LA PENDENZA DI UNA RETTA

Psicometria 1 - Lezione 14
Lucidi presentati a lezione

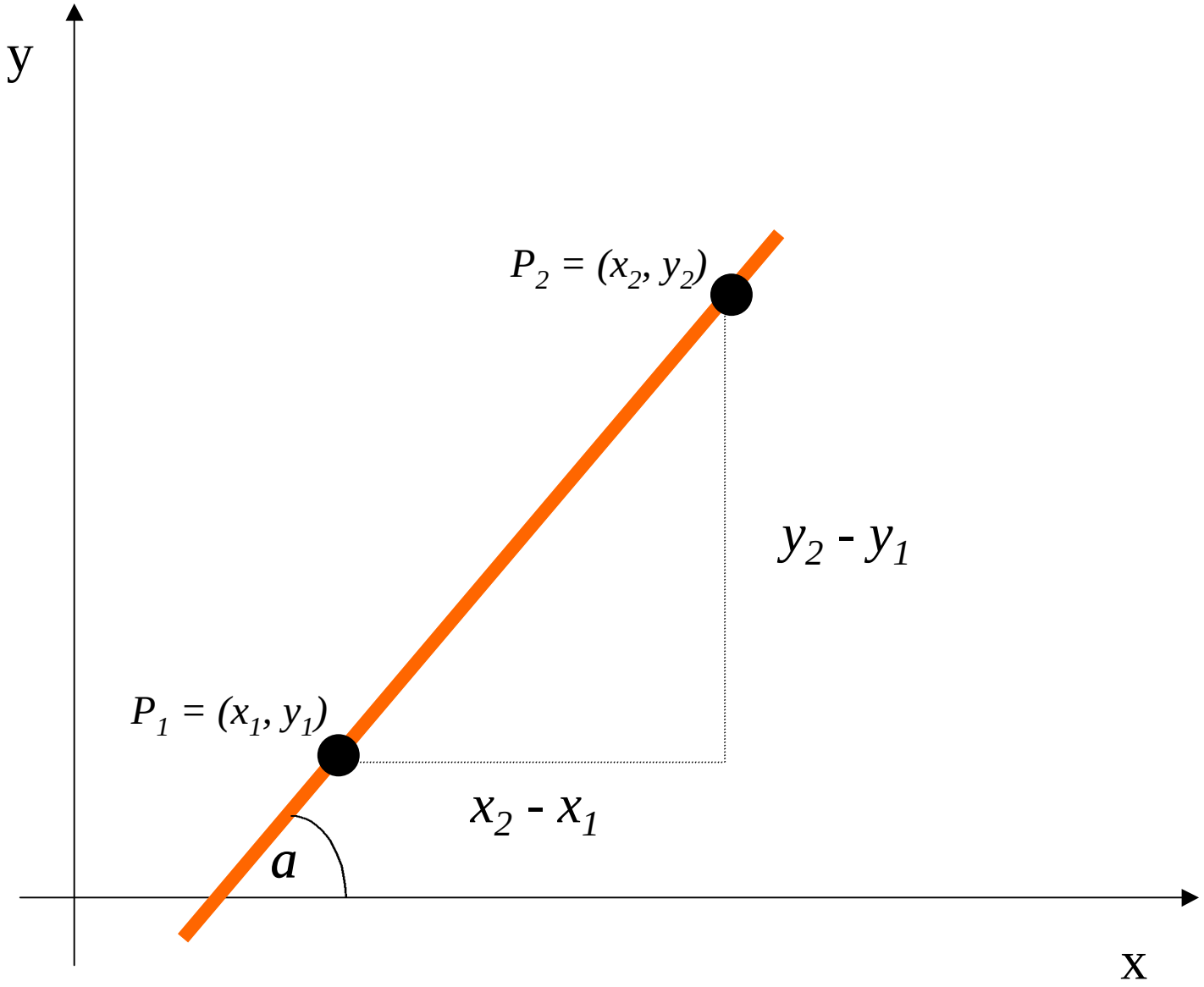
AA 2000/2001 dott. Corrado Caudek

A ciascuna retta non verticale possiamo associare un numero, chiamato *pendenza*, che ne specifica la direzione.

La pendenza di una retta è definita nel modo seguente.

Dati due punti $P_1 = (x_1, y_1)$ e $P_2 = (x_2, y_2)$ che appartengono alla retta, la pendenza della retta è definita come:

$$m = \frac{y_2 - y_1}{x_2 - x_1}$$


y
$P_2 = (x_2, y_2)$
$y_2 - y_1$
$P_1 = (x_1, y_1)$
$x_2 - x_1$
a
x

$P_1 = (2, 1)$ $P_2 = (4, 5)$

$x_2 - x_1 = 4 - 2 = 2$

$y_2 - y_1 = 5 - 1 = 4$

$m = (y_2 - y_1) / (x_2 - x_1) = 4/2 = 2$

$P_1 = (2, 1)$ $P_2 = (4, 5)$

$x_1 - x_2 = 2 - 4 = -2$

$y_1 - y_2 = 1 - 5 = -4$

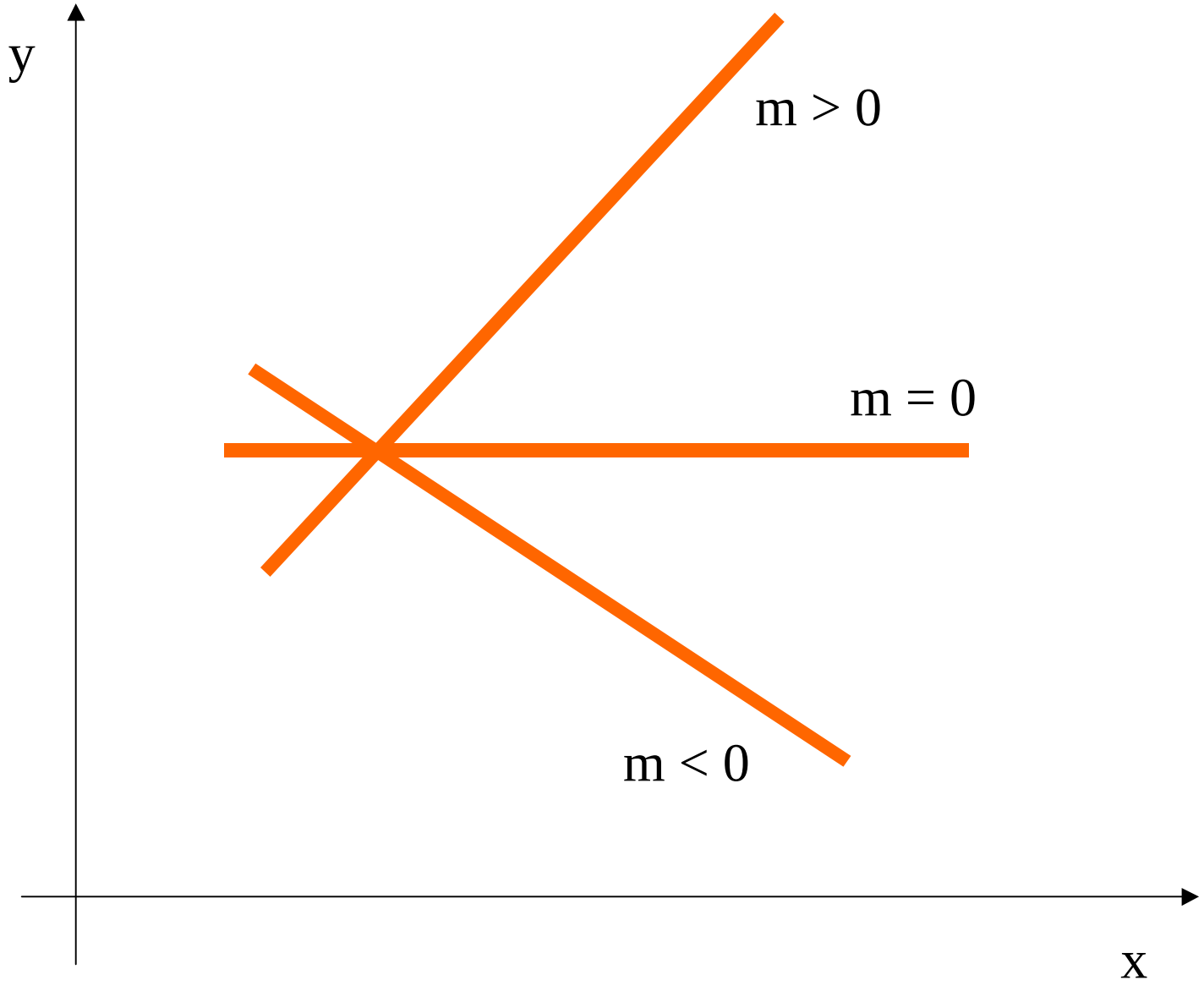
$m = (y_1 - y_2) / (x_1 - x_2) = (-4)/(-2) = 2$

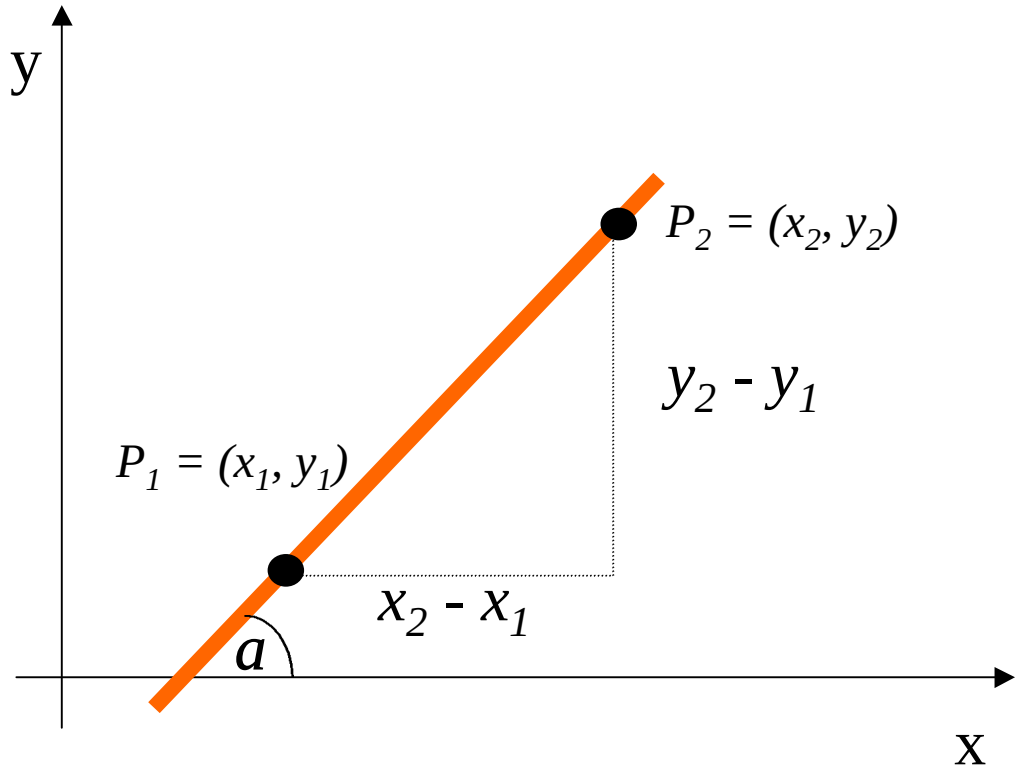
Il valore di *m* dipende soltanto dalla retta e rimane inalterato per qualsiasi coppia di punti appartenenti alla retta.

