# La predizione o regressione

Lezioni di Psicometria

Giovanni Battista Flebus

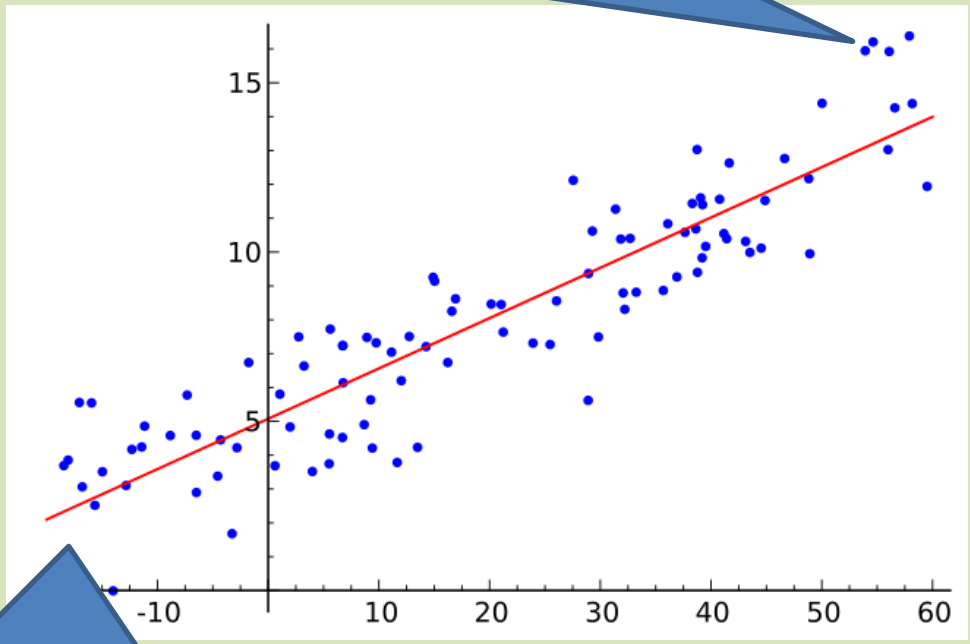
# Definizione di Predizione (1)

- Il termine di predizione in statistica e psicometria ha un significato molto limitato: si usa per indicare che una misurazione di un comportamento è usata per predire la misurazione di un altro comportamento.

- Le misurazioni sono generalmente dei test mentali (abilità, profitto, personalità, atteggiamenti, temperamenti) o dati fisici o altre rilevazioni comportamentali.

# Concetto della predizione statistica

A punteggi alti di un test (predittore) corrispondono punteggi alti di un altro test (comportamento da predire o stimare)




si può usare la prima misurazione per predire la seconda

a punteggi bassi del predittore corrispondono punteggi bassi del predetto,

# La predizione fa ricorso al concetto matematico di funzione

- Una funzione matematica è una regola che lega un insieme di numeri, usando costanti e variabili (e anche altre funzioni matematiche).

  Esempi di funzione:

- $y = k+x$
- $y = \log 10 (x)$
- $y = 3x^3 + 4x^2 + 11x + k$

Una funzione molto utile è quella che descrive la retta

$$y = x\,m + a$$

# Definizione di predizione (2)

- Dovremo trasformare il punteggio del test predittore con una equazione di una retta, che predica al meglio (ovvero commettendo meno errori possibili) il punteggio ottenuto dal soggetto nel test predetto.
- L'equazione per trasformare il punteggio è la seguente:

$$\hat{y}_i = x_i \cdot m + a$$

# Definizione di predizione (3)

Si deve tenere conto che le predizioni non sono precise, e quindi la funzione dovrebbe essere scritta sempre così

$$y = mx + a + e$$

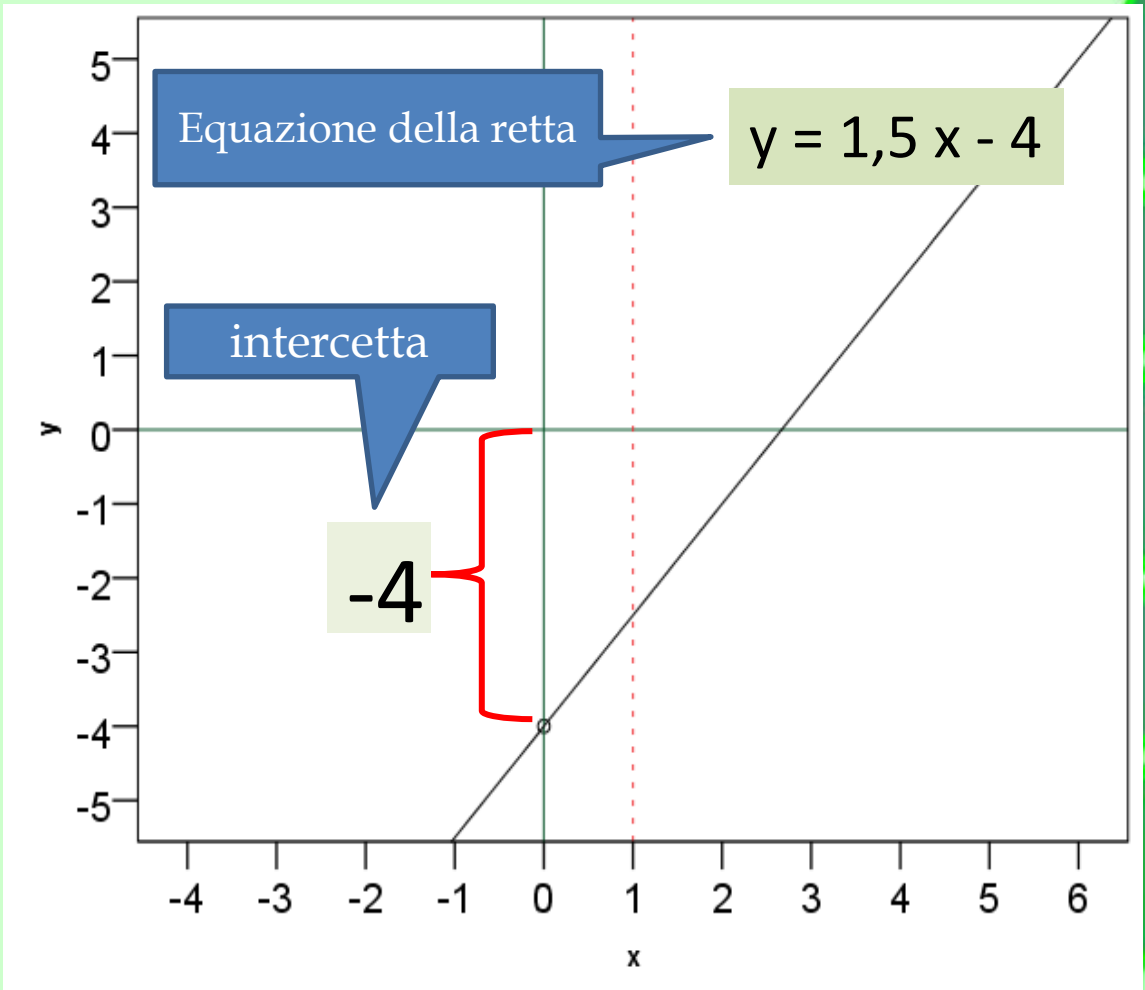
dove e indica la parte di errore della predizione.

