UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Il modello di regressione e la correlazione di Pearson

Capitolo 2 e 3

Marcello Gallucci

Milano-Bicocca

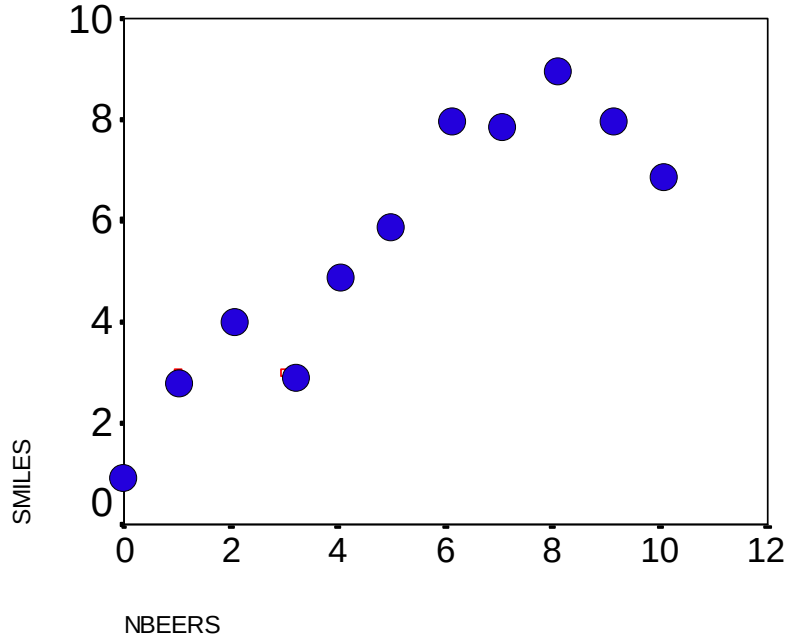
A
M
D

Lezione: II

Concetti fondamentali

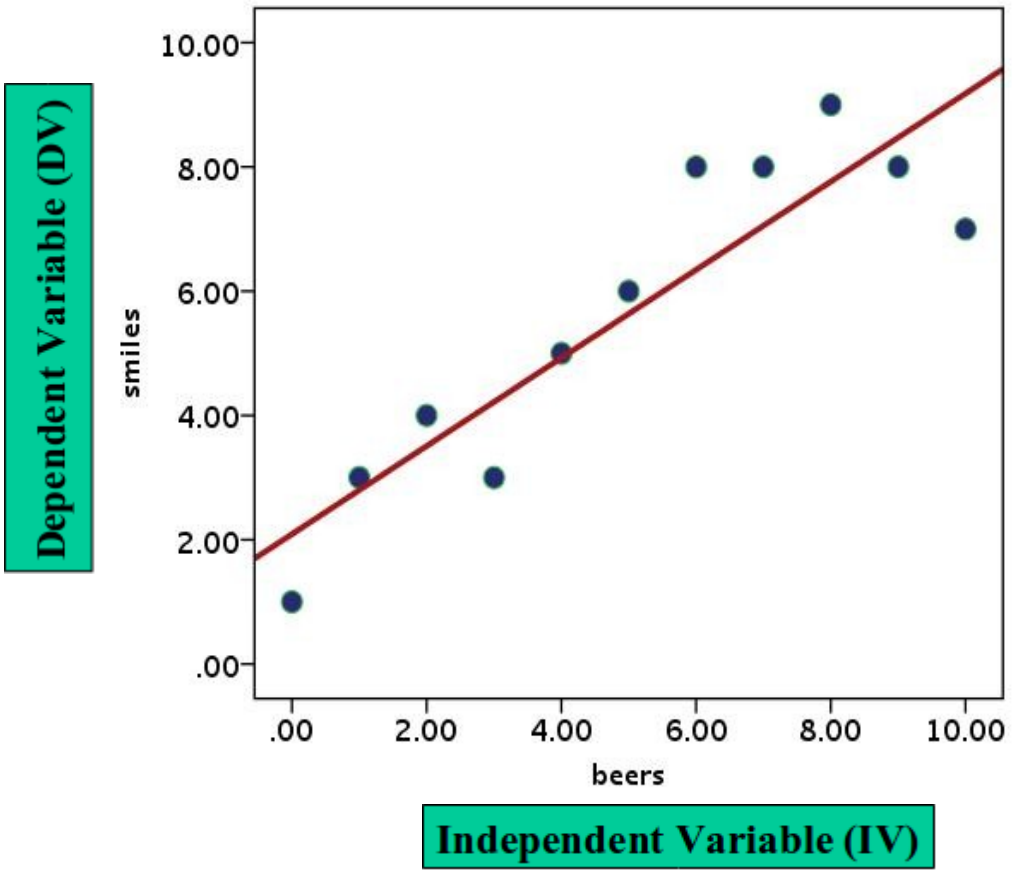
Consideriamo ora questa ipotetica ricerca: siamo andati in un pub ed abbiamo contato quanti sorrisi le persone ai tavoli producevano (ogni 10 minuti) e quante birre avevano bevuto fino a quel momento

Birre	Sorrisi
0	1
1	3
2	4
3	3
4	5
5	6
6	8
7	8
8	9
9	8
10	7



Concetti fondamentali

Lo scopo della retta di regressione è di rappresentare la relazione lineare tra la variabile indipendente e la dipendente



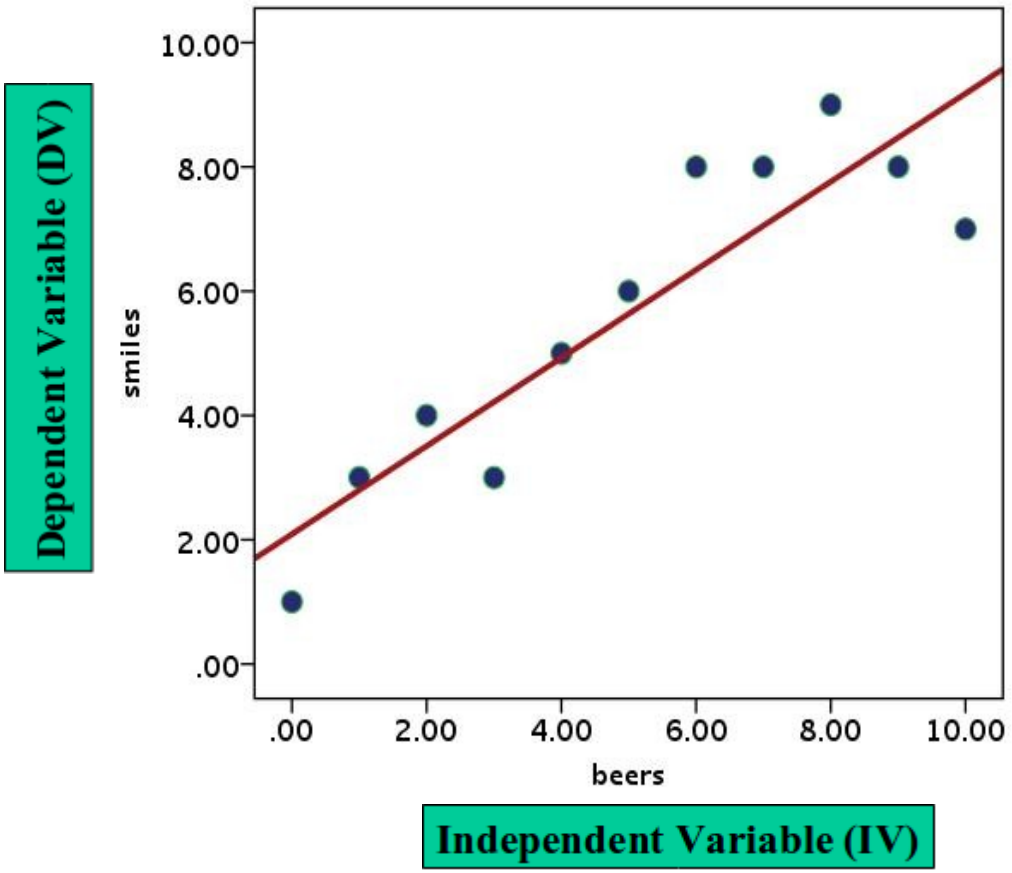
Nel caso più semplice, abbiamo una retta senplice

$$y_i = a + b \cdot y_i + e_i$$

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

Concetti fondamentali

La retta di regressione è la retta che meglio interpola la nuvola dei punti (minimizza la distanza con i dati)



Nel caso più semplice, abbiamo una retta senplice

$$y_i = a + b \cdot y_i + e_i$$

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

Concetti fondamentali

La retta può essere descritta mediante due coefficienti: il termine costante ed il coefficiente angolare

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

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.091	.684		3.057	.014
	NBEERS	.709	.116	.898	6.132	.000

a. Dependent Variable: SMILES

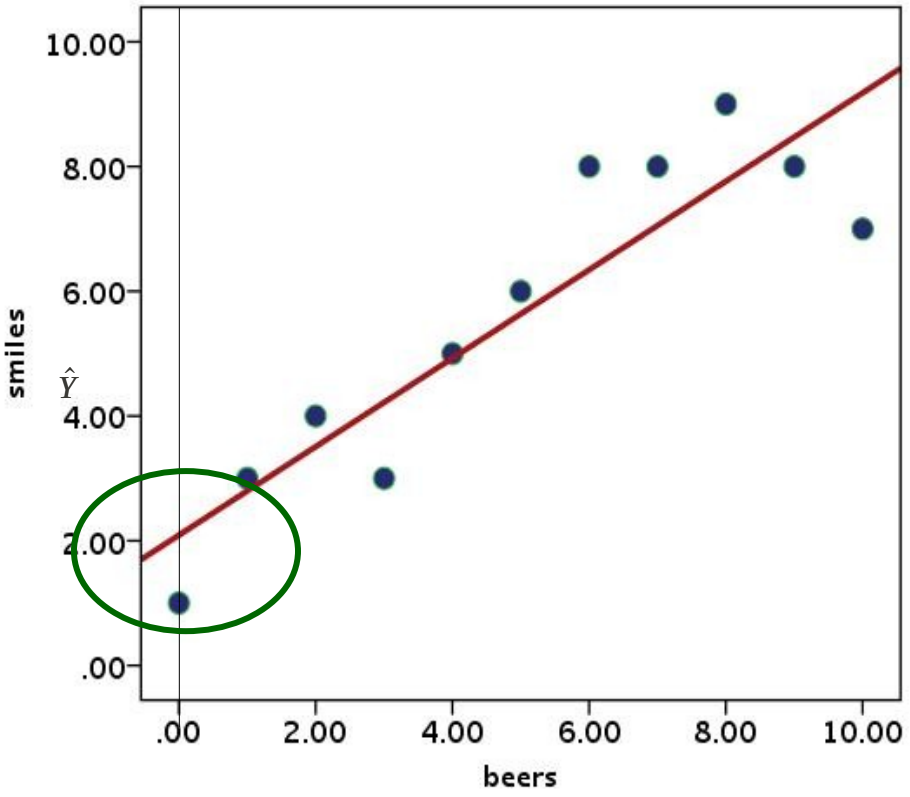
Termine constante (o intercetta)