Se la posizione di P_2 viene scelta in maniera tale che $x_2 - x_1 = 1$, allora

$$m = y_2 - y_1$$

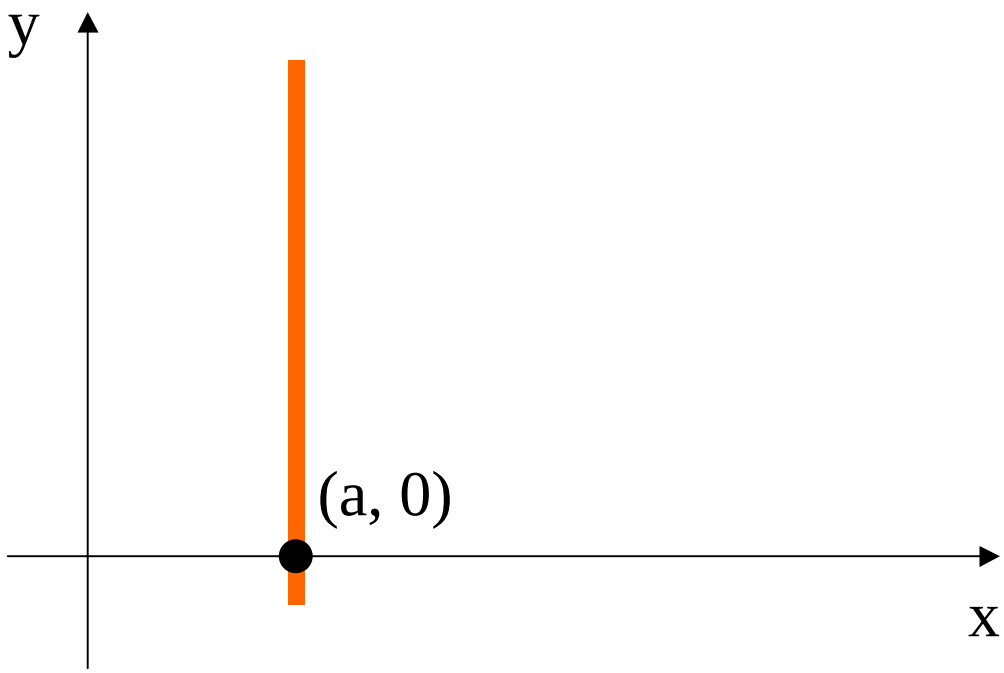
La pendenza della retta non è altro che l'entità del cambiamento in *y* corrispondente ad un cambiamento unitario in *x*.


y
m > 0
m = 0
m < 0
x



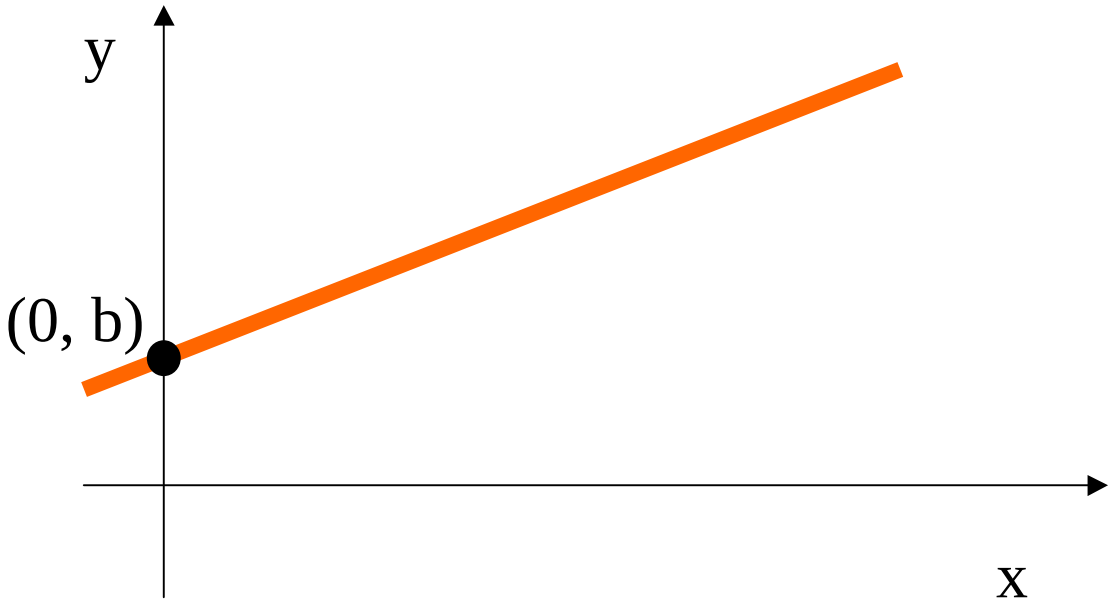
La pendenza della retta non è altro che la tangente dell'angolo *a*: $m = tan\ a$

EQUAZIONE DI UNA RETTA



Una retta verticale è caratterizzata dal fatto che tutti i punti sulla linea hanno la stessa coordinata *x*.

Se la retta interseca l'asse *x* in corrispondenza di del punto (*a*, 0), allora un generico punto *P* si troverà sulla retta se e solo se $x = a$.



Consideriamo ora una retta non verticale con una pendenza nota uguale a *m*.

Sia (0, b) il punto della linea che interseca l'asse y.

Sia (x, y) un secondo punto sul piano cartesiano.