Studieremo solo la relazione lineare

# Equazione di regressione

$$\hat{y}_i = x_i \cdot m + a$$

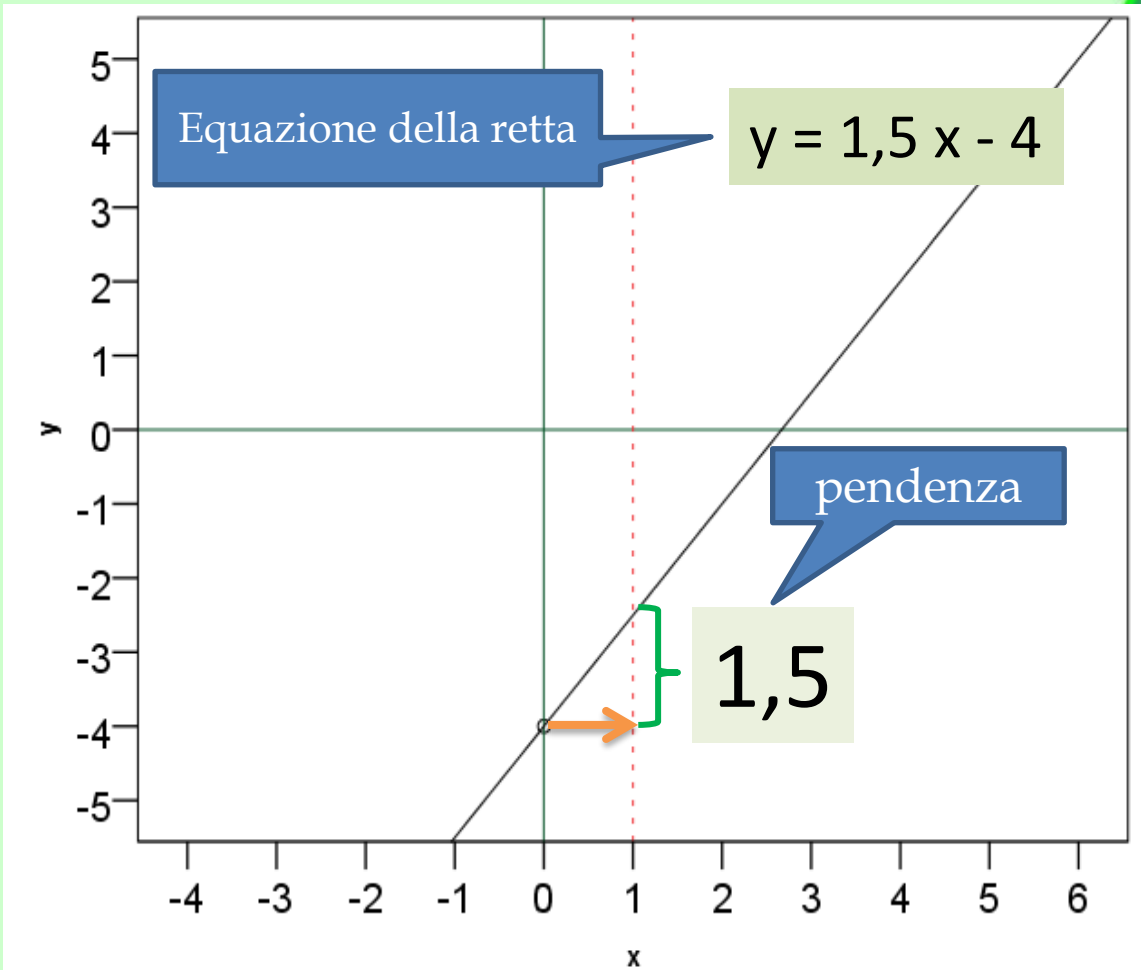
- La costante additiva *a* è chiamata intercetta. Rappresenta il punto in cui la retta incontra l'asse delle ordinate, ossia il valore che la predizione assume quando il predittore è uguale a zero



# Equazione di regressione

$$\hat{y}_i = x_i \cdot m + a$$

- La costante moltiplicativa *m* è chiamata pendenza o coefficiente angolare. Rappresenta il cambiamento in y all'aumentare di una unità in x.



# Esempi di predizione

- Un test di abilità verbale predice il profitto a scuola
- Una scala di Stima di sé è usata per predire il Senso di benessere e di salute psicofisica
- Il punteggio di Coscienziosità predice il livello di efficienza nel lavoro di gruppo.

# Piccolo esempio numerico

- Raccogliamo un piccolo numero di osservazioni:
  - Abilità verbale (un test  psicometrico)
  - Profitto scolastico (voto scolastico dato da insegnanti)
- Supponiamo che entrambe le misurazioni siano delle scale a intervalli

