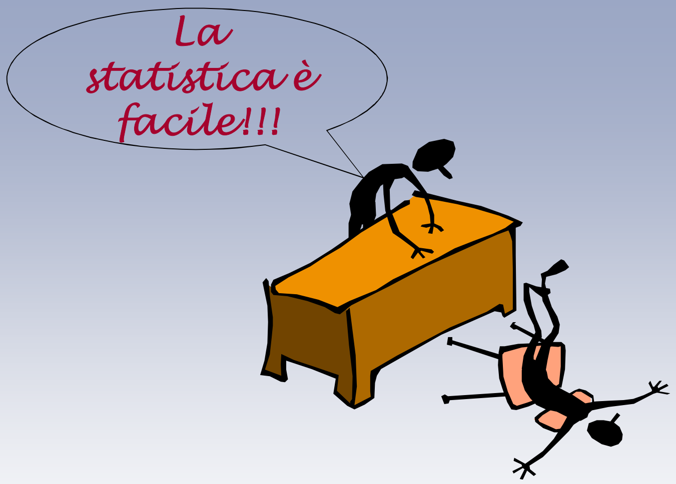
Analisi dei Dati



Regressione bivariata

REGRESSIONE

- Studiare la relazione tra due variabili significa descrivere in che modo una variabile "dipenda" da un'altra:

$$y_i = f(x_i)$$

- Descrizione di come una variabile **X** (variabile INDIPENDENTE = "causa") produce il variare di una variabile **Y** (variabile DIPENDENTE= "effetto")
  - ❖Un particolare tipo di relazione è quella di tipo lineare

$$\mathbf{Y_i = a + bx_i + e}$$

- Concettualmente identica all'ANOVA (e al t-test), ma applicata quando la **X è una misura continua.**

REGRESSIONE: obiettivi

- Ha due obiettivi:
  - misurare il ***grado*** e il ***verso*** dell'influenza della variabile indipendente X sulla variabile dipendente Y (più usato in psicologia)
  - ottenere un'equazione che permetta di ***prevedere*** il valore della variabile dipendente Y, conoscendo solo quello della indipendente X (meno usato in psicologia, e più usato dalle agenzie di assicurazione, o dagli economisti)

Che teoria indaghiamo…

- Se indaghiamo l'equazione $\mathbf{Y_i = a + bx_i + e}$
  - allora pensiamo che il coefficiente che lega la X alla Y (*b*) sia grande a sufficienza.
- Infatti, più è grande il *coefficiente di regressione*, più forte è il legame fra X e Y.
- Indaghiamo quindi una teoria che afferma l'esistenza di una relazione lineare.
  - La forza della relazione è rappresentata dal coefficiente di regressione (che dipende dalla associazione fra X e Y)

Quanto è buona la nostra previsione

- Se X è legata *perfettamente* ad Y, sapendo i valori di X possiamo prevedere quelli di Y (Y^), e osserveremo che la Y^ ricavata dalla X coincide con la Y osservata.

$$Y_i = \hat{Y}_i = a + bx_i$$

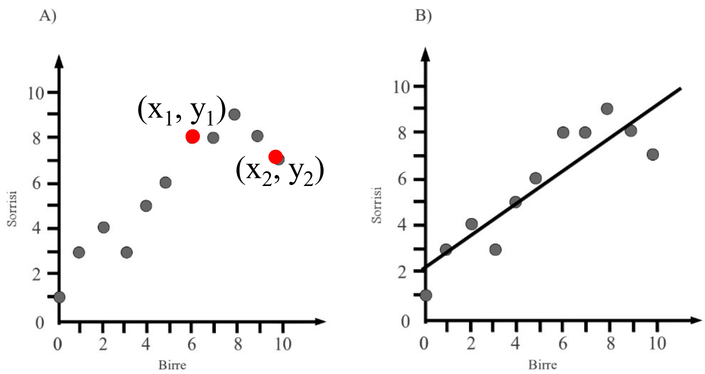
- Per legami deboli o inesistenti, o in genere per relazioni meno che perfette, invece la Y^ prevista sarà distante dalla Y osservata e l'errore *e* sarà più ampio.