Questo secondo punto apparterrà alla retta considerata se e solo se

$$\frac{y-b}{x-0}=m$$

$$\frac{y-b}{x-0}=m$$

$$y-b=mx$$

L’equazione di una retta non verticale diventa dunque

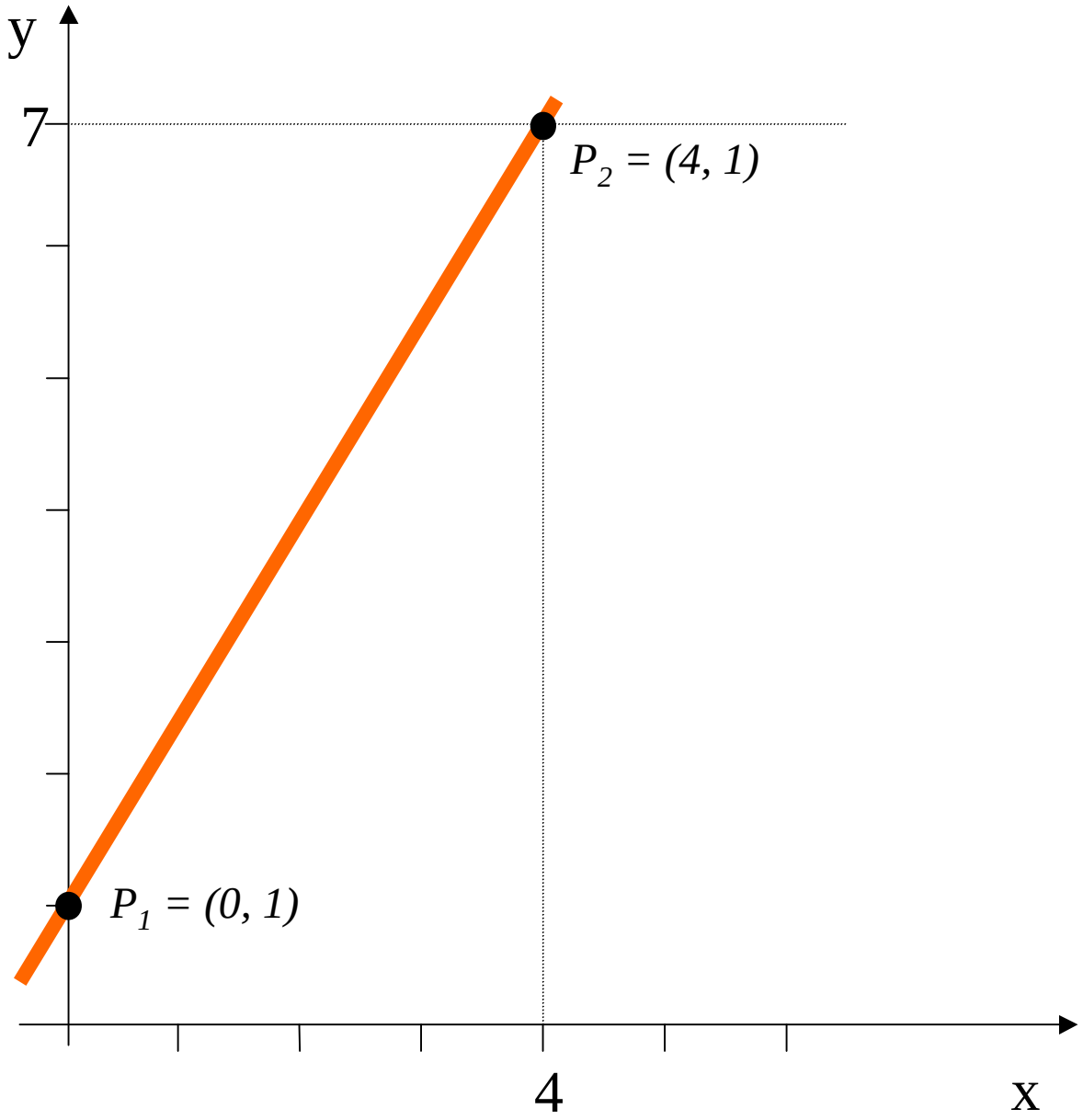
$$y=mx+b$$

Esempio. Sia *P* un punto che appartiene alla retta con pendenza $m = 3.5$. Sappiamo inoltre che la retta interseca l'asse y in corrispondenza del punto (0, 9).

Se la coordinata *x* del punto *P* è uguale a 2, quale è la coordinata *y* di *P*?

$$y = mx + b = 3.5 \times 2 + 9 = 16$$

Esempio. Disegnate la retta con pendenza $m = 1.5$ e passante per il punto (0, 1).


y
7
$P_2 = (4, 1)$
$P_1 = (0, 1)$
4
x

Analisi della regressione

Fino a questo punto abbiamo assunto che il valore atteso delle variabili aleatorie fosse costante, ovvero, non dipendesse dal valore di altre variabile.

Considereremo ora il caso in cui il valore atteso di una variabile aleatoria Y (chiamata *variabile dipendente*) è funzione di una variabile non aleatoria X (chiamata *variabile indipendente*).

Benché infinite funzioni diverse possano essere utilizzate per descrivere la relazione che intercorre tra il valore atteso della variabile dipendente e la variabile indipendente, esamineremo il caso in cui la relazione tra X e $E(Y)$ può essere descritta da una funzione lineare.

Il ***modello della regressione lineare*** che mette in relazione il valore atteso della variabile dipendente con la variabile indipendente è