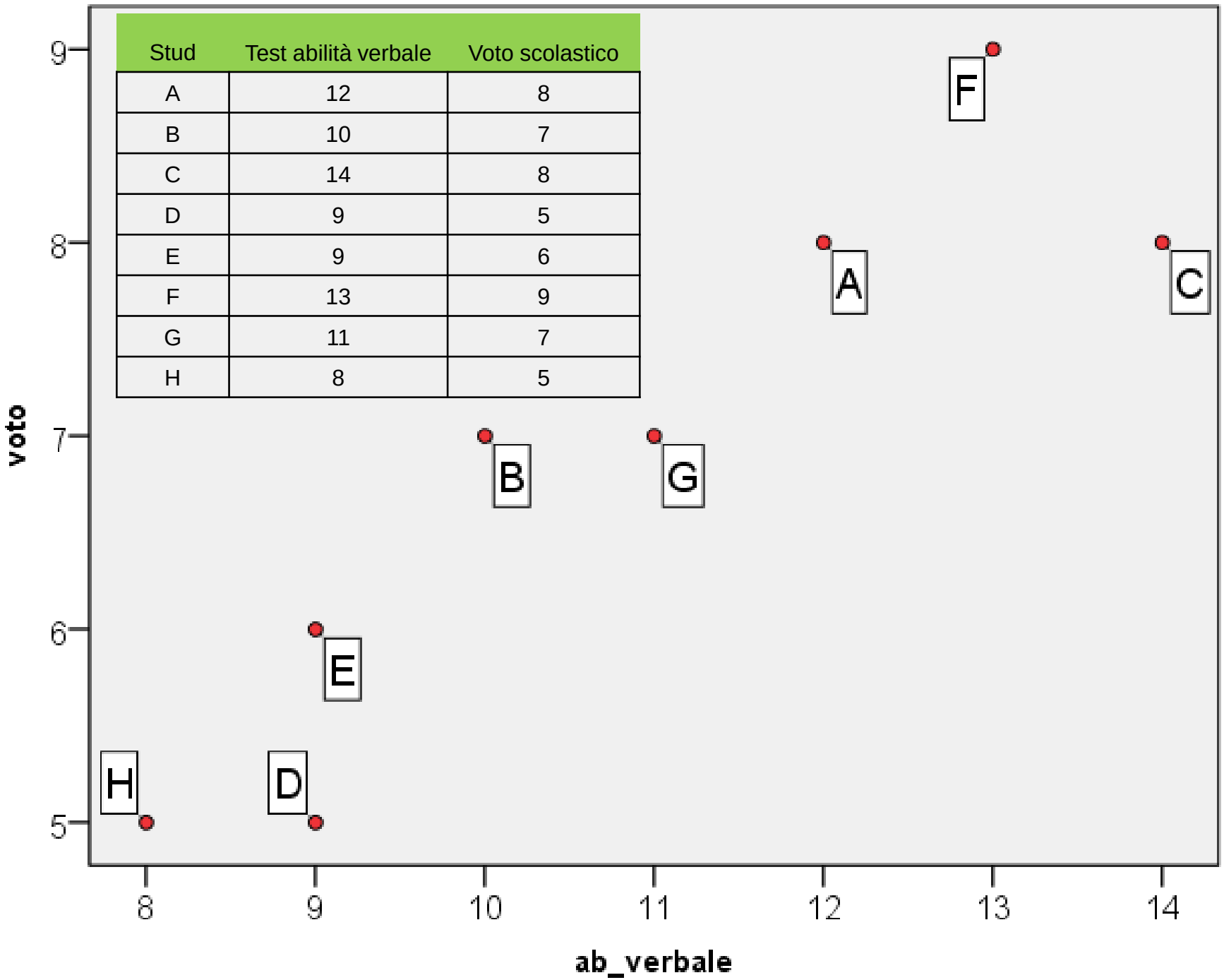
# Osservazioni per otto studenti

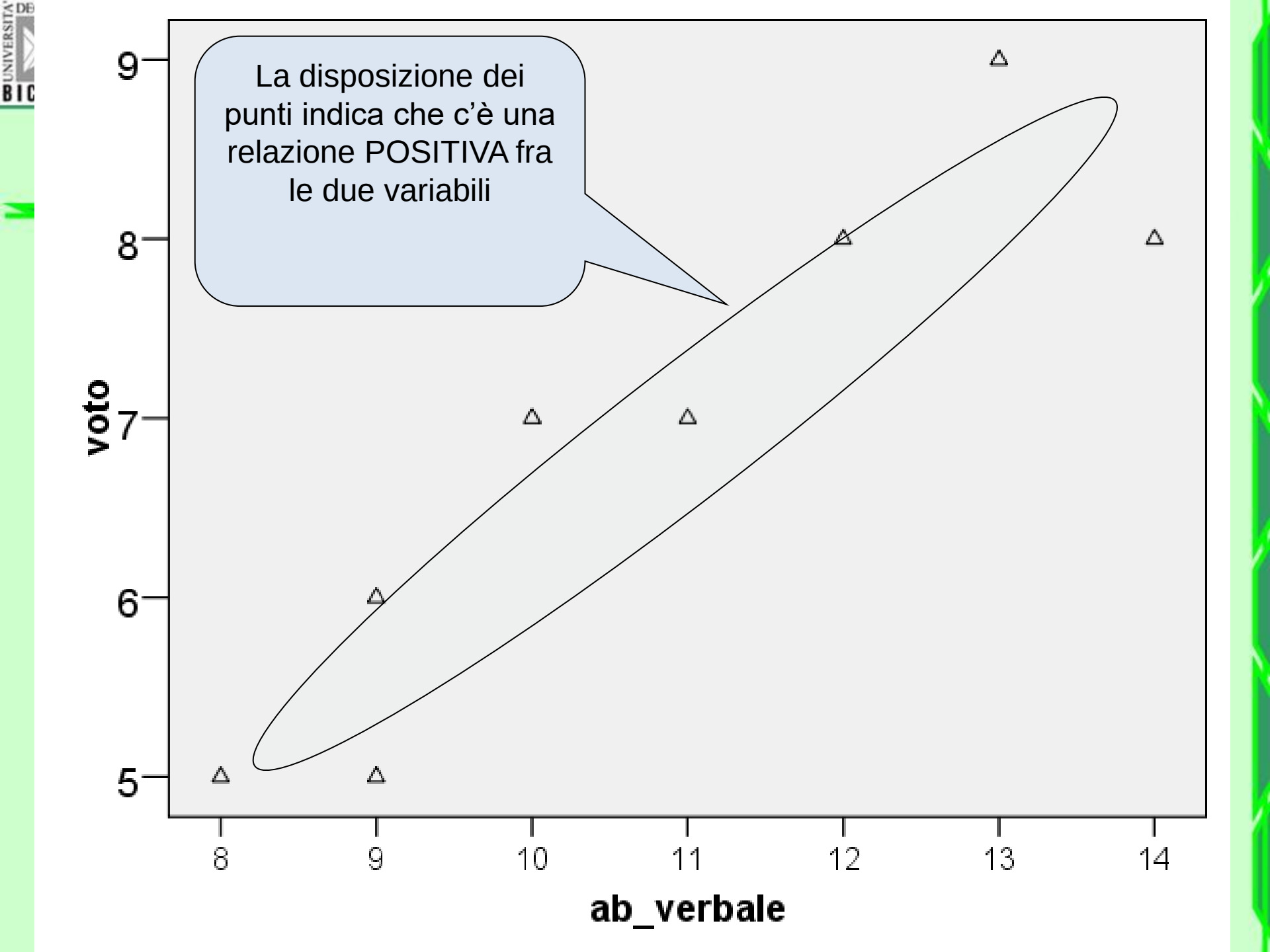
| Studente | Test abilità verbale | Voto scolastico |
|---|---|---|
| A | 12 | 8 |
| B | 10 | 7 |
| C | 14 | 8 |
| D | 9 | 5 |
| E | 9 | 6 |
| F | 13 | 9 |
| G | 11 | 7 |
| H | 8 | 5 |

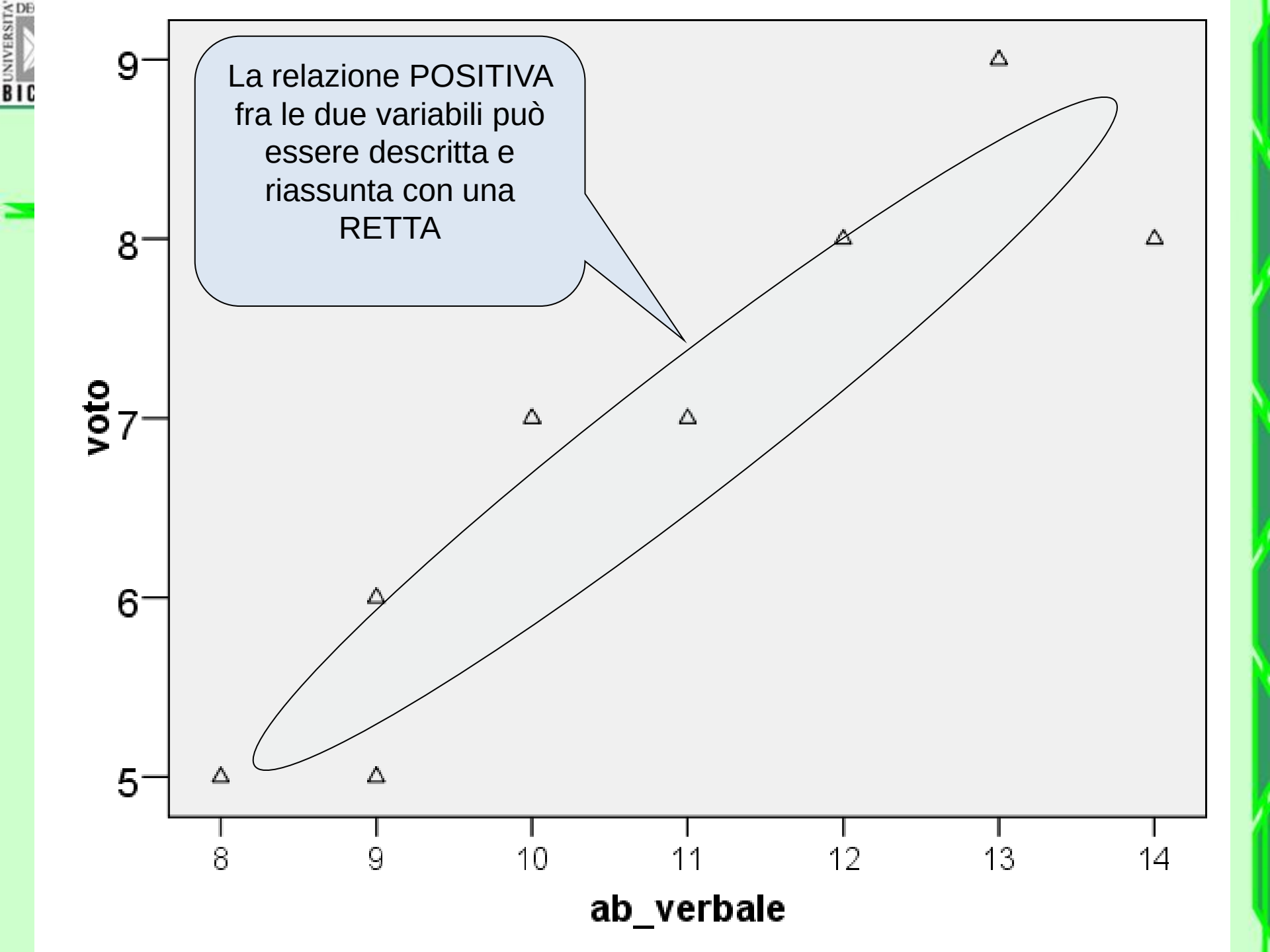
# Riportiamo in un grafico cartesiano le otto coppie di osservazioni

