Facoltà di Psicologia
Università di Padova
Anno Accademico 2010-2011

Regressione lineare

Corso di Psicometria - Modulo B

Dott. Marco Vicentini
marco.vicentini@unipd.it

Rev. 27/12/2010

Sommario (2/3)

- Regressione lineare
  - Modello geometrico
  - Cenni sul metodo dei Minimi Quadrati
  - Esercitazione

- Approfondimenti
  - Cristante et al. (2001). *Statistica per psicologi*, capitolo V.
  - Keppel et al. (2001). *Disegno sperimentale e analisi dei dati*, capitolo 15.

Equazione di regressione

- La regressione lineare consiste in una serie di procedure attraverso le quali si individua quella particolare retta che può essere usata per predire il punteggio di un soggetto su una variabile a partire dalla conoscenza del suo punteggio nell'altra.

- La regressione lineare ci aiuta a utilizzare la relazione tra due variabili a scopo predittivo.

Keppel, p. 445

Regressione lineare

- Caratteristiche distintive:

1. **Coefficiente angolare (la pendenza):** l'angolo che la retta forma con uno degli assi cartesiani
2. **Intercetta:** il punto in cui la retta incontra l'asse delle Y quando X=0

$$Y = a + bX + \varepsilon$$

3. Errore: la misura della bontà dell'adattamento del modello ai dati.

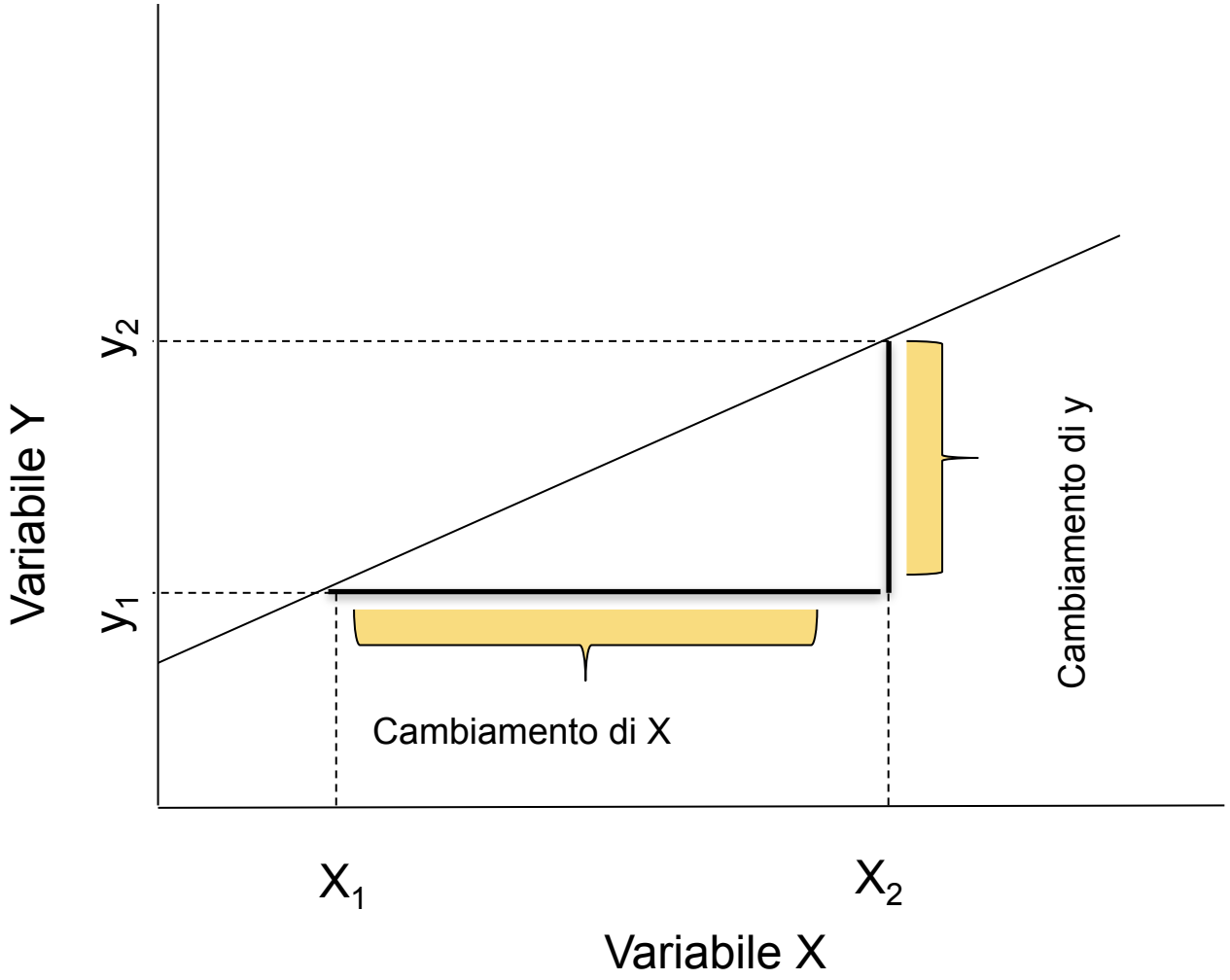
Pendenza

- Incremento di Y = $\Delta y = y_2 - y_1$
- Incremento di X = $\Delta x = x_2 - x_1$

