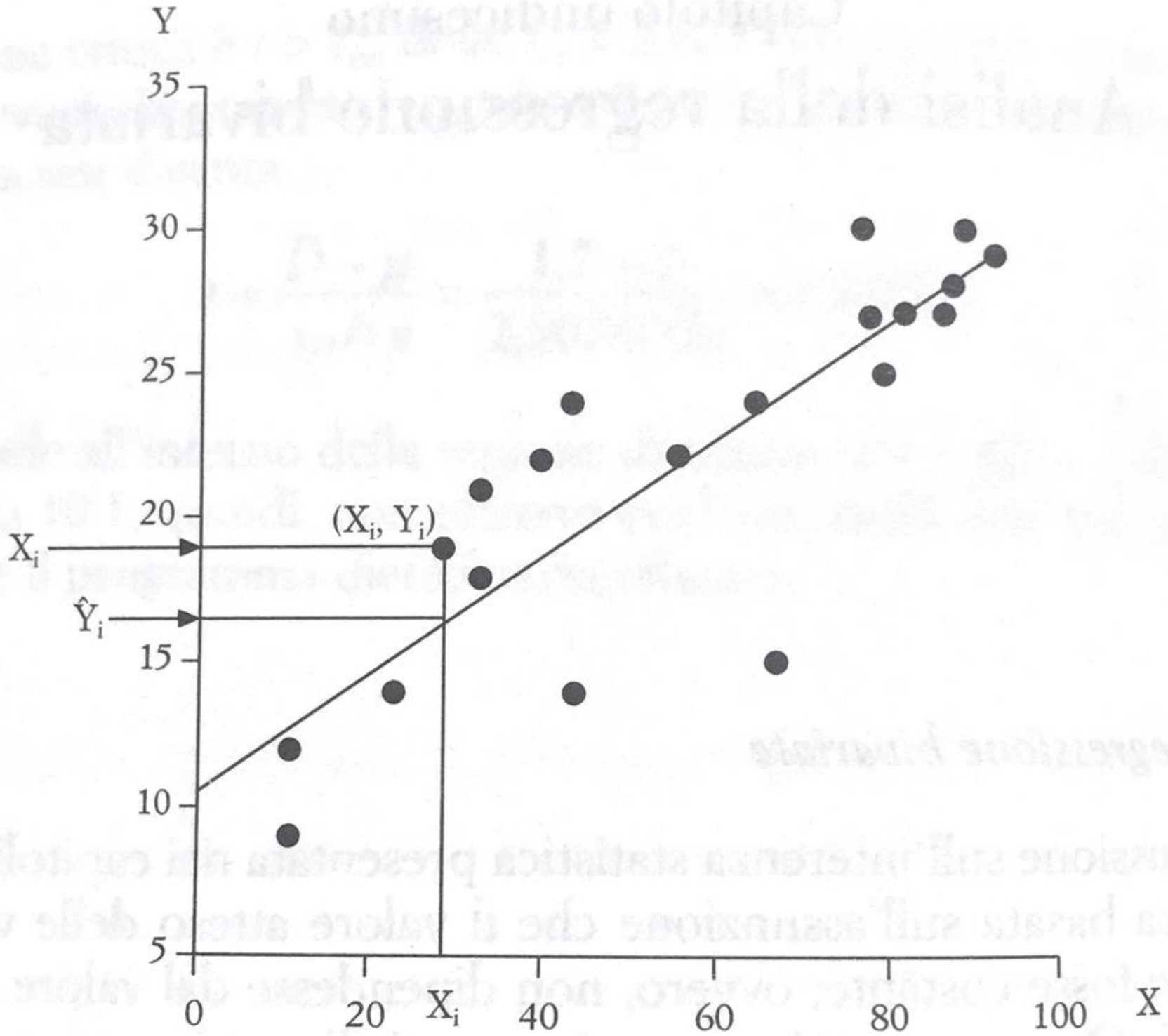


Figura 11.1. Diagramma di dispersione per i dati dell'esempio 11.1.