$$Y_i \neq \hat{Y}_i = a + bx_i;$$

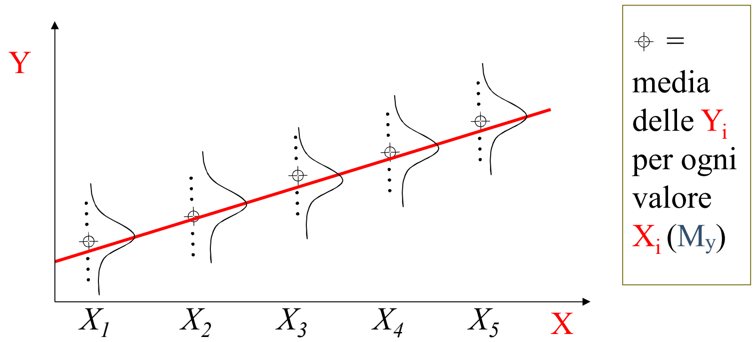
$$\hat{Y}_i - Y_i = e_i$$

REGRESSIONE

Dati osservati nel campione:



- Nella popolazione:



- La linea che unisce tutte le medie M_y è detta CURVA DI REGRESSIONE di Y su X

  ⇨ Le distribuzioni di Y_i per ogni X_i sono normali

  ⇨ Le varianze di Y_i per ogni valore X_i sono omogenee (*omoschedasticità*)

Il modello di regressione definisce dunque che per ogni variazione di X osserveremo una corrispondente variazione di Y.

Più formalmente

Modelli statistici lineari: esprimono una relazione tra i fenomeni osservati di tipo lineare tramite l'equazione di una retta

$$Y = \underbrace{\alpha + \beta X}_{\text{Componente deterministica}} + \varepsilon$$

Variabile casuale

Componente casuale

Componente deterministica

Intercetta, coefficiente, errore

$$Y = \alpha + \beta X + \varepsilon$$

- ✓α = ***intercetta***, punto in cui la retta incontra l'asse delle *Y,* rappresenta il ***valore previsto*** di Y in corrispondenza di X uguale zero;
- ✓β = ***coefficiente di regressione***, inclinazione della retta, parametro della popolazione, rappresenta ***l'incremento previsto* di Y per un *incremento unitario* di X**
- ✓ε = ***errore stocastico o residuo, o errore di previsione***

- Assunzioni teoriche:
  - gli errori hanno media 0 e varianza σ^2
  - gli errori sono ***indipendenti*** tra loro ⇨ $Cov(e_i, e_j) = 0$
  - gli errori sono ***indipendenti*** da X ⇨ $Cov(e, X) = 0$

Calcolo dei coefficienti: Coefficiente di regressione *b*

- Il coefficiente *b*, dipende dalla covarianza *xy*, e si ottiene dividendo quest'ultima per la varianza di *x*

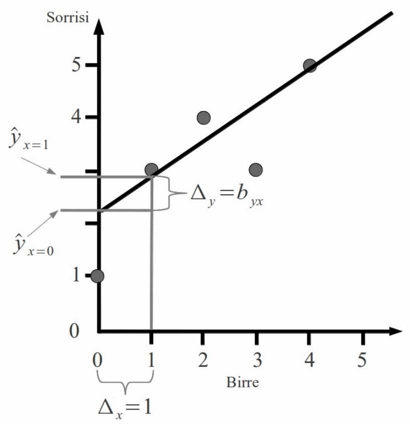
$$b = \frac{Covarianza_{xy}}{Varianza_{x}}$$

$$b = \frac{\mathrm{cov}_{xy}}{s_x^2}$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SQ_{xy}}{SQ_x}$$

- In questo modo, il coefficiente indica la variazione in *y*, per cambiamento unitario di *x*

REGRESSIONE



In questo grafico vediamo il senso geometrico del coefficiente *b*
Δx = 1 → Δy = *b*

Calcolo dei coefficienti: l'intercetta *a*

- Siccome la retta di regressione passa per il punto

$$(\bar{x}, \bar{y})$$

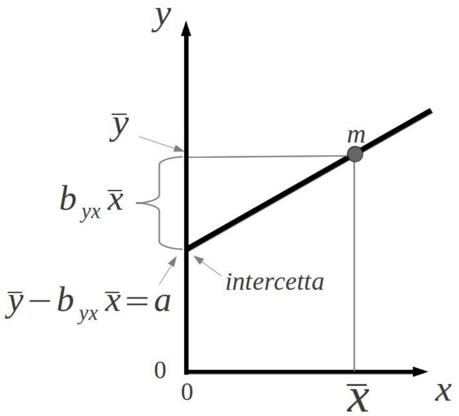
Allora:

$$\bar{y} = a + b\bar{x}$$

Ergo:

$$a = \bar{y} - b\bar{x}$$

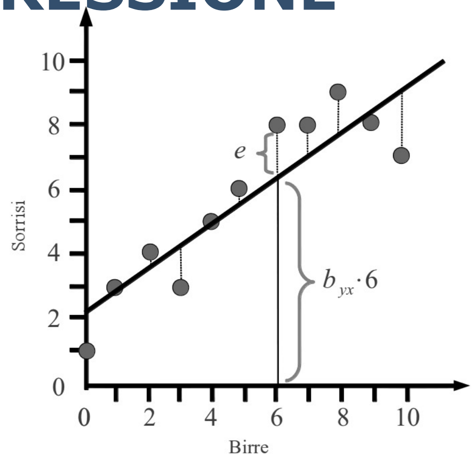
REGRESSIONE



In questo grafico vediamo il senso geometrico dell'intercetta *a*: $x = 0 \rightarrow y = a$

Notiamo che la media di x è collegata alla media di y dall'equazione (la retta passa per forza nel punto medio di x e y che è il "centro" della distribuzione)

REGRESSIONE



In questo grafico vediamo il senso geometrico dell'errore *e*: $y\hat{}_{x6} \neq y_{x6}$; $y_{x6} - y\hat{}_{x6} = e$

Il coefficiente di regressione standardizzato β

- Il coefficiente *b* è dipendente dalle scale di x e y
  - A volte è interpretabile (misure a rapporti, o fisiche) a volte no (misure a intervalli, scale arbitrarie test psicologici).
- La standardizzazione riporta ad una misura facile da interpretare come la deviazione standard

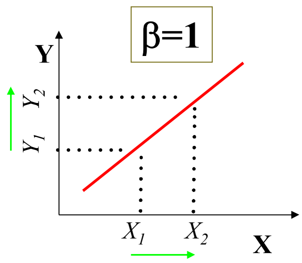
$$\beta = b \frac{\mathrm{cov}_{xy}}{s_x^2} \frac{s_y}{s_x} = \frac{\mathrm{cov}_{xy}}{s_x s_y} = r$$

- Nella regressione bivariata, il coefficiente *b* è identico al coefficiente di correlazione *r* di Pearson
  - Il coefficiente standardizzato si usa per confrontare tra di loro il contributo di ciascun predittore (x) quando ne abbiamo più di uno, quando essi hanno metriche diverse.

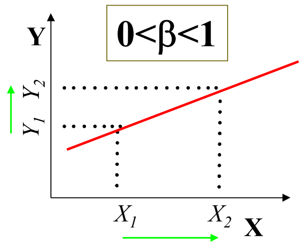
REGRESSIONE

- $\boxed{\beta > 0}$ = all'*aumentare* di X *aumenta* Y

Ad ogni variazione di X corrisponde un'***uguale*** variazione in Y



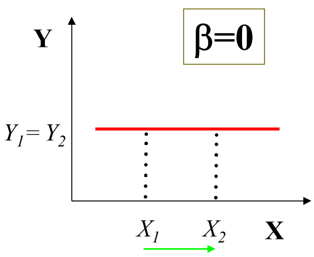
Ad ogni variazione di X corrisponde una variazione ***positiva*** in Y



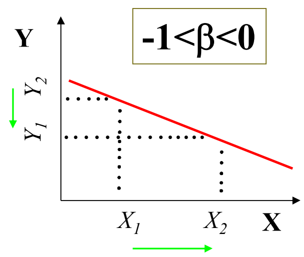
REGRESSIONE

- $\boxed{\beta \leq 0}$ = all'*aumentare* di X *diminuisce* Y, o Y rimane invariata.

Ad ogni variazione di X ***non*** corrisponde ***alcuna*** variazione in Y ⇨ **ASSENZA DI RELAZIONE**