$$E(Y) = a + bX$$

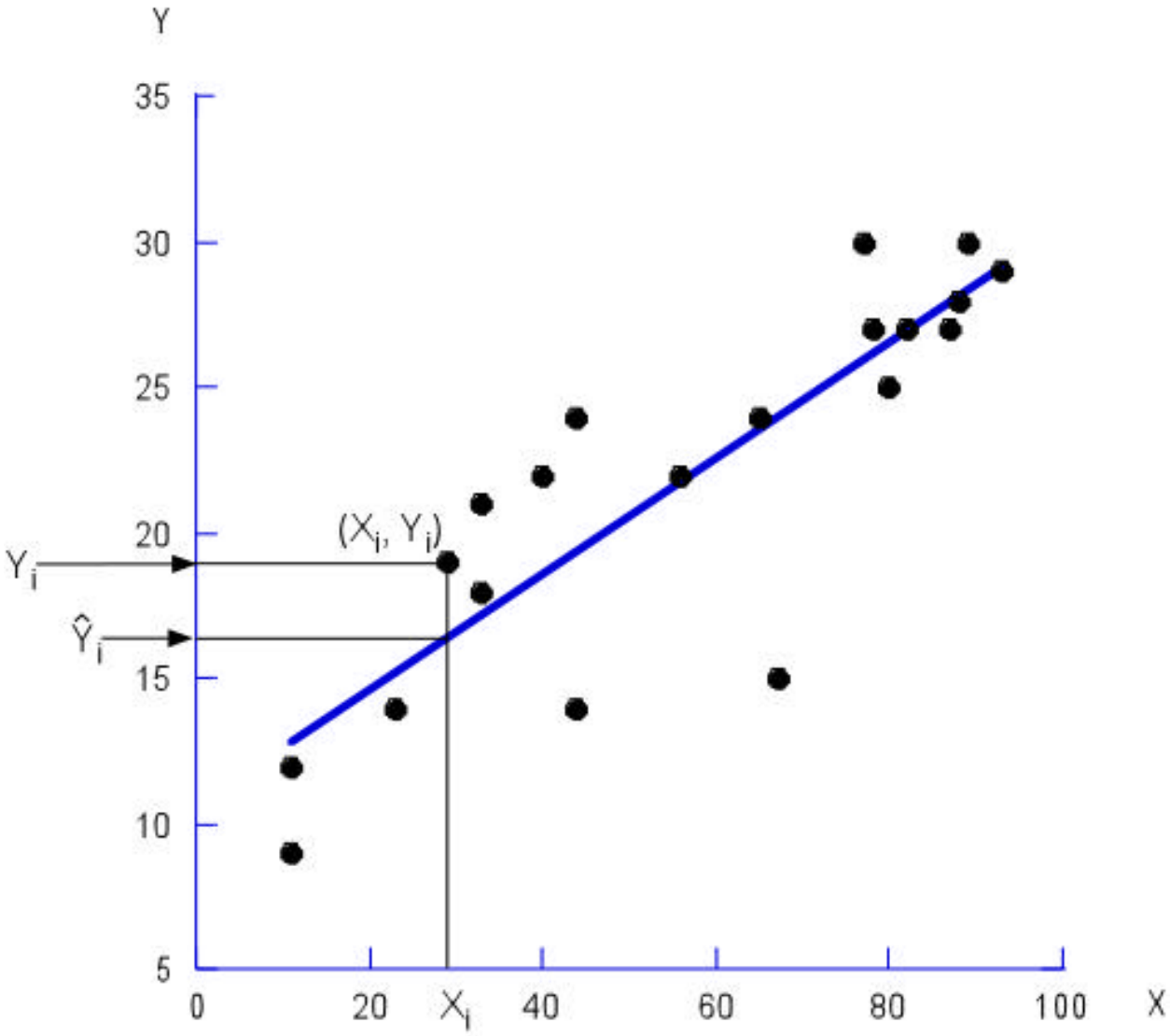
Il termine a denota l'*intercetta* e il termine b denota la *pendenza* della retta che mette in relazione $E(Y)$ con X.

In maniera equivalente, il modello della regressione lineare può essere scritto come

$$Y = a + bX + e$$

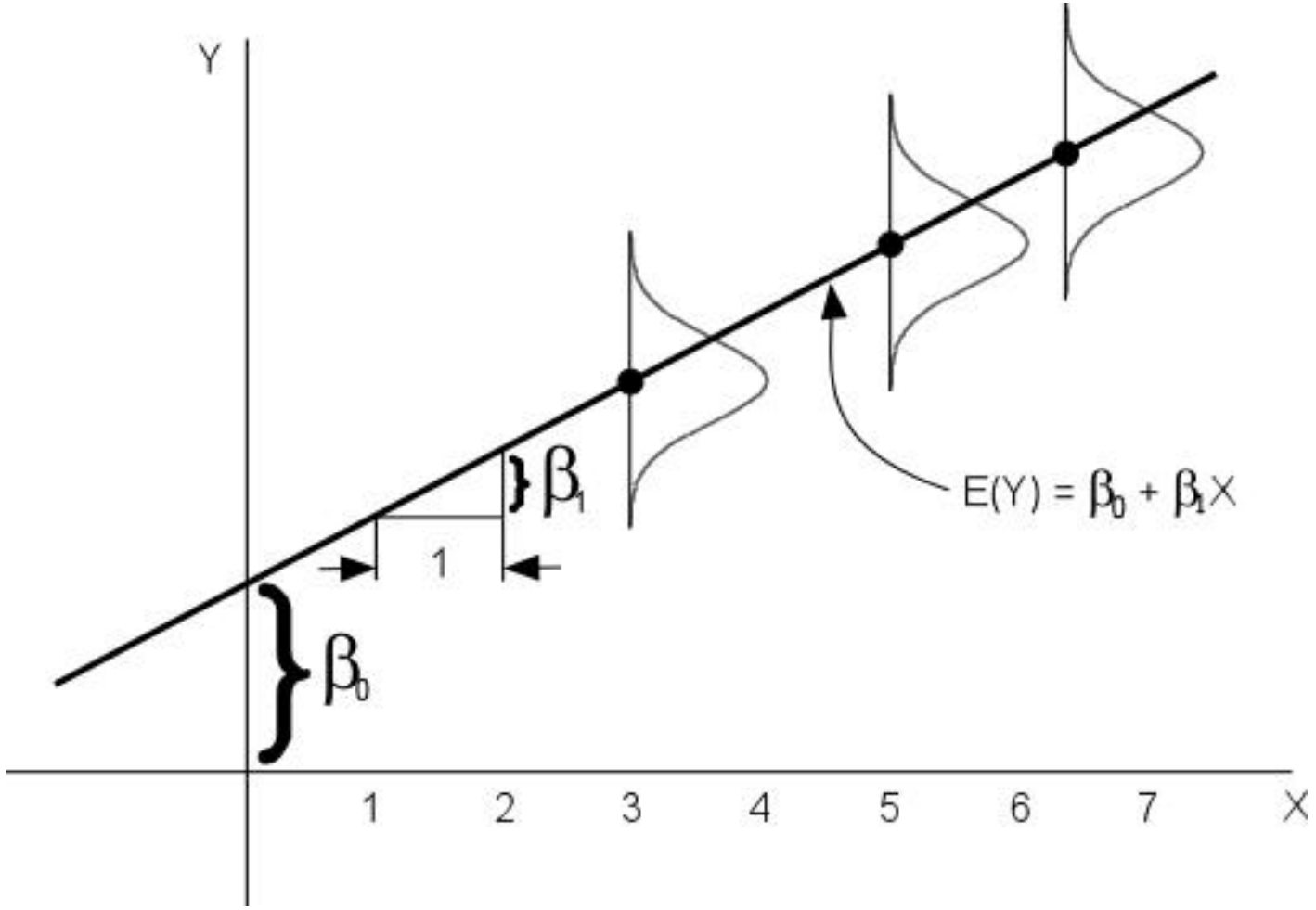
laddove e rappresenta una variabile aleatoria avente una specifica distribuzione di probabilità con media uguale a zero.