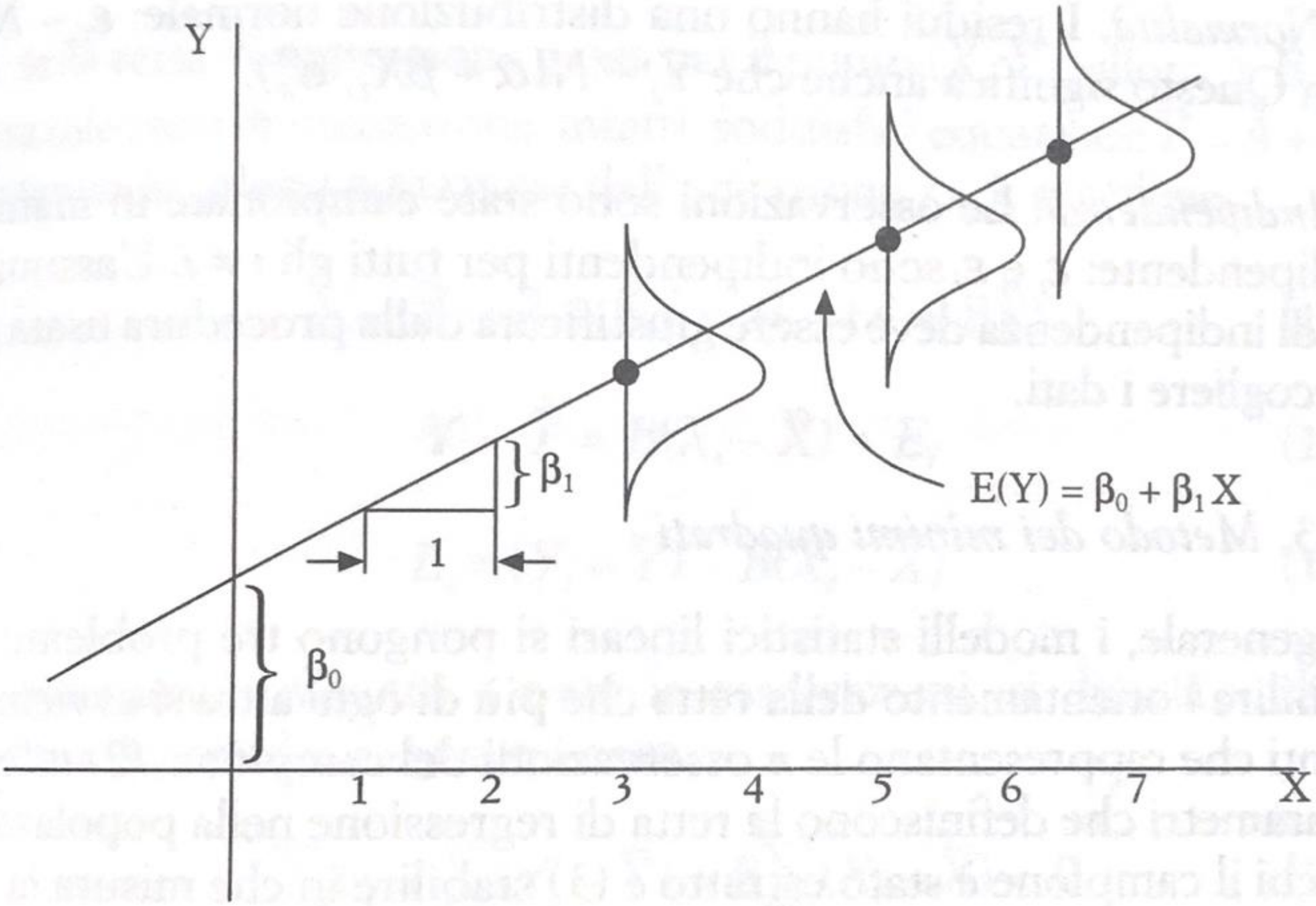


Figura 11.2. Rappresentazione grafica del modello probabilistico $Y = \alpha + \beta X + \varepsilon$.

$$SQ_{ERR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (11.8)$$

Una volta deciso di utilizzare la *somma dei quadrati degli errori* (SQ_{ERR}) come indice complessivo degli scostamenti tra le osservazioni del campione e la retta di regressione, il problema diventa quello di trovare la retta per la quale SQ_{ERR} assume il minore valore possibile. La procedura solitamente usata per raggiungere questo obiettivo va sotto il nome di *metodo dei minimi quadrati*.

