

# Psicometria

## 5-Regressione lineare semplice vers. 1.0

Germano Rossi<sup>1</sup>

`germano.rossi@unimib.it`

Giovanni Battista Flebus<sup>1</sup>

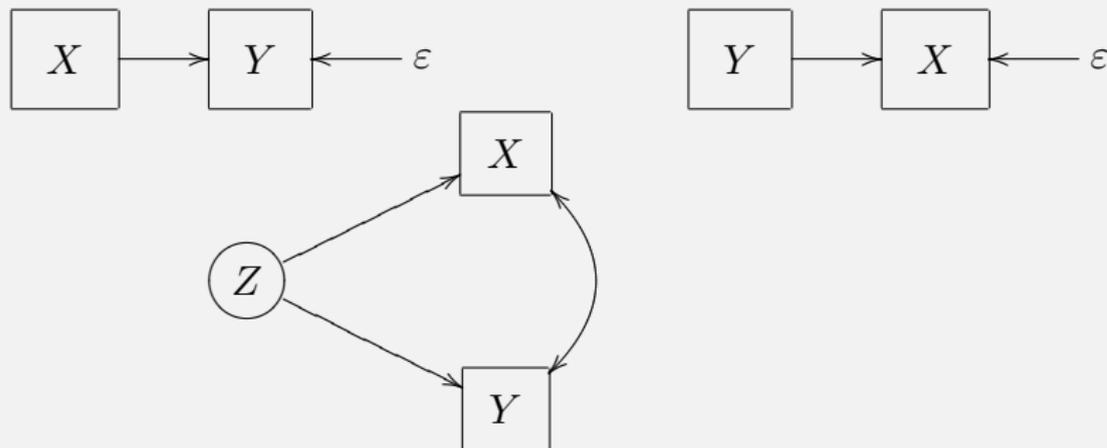
`giovannibattista.flebus@unimib.it`

<sup>1</sup>Dipartimento di Psicologia, Università di Milano-Bicocca

2008-2008

# Correlazione e causalità

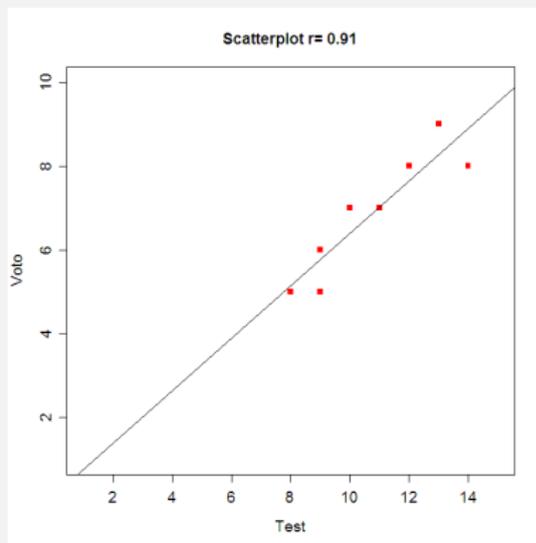
- La correlazione fra due variabili (X e Y) non implica causalità
- Vi sono diverse possibili spiegazioni
  - X causa Y [regr. semplice]
  - Y causa X [regr. semplice]
  - X e Y sono causati da Z (an. Fatt.)
  - X e Y sono causati da Z1, Z2, Z3... Zn (eq. strutturale)