$$pendenza = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

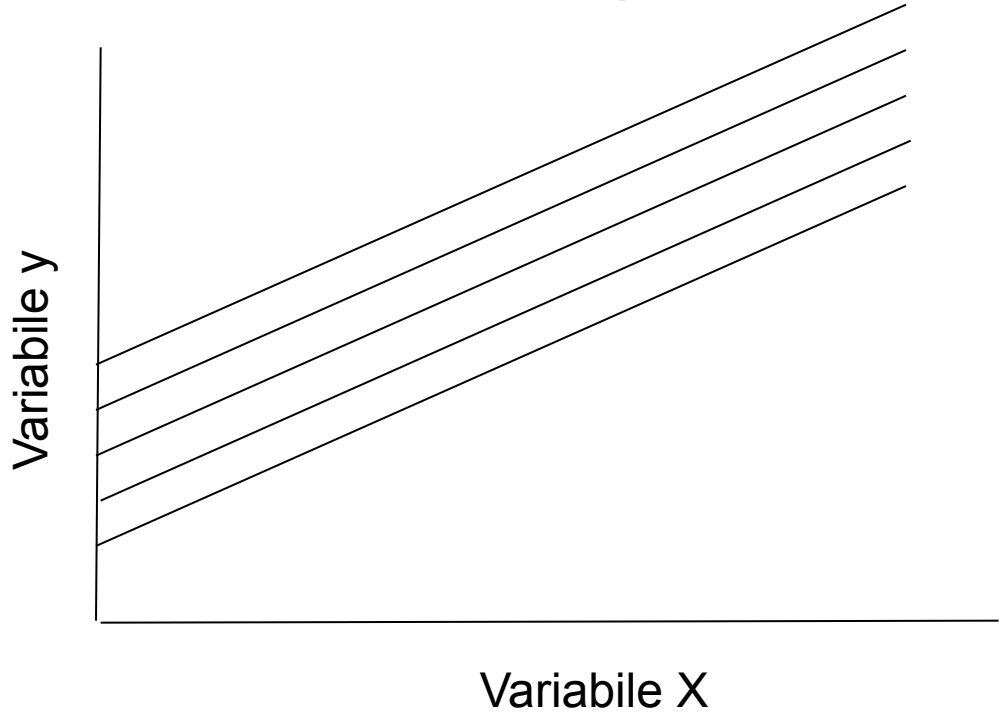
- La pendenza indica la frequenza con cui al variare di X cambia Y;
- il segno indica se i cambiamenti della variabile Y sono congruenti con quelli della variabile X.

Pendenza


Variabile Y
y_2
y_1
Cambiamento di y
Cambiamento di X
X_1
X_2
Variabile X

Intercetta

- **Intercetta**: numero infinito di rette parallele che condividono la stessa pendenza e che differiscono solo rispetto al punto di intersezione con l'asse delle ordinate quando X=0.