I valori predetti dalla retta di regressione sono

$$\hat{Y}_i = A + BX_i \qquad (11.9)$$

Possiamo dunque scrivere

$$SQ_{ERR} = \sum_{i=1}^{n} (Y_i - A - BX_i)^2 \qquad (11.11)$$

L'equazione 11.11 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 11.10 rispetto ad A e B. Sviluppando l'equazione 11.11 si ottiene

$$SQ_{ERR} = \sum_{i=1}^{n} (Y_i^2 + A^2 + B^2X_i^2 - 2AY_i - 2BY_iX_i + 2ABX_i) \qquad (11.12)$$

$$\frac{\partial SQ_{ERR}}{\partial A} = \sum_{i=1}^{n} (2A - 2Y_i + 2BX_i) \qquad (11.13)$$

$$= 2\left(\sum_{i=1}^{n} A - \sum_{i=1}^{n} Y_i + B\sum_{i=1}^{n} X_i\right) \qquad (11.14)$$

$$\frac{\partial SQ_{ERR}}{\partial B} = \sum_{i=1}^{n} (2BX_i^2 - 2Y_iX_i + 2AX_i) \qquad (11.15)$$

$$= 2\left(B\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} Y_iX_i + A\sum_{i=1}^{n} X_i\right) \qquad (11.16)$$

$$\begin{bmatrix} A \\ B \end{bmatrix} = \begin{bmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y_i \\ \sum Y_i X_i \end{bmatrix} \tag{11.20}$$

ovvero

$$A = \bar{Y} - B\bar{X} \tag{11.21}$$

$$B = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \tag{11.22}$$

Dato che la covarianza tra X e Y è

$$S_{XY} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n} \tag{11.23}$$

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

$$B = \frac{S_{XY}}{S_X^2} \tag{11.24}$$

Tabella 11.1. Punteggi nel test di pre-iscrizione e voto finale di 20 studenti.

Studenti	Voto nel test di pre-iscrizione	Voto finale nel corso		
	X_i	Y_i	$\hat{Y}_i$	E_i
1	80	25	26,5535	–1,5535
2	40	22	18,5476	3,4524
3	44	24	19,3482	4,6518
4	82	27	26,9537	0,0463
5	65	24	23,5512	0,4488
6	77	30	25,9530	4,0470
7	23	14	15,1451	–1,1451
8	87	27	27,9545	–0,9545
9	78	27	26,1532	0,8468
10	33	18	17,1465	0,8535
11	56	22	21,7499	0,2501
12	88	28	28,1546	–0,1546
13	89	30	28,3548	1,6452
14	67	15	23,9515	–8,9515
15	11	12	12,7433	–0,7433
16	29	19	16,3460	2,6540
17	93	29	29,1554	–0,1554
18	11	9	12,7433	–3,7433
19	44	14	19,3482	–5,3482
20	33	21	17,1465	3,8535

Per il primo studente, ad esempio, il valore predetto della prova finale è

$$\hat{Y}_1 = A + BX_1 = 10{,}5417 + 0{,}2001 \cdot 80 = 26{,}5535$$

e l'errore della predizione è

$$E_1 = Y_1 - \hat{Y}_1 = 25 - 26{,}5535 = -1{,}5535$$

Tabella 11.1. Punteggi nel test di pre-iscrizione e voto finale di 20 studenti.

Studenti	Voto nel test di pre-iscrizione X_i	Voto finale nel corso Y_i	$\hat{Y}_i$	E_i
1	80	25	26,5535	–1,5535
2	40	22	18,5476	3,4524

11.5. *Verifica di ipotesi relative ai coefficienti di regressione*

Come si esegue un test statistico a proposito dei coefficienti di regressione? Supponiamo, ad esempio, che l'ipotesi nulla sia $H_0 : \beta = \beta_0$, dove β_0 è una costante (spesso $\beta_0 = 0$). Se la varianza degli errori nella popolazione σ_ε^2 fosse conosciuta, l'ipotesi nulla potrebbe essere sottoposta a verifica usando la statistica

$$z = \frac{B - \beta_0}{\sqrt{\frac{\sigma_\varepsilon^2}{\sum (X_i - \bar{X})^2}}} \tag{11.40}$$

con $z \sim N(0,1)$. Solitamente, però, il parametro σ_ε^2 non è conosciuto e deve essere stimato sulla base delle osservazioni del campione. Può essere dimostrato che uno stimatore privo di errore sistematico di σ_ε^2 è fornito da

$$\frac{SQ_{ERR}}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} \tag{11.41}$$

laddove il valore 2 presente al denominatore dell'equazione 11.41 corrisponde al numero di parametri stimati dal modello (α, β).