Coefficiente di regressione (angolare)

Coefficiente costante

a l'intercetta della linea: indica il valore atteso (medio) della VD per la VI=0

$$\hat{y} = a + b \cdot 0$$

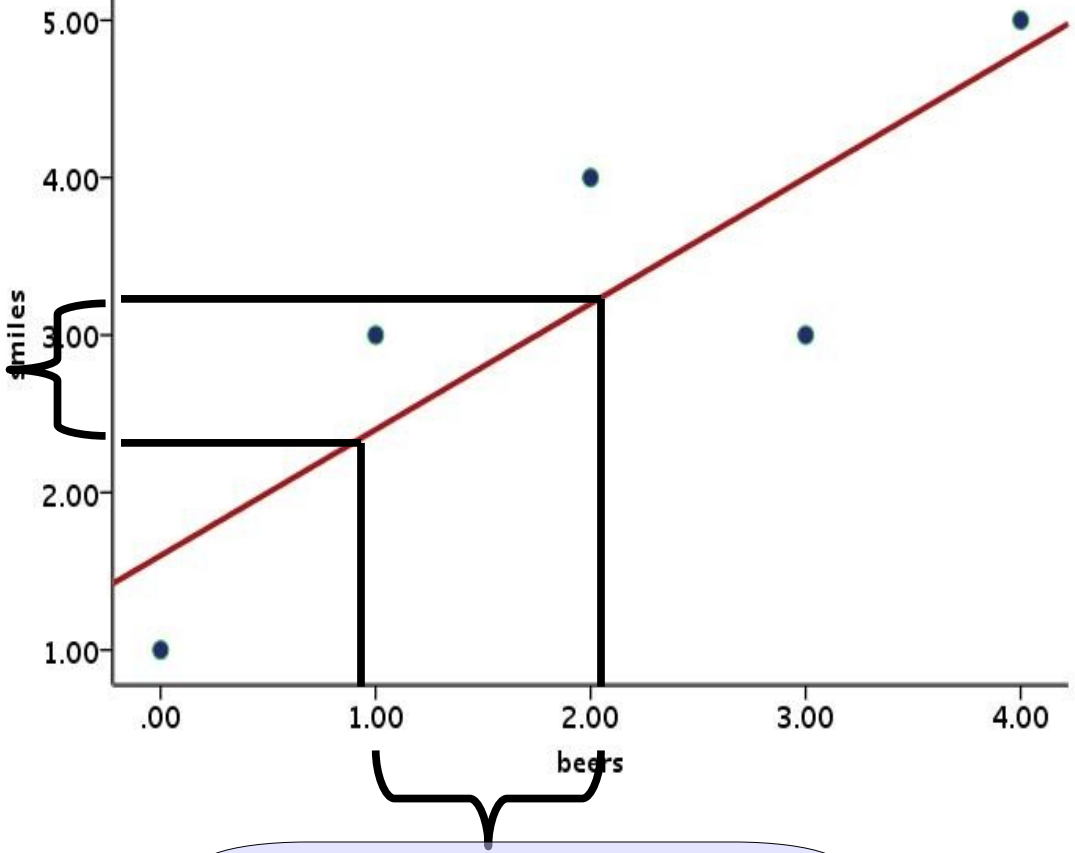


Quando un partecipante ha bevuto zero birre, mostra (in media) 2.09 sorrisi

Coefficiente di regressione

B è il coefficiente angolare della retta: indica il cambiamento atteso nella VD al variare di una unità della VI

I sorrisi aumentano di B unità

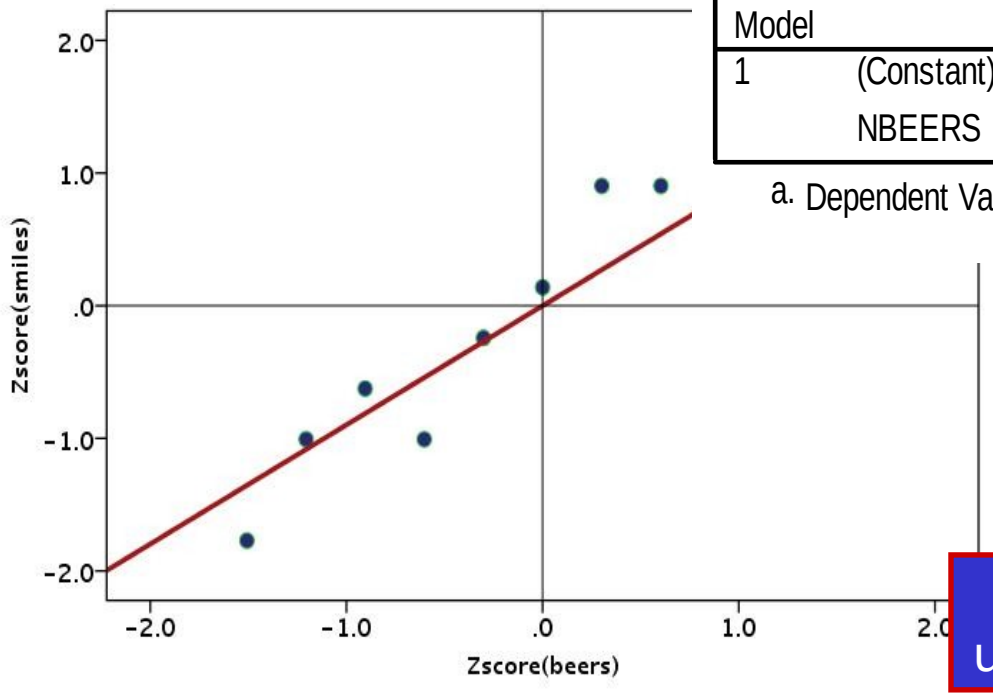


Per una unità in più della VI: una birra in più

Per ogni birra che si beve, i sorrisi aumentano in media di .709 unità

Coefficienti standardizzati

Il coefficiente **Beta** equivale al coefficiente di regressione calcolato dopo aver **standardizzato** tutte le variabili



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.091	.684		3.057	.014
	NBEERS	.709	.116	.898	6.132	.000

a. Dependent Variable: SMILES

Il coefficiente standardizzato è uguale al coefficiente r di Pearson

Standardizzazione delle variabili

- L'unita' di misura di una variabile si elimina standardizzando la variabile

Standardizzare serve per ridurre ogni variabile ad una stessa metrica

Variabile originale

Media di v

Punteggio z per ogni soggetto (caso)

Scarto di v

$$z_{iv} = \frac{v_i - M_v}{s_v} = \frac{d_i}{s_v}$$

Deviazione standard di v

- Le variabili standardizzate hanno tutte **media uguale a 0** e **varianza uguale ad 1**

Correlazione=beta

- Il beta è il **coefficiente di regressione standardizzato**

- In una regressione semplice con una variabile dipendente ed una indipendente, il coefficiente beta è identico al **coefficiente di correlazione di Pearson**

- Il coefficiente di correlazione di Pearson è **l'indice di associazione bivariata** più usato per quantificare la relazione tra due variabili continue