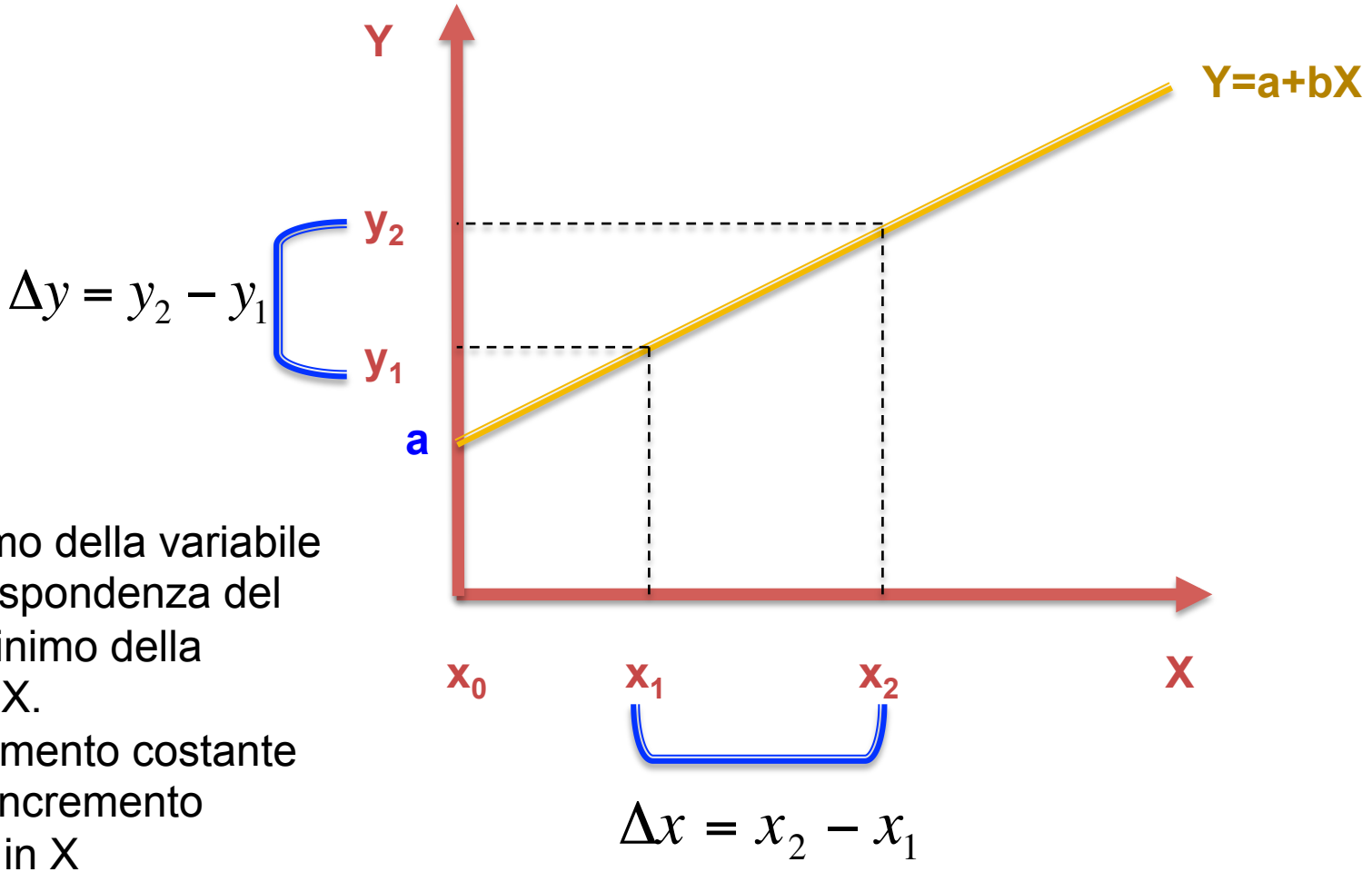


Equazione di regressione

$$Y' = a + bX$$

- dove:
  - Y′ = gli elementi stimati della variabile Y dipendenti dai valori della variabile X
  - *a* = intercetta
  - *b* = pendenza della retta o coefficiente angolare
  - X = i valori della variabile X

Equazione di regressione



- **a** = minimo della variabile Y in corrispondenza del valore minimo della variabile X.
- **b** = incremento costante in Y per incremento costante in X

Il metodo dei Minimi quadrati

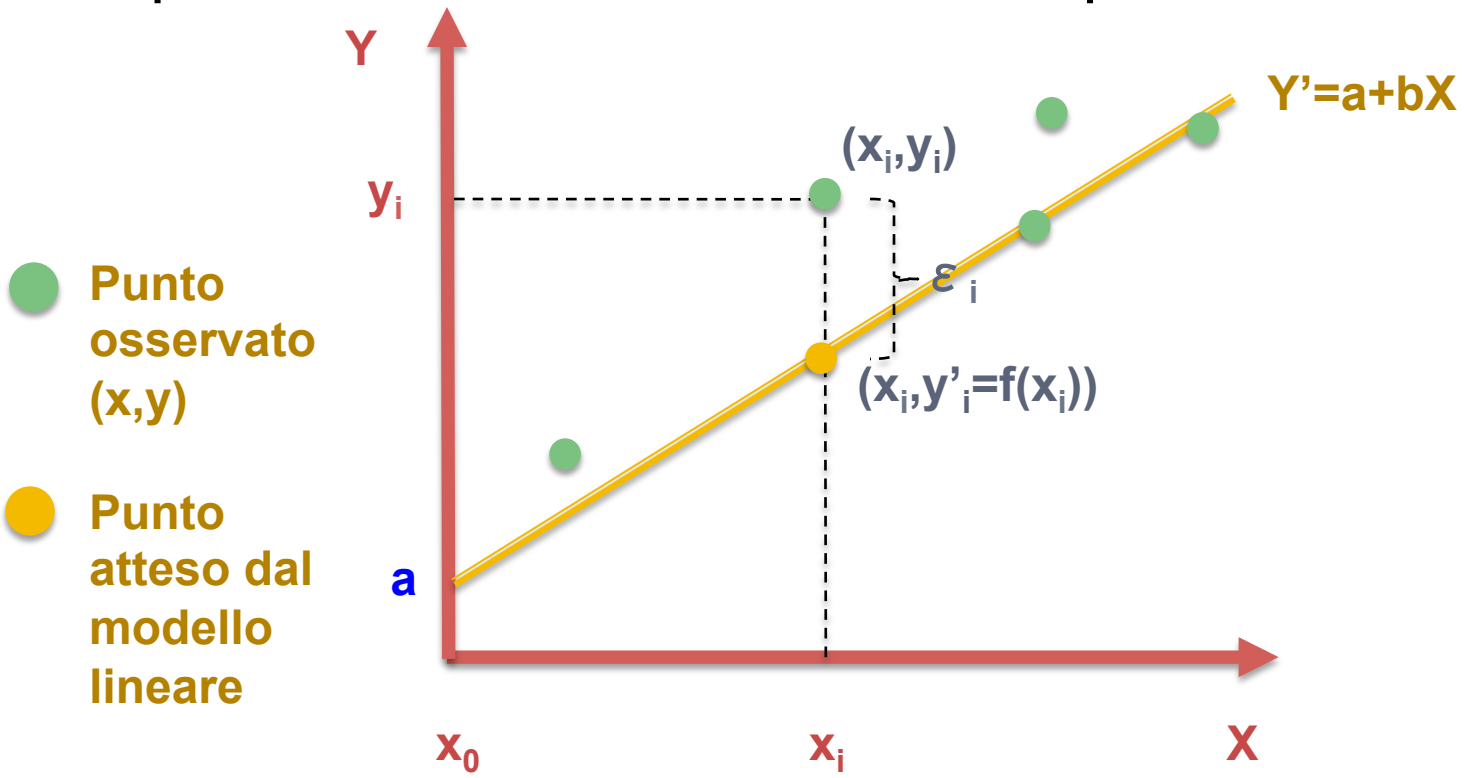
- Per arrivare a calcolare i parametri che definiscono l'equazione di regressione si utilizza generalmente il metodo dei "Minimi quadrati" (Least Squares)
  - Con questo metodo la retta individuata è quella per cui la sommatoria degli scarti quadratici fra i valori di Y calcolati sulla base della formula f(x) e quelli reali è minima:

$$LS = \min\left(\sum\left(y - f(x)\right)^2\right)$$

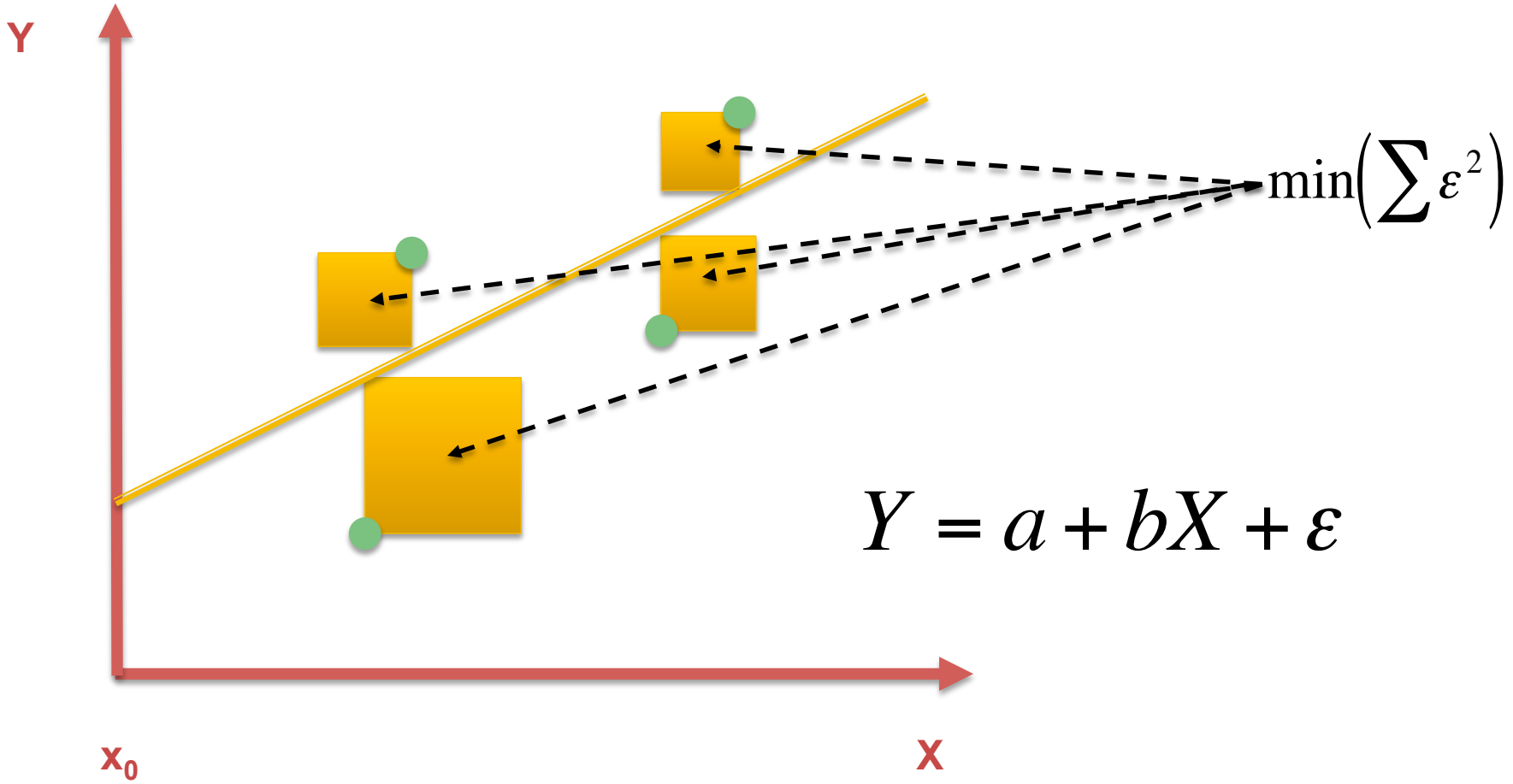
- Conoscendo il punteggio X di un soggetto e la relazione f(x) che lega la variabile X e la variabile Y è possibile predire il punteggio atteso Y' di quel soggetto.