$$\frac{SQ_{ERR}}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} \tag{11.41}$$

laddove il valore 2 presente al denominatore dell'equazione 11.41 corrisponde al numero di parametri stimati dal modello (α, β). Se la varianza degli errori σ_ε^2 viene stimata sulla base dei dati del campione, dunque, l'ipotesi nulla $H_0 : \beta = \beta_0$ può essere sottoposta a verifica usando la statistica

$$t = \frac{B - \beta_0}{\sqrt{\dfrac{\sum(Y_i - \hat{Y}_i)^2}{(n-2)\sum(X_i - \bar{X})^2}}} \tag{11.42}$$

distribuita come t con $(n-2)$ gradi di libertà. Notate che il valore della statistica 11.42 cresce quando: (1) la varianza dei residui è piccola, (2) la varianza di X è grande e (3) le dimensioni del campione sono grandi. Una procedura simile a quella sopra descritta viene usata per le inferenze riguardanti l'intercetta.

Esempio 11.2. Si usino i dati della tabella 11.1 per sottoporre a verifica l'ipotesi nulla $H_0 : \beta = 0$ contrapposta all'ipotesi sostantiva $H_1 : \beta > 0$.

$H_1 : \beta > 0$. L'ipotesi nulla afferma che, nella popolazione da cui il campione è stato estratto, il punteggio nel test di pre-iscrizione non consente di predire il punteggio finale del corso. In base all'ipotesi sostantiva, invece, nella popolazione vi è una significativa associazione tra il punteggio finale e il punteggio del test di pre-iscrizione.

La somma dei quadrati degli errori è uguale a

204,2297

Una stima della varianza degli errori nella popolazione è data da

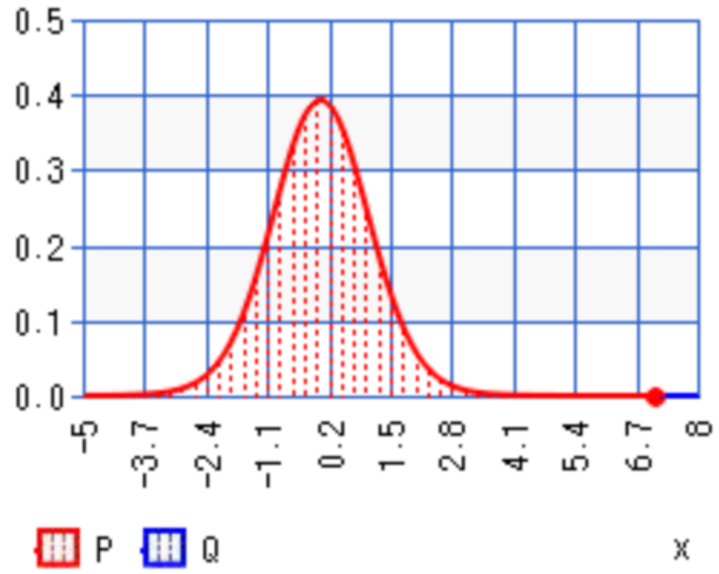
$$S_E^2 = \frac{SQ_{ERR}}{n-2} = \frac{204{,}2297}{18} = 11{,}3461$$

L'errore standard stimato di B è quindi uguale a

$$S_B = \sqrt{\frac{S_E^2}{\sum (X_i - \bar{X})^2}} = \sqrt{\frac{11{,}3461}{14287}} = 0{,}0282$$

$$t_0 = \frac{B - \beta}{S_B} = \frac{0{,}2001 - 0}{0{,}0282} = 7{,}1023$$

e si distribuisce come t con $20 - 2 = 18$ gradi di libertà. Siccome l'ipotesi sostantiva è monodirezionale, è necessario calcolare l'area sottesa alla funzione di densità di probabilità $t(18)$ nella sola coda destra della distribuzione, $P(t > t_0)$. Dato che il valore di probabilità così trovato è minore di 0,05, l'ipotesi nulla può essere rigettata. È quindi possibile concludere che i punteggi ottenuti alla fine del corso aumentano in maniera significativa all'aumentare del punteggio ottenuto nel test di pre-iscrizione.