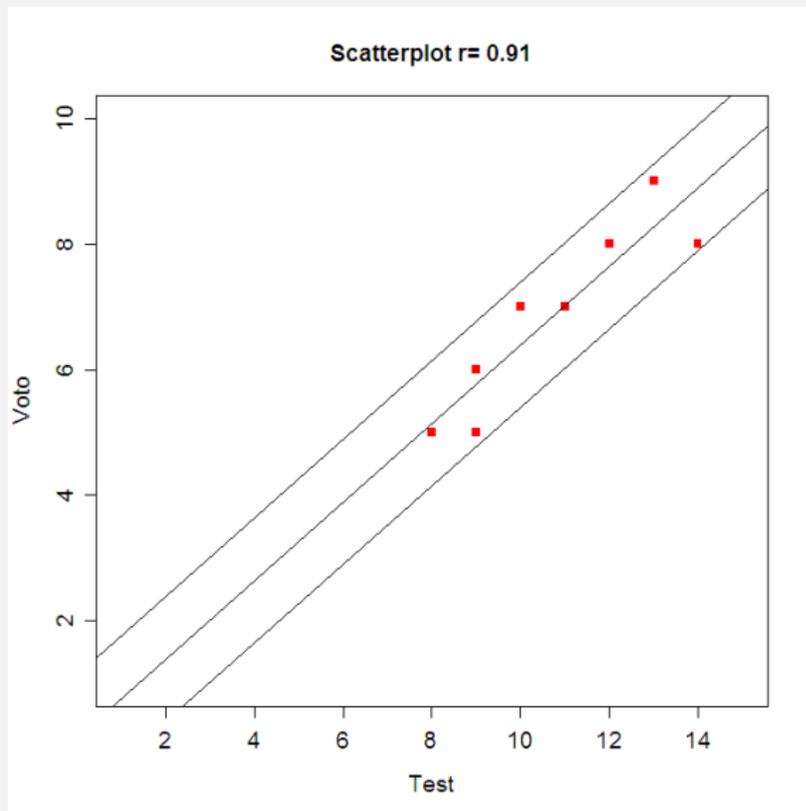
# Cos'è la regressione semplice

- Con la regressione lineare semplice cerchiamo di vedere se la (cor)relazione fra due variabili (ad es. Test di metà anno e Voto finale) può essere spiegata tramite l'equazione di una retta
- Quale variabile sia la dipendente e quale l'indipendente, è una scelta teorica (ipotizzo che il test di metà anno spieghi il voto finale)

	Test	Voto
A	12	8
B	10	7
C	14	8
D	9	5
E	9	6
F	13	9
G	11	7
H	8	5

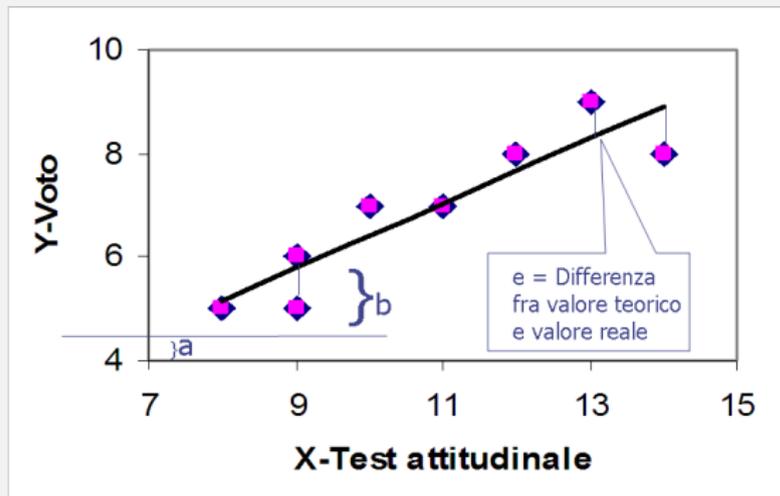


# Scatterplot test-voto



- Cercando una funzione che “approssimi” i punti, posso usare una retta
- ci sono però diverse rette che posso usare
- ciascuna è più vicina a certi punti rispetto alle altre
- qual è la migliore?
- quella che è alla minor distanza possibile da tutti i punti osservati

# Grafico retta



- La formula della retta è  $Y_i = a + bX_i$
- dove X e Y sono le variabili misurate
- b è la pendenza della retta
- a è l'intercetta sull'asse delle ordinate