Nel modello della regressione bivariata la variabile aleatoria *Y* viene concepita come la somma di una componente deterministica, *E(Y)* (completamente predicibile dalla variabile indipendente), e di una componente aleatoria, *e*.


Y
35
30
25
20
15
10
5
(X_i, Y_i)
Y_i
$\hat{Y}_i$
0
20
X_i
40
60
80
100
X

Modello probabilistico della regressione bivariata

Consideriamo il modello probabilistico $Y = a + bX + e$.



In corrispondenza di $X = 3$, ad esempio, c'è una popolazione di possibili valori Y. Questa popolazione ha media $a + b(3)$ e varianza σ^2.

In corrispondenza di $X = 5$ avremo un'altra popolazione di possibili valori Y. Questa seconda popolazione avrà la stessa forma e varianza della precedente ma, quando $X = 5$, la distribuzione di Y avrà media uguale a $a + b(5)$.

Assunzioni del modello probabilistico della regressione bivariata

Linearità. Le variabili Y e X sono linearmente associate:

$$E(Y_i|X_i) = m_i = a + bX_i .$$

Da questo deriva che $E(e) = 0$.

Dato che $e_i \equiv Y_i - m_i$, infatti,

$$E(e) = E(Y_i - m_i) = E(Y_i) - m_i = m_i - m_i = 0 .$$

Omoschedasticità. La variabilità attorno alla retta della regressione nella popolazione è costante: $s^2_{e_i} = s^2_{e_j}$ per tutti gli *i* e *j*.

Dato che $e_i \equiv Y_i - m_i$, l'errore ha dunque una distribuzione identica a quella di Y_i eccetto per il suo valore atteso.

Normalità. I residui hanno una distribuzione normale:

$$e_i \sim N(0, s_e^2).$$

Questo significa anche che $Y_i \sim N\left(a + bX_i, s_e^2\right)$.

Indipendenza. Le osservazioni sono state campionate in maniere indipendente: e_i e e_j sono indipendenti per tutti gli $i \neq j$.

L'assunzione di indipendenza deve essere giustificata dalla procedura usata per raccogliere i dati.

Metodo dei minimi quadrati

I modelli statistici lineari si pongono tre problemi:

(i) stabilire l'orientamento della retta che più di ogni altra si avvicina ai punti X_i, Y_i che rappresentano le *n* osservazione del campione;

(ii) inferire i parametri che definiscono la retta di regressione nella popolazione da cui il campione è stato estratto;

(iii) stabilire in che misura la retta di regressione si approssima ai dati.

Consideriamo il primo di questi 3 problemi.

Per l'*i*-esima delle *n* osservazioni di un campione, il modello della regressione lineare è

$$Y_i = A + BX_i + E_i$$

Se indichiamo con $\hat{Y}_i$ il valore predetto dal modello lineare, allora l'errore della predizione sarà:

$$E_i \equiv Y_i - \hat{Y}_i$$

L'errore della predizione rappresenta la porzione della variabile dipendente che non può essere predetta dalla variabile indipendente; E_i è anche chiamato residuo.

Per trovare la retta che giunge il più vicino possibile a *n* osservazioni di un campione è necessario decidere anzitutto come misurare la distanza tra ciascuna delle osservazioni e la retta.

Una misura di questa distanza è data dal termine d'errore E_i .

E_i rappresenta infatti la distanza verticale tra la retta di regressione e l'*i*-esima osservazione.

Diventa poi necessario trovare un indice che fornisca una misura complessiva di tutti gli scostamenti tra le osservazioni del campione e la retta di regressione.

L'indice più semplice a questo proposito è la somma dei residui, $\sum_{i=1}^{n} E_i$.

Questo indice, però, è di poca utilità in quanto i residui possono essere sia positivi che negativi e la loro somma può essere molto prossima allo zero anche per differenze molto grandi tra le osservazioni e la retta di regressione.

In particolare, se la retta di regressione passa per il punto $(\overline{X}, \overline{Y})$, allora $\sum E_i = 0$.

Una tale retta di regressione soddisfa l'equazione $\overline{Y} = A + B\overline{X}$.

$$Y_i - \overline{Y} = A + BX_i + E_i - \left(A + B\overline{X}\right)$$

$$Y_i - \overline{Y} = B\left(X_i - \overline{X}\right) + E_i$$

$$E_i = \left(Y_i - \overline{Y}\right) - B\left(X_i - \overline{X}\right)$$

$$\sum E_i = \sum \left(Y_i - \overline{Y}\right) - B\sum \left(X_i - \overline{X}\right) = 0$$

La somma dei residui non consente quindi di stabilire in che misura la retta di regressione si approssima ai dati dal momento che tutte le rette passanti per il punto $(\overline{X}, \overline{Y})$ rendono questo indice uguale a zero.

Un indice migliore si ottiene elevando al quadrato i residui in modo tale che abbiano sempre valore positivo:

$$SQ_{ERR} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

La procedura che consente di trovare la retta per la quale SQ_{ERR} assume il minore valore possibile va sotto il nome di *metodo dei minimi quadrati*.

