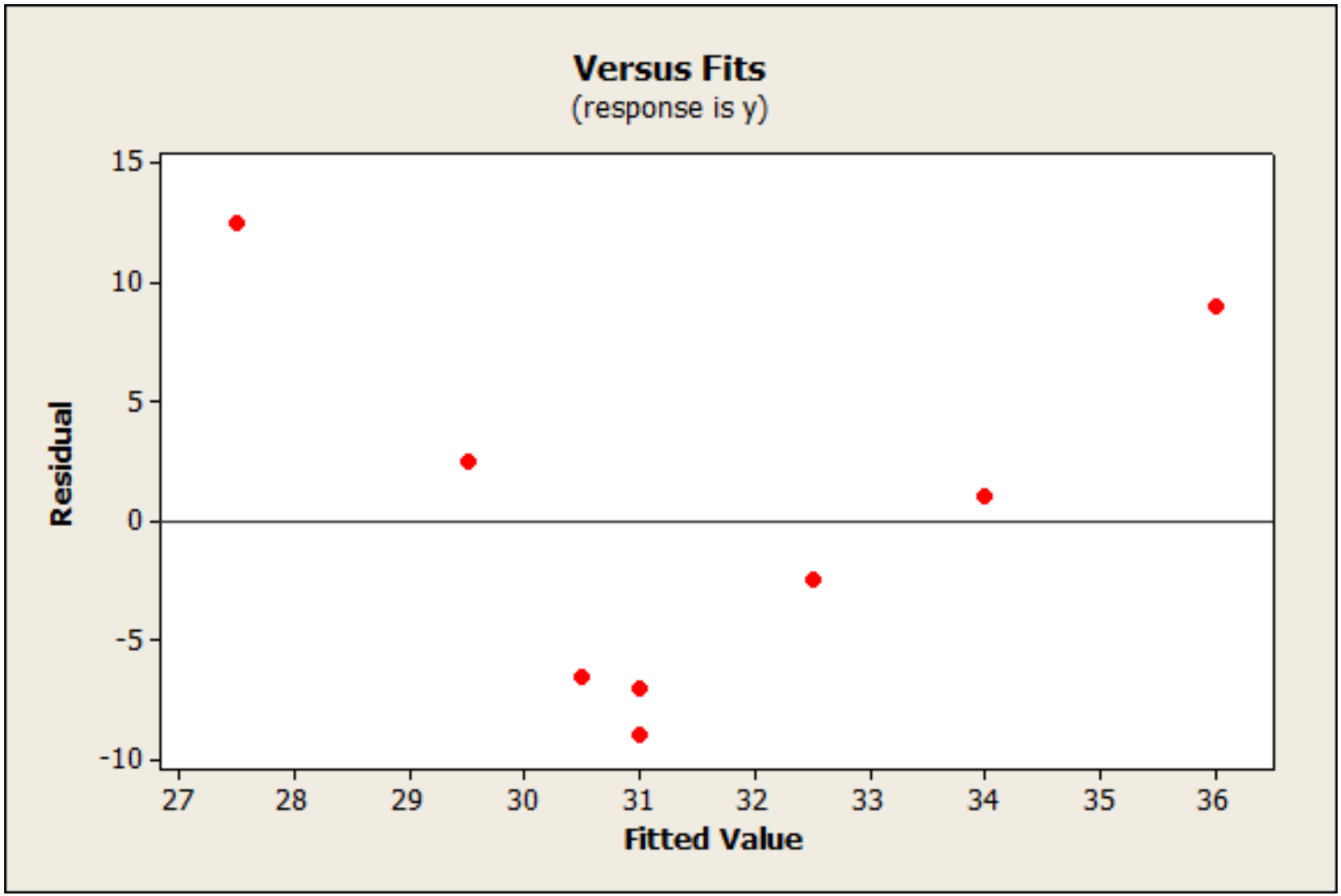


Grafico dei residui relativo all'es. precedente



La seguente tabella riporta le temperature medie giornaliere in gradi Fahrenheit e le corrispondenti precipitazioni piovose misurate in pollici, negli anni compresi tra 1975 e il 1984, in una città degli US

1) riportate il coefficiente di correlazione e la sua unità di misura
2) quale proporzione di variabilità dei valori di y è spiegata dal modello di regressione di y su x?
3) Il modello di regressione lineare descrive bene i dati osservati?

Si può affermare che esiste una relazione di causa-effetto tra le due variabili?

Temp (x)	**Precipit (y)**
78.1	6.23
71.8	3.64
75.6	3.42
72.7	2.84
75.3	1.83
73.6	2.82
75.1	4.04
75.3	2.56
73.8	1.18
70.4	4.19

SOLUZIONI

Per questi dati riportate il coefficiente di correlazione e la sua unità di misura

0.262; adimemsionale

ii) Calcolate il residuo corrispondente alla temperatura 73.6

-0.36

iii) quale proporzione di variabilità dei valori di y è spiegata dal modello di regressione di y su x?