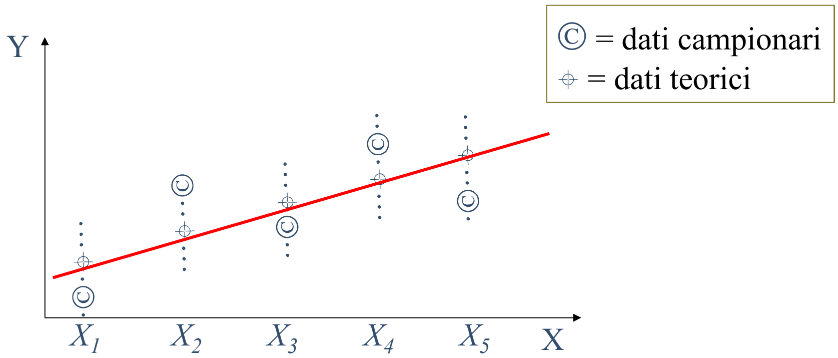


Ad ogni variazione di X corrisponde una variazione ***Negativa*** in Y

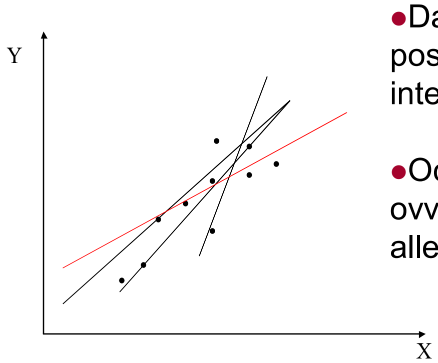


Riassumendo:

- Raccogliamo i dati su un campione, su *y* e su *x*:



REGRESSIONE



- Data una nube di punti si possono tracciare infinite rette interpolanti

- Occorre trovare retta ottimale, ovvero quella più vicina possibile alle osservazioni

⇨ ***Stima*** della retta di regressione: $Y = a + bX$

dove: ***a*** = stima di α

b = stima di β

METODO DEI MINIMI QUADRATI: per trovare la retta che rende ***minima*** la somma degli scarti tra ***valori stimati*** o ***teorici*** (⌖) e ***valori empirici*** o ***osservati*** (©)

$\hat{Y}_i$ = ***Valore predetto*** dal modello lineare

$e_i = Y_i - \hat{Y}_i$ = ***Errore della predizione***, ovvero la porzione della variabile dipendente non spiegata dalla indipendente (***Residuo***); $\sum e_i = \mathbf{zero}$

$\hat{Y}_i - \bar{Y}_i$ = ***Differenza fra la predizione offerta dalla teoria (Y dipende da X), e predizione in assenza di teoria (y_i=M_y)***

⇩

$SQ_{err} = \Sigma\left(Y_i - \hat{Y}_i\right)^2$ = *minima*

REGRESSIONE

- Attraverso il METODO DEI MINIMI QUADRATI si ricavano ***b*** e quindi ***a***:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{SQ_{xy}}{SQ_x}$$

CODEVIANZA: Misura di come ***x*** e ***y*** variano insieme. Dividendo per *n* si ottiene la **COVARIANZA**

DEVIANZA di ***x*** (somma di tutti gli scarti dalla media). Dividendo per *n* si ottiene la **VARIANZA di x**

$$a = \bar{y} - b\bar{x}$$

Esempio

- ***Vogliamo prevedere il numero di incubi (y) in base allo stress (x)***
- ***M_x**=5, **M_y**=4, **n**=6*

Sogg.	x_i	y_i	$(x_i\text{-}M_x)$	$(y_i\text{-}M_y)$	$(x_i\text{-}M_x)(y_i\text{-}M_y)$	$(x_i\text{-}M_x)^2$
1	1	2	-4	-2	8	16
2	3	2	-2	-2	4	4
3	4	4	-1	0	0	1
4	6	4	1	0	0	1
5	7	5	2	1	2	4
6	9	7	4	3	12	16
	30	**24**	**0**	**0**	**26**	**42**

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b = \frac{26}{42} = .62$$

$$y = .90 + .62x$$

$$a = \bar{y} - b\bar{x}$$

$$a = 4 - .62(5) = .90$$

Esempio

Per disegnare la retta occorrono due punti:

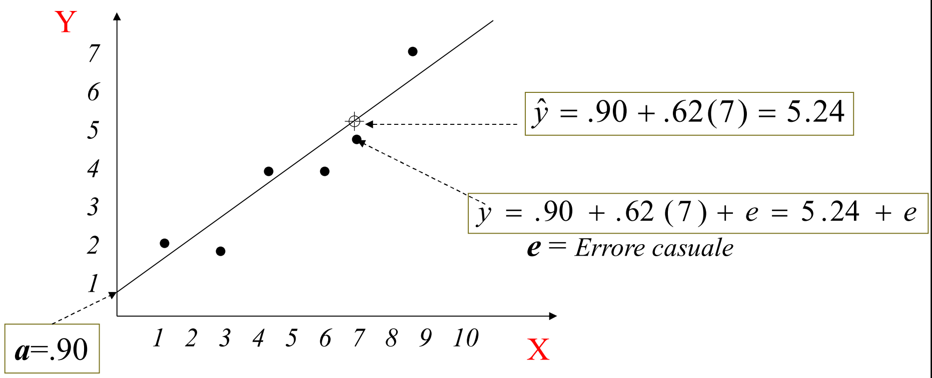
1. l'intercetta ***a*** =.90 ⇨ P_1(.90; 0)