I valori predetti dalla retta di regressione sono uguali a

$$\hat{Y}_i = A + BX_i$$

e i valori osservati in funzione dei coefficienti A e B sono

$$Y_i = A + BX_i + E_i$$

$$SQ_{ERR} = \sum_{i=1}^{n} \left(Y_i - A - BX_i \right)^2$$

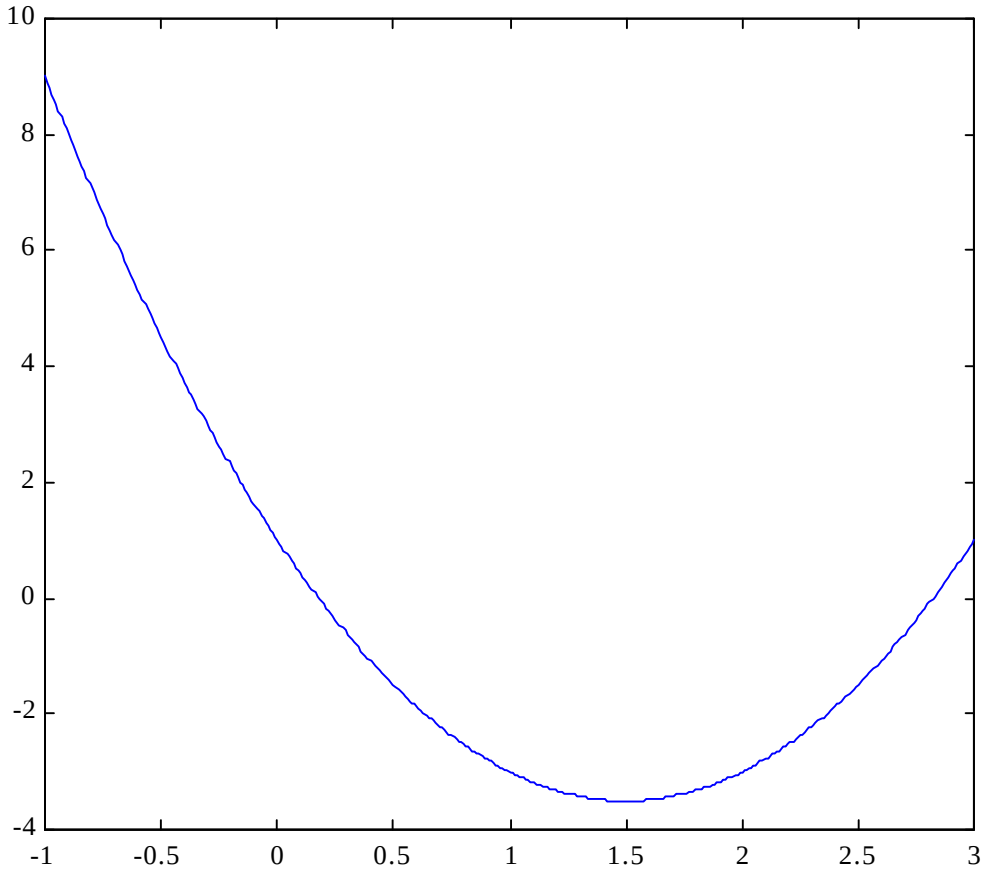
$$SQ_{ERR} = \sum_{i=1}^{N} \left(Y_i^2 + A^2 + B^2 X_i^2 - 2AY_i - 2BY_i X_i + 2ABX_i \right)$$

L'equazione precedente è un'equazione quadratica in A e B e può essere rappresentata da una parabola.

Il minimo della parabola si calcola ponendo uguali a zero le derivate dell'equazione rispetto a A e B.

Esempio

$$y = 2x^2 - 6x + 1$$



La derivata prima dell'equazione $y = 2x^2 - 6x + 1$ è

$$(2 \acute{} 2)\, x - 6$$

Ponendo la derivata uguale a zero otteniamo:

$$4x - 6 = 0$$

$$x = 6/4 = 1.5$$

Il che corrisponde al minimo dell'equazione $y = 2x^2 - 6x + 1$

$$SQ_{ERR} = \sum_{i=1}^{N} \left(Y_i^2 + A^2 + B^2 X_i^2 - 2AY_i - 2BY_i X_i + 2ABX_i \right)$$

$$\frac{\P SQ_{ERR}}{\P B} = \sum_{i=1}^{n} \left(2BX_i^2 - 2Y_i X_i + 2AX_i \right)$$

$$= 2\left(B\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} Y_i X_i + A\sum_{i=1}^{n} X_i \right)$$

$$\sum_{i=1}^{n} Y_i X_i = A\sum_{i=1}^{n} X_i + B\sum_{i=1}^{n} X_i^2$$

$$SQ_{ERR} = \sum_{i=1}^{N} \left(Y_i^2 + A^2 + B^2 X_i^2 - 2AY_i - 2BY_i X_i + 2ABX_i \right)$$

$$\frac{\P SQ_{ERR}}{\P A} = \sum_{i=1}^{n} \left(2A - 2Y_i + 2BX_i \right)$$

$$= 2\left(\sum_{i=1}^{n} A - \sum_{i=1}^{n} Y_i + B \sum_{i=1}^{n} X_i \right)$$

$$\sum_{i=1}^{n} Y_i = nA + B \sum_{i=1}^{n} X_i$$

Si ottiene così un sistema di due equazioni in due incognite:

$$\sum_{i=1}^{n} Y_i X_i = A\sum_{i=1}^{n} X_i + B\sum_{i=1}^{n} X_i^2$$

$$\sum_{i=1}^{n} Y_i = nA + B\sum_{i=1}^{n} X_i$$

Risolvendo, otteniamo:

$$A = \overline{Y} - B\overline{X}$$

$$B = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2}$$

Esempio

X	Y
1	1.2000
2	3.8000
3	1.8000
4	4.6000
5	4.1000
6	7.0000

$$\overline{X} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\overline{Y} = \frac{1.2+3.8+1.8+4.6+4.1+7}{6} = 3.75$$