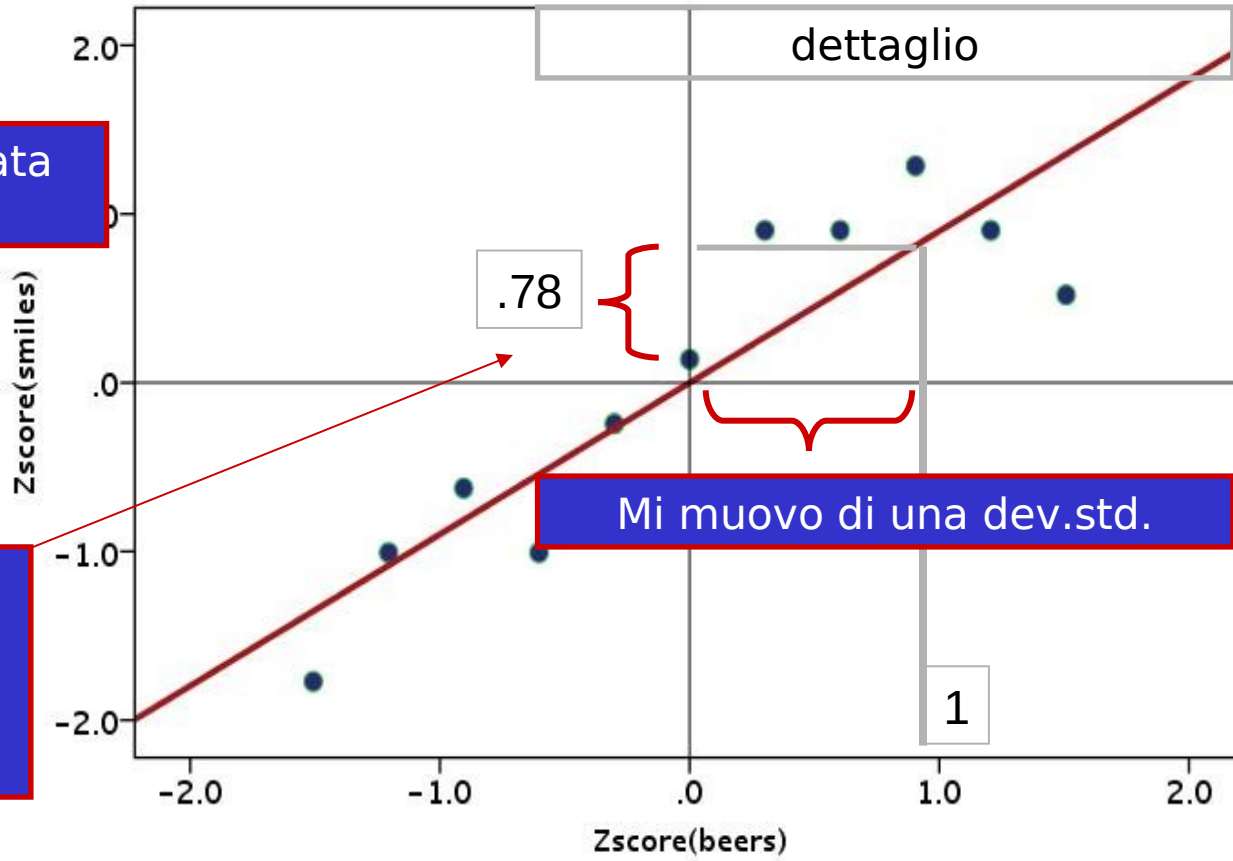
Correlazione: Interpretazione

- La correlazione indica di quante **deviazioni standard varia v**, al variare di **x** di **una deviazione standard**

Regressione standardizzata
r=0.78

$$\hat{v}_z = r_{xv} x_z$$

Mi aspetto una scostamento pari a 78% della dev.std di **v**



Correlazione: Interpretazione

- Il coefficiente di correlazione varia tra -1 e 1

  r = 1 indica che le variabili sono perfettamente proporzionali (sono uguali al netto dell'unità di misura)

  r = -1 indica che le variabili sono perfettamente inversamente proporzionali

- Il coefficiente di correlazione è positivo quando c'è concordanza tra i punteggi delle due variabil

- è negativo quando c'è disconcordanza tra i punteggi

- è pressoché nullo (quasi zero) quando le variabili non sono associate linearmente (per alcuni casi c'è concordanza, per altri discordanza)

Correlazione: Interpretazione

a1 e a2 sono praticamente indipendenti

a3 e a4 sono fortemente associate

Correlazioni

		a1	a2	a3	a4
a1	Correlazione di Pearson	1	.084	.154	.242*
	Sig. (2-code)		.409	.126	.015
	N	100	100	100	100
a2	Correlazione di Pearson	.084	1	.514**	.231*
	Sig. (2-code)	.409		.000	.021
	N	100	100	100	100
a3	Correlazione di Pearson	.154	.514**	1	.588**
	Sig. (2-code)	.126	.000		.000
	N	100	100	100	100
a4	Correlazione di Pearson	.242*	.231*	.588**	1
	Sig. (2-code)	.015	.021	.000	
	N	100	100	100	100

*. La correlazione è significativa al livello 0,05 (2-code).

**. La correlazione è significativa al livel

a2 e a4 sono debolmente associate

Valori guida in pratica

Interpretazione	r	R2	% condivisa
Troppo Alta	1,0	1,0	100%
Alta	0,9	0,8	81%
	0,8	0,6	64%
	0,7	0,5	49%
	0,6	0,4	36%
	0,5	0,3	25%
Media	0,4	0,2	16%
	0,3	0,1	9%
	0,2	0,0	4%
Bassa/assente	0,1	0,0	1%
	0,0	0,0	0%
	-0,1	0,0	1%
Media	-0,2	0,0	4%
	-0,3	0,1	9%
	-0,4	0,2	16%
Alta	-0,5	0,3	25%
	-0,6	0,4	36%
	-0,7	0,5	49%
	-0,8	0,6	64%
	-0,9	0,8	81%
Troppo Alta	-1,0	1,0	100%

Test inferenziale

I coefficenti vengono testati per la loro significatività statistica mediante il t-test **t test**

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.091	.684		3.057	.014
	NBEERS	.709	.116	.898	6.132	.000

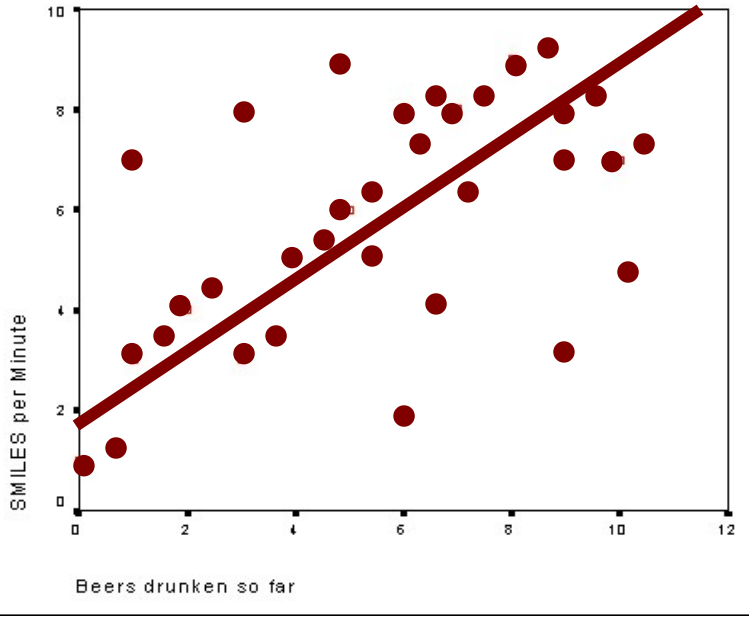
a. Dependent Variable: SMILES

Se Sig. < 0.05, diremo che B (e β) sono significativamente diversi da zero

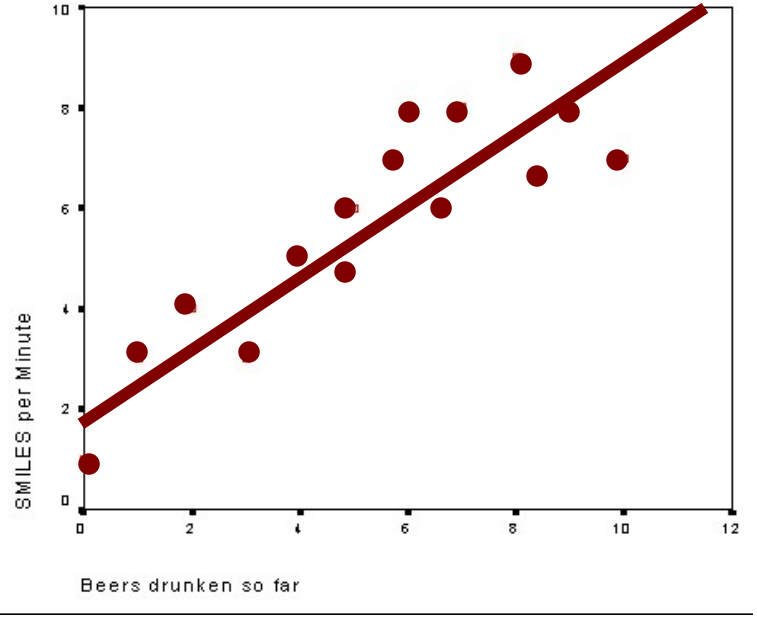
Bonta’ di adattamento

Non tutte le rette di regressione hanno lo stesso potere predittivo, cioè la stessa capacità di adattarsi ai dati osservati

bassa



alta



Errore di regressione

Notiamo che la predizione non corrisponde di norma ai valori osservati

$$\hat{y}_i = a + b_{yx} x_i$$

predetti

errore

$$y_i - \hat{y}_i = y_i - (a + b_{yx} x_i)$$

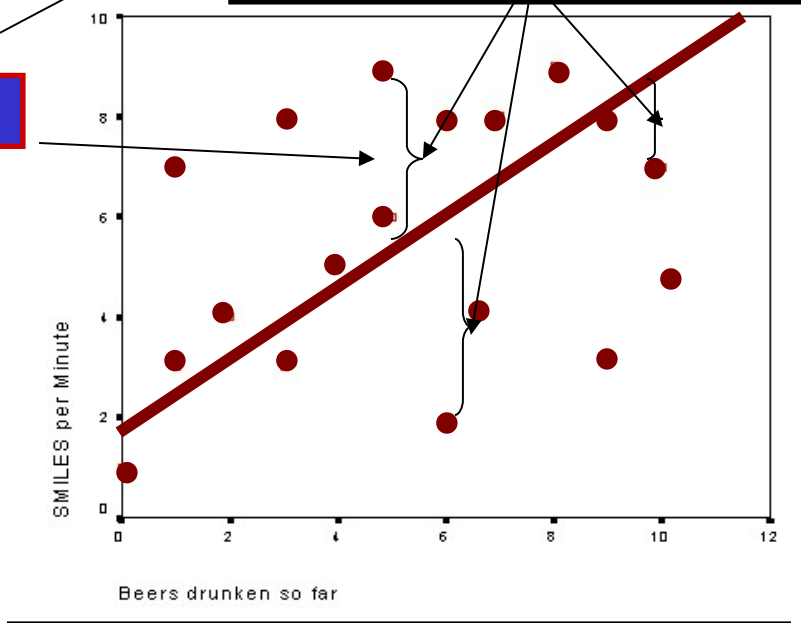
Dunque i valori osservati di Y possono essere espressi come somma dei valori predetti e l'errore

$$y_i = (a + b_{yx} x_i) + (y_i - \hat{y}_i)$$

retta

errore

Discrepanza osservati-predetti



Quanto e' grande l'errore di regressione

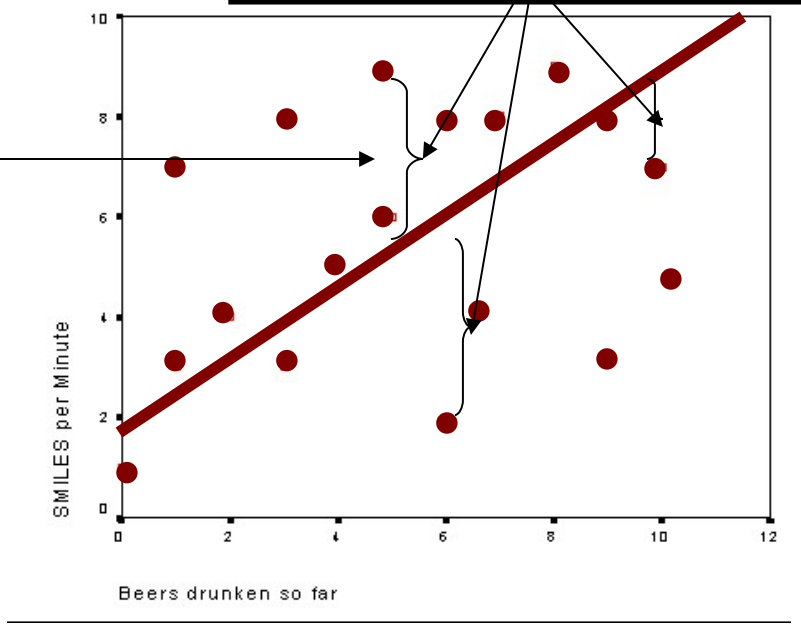
Calcoliamoci la distanza media tra i punti osservati e la retta