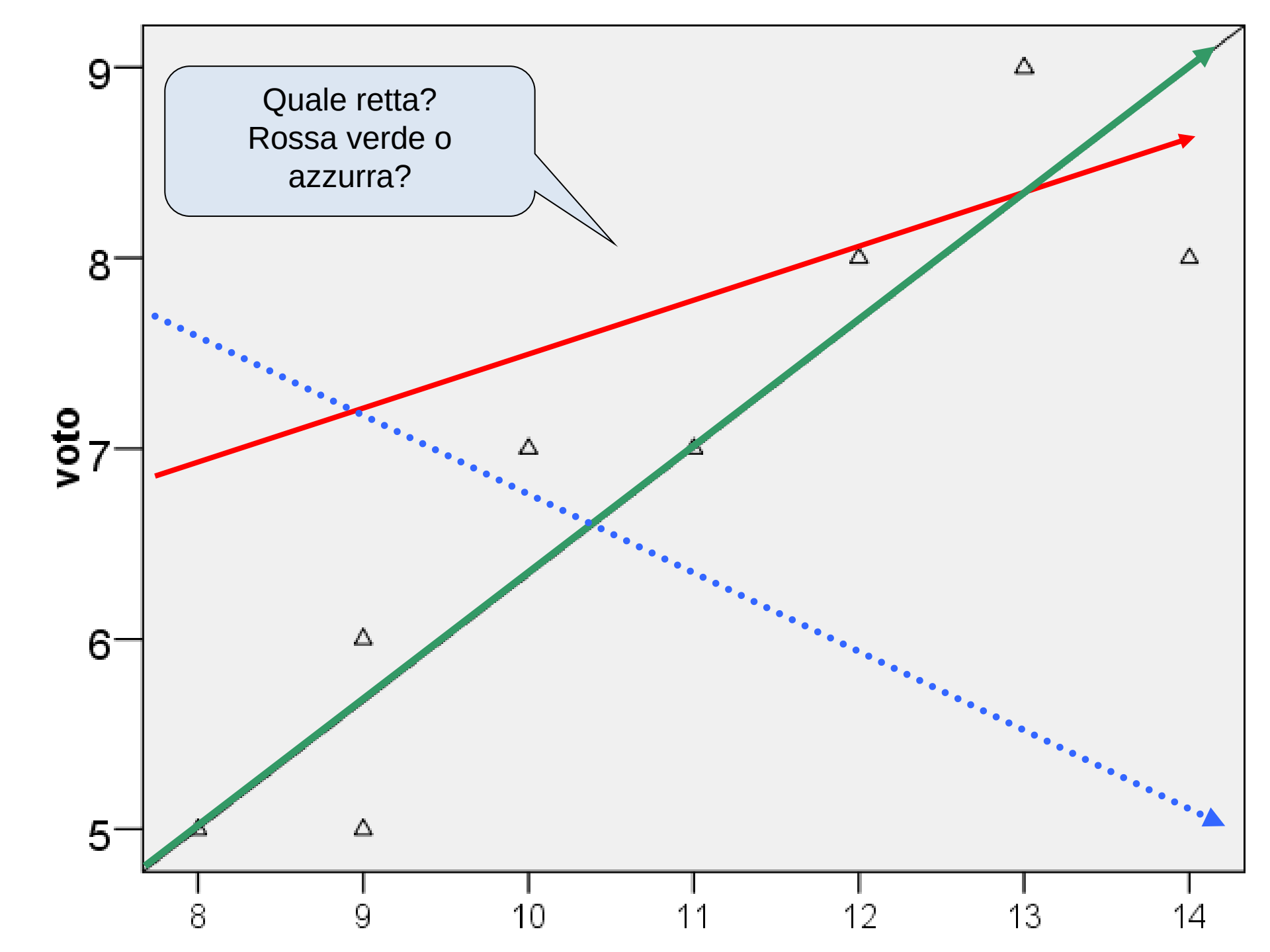
- In ascissa indichiamo la variabile indipendente (Abilità verbale)
- In ordinata riportiamo il valore della variabile dipendente (Voto scolastico)
- Osserviamo la distribuzione dei punteggi

| Stud | Test abilità verbale | Voto scolastico |
|---|---|---|
| A | 12 | 8 |
| B | 10 | 7 |
| C | 14 | 8 |
| D | 9 | 5 |
| E | 9 | 6 |
| F | 13 | 9 |
| G | 11 | 7 |
| H | 8 | 5 |




La disposizione dei punti indica che c'è una relazione POSITIVA fra le due variabili
voto
9
8
7
6
5
8
9
10
11
12
13
14
ab_verbale


La relazione POSITIVA fra le due variabili può essere descritta e riassunta con una RETTA
voto
9
8
7
6
5
8
9
10
11
12
13
14
ab_verbale