https://keisan.casio.com/exec/system/1180573203

student's t-distribution	value
● probability density **f**	1.21441E-6
lower cumulative **P**	0.999999
upper cumulative **Q**	6.3909E-7

percentile x 7.1023

degree of freedom v 18 v > 0

Execute Clear Store/Read Print 6digit

kelsan
Online Calculator

Welcome, Guest

Service · How to use · Sample calculation

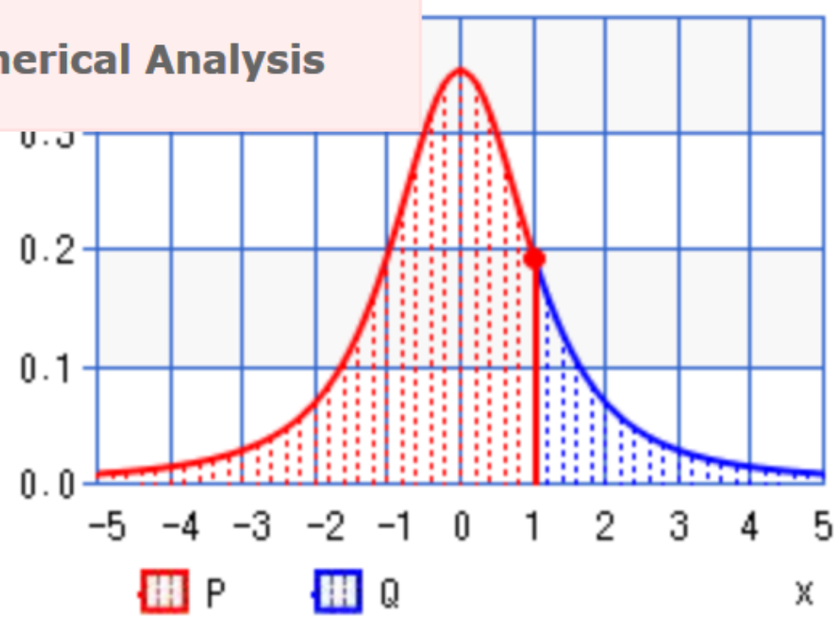
Life | Education | Professional | Shared | Private | Column | Advanced Cal

- Probablility Function
- Special Function
- Numerical Integration
- Linear Algebra
- Numerical Analysis

Student's t-distr

Home / Probability Function

Calculates the probabilit wer and upper cumulative distribution functions of the student's t-distributi



percentile x 1

degree of freedom v 2 v > 0

Coefficiente di correlazione di Pearson

$$r_{XY} \equiv \frac{\sum z_{x_i} z_{y_i}}{n} \tag{11.59}$$

Sviluppando l'equazione 11.59, si ottiene

$$r_{XY} = \frac{\sum (X_i - \bar{X}) \sum (Y_i - \bar{Y})}{n S_X S_Y} \tag{11.60}$$