Le distanze si calcolano mediante le differenze al quadrato

$$\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-1} = s_e^2$$

Notiamo che questa e' una varianza, che chiameremo varianza di errore

Discrepanza osservati-predetti



Proporzione riduzione errore

Il modello si adatterà ai dati tanto più riduce l'errore di predizione rispetto a non usare tale modello

- La logica è di confrontare due casi:
  - L'errore calcolato per la regressione data
  - L'errore associato alla media, cioè errore associato a non utilizzare la regressione

Proporzione riduzione errore

Senza regressione l'unica predizione plausibile di Y e' la media di Y

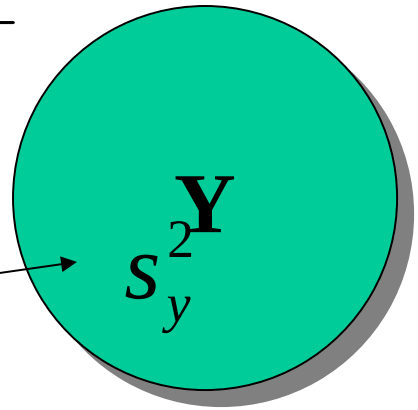
Predizione senza regressione

$$\hat{y}_i = M_y$$

Errore senza regressione

$$s_y^2 = \frac{\sum (y_i - M)^2}{n - 1}$$

Le deviazioni dalla media (la varianza) non siamo in grado di spiegarle



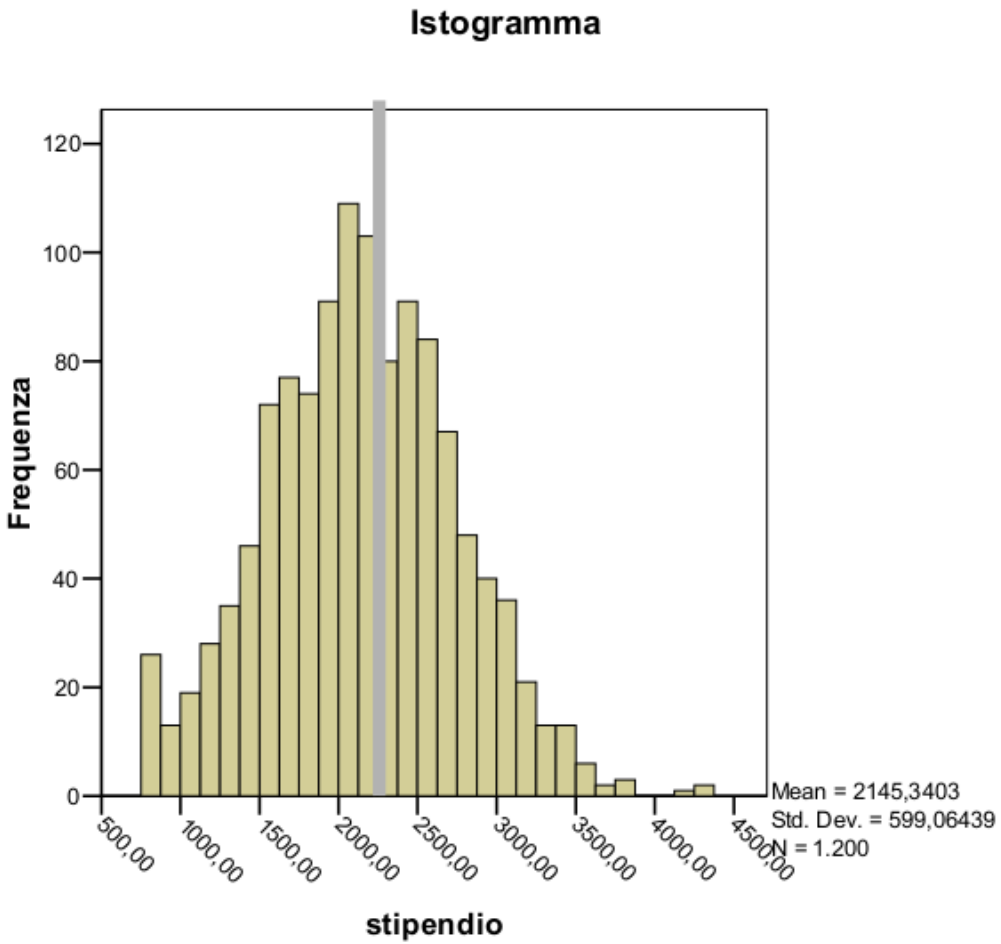
Predizione senza regressione

- Ricordiamo che in assenza di ogni ulteriore informazione, la miglior predizione che si può fare de[illegible] valore medio

Quale è lo stipendio più probabile di un accademico?

$$\hat{y}_i = M_y$$

Media=2145
Varianza=599



Varianza ed errore di predizione

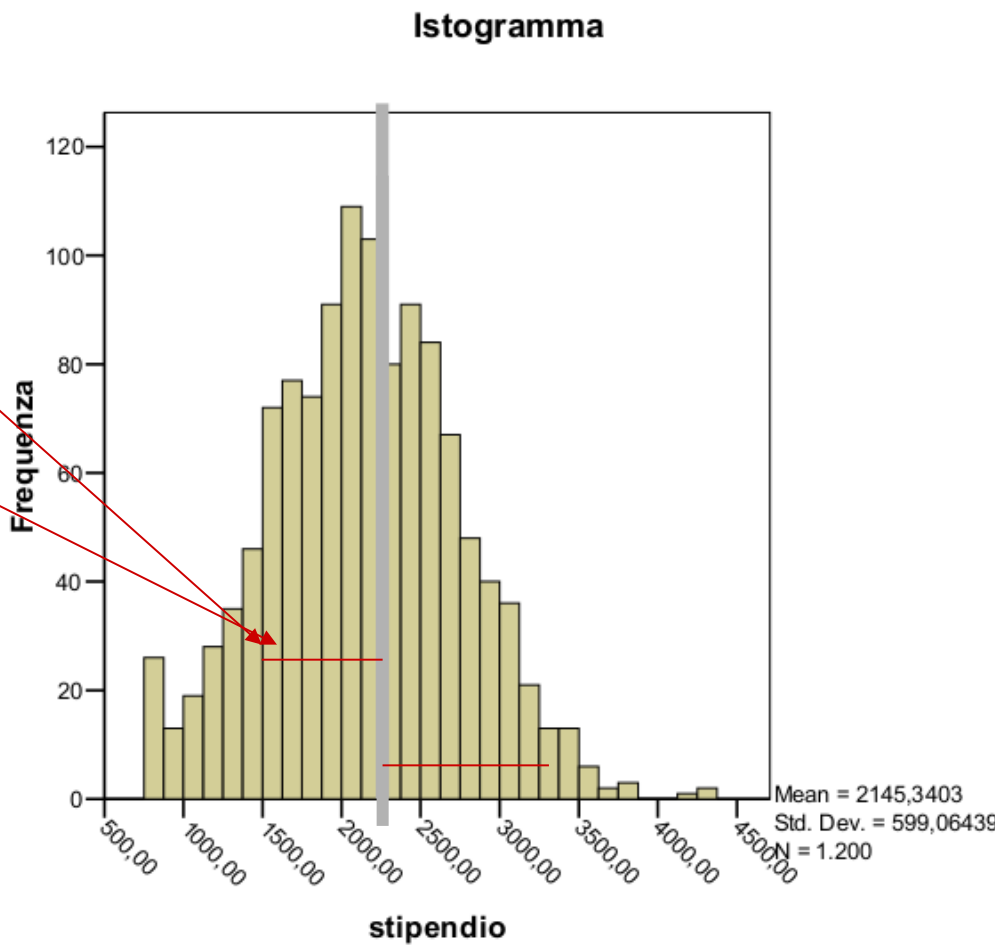
- Se predicessimo che tutti hanno un punteggio pari al valore medio, quale sarebbe il nostro errore?