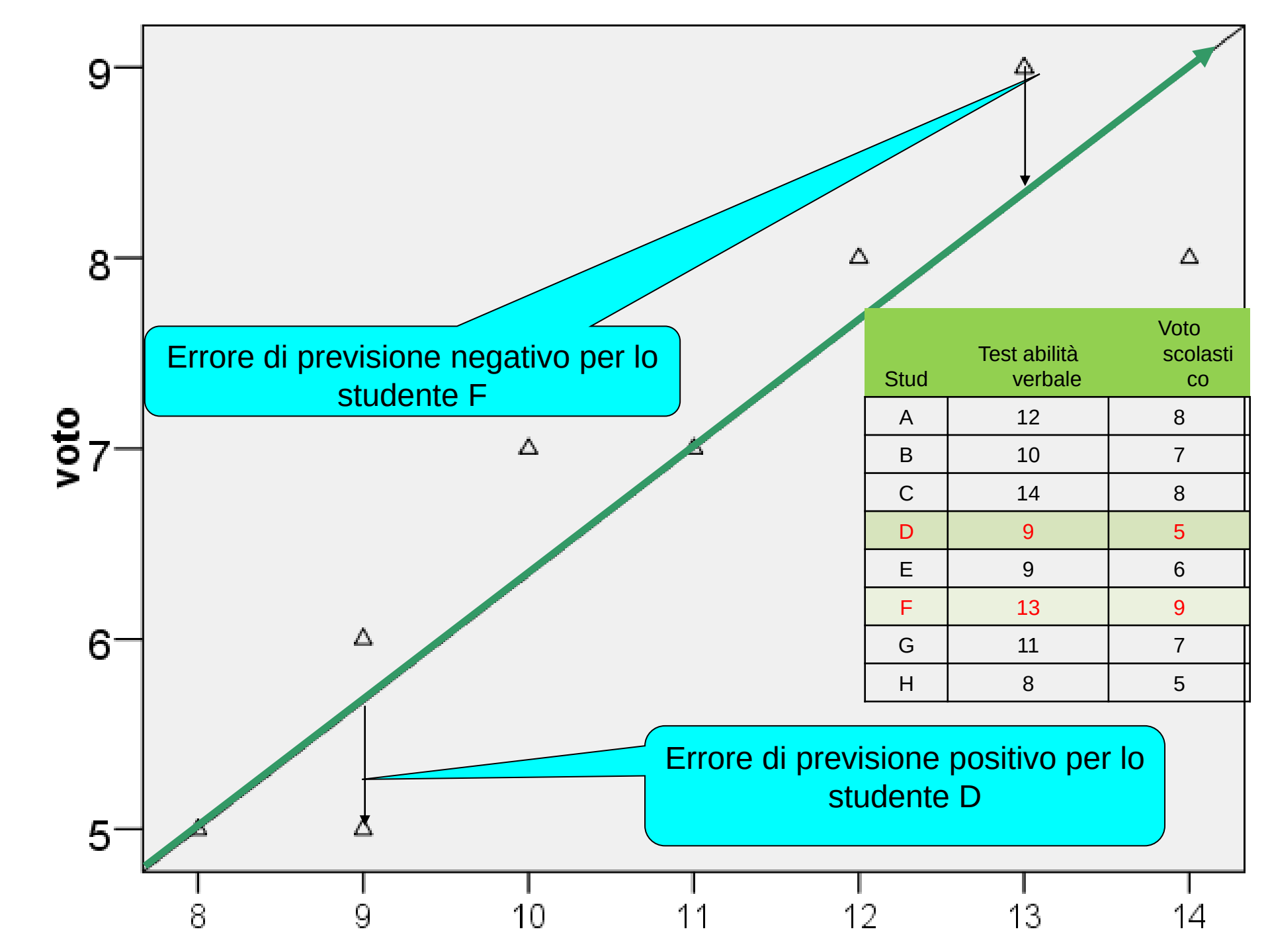

Quale retta?
Rossa verde o
azzurra?
voto
9
8
7
6
5
8
9
10
11
12
13
14


Quella che è più vicina a tutti i punti è la migliore
voto
9
8
7
6
5
8
9
10
11
12
13
14

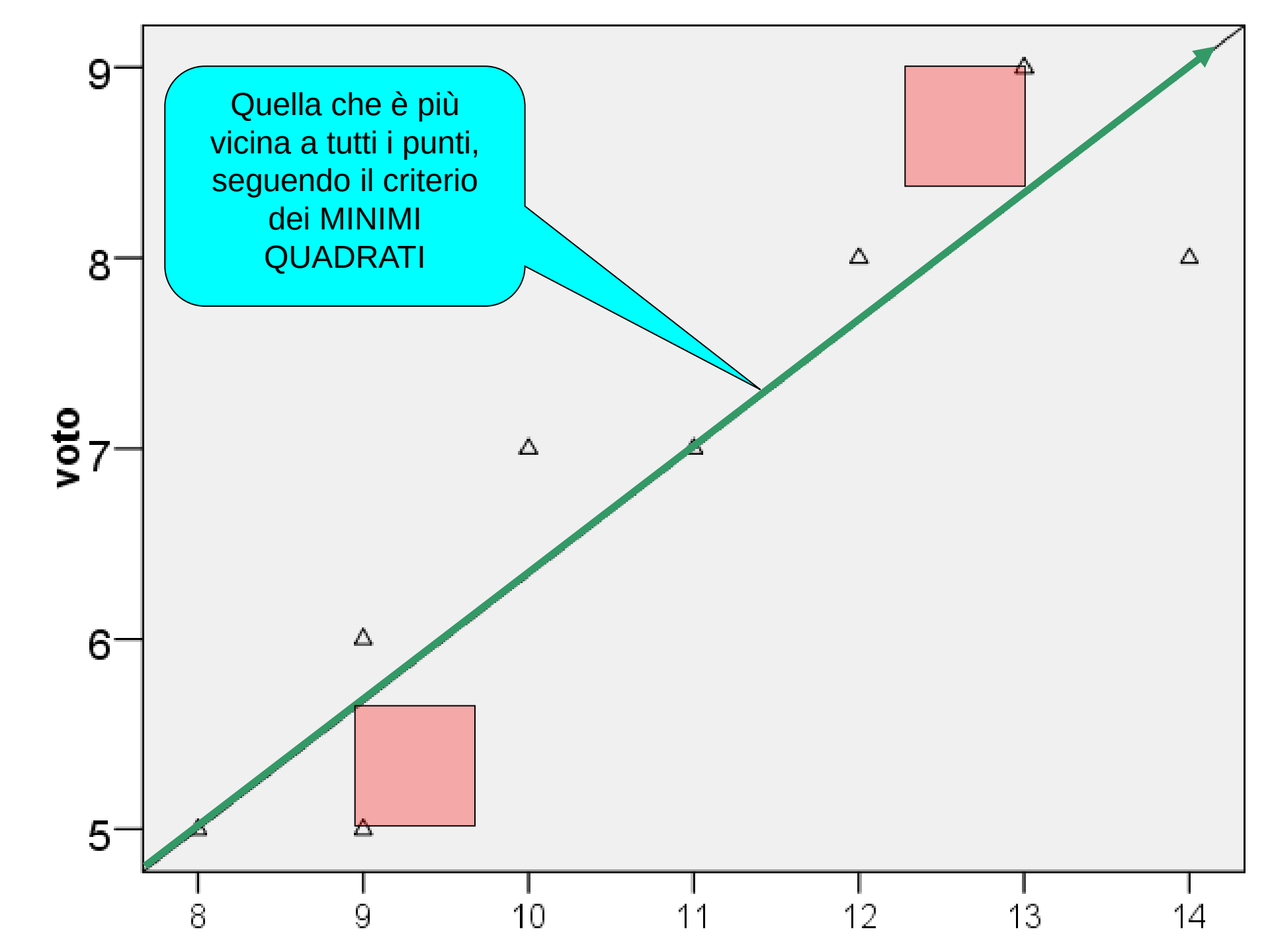
# Come stabilire i parametri della retta di predizione? Che criterio si può seguire?