7.2%

iv) Il modello di regressione lineare descrive bene i dati osservati

VERO FALSO

v) Si può affermare che esiste una relazione di causa-effetto tra le due variabili?

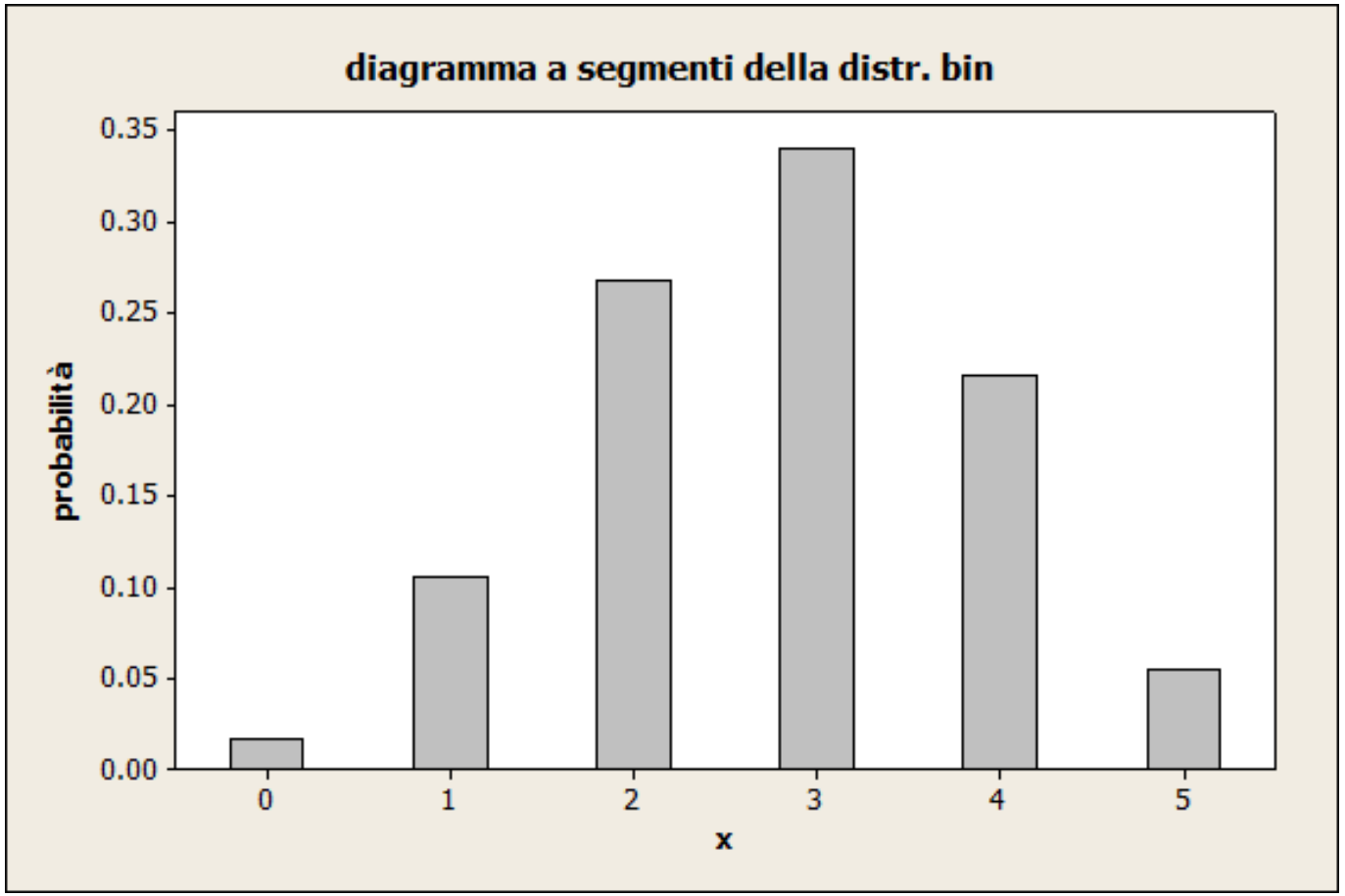
VERO FALSO

Esempio

In una pop. di soggetti affetti da tumore cerebrale il 56% dei malati non presenta crisi epilettiche come primo sintomo.
Si devono esaminare 5 nuovi soggetti e ci si chiede quale sia la prob. che 3 dei 5 non presentino una crisi epilettica come primo sintomo.

Quale v. a.? Come modellizzare il problema? Quali parametri?