Tutto ciò che si distanzia dalla media

$$y_i - \hat{y}_i = y_i - M_y$$

$$s^2 = \frac{\sum (y_i - M_y)^2}{(n-1)}$$

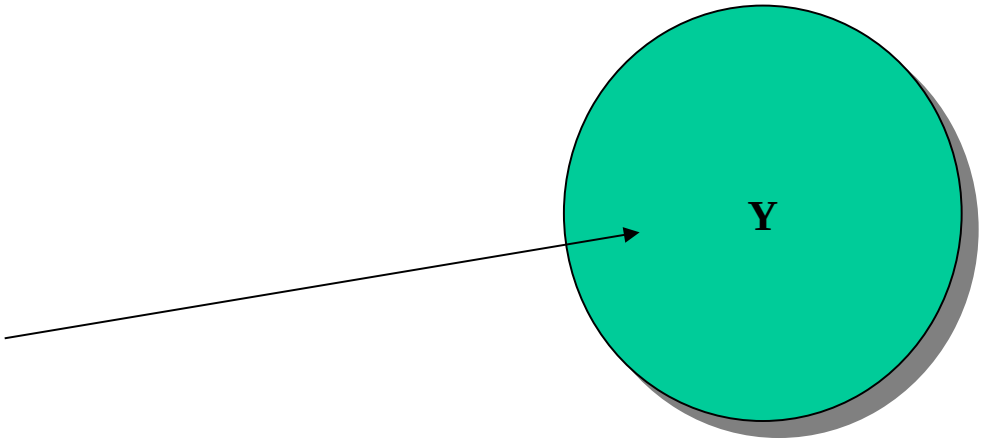
Media=16.14
Varianza=20.38



Varianza ed errore di predizione

- La varianza della variabile da predire rappresenta sia l'errore che commettiamo nell'usare la media come predittore, sia tutta l'informazione che possiamo spiegare se usassimo un predittore migliore della media

$$s^2 = \frac{\sum (y_i - M_y)^2}{(n-1)}$$



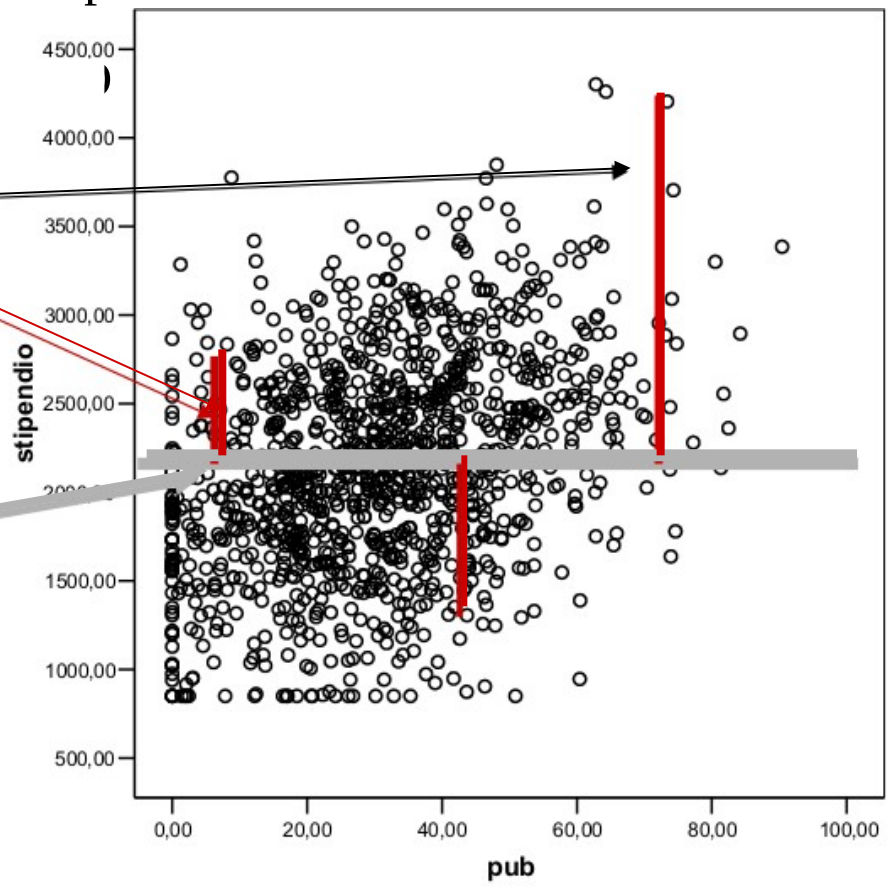
Varianza ed errore di predizione

- Consideriamo il diagramma di dispersione tra la nostra variabile dipendente ed una altra variabile, sempre nel caso volessimo usare il valore medio come predittore de

Errore di predizione: Tutto ciò che si distanzia dalla media

$$y_i - M_y$$

$$s^2 = \frac{\sum (y_i - M_y)^2}{(n-1)}$$

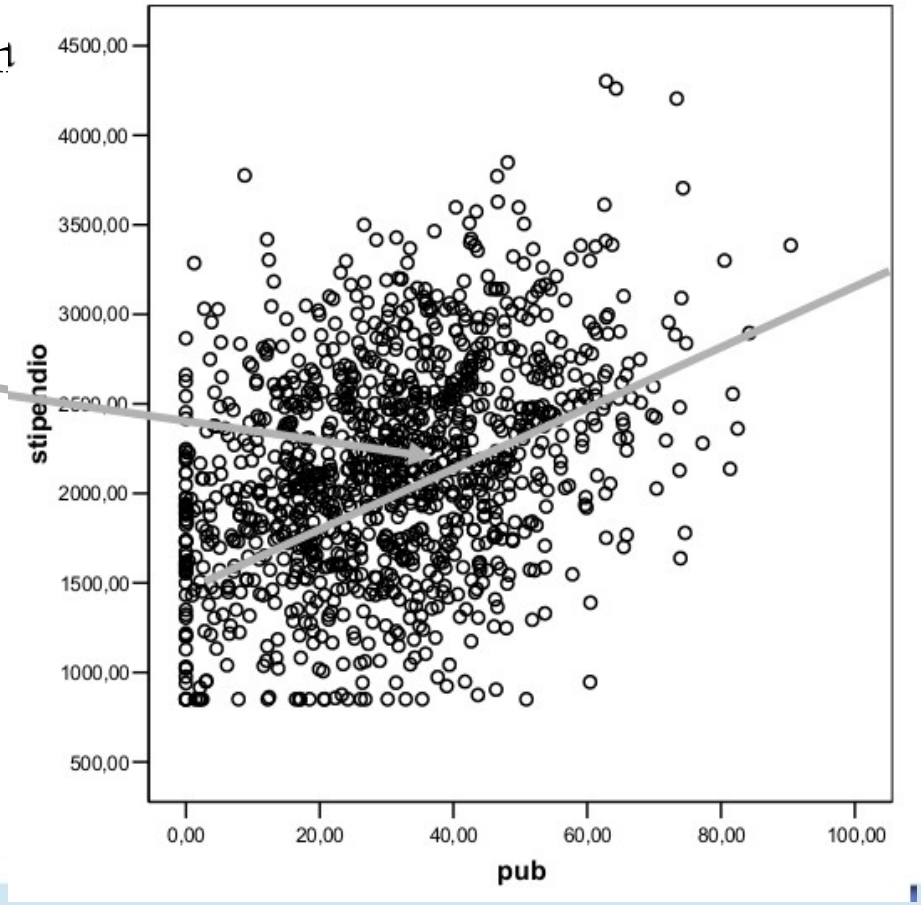


Regressione

- Se ora usiamo i valori di una variabile indipendente, pesati per i coefficienti di regressione, come predittori, il nostro punteggio predetto sarà generalmente diverso da prim

Valori predetti

$$\hat{y}_i = a + b_{yx} x_i$$



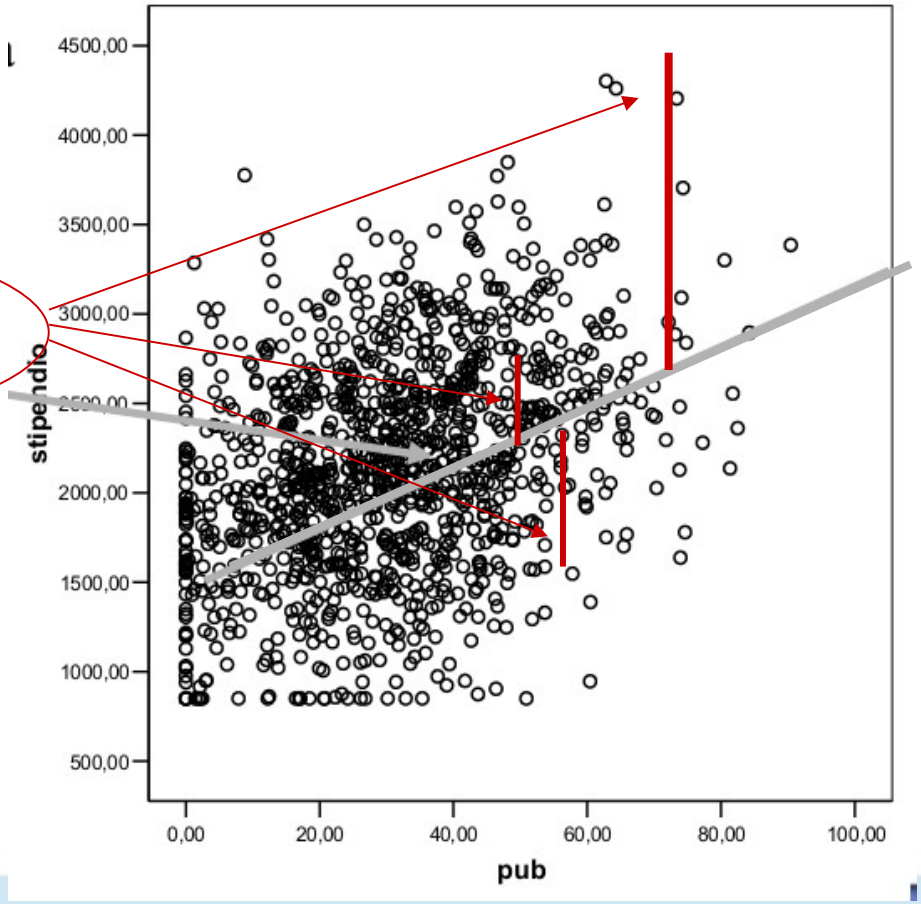
Errore della Regressione

- Anche la predizione fatta con la regressione commetterà degli errori, cioè il valore predetto non coinciderà perfettamente con il valore osservato