- Si stabilisce  che il criterio fondamentale è quello di cercare la retta che rende minimi gli errori.
- Tuttavia i metodi derivanti dall'analisi matematica risultano più efficace se si considerano non gli errori ma i loro quadrati.
- Quindi saranno gli errori elevati al quadrato il criterio da minimizzare e l'equazione che si otterrà si chiama appunto

  equazione dei minimi quadrati.




| Stud | Test abilità verbale | Voto scolastico |
|---|---|---|
| A | 12 | 8 |
| B | 10 | 7 |
| C | 14 | 8 |
| D | 9 | 5 |
| E | 9 | 6 |
| F | 13 | 9 |
| G | 11 | 7 |
| H | 8 | 5 |


Quella che è più vicina a tutti i punti, seguendo il criterio dei MINIMI QUADRATI
voto
9
8
7
6
5
8
9
10
11
12
13
14

# Il criterio può essere espresso con la formula

$$\sum_{i=1}^{N} \varepsilon_i^2 = \sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2 = \min$$

# La formula completa

$$Y = m X + a + \varepsilon$$

Variabile dipendente, spiegata, valore osservato

Pendenza

Intercetta

Variabile indipendente o predittore

errore

Stima di y, valore predetto

$$Y = m X + a$$

# La formula più precisa

$$Y = mX + a$$

Segno della stima

$$\hat{Y}_i = mX_i + a$$

Pedice che indica per un valore i fra quelli osservati

# Formula di calcolo

$$m = r\frac{s_y}{s_x} =$$

$$\frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

$$a = \bar{Y} - m\bar{X}$$

# Applicazione della formule

$$m = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - (\sum X)^2}$$

| somma X | somma Y | quad X | quad Y | prod XY |
|---|---|---|---|---|
| 150 | 95 | 1668 | 673 | 1056 |

$$m = \frac{8*611 - 86*55}{8*956 - (86)^2} =$$

Coefficiente angolare o moltiplicativo

$$\frac{4888 - 4730}{7648 - 7396} = \frac{158}{252} = 0{,}626$$

# Applicazione della formule

| somma X | somma Y | quad X | quad Y | prod XY |
|---|---|---|---|---|
| 150 | 95 | 1668 | 673 | 1056 |

$$a = \bar{Y} - m\bar{X}$$

Coefficiente additivo o intercetta

$$a = 6{,}875 - 0{,}626 * 10{,}75 = 0{,}135$$

$$\hat{Y} = 0{,}135 + 0{,}626 * X$$

Equazione di regressione

$$\hat{Y} = 0{,}135 + 0{,}626 * X$$

Equazione di regressione

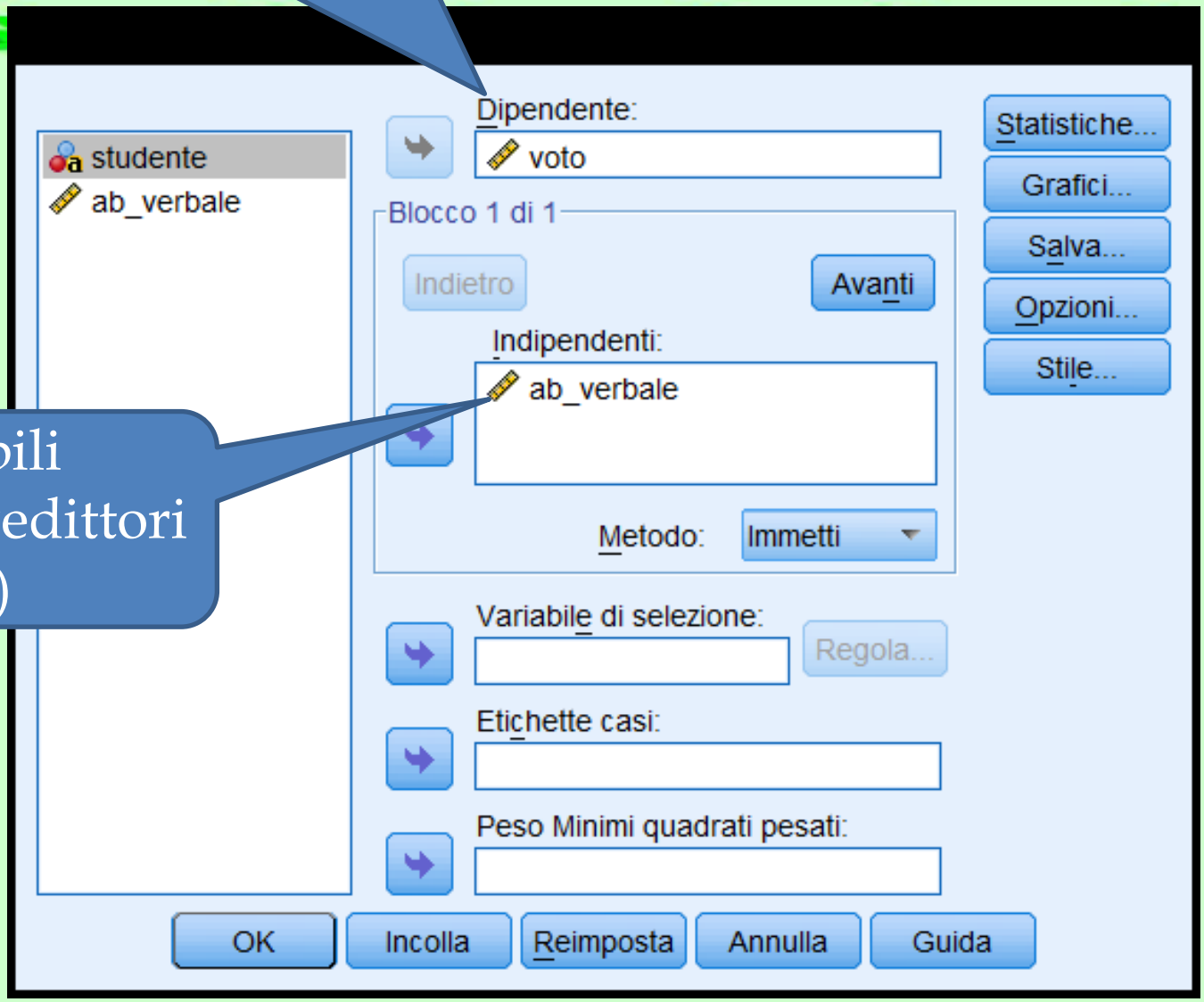
| | abilità | voto | Voto predetto |
|---|---|---|---|
| | 8 | 5 | 5,15 |
| | 9 | 5 | 5,78 |
| | 9 | 6 | 5,78 |
| | 10 | 7 | 6,4 |
| | 11 | 7 | 7,03 |
| | 12 | 8 | 7,66 |
| | 13 | 9 | 8,29 |
| | 14 | 8 | 8,91 |
| somma | 86 | 55 | 55 |
| media | 10,75 | 6,875 | 6,875 |

# Regressione con SPSS...

- Menu Analizza
- Poi Regressione
- Poi Lineare
- Compare la finestra di scelta
- Per il momento non facciamo niente altro e diamo l'OK
- Fra i  risultati, solo alcuni sono immediatamente comprensibili

Qui la variabile dipendente (una sola)

Qui le variabili indipendenti o predittori (una o più)



**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Errore std. | Beta | | |
| 1 | (Costante) | ,135 | 1,255 | | ,107 | ,918 |
| | ab_verbale | ,627 | ,115 | ,912 | 5,460 | ,002 |

a. Variabile dipendente: Voto_scolastico

Ecco la costante moltiplicativa: è il valore che moltiplica il punteggio dell'abilità verbale

$$\hat{Y} = 0{,}135 + 0{,}626 * X$$

**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Errore std. | Beta | | |
| 1 | (Costante) | ,135 | 1,255 | | ,107 | ,918 |
| | ab_verbale | ,627 | ,115 | ,912 | 5,460 | ,002 |

a. Variabile dipendente: Voto_scolastico

Costante additiva. E' il valore della VD quando la VI è uguale a zero. In psicologia ha un senso relativo, dovuto all'arbitrarietà delle unità di misura (per i test mentali)

$$\hat{Y} = 0{,}135 + 0{,}626 * X$$

# Ricordiamo la correlazione fra le due misurazioni

**Correlazioni**

| | | ab_verbale | Voto_scolastico |
|---|---|---|---|
| ab_verbale | Correlazione di Pearson | 1 | ,912** |
| | Sig. (2-code) | | ,002 |
| | N | 8 | 8 |
| Voto_scolastico | Correlazione di Pearson | ,912** | 1 |
| | Sig. (2-code) | ,002 | |
| | N | 8 | 8 |

**. La correlazione è significativa al livello 0,01 (2-code).

**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Errore std. | Beta | | |
| 1 | (Costante) | ,135 | 1,255 | | ,107 | ,918 |
| | ab_verbale | ,627 | ,115 | ,912 | 5,460 | ,002 |

a. Variabile dipendente: Voto_scolastico

Coefficiente beta standardizzato: con una sola VI, è uguale a r.
Indica l'ammontare di cambiamento della VD per ogni unità della VI, se entrambe le variabili sono standardizzate.