Esempio segue



Parametri n=5
p= 0.56

X	p(x)
0	0.016492
1	0.104947
2	0.267137
3	0.339993
4	0.216359
5	0.055073

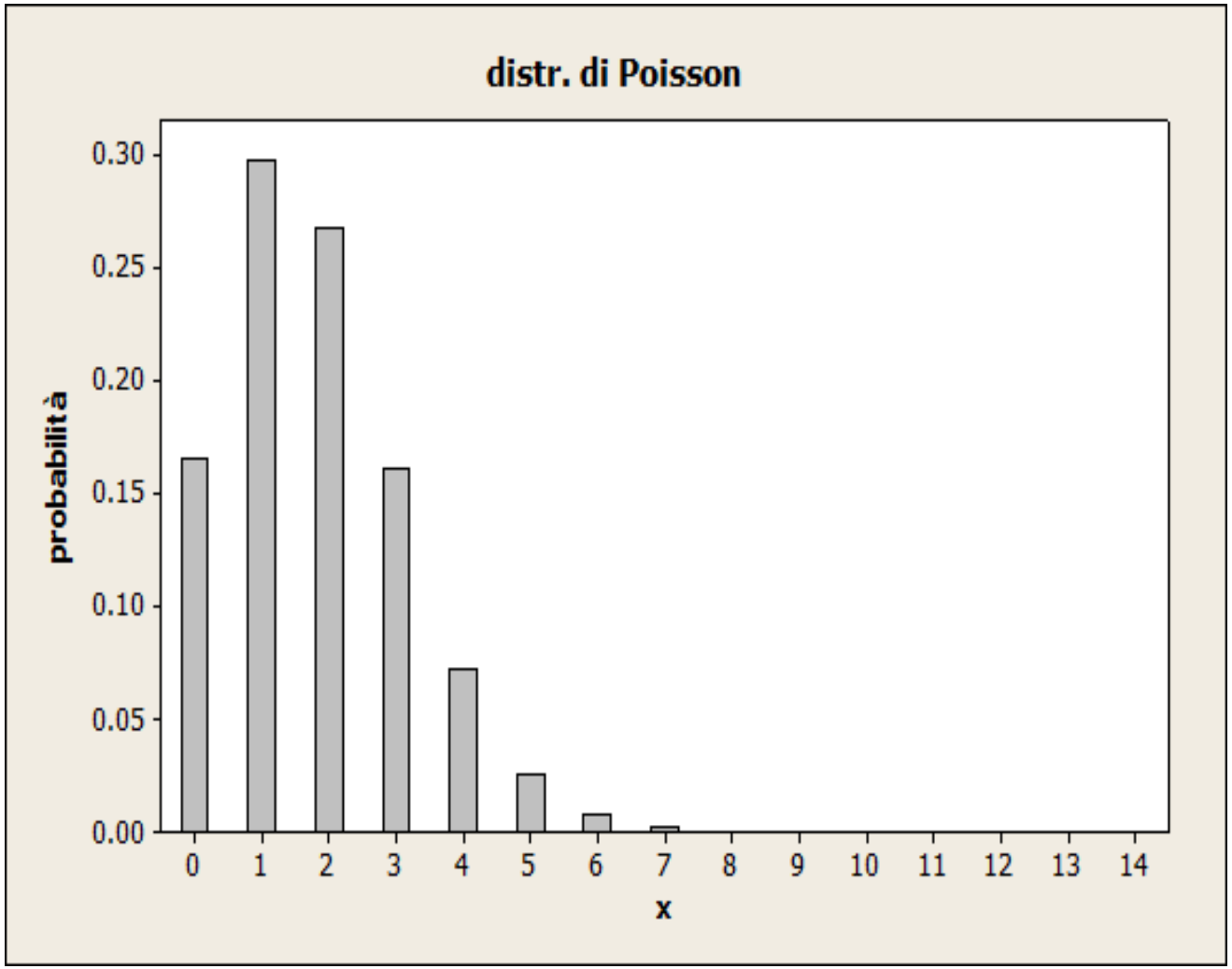
E' molto importante che x e p si riferiscano all'esito identificato come "successo"

Esempio

In un ospedale le nascite avvengono casualmente e ci sono mediamente 1.8 nascite all'ora.
Qual è la prob di osservare 4 nascite fra le 21 e le 22 di un qualsiasi giorno?

Quale v.a.? Quale modello? Quali parametri?

Esempio segue



parametro λ =1.8

x	p(x)
0	0.165299
1	0.297538
2	0.267784
3	0.160671
4	0.072302
5	0.026029
6	0.007809
7	0.002008
8	0.000452
9	0.000090

Esempio

- Per analizzare le tracce delle bombe V-I della seconda guerra mondiale, la zona meridionale di Londra è stata suddivisa in 576 regioni , ognuna delle quali di area 0.25 Km^2. Un totale di 535 bombe ha colpito l'area delle 576 regioni.

  Qual è la prob che, una regione scelta a caso, sia stata colpita 2 volte?

Quale v.a.? Quale modello? Quale parametro?