# Equazione

- In teoria:

$$Y_i = a + bX_i$$

- ma sarebbe vero se i dati fossero perfettamente posizionati sulla retta, mentre invece non lo sono affatto
- questa è l'equazione esatta, perché considera anche l'errore che permette di aggiustare i dati:

$$Y_i = bX_i + a + e$$

# Riepilogo

$$Y_i = bX_i + a + \varepsilon_i$$

Variabile dipendente, spiegata, valore osservato

inclinazione

variabile indipendente

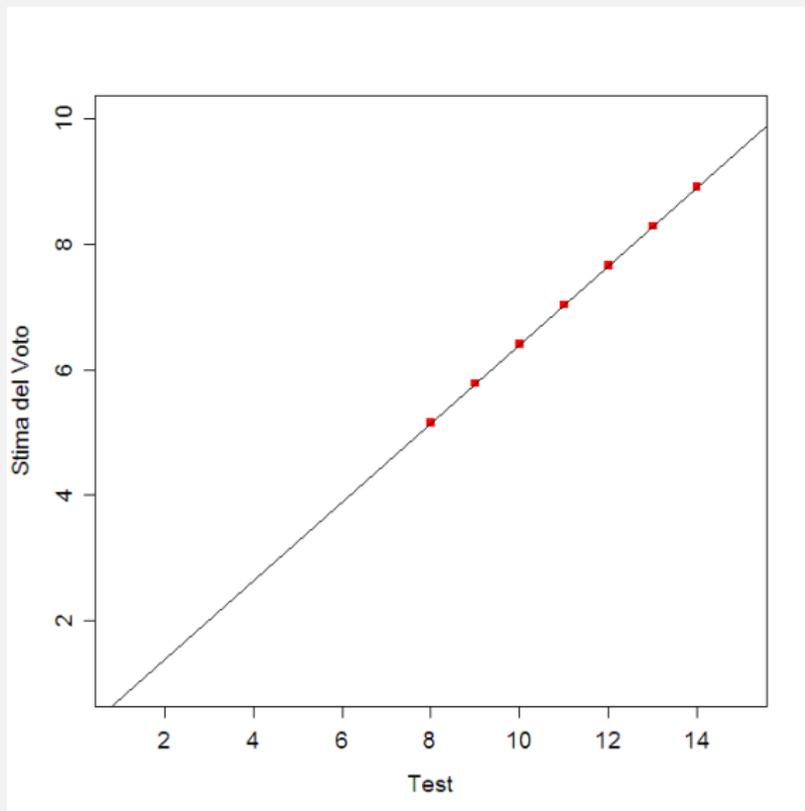
intercetta

errore

Stima di y, valore predetto

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

# Grafico fra $x$ e $y'$ (stime basate sulla retta)



- Se stimo i valori  $y$  usando la retta
- i punti cadono esattamente sulla retta

# Miglior retta

- Fra tutte le possibili rette, selezioniamo la migliore, ovvero quella che è alla minor distanza possibile da tutti i punti osservati
- ovvero, in cui la somma degli errori è la minima possibile
- Obiettivo: minimizzare  $\sum e_i$
- $e_i = Y_i - \hat{Y}_i = Y_i - bX_i + a$
- Bisogna che questi errori siano i più piccoli possibili e quindi usiamo il “metodo dei minimi quadrati”

# Formule algebriche

$$b = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{cov(X, Y)}{var(X)} = \frac{s_{xy}}{s_x^2} = r \frac{s_y}{s_x}$$
$$= \frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{N \sum X_i^2 - (\sum X_i)^2}$$

$$a = \bar{Y} - b\bar{X} = \frac{\sum Y}{N} - b \frac{\sum X}{N}$$

# Esempio

	X-Test	Y-Voto	$X^2$	$Y^2$	XY
A	12	8	144	64	96
B	10	7	100	49	70
C	14	8	196	64	112
D	9	5	81	25	45
E	9	6	81	36	54
F	13	9	169	81	117
G	11	7	121	49	77
H	8	5	64	25	40
Somma	86	55	956	393	611
Media	10,75	6,875			

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

$$a = 6.875 - 0,627$$

$$\times 10.75 = 0.135$$

$$b =$$

$$\frac{N \sum X_i Y_i - \sum X_i \sum Y_i}{N \sum X_i^2 - (\sum X_i)^2}$$

$$b = \frac{8 \times 611 - 86 \times 55}{8 \times 956 - (86)^2} = \frac{4888 - 4730}{7648 - 7396} = \frac{158}{252} = 0,627$$