$$= \frac{S_{XY}}{S_X S_Y} \tag{11.61}$$

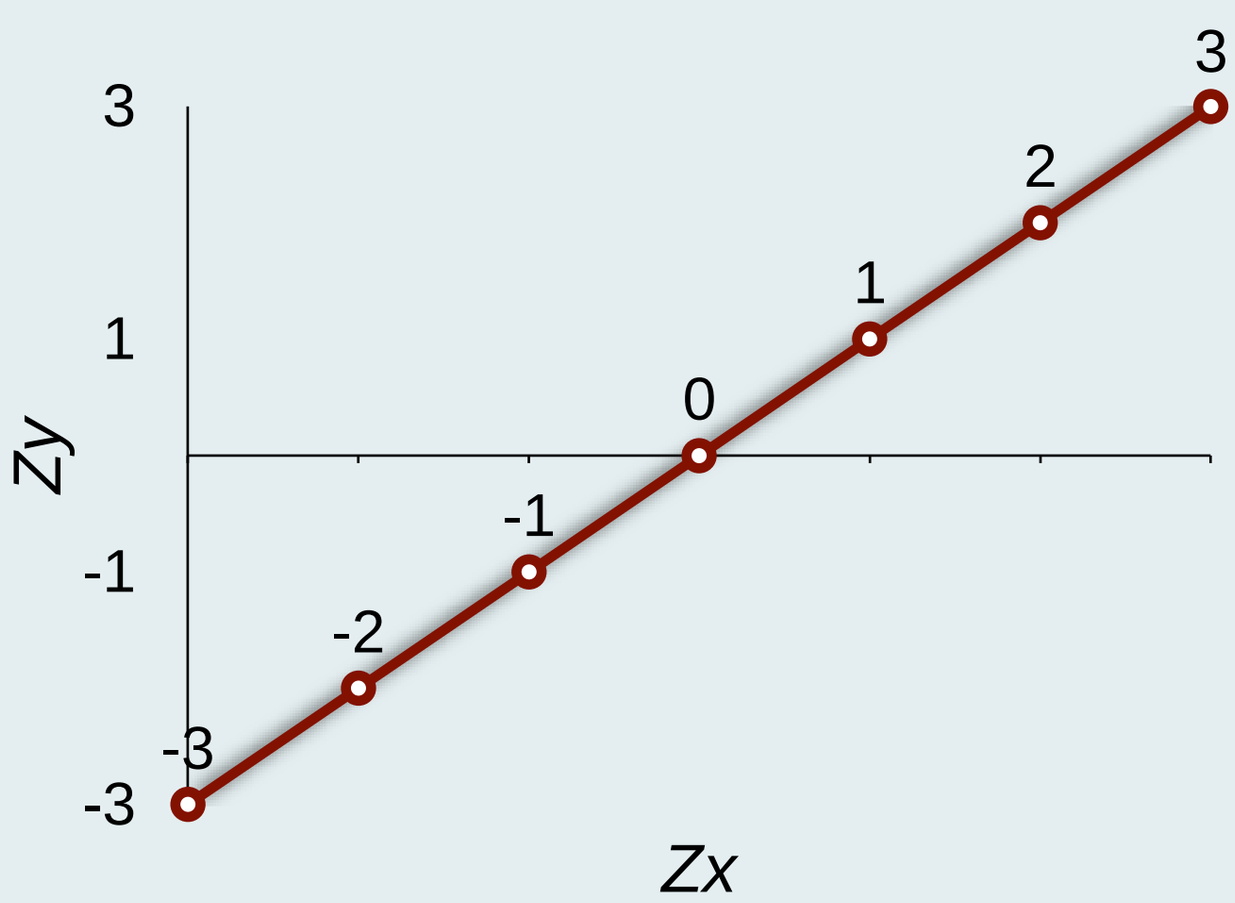
$$B = \frac{S_{XY}}{S_X^2} \tag{11.62}$$

Sostituendo $S_{XY} = r_{XY} S_X S_Y$ nella formula 11.62, il coefficiente di regressione B diventa uguale al prodotto tra il coefficiente di correlazione r_{XY} e il rapporto tra la deviazione standard della variabile dipendente, S_Y, e la deviazione standard della variabile indipendente, S_X:

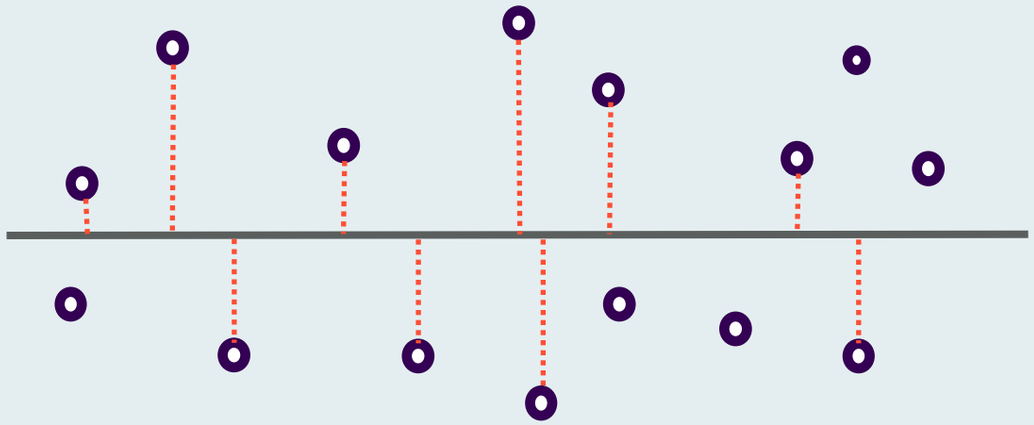
$$B = \frac{r_{XY} S_X S_Y}{S_X^2} = r_{XY} \frac{S_Y}{S_X} \tag{11.63}$$