**Riepilogo del modello**

| Modello | R | R-quadrato | R-quadrato corretto | Errore std. della stima |
|---|---|---|---|---|
| 1 | ,912[a] | ,832 | ,805 | ,644 |

a. Stimatori: (Costante), ab_verbale

R multiplo: indica la precisione della predizione.
Importante nella regressione multipla.
In quella semplice, R = r.
È un valore sempre positivo, anche quando r è negativo.

**Riepilogo del modello**

| Modello | R | R-quadrato | R-quadrato corretto | Errore std. della stima |
|---|---|---|---|---|
| 1 | ,912[a] | ,832 | ,805 | ,644 |

a. Stimatori: (Costante), ab_verbale

Quadrato di R multiplo. Se moltiplicato per 100, dà la percentuale di varianza spiegata dalla VI

**Riepilogo del modello**

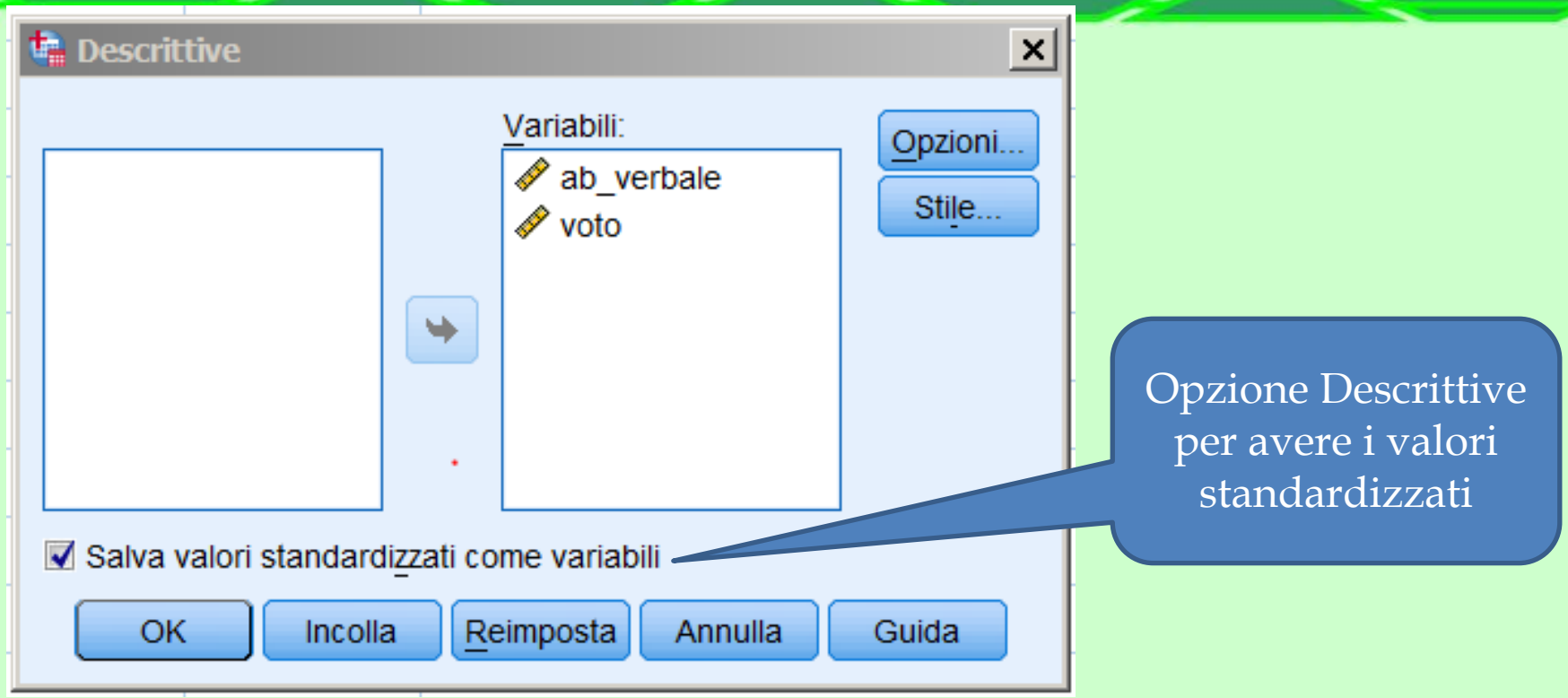
| Modello | R | R-quadrato | R-quadrato corretto | Errore std. della stima |
|---|---|---|---|---|
| 1 | ,912[a] | ,832 | ,805 | ,644 |

a. Stimatori: (Costante), ab_verbale

R quadrato corretto: dà una stima del possibile coefficiente ripetuto su un nuovo campione. Si abbassa molto se è calcolato su pochi casi e su molte variabili.

# Predizione usando i punti standardizzati

# Trasformiamo le due serie di dati in valori standardizzati




- Poi applichiamo la regressione e guardiamo i risultati

# Regressione con i valori standardizzati

**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | | |
|---|---|---|---|---|---|---|
| | | T | Errore std | Beta | t | Sign. |
| 1 | (Costante) | ,000 | ,156 | | ,000 | 1,000 |
| | Z_ab_verbale | ,912 | ,167 | ,912 | 5,460 | ,002 |

a. Variabile dipende Z_voto

La costante moltiplicativa è uguale a rxy

La costante additiva è uguale a zero

| Coefficienti standardizzati | | |
|---|---|---|
| Beta | t | Sig. |
| | ,107 | ,918 |
| ,912 | 5,460 | ,002 |

# L'equazione si semplifica…

$$\hat{z}_{yi} = z_{xi} \cdot r_{xy}$$

$\hat{z}_{yi}$ = zeta predetto
$z_{xi}$ = zeta predittore
$r_{xy}$ = coefficiente di correlazione

# Notiamo alcuni elementi…

**Statistiche descrittive**

| | Media | Deviazione std. | Varianza |
|---|---|---|---|
| ab_verbale | 10,7500 | 2,1213 | 4,5000 |
| voto | 6,8750 | 1,4577 | 2,1250 |
| Predetti Grezzi | 6,8750 | 1,3300 | 1,7690 |
| Predetti standardizzati | ,0000 | ,9124 | ,8325 |

Le medie dei valori predetti e dei valori osservati sono uguali

# Notiamo alcuni elementi…

**Statistiche descrittive**

| | Media | Deviazione std. | Varianza |
|---|---|---|---|
| ab_verbale | 10,7500 | 2,1213 | 4,5000 |
| voto | 6,8750 | 1,4577 | 2,1250 |
| Predetti Grezzi | 6,8750 | 1,3300 | 1,7690 |
| Predetti standardizzati | ,0000 | ,9124 | ,8325 |