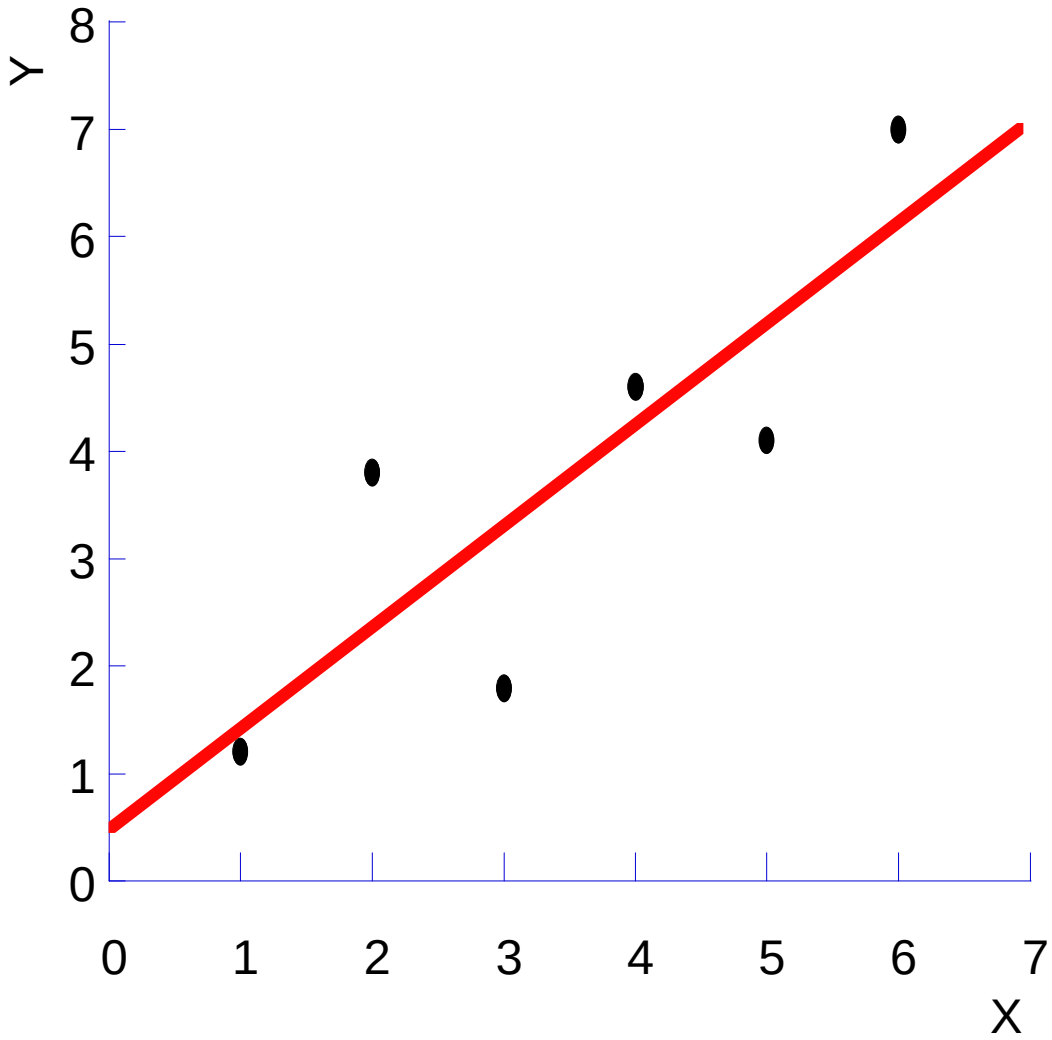
$$B = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2}$$

S_{xy} = ((1-3.5)(1.2-3.75)+(2-3.5)(3.8-3.75)+ … +(6-3.5)(7-3.75))/6
= 2.725

S^2_x = ((1-3.5)(1-3.5) + (2-3.5)(2-3.5) + … + (6-3.5)(6-3.5))/6
= 2.9167

$$B = \frac{\sum \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum \left(X_i - \overline{X}\right)^2} = \frac{2.725}{2.9167} = .9347$$

$$A = \overline{Y} - B\overline{X} = 3.65 - .9347 \times 3.5 = .48$$


Y
8
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6
7
X

L'equazione relativa al coefficiente A implica che la retta di regressione passa per il punto $(\overline{X}, \overline{Y})$.

Di conseguenza, la somma dei residui della retta di regressione calcolata con il metodo dei minimi quadrati è uguale a zero.

Questo implica inoltre che il valore atteso dei residui calcolati con il metodo dei minimi quadrati è uguale a zero.

Interpretazione
dei coefficienti di regressione

Sia la covarianza tra le variabili aleatorie X e Y uguale a

$$S_{XY} = \frac{\sum \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right)}{n}$$

Il coefficiente di regressione B calcolato con il metodo dei minimi quadrati è dunque uguale al rapporto tra la covarianza di X e Y e la varianza di X:

$$B = \frac{S_{XY}}{S_X^2}$$

Ai coefficienti di regressione si può assegnare la seguente interpretazione.

Il coefficiente *A* rappresenta il valore predetto di *Y* in corrispondenza di $X = 0$.

Il coefficiente *B* rappresenta l'incremento predetto del valore atteso della variabile dipendente per un incremento unitario della variabile indipendente.

Calcolo degli errori della predizione

$$E_i = Y_i - \hat{Y}_i$$

$$\hat{Y}_i = A + BX_i$$

$$\hat{Y}_1 = .48 + 0.9343 \times 1 = 1.4143$$

$$E_1 = Y_1 - \hat{Y}_1 = 1.2 - 1.4143 = -0.2143$$

Y	Y_{pred}	E
1.2000	1.4143	-0.2143
3.8000	2.3486	1.4514
1.8000	3.2829	-1.4829
4.6000	4.2171	0.3829
4.1000	5.1514	-1.0514
7.0000	6.0857	0.9143

$$\sum E_i = 0$$