Errore che commettiamo

$$y_i - \hat{y}_i = y_i - (a + b_{yx} x_i)$$

$$s_e^2 = \frac{\sum [y_i - (a + b_{yx} x_i)]^2}{(n-1)}$$

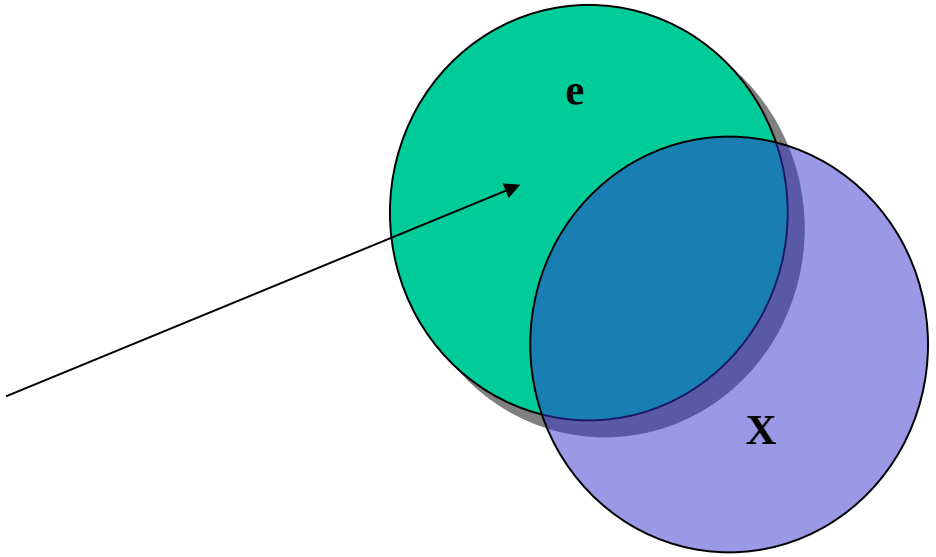


Varianza di errore

- Questa varianza, detta di errore, indica la parte della varianza della VD che non è predicibile mediante i punteggi della VI

Media degli errori di regressione

$$s_e^2 = \frac{\sum [y_i - (a + b_{yx} x_i)]^2}{(n-1)}$$



% Varianza di errore

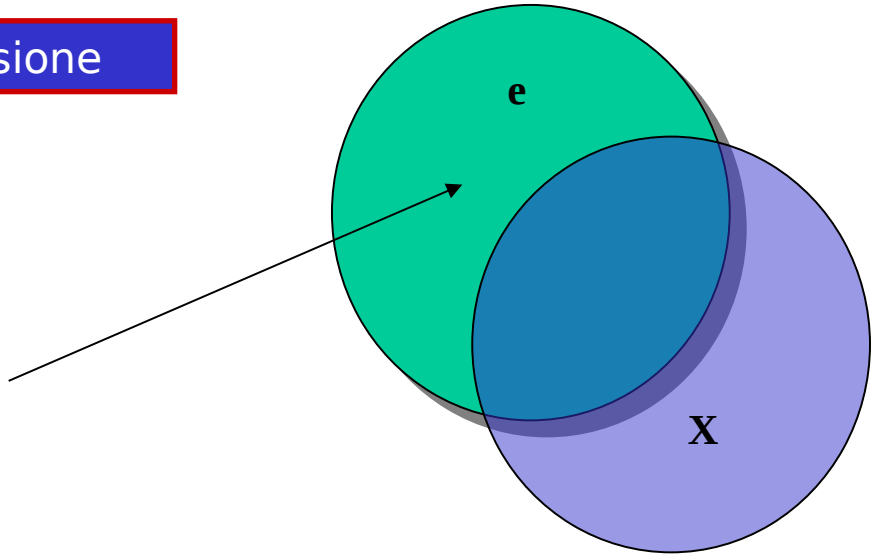
- Rapportando tutto a 1 (standardizzando) otteniamo la percentuale di errore

% di errore di regressione

errore di regressione

$$\frac{s_e^2}{s_y^2} = \frac{\sum [y_i - (a + b_{yx} x_i)]^2}{\sum (y_i - M_y)^2}$$

massimo errore totale

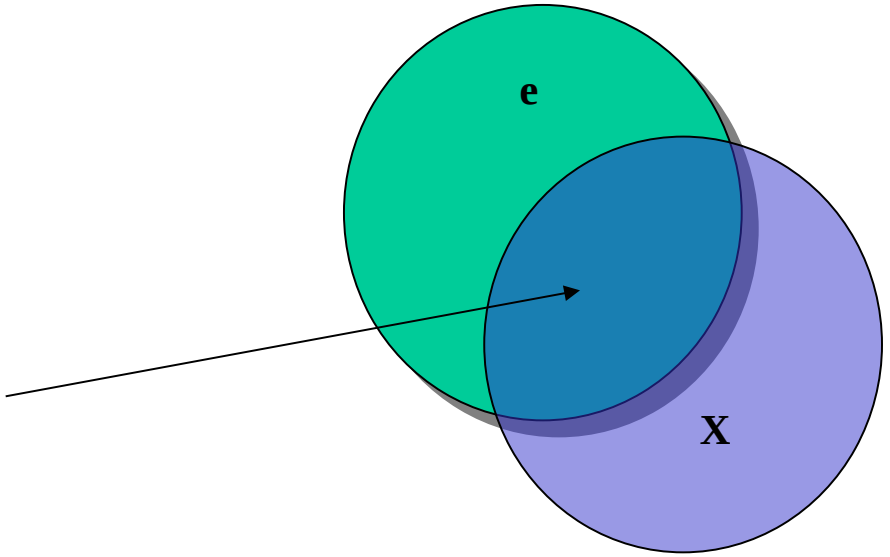


Riduzione dell'errore

- Potremo dire che l'errore di predizione si è ridotto, al confronto con l'errore che facevamo senza usare la regressione (usando cioè la media di Y come valore predetto)

% di riduzione

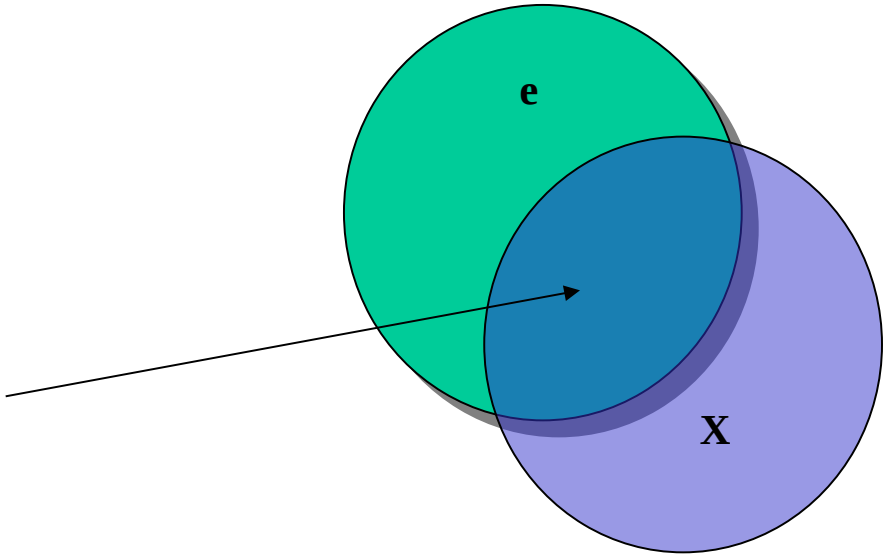
$$\frac{s_y^2 - s_e^2}{s_y^2} = R^2$$



Varianza spiegata

- Quella parte della varianza che non è di errore, sarà varianza che possiamo spiegare (predire) grazie all'uso della regressione

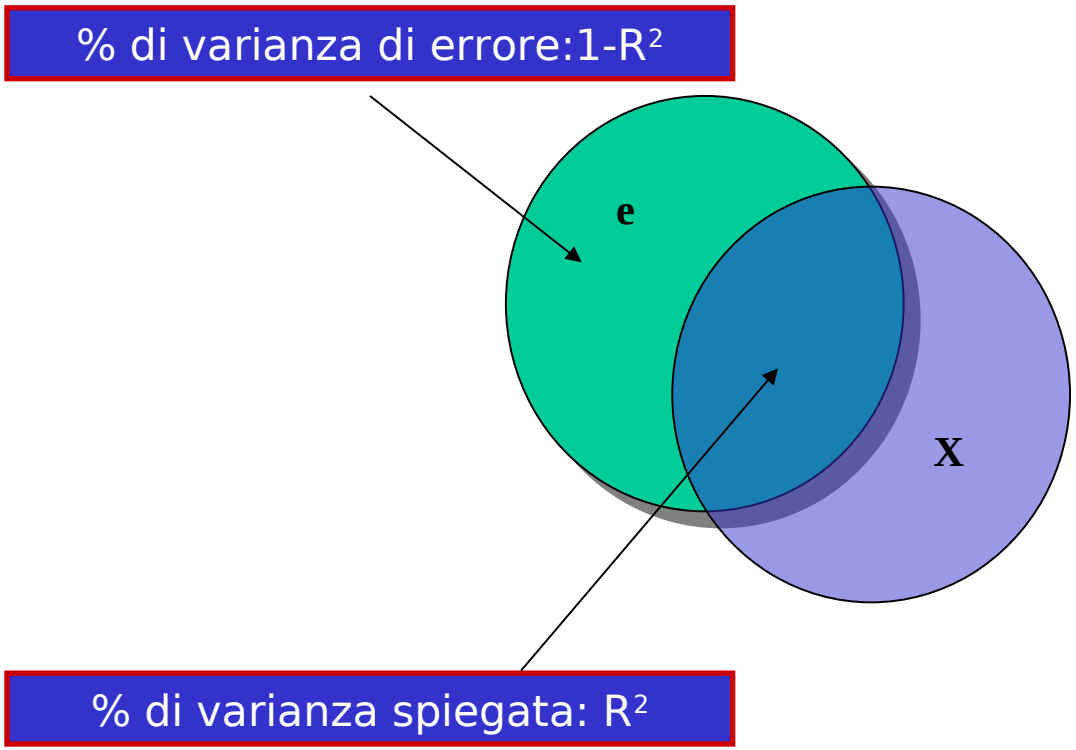
Chiamiamo tale % di varianza:
R^2



$$\frac{s_y^2 - s_e^2}{s_y^2} = R^2$$

Decomposizione della Varianza

- Dunque la varianza di errore iniziale, cioè la varianza della y, dopo la regressione si può decomporre in