Il metodo dei minimi quadrati

- Il metodo dei minimi quadrati stima i parametri dell'equazione di regressione minimizzando il quadrato delle distanza ε dei punti dalla retta.



Il metodo dei minimi quadrati



Stima dei parametri

- La pendenza della retta di regressione, chiamata anche coefficiente di regressione, si ottiene tramite:

$$b = \frac{\mathrm{cov}_{xy}}{\mathrm{var}_x} = \frac{\sum XY - \sum X \sum Y \Big/ n}{\sum X^2 - \left(\sum X\right)^2 \Big/ n}$$

- Il valore dell'intercetta è definito tramite:

$$a = \bar{Y} - b\bar{X} = \sum Y \Big/ n - b \sum X \Big/ n$$

Significatività della stima di b

- È possibile verificare la significatività del parametro *b* stimato, ovverosia se il suo valore reale è differente da 0, sapendo che:

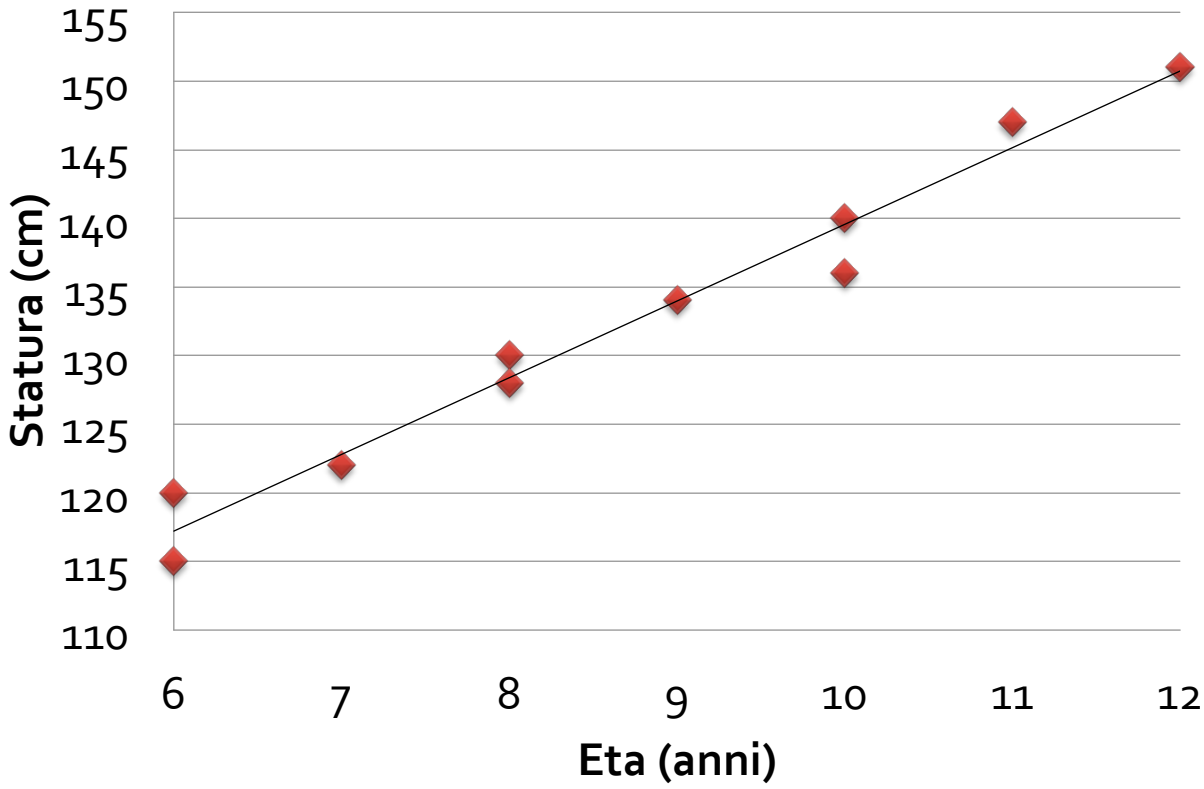
$$t_{n-2} = \frac{b}{ES_b}$$

$$ES_b = \frac{\sqrt{\sum \varepsilon^2 / (n-2)}}{\sqrt{\sum (x-\bar{x})^2}} = \frac{\sqrt{\sum (y-\hat{y})^2 / (n-2)}}{\sqrt{\sum (x-\bar{x})^2}}, \hat{y} = a + bx$$

Un esempio numerico

- Supponiamo di avere misurato la statura di 10 bambini di età compresa tra 6 e 12 anni e di riportare i dati su una tabella:

Età	Statura
6	115
6	120
7	122
8	130
8	128
9	134
10	136
10	140
11	147
12	151



Un esempio numerico

- Avevamo precedentemente calcolato il coefficiente di correlazione lineare tra X e Y:

Età (X)	Statura (Y)	XY	X^2	Y^2
6	115	690	36	13225
6	120	720	36	14400
7	122	854	49	14884
8	130	1040	64	16900
8	128	1024	64	16384
9	134	1206	81	17956
10	136	1360	100	18496
10	140	1400	100	19600
11	147	1617	121	21609
12	151	1812	144	22801

$$\sum X = 87$$

$$\sum Y = 1323$$

$$\sum X^2 = 795$$

$$\sum Y^2 = 176255$$

$$\sum XY = 11723$$

$$r_{XY} = 0.986$$

Un esempio numerico

$\sum X = 87$

$\sum Y = 1323$

$\sum X^2 = 795$

$\sum Y^2 = 176255$

$\sum XY = 11723$

- Partendo dalle sommatorie precedentemente calcolate, procediamo ora con il calcolo dei parametri della retta di regressione:

$$b = \frac{\sum XY - \sum X \sum Y / n}{\sum X^2 - \left(\sum X\right)^2 / n} = \frac{11723 - 87 \cdot 1323/10}{795 - 87^2/10} = 5.59$$

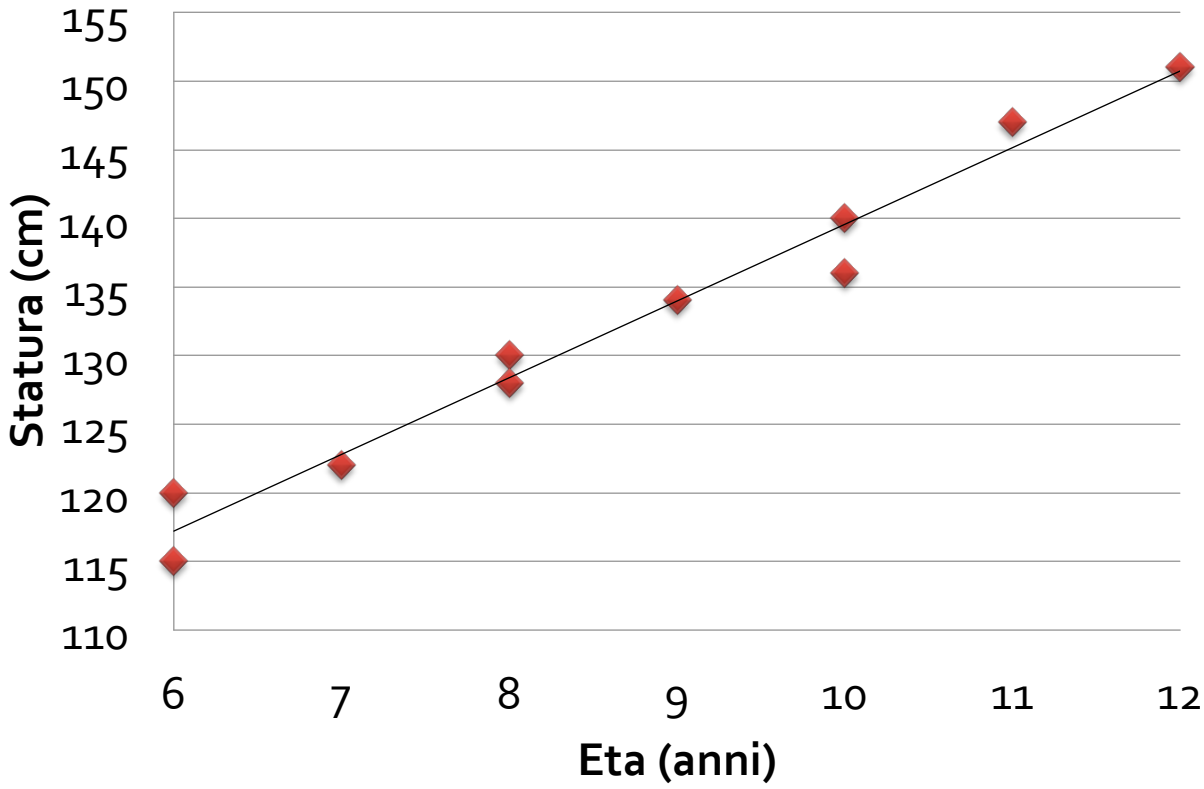
$$a = \sum Y / n - b \sum X / n = \frac{1323}{10} - b\frac{87}{10} = 83.69$$

Un esempio numerico

- Avendo calcolato i valori dei parametri a e b possiamo formulare l'equazione della retta di regressione:

$$Y' = 83.69 + 5.59X$$

- Si noti come il coefficiente angolare permetta indichi di la variazione unitaria in Y all'occorrere di una variazione unitaria in X (ovvero di quanti cm si cresce ogni anno).