## Stime di Y e residui

	Test $X$	Voto $Y$	Eq.	Stimati $\hat{Y}$	Residui $e = Y - \hat{Y}$
A	12	8	$0.135+0.627(12)$	7.659	0.341
B	10	7	$0.135+0.627(10)$	6.405	0.595
C	14	8	$0.135+0.627(14)$	8.913	-0.913
D	9	5	$0.135+0.627(9)$	5.778	-0.778
E	9	6	$0.135+0.627(9)$	5.778	0.222
F	13	9	$0.135+0.627(13)$	8.286	0.714
G	11	7	$0.135+0.627(11)$	7.032	-0.032
H	8	5	$0.135+0.627(8)$	5.151	-0.151

# Regressione semplice matriciale

Immaginiamo di avere due variabili con 5 osservazioni ciascuna (prime 2 colonne) e chiamiamo l'intercetta  $b_0$  e la pendenza  $b_1$ , per cui l'equazione  $Y = a + bX + e$  diventa:  $Y = b_0 + b_1X + e$  e applichamola ad ogni valore di X e di Y

Y	X	$X^2$	XY	$Y = b_0 + b_1X + e$	
3	2	4	6	$3 = b_0 + b_1 \cdot 2 + e_1$	$3 = b_0 \mathbf{1} + b_1 \cdot 2 + e_1$
2	3	9	6	$2 = b_0 + b_1 \cdot 3 + e_2$	$2 = b_0 \mathbf{1} + b_1 \cdot 3 + e_2$
4	5	25	20	$4 = b_0 + b_1 \cdot 5 + e_3$	$4 = b_0 \mathbf{1} + b_1 \cdot 5 + e_3$
5	7	49	35	$5 = b_0 + b_1 \cdot 7 + e_4$	$5 = b_0 \mathbf{1} + b_1 \cdot 7 + e_4$
8	8	64	64	$8 = b_0 + b_1 \cdot 8 + e_5$	$8 = b_0 \mathbf{1} + b_1 \cdot 8 + e_5$
22	25	151	131	← somma	
4.4	5			← media	

# Regressione semplice matriciale

Dall'espressione algebrica, passiamo ad un'espressione matriciale, perché  $b_01 + b_1x_i$  può essere considerato una combinazione lineare della matrice  $\mathbf{X}$  dei dati per il vettore  $\mathbf{b}$  dei pesi

$$Y = b_01 + b_1X + e$$

$$3 = b_01 + b_12 + e_1$$

$$2 = b_01 + b_13 + e_2$$

$$4 = b_01 + b_15 + e_3$$

$$5 = b_01 + b_17 + e_4$$

$$8 = b_01 + b_18 + e_5$$

$$\begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

Poiché per noi la nostra incognita è il vettore  $\mathbf{b}$ , dobbiamo ri-esprimerla in modo da ottenere le stime di  $\mathbf{b}$

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

# Calcolo

$$\left( \begin{array}{c} [1 \ 1 \ 1 \ 1 \ 1] \\ [2 \ 3 \ 5 \ 7 \ 8] \end{array} \begin{array}{c} \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & 5 \\ 1 & 7 \\ 1 & 8 \end{bmatrix} \end{array} \right)^{-1} \begin{array}{c} [1 \ 1 \ 1 \ 1 \ 1] \\ [2 \ 3 \ 5 \ 7 \ 8] \end{array} \begin{array}{c} \begin{bmatrix} 3 \\ 2 \\ 4 \\ 5 \\ 8 \end{bmatrix} \end{array}$$

Se ricordiamo i prodotti scalari,  $\mathbf{X}'\mathbf{X}$  e  $\mathbf{X}'\mathbf{y}$  vengono espressi come somme, somme al quadrato e coprodotti.

$$\begin{bmatrix} 5 & 25 \\ 25 & 151 \end{bmatrix}^{-1} \begin{bmatrix} 22 \\ 131 \end{bmatrix} \quad \begin{bmatrix} N & \sum X \\ \sum X & \sum X^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum Y \\ \sum XY \end{bmatrix}$$