Predizione e Spiegazione

- All’aumentare della correlazione, aumenta la nostra capacità di predire il cambiamento di una variabile usando la variabilità dell’altra

- All’aumentare della correlazione, aumenta la nostra capacità di spiegare la variabilità una variabile usando la variabilità dell’altra

- In sostanza, predire una variabile mediante un’altra ci consente di spiegarne la variabilità. Migliore è l’adeguatezza della nostra predizione, migliore è la capacità esplicativa

Spiegazione e Causalità

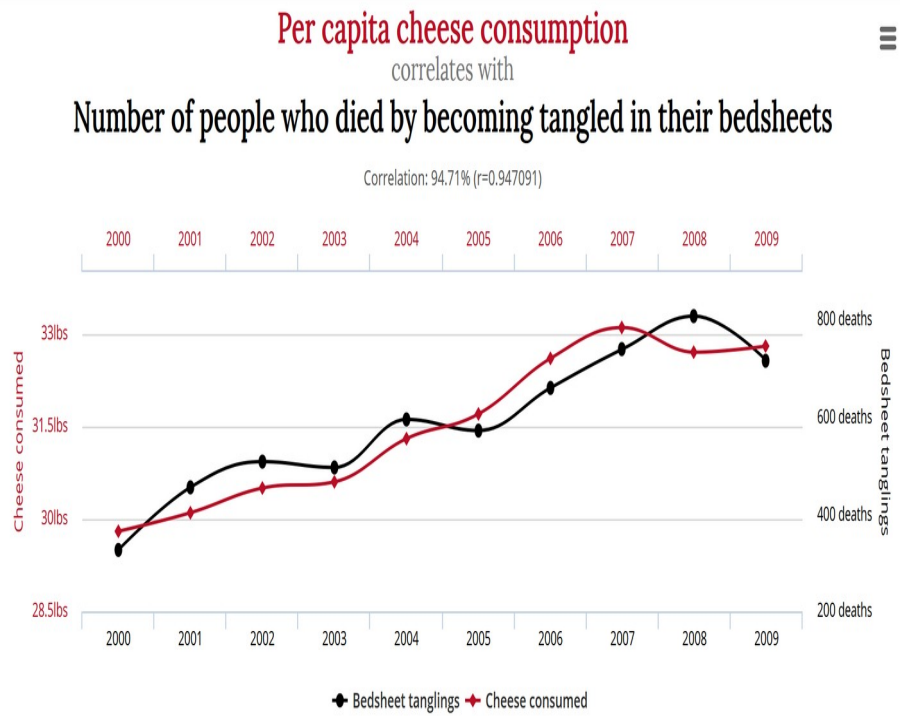
- Spiegare la variabilità statistica di una variabile non significa spiegare le cause del fenomeno che la variabile misura

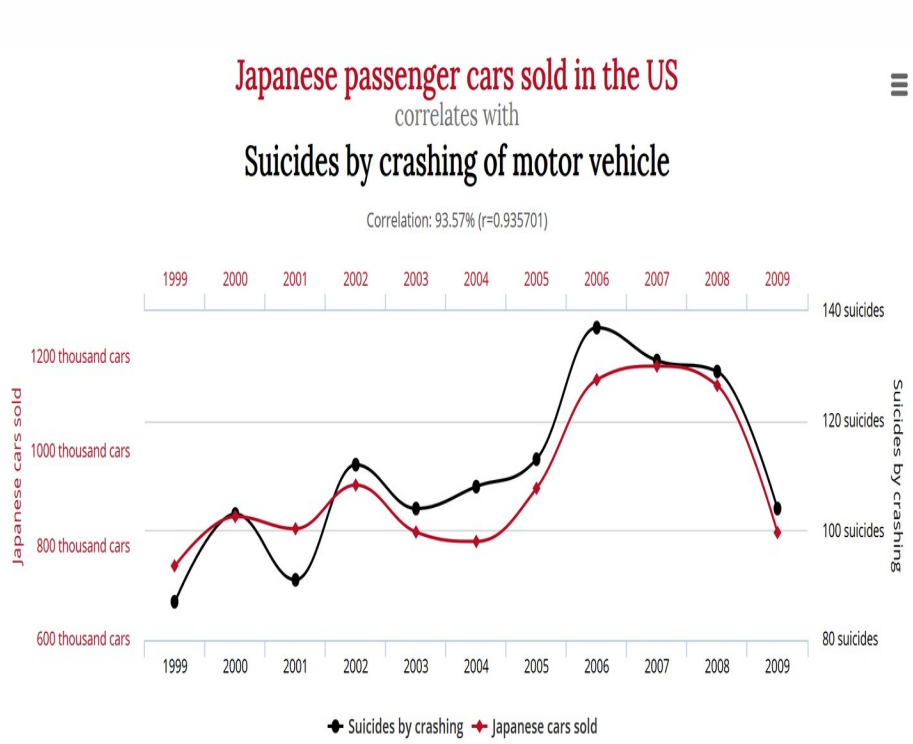
- La spiegazione statistica dipende dalla bontà del modello statistico e dall'associazione fra variabili

- La spiegazione **causale** non dipende dalla spiegazione statistica, ma dalla fondatezza teorica del modello utilizzato

Correlazione non è Causalità

- “*correlation is not causation*”

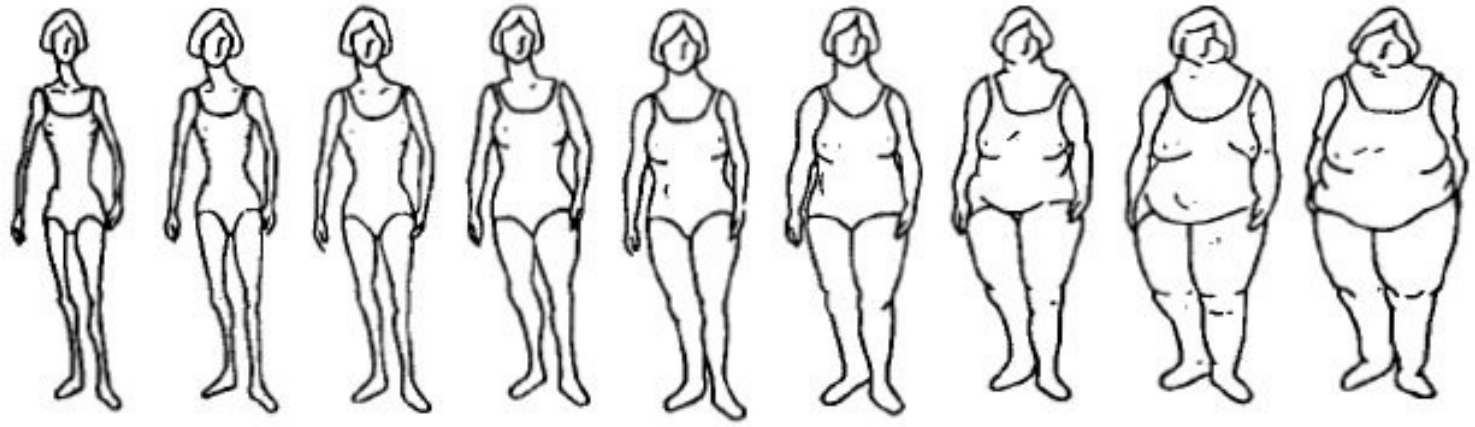




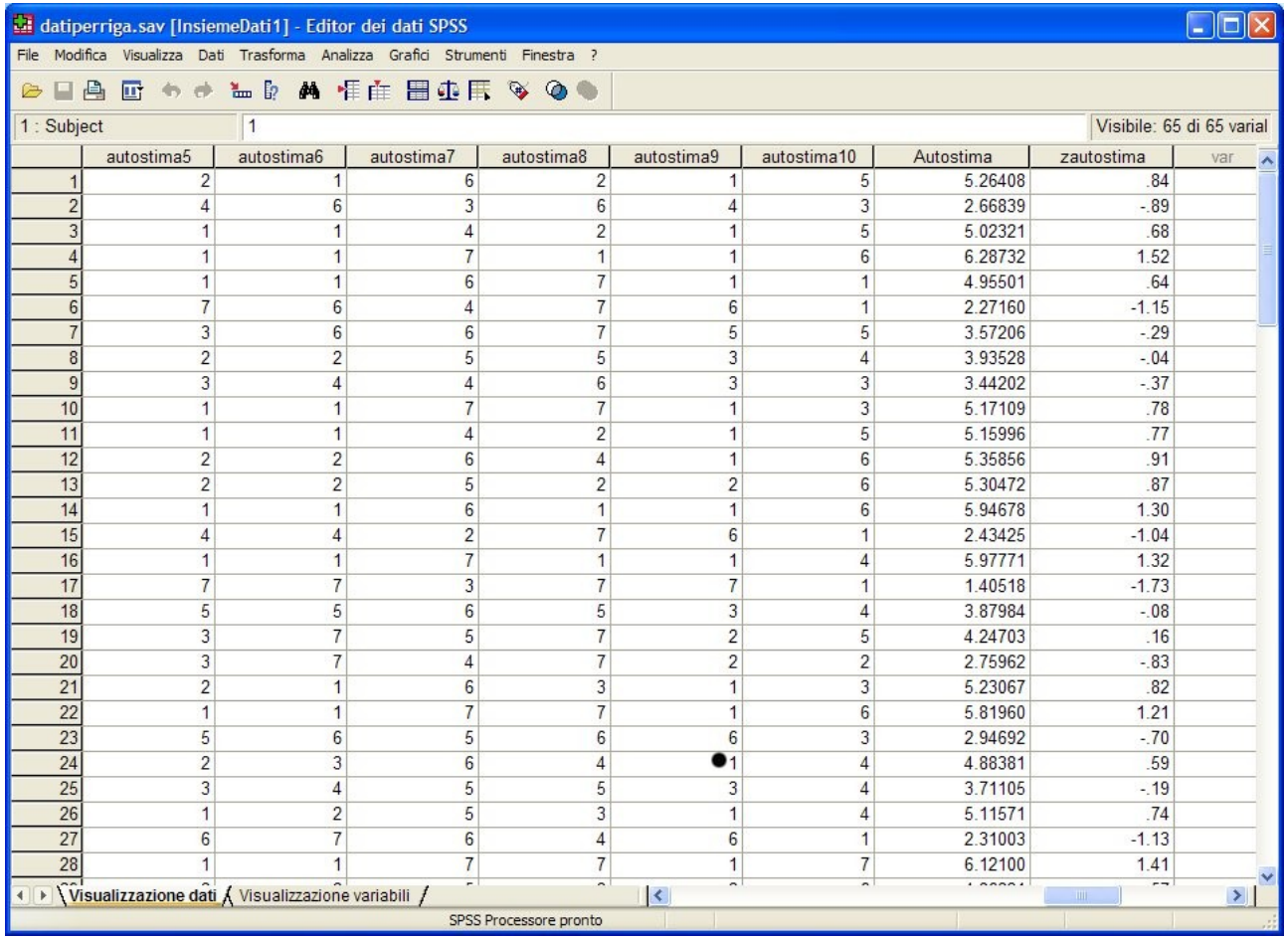
Esempio di regressione semplice

- In ricerca sull'anoressia sono state misurate su un campione di 85 donne la propria "figura reale", la "figura ideale" e l'autostima.