2. un qualsiasi altro punto con coordinate $(x_i;y_i)$, ad esempio: se x=7 sostituendo nella retta di regressione ottengo ⇨ y=5.24 (valore stimato) ⇨P_2(7; 5.24)

Con P_1 e P_2 è possibile tracciare la retta che descrive, essendo ***b positivo***, come i valori di Y crescono al crescere di X

Esempio

Relazione lineare ***positiva***: ↑X⇨Y↑



Fit del modello: Stimare i punteggi secondo una teoria

- Se non conosciamo ogni punteggio ***y*** di ogni soggetto, la migliore stima che di esso abbiamo è il punteggio medio in **y** di tutto un campione:

$$y_i = \bar{y} + e$$

- Se noi però supponiamo che il punteggio dipende dal punteggio ***x*** del soggetto, la nostra teoria afferma:

$$y_i = a + bx + e$$

- Se questa teoria è vera, allora la media di ***y*** differisce da ognuno dei valori previsti di ***y*** a causa delle influenze di ***x***

Bontà del modello statistico e significatività della previsione

- Le fonti di variazione dietro la variabilità di *y* sono:

$y_i - \bar{y}$ = differenza tra un'osservazione e la media della distribuzione corrispondente

$y_i - \hat{y}_i$ = discrepanza tra osservazione e valore atteso (errore)

$\hat{y}_i - \bar{y}$ = porzione del valore atteso attribuibile alla relazione lineare tra X e Y

Bontà del modello statistico e significatività della previsione

- Le quantità che esprimono il peso di queste fonti di variazione sono:

$y_i - \bar{y} \longrightarrow SQ_y = SQ_{tot} = \sum (y_i - \bar{y})^2$

$y_i - \hat{y}_i \longrightarrow SQ_e = \sum \left(y_i - \hat{y}_i \right)^2$

$\hat{y}_i - \bar{y} \longrightarrow SQ_{reg} = \sum \left(\hat{y}_i - \bar{y} \right)^2$

Stimare i punteggi secondo una teoria e bontà dei modelli

- Per scegliere fra le due teorie ("regressione" vs "solo media"), vediamo quale modello (statistico) delle due teorie si associa ad un errore minore

$$y_i = \bar{y} + e \longrightarrow e = \sum (y_i - \bar{y})^2 = SQ_y$$

$$y_i = a + bx + e \longrightarrow e = \sum (y_i - \hat{y})^2 = SQ_e$$

$$SQ_e > SQ_y;$$

$$SQ_y > SQ_e$$

Riduzione dell'errore come proporzione di varianza spiegata

- È possibile utilizzare le quote di errore del modello di regressione e del modello "solo media" per rappresentare quanta varianza ci consente di spiegare il modello di regressione: R^2.

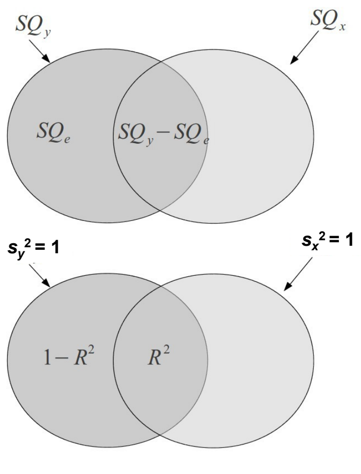
$$R^2 = \frac{SQ_y - SQ_e}{SQ_y} = \frac{SQ_{reg}}{SQ_y}$$

$$1 - R^2 = \frac{SQ_e}{SQ_y}$$

- $1-R^2$ è la proporzione di errore (ciò che il modello NON spiega), e vien detto coefficiente di alienazione

Riduzione dell'errore come proporzione di varianza spiegata

Diagrammi di Venn



Riduzione dell'errore come proporzione di varianza spiegata

$$0 \le R^2 \le 1$$

$R^2 = 1$ Tutte le osservazioni cadono sulla retta di regressione :
$SQ_{tot} = SQ_{reg} \leftarrow SQ_{err} = 0$

$R^2 = 0$ Massima dispersione delle osservazioni attorno alla retta: non vi è associazione fra X e Y :
$SQ_{tot} = SQ_{err} \leftarrow SQ_{reg} = 0$