# Calcoliamo

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 151/130 & -25/130 \\ -25/130 & 5/130 \end{bmatrix}$$

$$\begin{bmatrix} 151/130 & -25/130 \\ -25/130 & 5/130 \end{bmatrix} \begin{bmatrix} 22 \\ 131 \end{bmatrix} = \begin{bmatrix} \frac{151 * 22 - 25 * 131}{130} \\ \frac{-25 * 22 + 5 * 131}{130} \end{bmatrix} = \begin{bmatrix} 0.362 \\ 0.808 \end{bmatrix}$$

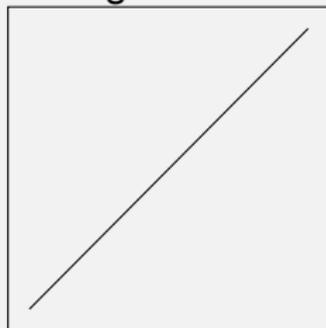
$$\begin{bmatrix} 0.362 \\ 0.808 \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \text{intercetta} \\ \text{pendenza} \end{bmatrix}$$

$$\hat{Y}_i = 0.362 + 0.808X_i$$

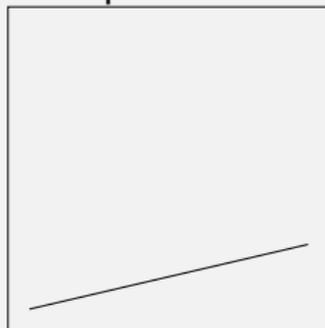
# Confronto delle pendenze

- Il coefficiente angolare dipende dal modo in cui è espressa la variabile X e non si può dire se sia piccolo o grande se non conoscendo la gamma (differenza fra massimo e minimo) di X oppure facendo una rappresentazione grafica

grande



piccola



# Con dati standardizzati

- Se usiamo X e Y trasformati in punti z, la formula della retta cambia in

$$z_{\hat{y}} = r z_x$$

perché tutti dati sono espressi con media 0 e dev.st 1, quindi

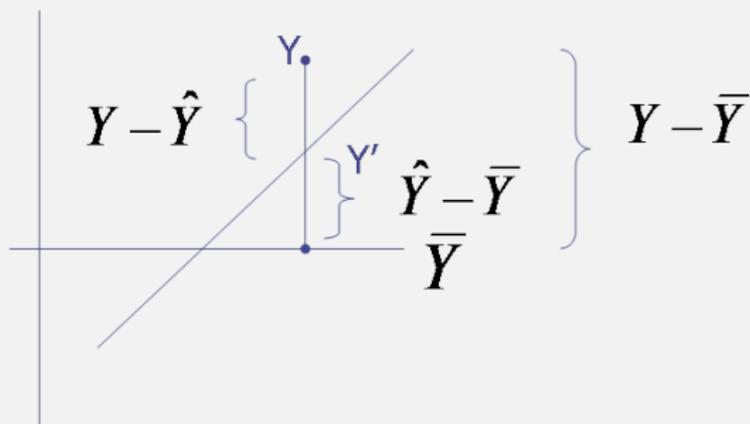
$$b = r \frac{s_y}{s_x}$$

diventa

$$b^* = r$$

- $b^*$  è la pendenza standardizzata
- l'intercetta è 0 perché  $a = \bar{Y} - b\bar{X}$  e le medie sono 0

# I residui e le loro sommatorie



- Per ogni singolo valore, se non ci fossero interferenze esterne, la  $Y$  corrisponderebbe al suo valore atteso  $E(Y) = \bar{Y}$
- Il segmento corrispondente a  $Y - \bar{Y}$  può essere diviso in due parti
- L'introduzione di  $X$ , giustifica la parte  $\hat{Y} - \bar{Y}$
- Mentre non abbiamo idea di cosa produca la parte restante:  $Y - \hat{Y}$
- Ma possiamo dire che lo scarto dal valore atteso è divisibile in una parte spiegata da  $X$  e in una parte non spiegata (il residuo)