Pictorial Body Image Scale



Regressione con SPSS



	autostima5	autostima6	autostima7	autostima8	autostima9	autostima10	Autostima	zautostima	var
1	2	1	6	2	1	5	5.26408	.84	
2	4	6	3	6	4	3	2.66839	-.89	
3	1	1	4	2	1	5	5.02321	.68	
4	1	1	7	1	1	6	6.28732	1.52	
5	1	1	6	7	1	1	4.95501	.64	
6	7	6	4	7	6	1	2.27160	-1.15	
7	3	6	6	7	5	5	3.57206	-.29	
8	2	2	5	5	3	4	3.93528	-.04	
9	3	4	4	6	3	3	3.44202	-.37	
10	1	1	7	7	1	3	5.17109	.78	
11	1	1	4	2	1	5	5.15996	.77	
12	2	2	6	4	1	6	5.35856	.91	
13	2	2	5	2	2	6	5.30472	.87	
14	1	1	6	1	1	6	5.94678	1.30	
15	4	4	2	7	6	1	2.43425	-1.04	
16	1	1	7	1	1	4	5.97771	1.32	
17	7	7	3	7	7	1	1.40518	-1.73	
18	5	5	6	5	3	4	3.87984	-.08	
19	3	7	5	7	2	5	4.24703	.16	
20	3	7	4	7	2	2	2.75962	-.83	
21	2	1	6	3	1	3	5.23067	.82	
22	1	1	7	7	1	6	5.81960	1.21	
23	5	6	5	6	6	3	2.94692	-.70	
24	2	3	6	4	1	4	4.88381	.59	
25	3	4	5	5	3	4	3.71105	-.19	
26	1	2	5	3	1	4	5.11571	.74	
27	6	7	6	4	6	1	2.31003	-1.13	
28	1	1	7	7	1	7	6.12100	1.41	

Regressione

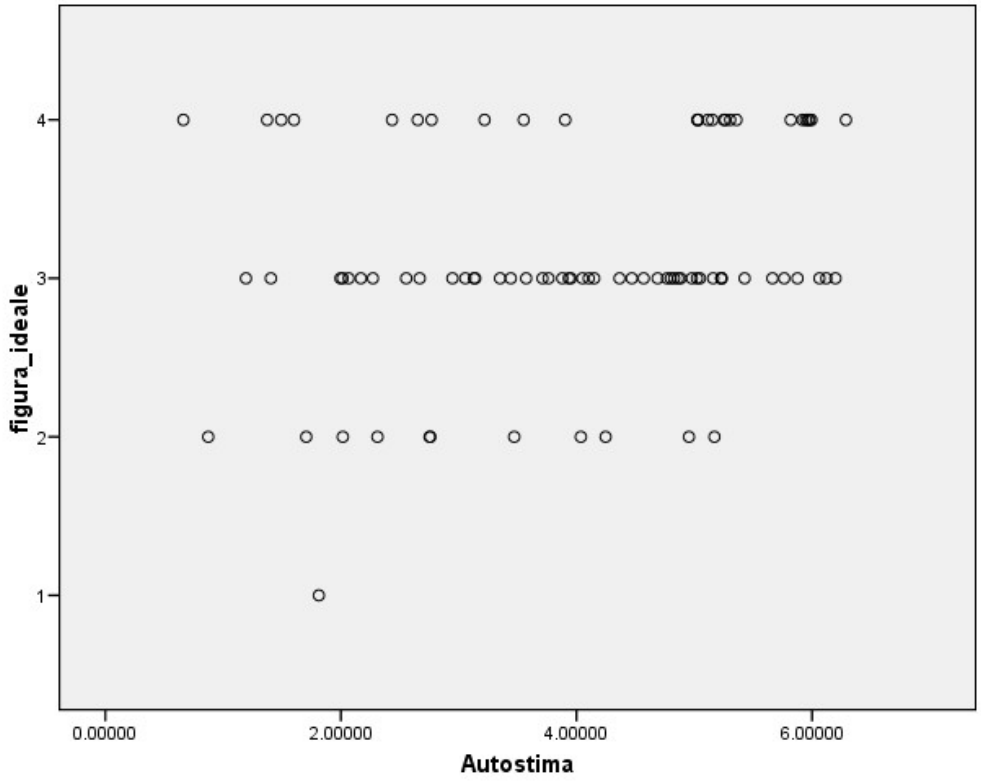
- Ci proponiamo di capire se la propria corporatura ideale (variabile figura_ideale) dipenda (cioè sia influenzata) dalla autostima (media di dieci items su scala da 0 a 6).

Statistiche descrittive

	N	Minimo	Massimo	Media	Deviazione std.
figura_ideale	85	1	4	3.15	.681
Autostima	85	.66172	6.00000	4.0000000	1.50000000
Validi (listwise)	85				

Regressione

- Ci proponiamo di capire se la propria corporatura ideale (variabile figura_ideale) dipenda (cioè sia influenzata) dalla autostima.



Output

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	2.683	.206		13.048	.000
	Autostima	.117	.048	.258	2.437	.017

a. Variabile dipendente: figura_ideale

Output

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	2.683	.206		13.048	.000
	Autostima	.117	.048	.258	2.437	.017

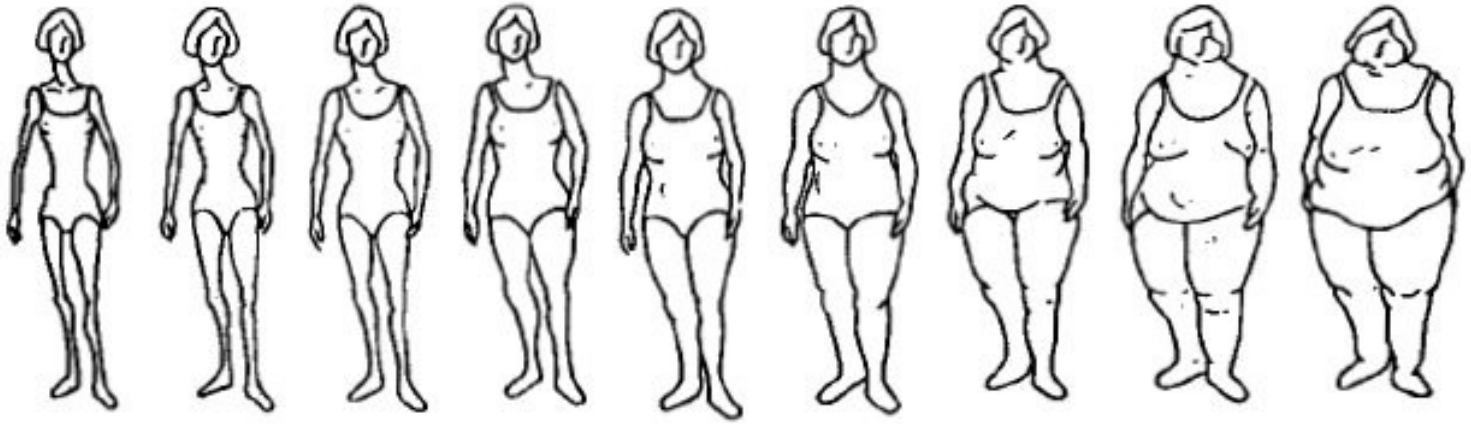
a. Variabile dipendente: figura_ideale

Per ogni punto in più di autostima, la figura ideale aumenta di .117

Per autostima molto bassa (x=0) si preferisce una figura molto magra (2.6)

Cioè

Media attesa per autostima molto bassa



Aumentando l’autostima…

Output

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	2.683	.206		13.048	.000
	Autostima	.117	.048	.258	2.437	.017

a. Variabile dipendente: figura_ideale

In termini di correlazione

Ad una devizione standard sopra la media di autostima, corrisponde un incremento della “figura” di .25 deviazioni standad

Interpretazione

- C’è dunque una relazione tra figura ideale ed autostima, nel senso che per minori livelli di autostima si tende ad una figura ideale più magra

- Quanto è forte questa relazione?

Interpretazione

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima
1	.258[a]	.067	.056	.662

a. Predittori: (Costante), Autostima

Anova[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	2.605	1	2.605	5.938	.017[a]
	Residuo	36.407	83	.439		
	Totale	39.012	84			

a. Predittori: (Costante), Autostima
b. Variabile dipendente: figura_ideale

Corrispondenze

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Errore std.	Beta		
1	(Costante)	2.683	.206		13.048	.000
	Autostima	.117	.048	.258	2.437	.017

a. Variabile dipendente: figura_ideale

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima
1	.258[a]	.067	.056	.662

a. Predittori: (Costante), Autostima

Anova[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	2.605	1	2.605	5.938	.017[a]
	Residuo	36.407	83	.439		
	Totale	39.012	84			

a. Predittori: (Costante), Autostima
b. Variabile dipendente: figura_ideale

Fine della Lezione II