$$z_{Y_i} = r_{XY} z_{X_i} + E_i^* \tag{11.70}$$

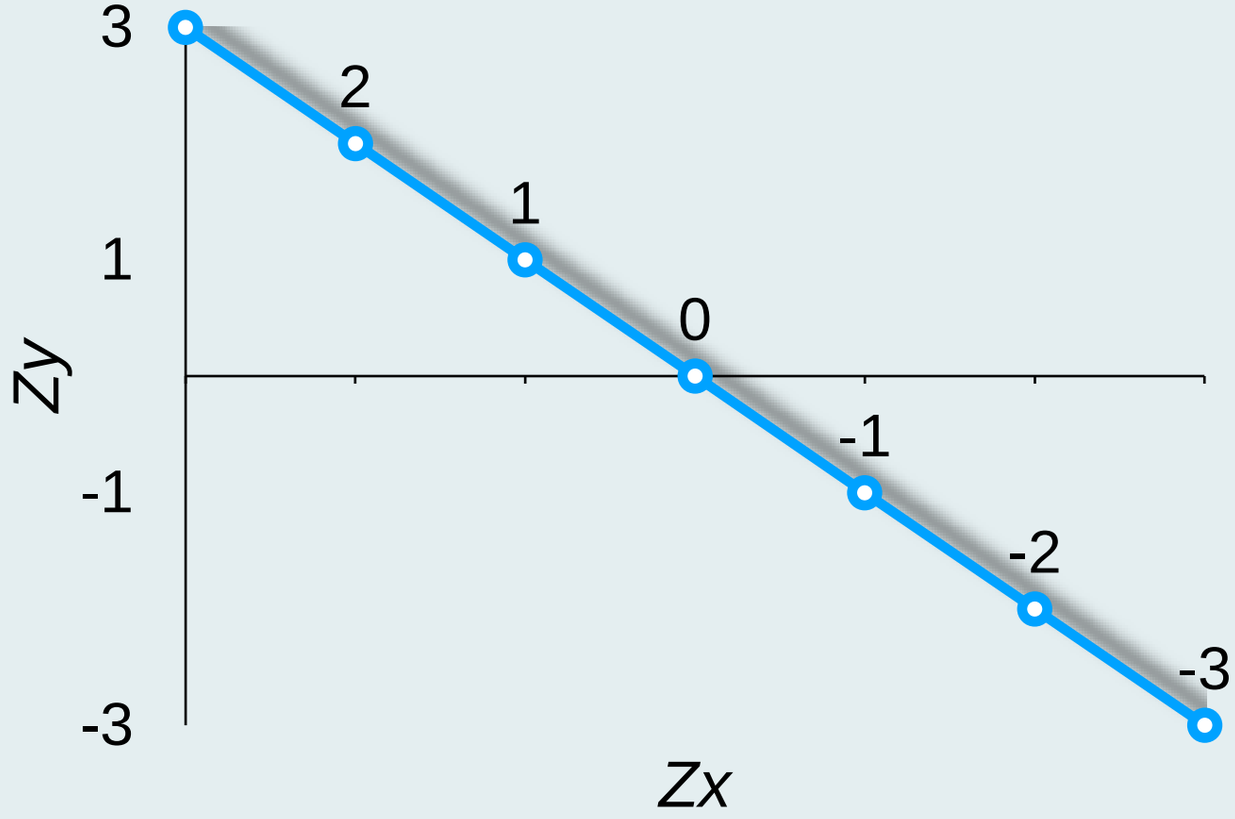
È dunque possibile concludere che, standardizzando le variabili X e Y, il coefficiente di correlazione diventa uguale al coefficiente di regressione. Quando il coefficiente di correlazione è uguale a 0, z_{Y_i} non può essere in nessun modo predetto conoscendo z_{X_i} dato che la retta di regressione è piatta. Quando il coefficiente di correlazione è uguale a 1, z_{Y_i} è completamente predetto da z_{X_i} dato che tutte le osservazioni giacciono sulla retta di regressione.