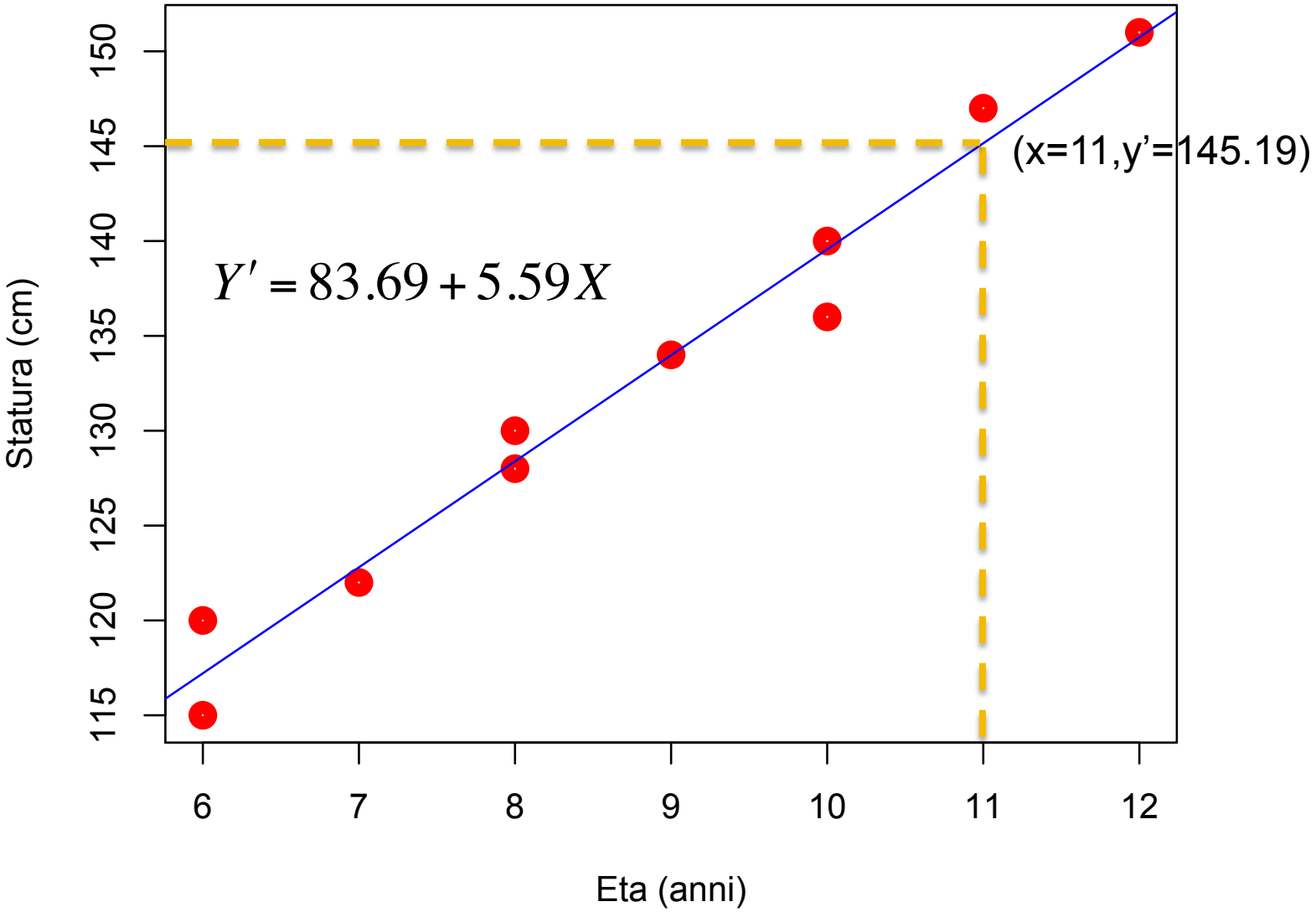
Un esempio numerico

- Conoscendo i parametri della equazione di regressione, è quindi possibile predire il valore atteso Y′ dato un qualsiasi valore appartenente al campo di esistenza di X.
  - È da notare che se la regressione è stata effettuata in un intervallo di $X=\{x_{min}, x_{max}\}$, i valori permissibili di X in f(X) devono appartenere a tale intervallo.

$$Y' = 83.69 + 5.59X$$

$$\begin{cases} x_i = 11 \\ y' = 83.69 + 5.59 \cdot 11 = 145.18 \end{cases}$$

Un esempio numerico



Un esempio numerico

- Conoscendo i parametri della regressione, è altresì possibile calcolare la significatività di b, tramite il calcolo dell'errore standard del parametro:

Età (X)	Statura (Y)	XY	X^2	Y'	Y-Y'	$(Y-Y')^2$
6	115	690	36	117.21	-2.21	4.88
6	120	720	36	117.21	2.79	7.78
7	122	854	49	122.80	-0.80	0.64
8	130	1040	64	128.39	1.61	2.59
8	128	1024	64	128.39	-0.39	0.15
9	134	1206	81	133.98	0.02	0.00
10	136	1360	100	139.57	-3.57	12.74
10	140	1400	100	139.57	0.43	0.18
11	147	1617	121	145.16	1.84	3.39
12	151	1812	144	150.75	0.25	0.06

$$\sum \varepsilon^2$$

Un esempio numerico

$$ES_b = \frac{\sqrt{\sum (y-\hat{y})^2 / (n-2)}}{\sqrt{\sum (x-\bar{x})^2}} = \frac{\sqrt{32.41/8}}{\sqrt{795 - 87^2/10}} = 0.326$$

$$t_{n-2} = \frac{b}{ES_b} = \frac{5.59}{0.326} = 17.13$$

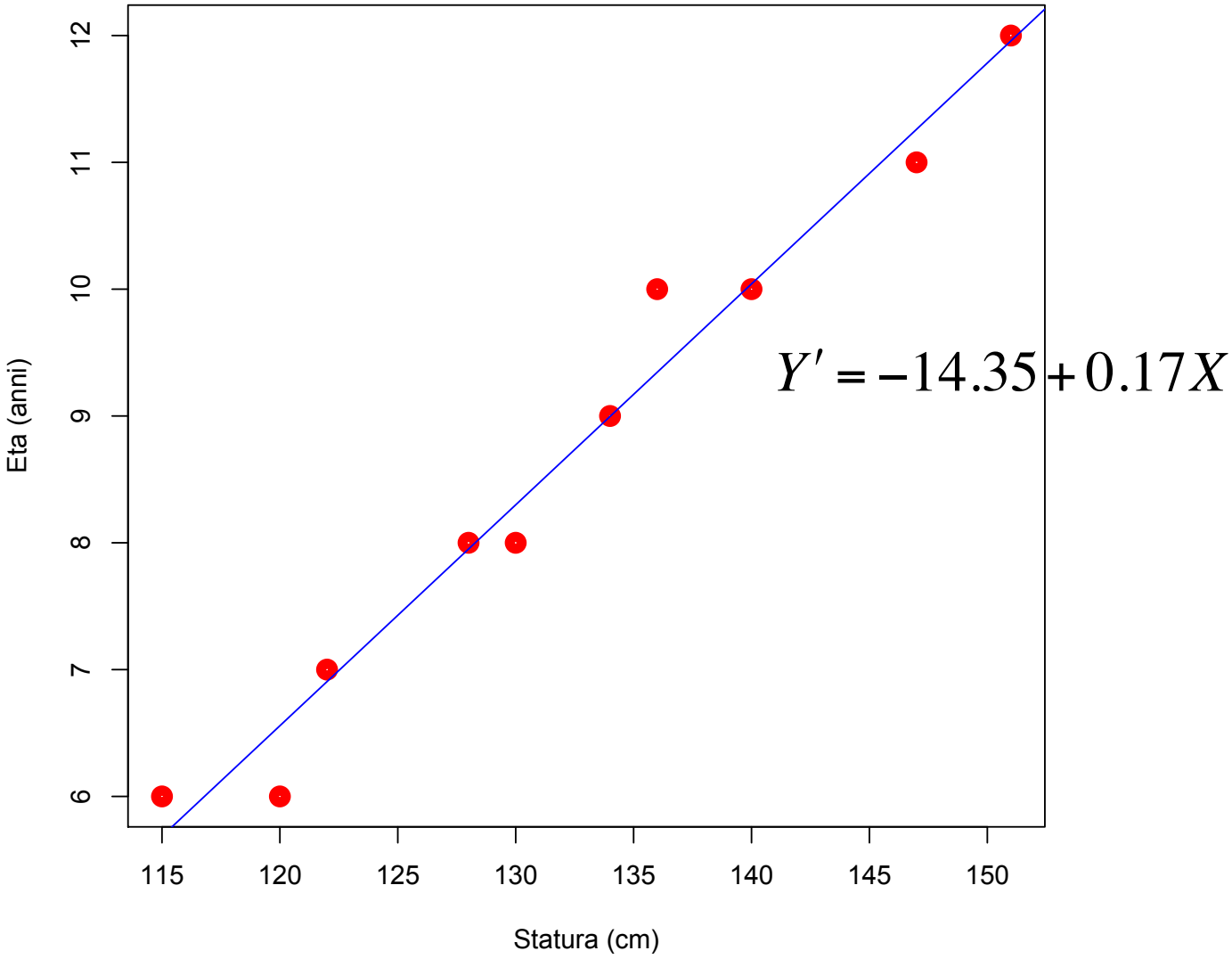
- Si osserva come al parametro b è associato il valore t=17.13, p(t)<0.001, molto inferiore al valore critico t_{crit}=2.31

Un esempio numerico

$$\sum X = 87$$
$$\sum Y = 1323$$
$$\sum X^2 = 795$$
$$\sum Y^2 = 176255$$
$$\sum XY = 11723$$

- Ovviamente se vogliamo calcolare l'equazione della retta di regressione di Y su X
  (non tenendo qui in considerazione il suo eventuale significato)
  otterremo risultati differenti:

$$b = \frac{\sum XY - \sum X \sum Y / n}{\sum X^2 - \left(\sum X\right)^2 / n} = \frac{11723 - 1323 \cdot 87/10}{176255 - 1323^2/10} = 0.17$$

$$a = \sum Y / n - b \sum X / n = \frac{87}{10} - b\frac{1323}{10} = -14.35$$

Un esempio numerico