Le medie dei predetti standardizzati è uguale a zero

# Notiamo alcuni elementi…

**Statistiche descrittive**

| | Media | Deviazione std. | Varianza |
|---|---|---|---|
| ab_verbale | 10,7500 | 2,1213 | 4,5000 |
| voto | 6,8750 | 1,4577 | 2,1250 |
| Predetti Grezzi | 6,8750 | 1,3300 | 1,7690 |
| Predetti standardizzati | ,0000 | ,9124 | ,8325 |

La deviazione standard dei valori predetti è uguale al coefficiente di correlazione

# Notiamo alcuni elementi…

**Statistiche descrittive**

| | Media | Deviazione std. | Varianza |
|---|---|---|---|
| ab_verbale | 10,7500 | 2,1213 | 4,5000 |
| voto | 6,8750 | 1,4577 | 2,1250 |
| Predetti Grezzi | 6,8750 | 1,3300 | 1,7690 |
| Predetti standardizzati | ,0000 | ,9124 | ,8325 |

La varianza dei valori predetti prende il nome di varianza spiegata dalla regressione

# Notiamo alcuni elementi…

**Statistiche descrittive**

| | Media | Deviazione std. | Varianza |
|---|---|---|---|
| ab_verbale | 10,7500 | 2,1213 | 4,5000 |
| voto | 6,8750 | 1,4577 | 2,1250 |
| Predetti Grezzi | 6,8750 | 1,3300 | 1,7690 |
| Predetti standardizzati | ,0000 | ,9124 | ,8325 |

Il rapporto fra varianza spiegata e varianza totale è pari a $\frac{1,7690}{2,1250} = 0,8325$ , ossia varianza dei predetti standardizzati, uguale al coefficiente di determinazione

# La varianza dei punteggi predetti dipende dal coefficiente di correlazione

Correlazione elevata

- Buona predizione
- Stime vicine ai valori osservati

Correlazione bassa

- Cattiva predizione
- Stime dei valori osservati attorno alla media della variabile dipendente

# Casi estremi: r= 1

$$Y = mX + a$$

Correlazione perfetta

- Nessun errore

$$Y = mX + a + 0err \Rightarrow Y = mX + a$$

# Casi estremi: r= 0

Correlazione nulla

- Predizione assente
- Stime dei valori osservati sempre uguali alla media

$$Y = mX + a$$

$$Y = 0 + a$$

**Statistiche descrittive**

| | Media | Deviazione std. | Varianza |
|---|---|---|---|
| ab_verbale | 10,7500 | 2,1213 | 4,5000 |
| voto | 6,8750 | 1,4577 | 2,1250 |
| Predetti Grezzi | 6,8750 | 1,3300 | 1,7690 |

# Perché stimare dei valori che abbiamo già in realtà?

- Per testare le capacità del test di predizione, per poterlo poi usare in situazioni reali, dove non si conosce il punteggio da predire.

# Parametri

- Le rilevazioni eseguite su un campione forniscono dei riassunti (variabili casuali) che stimano i parametri della popolazione.
- I parametri della popolazione possono essere uguali a zero (e non influenzano la regressione) o diversi da zero (e allora la influenzano).
- La distribuzione campionaria dei due coefficienti (angolare e additivo) seguono la curva normale in caso di ipotesi nulla e quindi si possono applicare le procedure che abbiamo già conosciuto per il coefficiente di correlazione
- La significatività è il valore di probabilità di trovare un coefficiente nullo, sotto l'ipotesi nulla.

# Riassumendo

- La regressione statistica permette di stimare (o predire) il punteggio di un test (o di un'altra misurazione).
- Nella predizione del singolo caso non è mai possibile sapere se la predizione è esatta o molto sballata.
- Si può quantificare la predizione totale, fatta su tutti i casi (presenti e futuri): la quota di varianza spiegata ($r^2$) è un utile indice per definire la precisione della predizione.

**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Errore std. | Beta | | |
| 1 | (Costante) | ,135 | 1,255 | | ,107 | ,918 |
| | ab_verbale | ,627 | ,115 | ,912 | 5,460 | ,002 |

a. Variabile dipendente: Voto_scolastico

I due parametri della regressione, come li abbiamo conosciuti

**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Errore std. | Beta | | |
| 1 | (Costante) | ,135 | 1,255 | | ,107 | ,918 |
| | ab_verbale | ,627 | ,115 | ,912 | 5,460 | ,002 |

a. Variabile dipendente: Voto_scolastico

Due parametri richiedono una doppia ipotesi nulla:

H0: il parametro additivo è uguale a zero e non aiuta a migliorare la predizione

H1: il parametro è diverso da zero e serve a migliorare la predizione

**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Errore std. | Beta | | |
| 1 | (Costante) | ,135 | 1,255 | | ,107 | ,918 |
| | ab_verbale | ,627 | ,115 | ,912 | 5,460 | ,002 |

a. Variabile dipendente: Voto_scolastico

Errore standard della distribuzione campionaria dei due coefficienti, servono a calcolare il t di Student

**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Errore std. | Beta | | |
| 1 | (Costante) | ,135 | 1,255 | | ,107 | ,918 |
| | ab_verbale | ,627 | ,115 | ,912 | 5,460 | ,002 |

a. Variabile dipendente: Voto_scolastico

il t di Student ci informa sulla rarità di un tale parametro sotto l'ipotesi nulla di mancanza di effetto nell'equazione di regressione

**Coefficienti[a]**

| Modello | | Coefficienti non standardizzati | | Coefficienti standardizzati | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Errore std. | Beta | | |
| 1 | (Costante) | ,135 | 1,255 | | ,107 | ,918 |
| | ab_verbale | ,627 | ,115 | ,912 | 5,460 | ,002 |

a. Variabile dipendente: Voto_scolastico

Significatività del t di Student dei due parametri: se la probabilità di ottenere quel valore è molto bassa, si rifiuta l'ipotesi nulla di assenza di effetto e si accetta l'ipotesi che il parametro è diverso da zero e serve nella predizione

# Riepilogo del modello

**Riepilogo del modello[b]**

| Modello | R | R-quadrato | R-quadrato adattato | Errore standard della stima |
|---|---|---|---|---|
| 1 | ,912[a] | ,832 | ,805 | ,644 |

a. Predittori: (costante), ab_verbale

b. Variabile dipendente: voto

Il coefficiente di correlazione multiplo (R maiuscolo) è la correlazione fra il predittore e il predetto. E' sempre positivo

# Riepilogo del modello

**Riepilogo del modello[b]**

| Modello | R | R-quadrato | R-quadrato adattato | Errore standard della stima |
|---|---|---|---|---|
| 1 | ,912[a] | ,832 | ,805 | ,644 |

a. Predittori: (costante), ab_ve ble

b. Variabile dipendente: vo

Il coefficiente quadrato di R è la quota di varianza predetta

# Riepilogo del modello

**Riepilogo del modello[b]**

| Modello | R | R-quadrato | R-quadrato adattato | Errore standard della stima |
|---|---|---|---|---|
| 1 | ,912[a] | ,832 | ,805 | ,644 |

a. Predittori: (costante), ab_verbale

b. Variabile dipendente: voto

Il coefficiente quadrato di R adattato è una stima di quanto si ridurrebbe il coefficiente di correlazione multiplo su un nuovo campione

# Riepilogo del modello

**Riepilogo del modello[b]**

| Modello | R | R-quadrato | R-quadrato adattato | Errore standard della stima |
|---|---|---|---|---|
| 1 | ,912[a] | ,832 | ,805 | ,644 |

a. Predittori: (costante), ab_verbale

b. Variabile dipendente: voto

L'errore standard della stima di R serve a stimare l'intervallo di fiducia entra cui ricade R nella popolazione