## Residui e varianze

- Dal momento che la somma al quadrato degli scarti dalla media corrisponde alla varianza... possiamo trasformare la relazione

$$Y - \bar{Y} = (Y - \hat{Y}) + (\hat{Y} - \bar{Y})$$

in

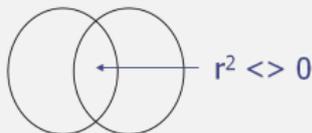
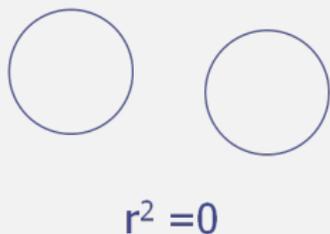
$$\sum_{\text{totale}} (Y - \bar{Y})^2 = \sum_{\text{non spiegata}} (Y - \hat{Y})^2 + \sum_{\text{spiegata}} (\hat{Y} - \bar{Y})^2$$

- facendo il rapporto fra la varianza spiegata e quella totale, la possiamo esprimere come **proporzione di varianza spiegata**

$$r^2 = (r)^2 = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{\sum (Y - \bar{Y})^2 - \sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$$

# Proporzione di varianza spiegata

- La proporzione di varianza spiegata è anche chiamata “Coefficiente di determinazione”, “r quadro” oppure “varianza comune”
- La parte complementare è chiamata “Coefficiente di indeterminazione” ( $1 - r^2$ )



Proporzione di varianza comune a due variabili

# Errore standard delle stime

- Varianza degli errori previsti

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

- e la relativa deviazione standard

$$s_{y.x} = \sqrt{\frac{\sum(Y - \hat{Y})^2}{N}} = s_y \sqrt{1 - r^2}$$

- Ne consegue che se  $r = 1$ , va 0 (nessun errore)
- se  $r = 0$ , va a  $s_y$  (massimo errore)

# Errore standard delle stime

A cosa serve l'errore standard delle stime?

- Essendo una deviazione standard, e presumendo che X e Y siano distribuite normalmente, possiamo stimare che il 95% dei valori Y stimati a partire da un certo valore X sarà compreso fra:

$$\hat{Y} - 1.96s_{y.x} \quad \text{e} \quad \hat{Y} + 1.96s_{y.x}$$

- dove 1.96 è il punto z corrispondente all'area 95% attorno alla media

# In SPSS

- Analizza | Regressione | Lineare...
- Trasferite in Dipendente la variabile che volete spiegare (la Y)
- Trasferite in Indipendenti la variabile che volete usare per spiegare (la X)
- Date l'OK

Coefficienti<sup>a</sup>

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	78,531	1,176		66,785	,000
	politica	1,801	,210	,360	8,591	,000

a. Variabile dipendente: Fundament

# In SPSS

Riepilogo del modello<sup>b</sup>

Modello	R	R-quadrato	R-quadrato corretto	Errore std. della stima
1	,360 <sup>a</sup>	,130	,128	11,35156

a. Stimatori: (Costante), politica

b. Variabile dipendente: Fundament

Statistiche dei residui<sup>a</sup>

	Minimo	Massimo	Media	Deviazione std.	N
Valore atteso	80,332	96,5363	87,639	4,37420	498
Residuo	-38,54	57,8653	,00000	11,34013	498
Valore atteso std.	-1,670	2,034	,000	1,000	498
Residuo std.	-3,395	5,098	,000	,999	498

a. Variabile dipendente: Fundament

# Riassunto terminologico

- Regressione lineare semplice = regressione bivariata = predizione bivariata
- $X$  = variabile indipendente, v. predittiva
- $Y$  = variabile dipendente, v. predetta, v. criterio
- $Y'$ ,  $\hat{Y}$  = valore stimato, v. previsto
- $a$ ,  $b_0$  = intercetta, costante
- $b$ ,  $b_1$  = coeff. angolare, coeff. di regressione, pendenza, parametro di regressione
- $\beta$ ,  $b^*$  = coeff. angolare standardizzato