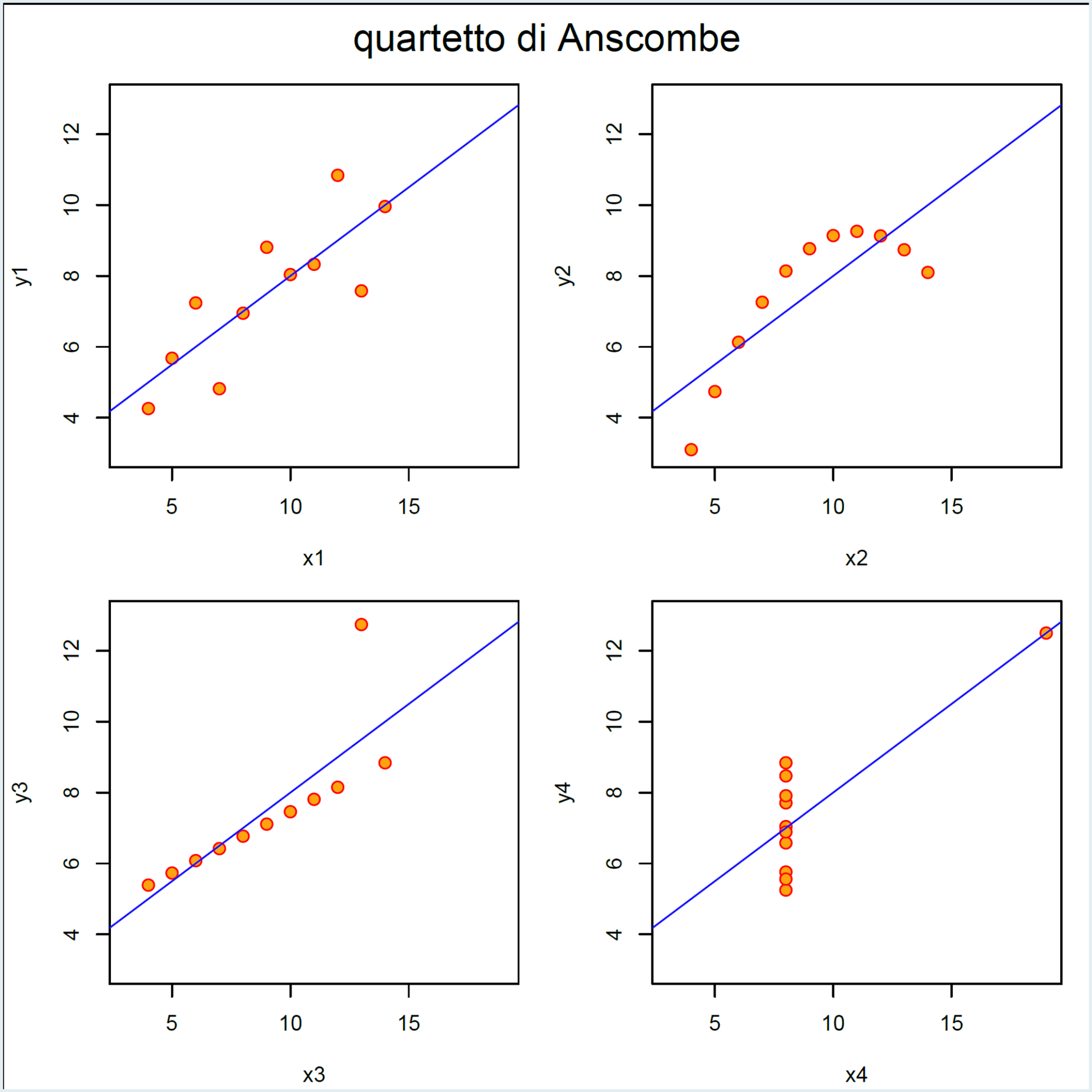
$$z_{Y_i} = 1 \times z_{X_i}$$



$$z_{Y_i} = 0 \times z_{X_i} + E_i^*$$



$$z_{Y_i} = -1 \times z_{X_i}$$


quartetto di Anscombe
y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
5
10
15