Esempi

- $R^2 = 1 \Rightarrow (1-r^2) = 0$

Il 100% della varianza di Y è spiegata da X
⇨ previsione perfetta

- $R^2 = 0 \Rightarrow (1-r^2) = 1$

La varianza di Y **non** è spiegata da X
⇨ nessuna previsione

- $R^2 = .72 \Rightarrow (1-r^2) = .28$

Il 72% della varianza di Y è spiegato da X
⇨ il 28% non è predicibile con la relazione lineare

Esempi

- R^2 = .25 ⇨ $(1-r^2)$ = .75

Soltanto il 25% della varianza di Y è spiegata da X, il 75% resta da spiegare

⇨ il 75% non è predicibile attraverso la relazione lineare

- R^2 = .50 ⇨ $(1-r^2)$ = .50

Il 50% della varianza di Y è spiegata da X, l'altro 50% resta da spiegare

⇨ soltanto il 50% è predicibile attraverso la relazione lineare

Torniamo alla significatività della previsione

SQ_y = DEVIANZA TOTALE

SQ_{reg} = DEVIANZA SPIEGATA dalla regressione

$$\sum\left(y_i-\bar{y}\right)^2=\sum\left(\hat{y}_i-\bar{y}\right)^2+\sum\left(y_i-\hat{y}_i\right)^2$$

SQ_e = DEVIANZA NON SPIEGATA o RESIDUA (somma di e)

La somma dei quadrati ***totale*** (SQ_y) è data da una componente di ***errore*** (SQ_e) e da una componente ***spiegata dalla regressione*** (SQ_{reg})

Significatività della previsione: **Scomposizione Devianza (Somma dei quadrati)**

$$\sum\left(y_i-\bar{y}\right)^2=\sum\left(\hat{y}_i-\bar{y}\right)^2+\sum\left(y_i-\hat{y}_i\right)^2$$

$$\Downarrow$$

$$SQ_y=SQ_{reg}+SQ_e$$

In termini di R^2, possiamo scrivere: $\Downarrow$

$$1=R^2+(1-R^2)$$

Possiamo testare se R^2 è significativamente più grande di 1- R^2

Significatività della previsione: *F*

- Per verificare se la previsione è significativa R^2 deve essere maggiore della varianza di errore
- Bisogna quindi dividere R^2 per il suo errore standard.

$$\sigma^2_{R^2}=\frac{1-R^2}{N-2}$$

- Il rapporto R^2/σ^2 è una F di Fisher
  - Gdl: *k* (numero delle variabili *x*), *N*-*k*-1

$$F_{(k,N-k-1)}=\frac{R^2}{\sigma_{R^2}^{\ 2}}=\frac{R^2}{1-R^2}\cdot N-2$$

Significatività della previsione: *F*

Un modo alternativo di rappresentare la stessa *F* è:

$$F = \frac{Var_{reg}}{Var_e} = \frac{\frac{SQ_{reg}}{k}}{\frac{SQ_e}{N-k-1}}$$

Esempio di R^2 e di *F*

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima
1	,946[a]	,894	,868	,69007

a. Predittori: (Costante), x

Anova[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	16,095	1	16,095	33,800	,004[a]
	Residuo	1,905	4	,476		
	Totale	18,000	5			

a. Predittori: (Costante), x

b. Variabile dipendente: y

Provate a calcolare la F a partire da R^2

Significatività coefficiente di regressione

- La *F* valuta se R^2 è diverso da 0
- R^2 valuta la precisione della regressione (tutto ciò che non è errore), e quindi la precisione di previsione offerta dal coefficiente di regressione *b* *(β, standardizzato)*.
- La significatività di β si ottiene dividendo il coefficiente per il suo errore standard:

$$\sigma_{\beta}^{2} = \sqrt{\frac{1-R^2}{N-2}}$$

$$t = \frac{\beta \cdot \sqrt{N-2}}{\sqrt{1-R^2}}$$

- La *t* risultante si distribuisce per *N-2 gdl*
- Per il coefficiente *b* (non standardizzato):

$$\sigma_{b}^{2} = \sqrt{\frac{1-R^2}{N-2}} \cdot \frac{s_y}{s_x}$$

Esempio e riassunto: *b* e β, significatività

- Il coefficiente non standardizzato è meno utilizzato in psicologia
  - Perché spesso abbiamo scale con unità di misura arbitrarie
- Il coefficiente standardizzato è di più immediata interpretazione qualora non si dia grande importanza all'unità di misura originaria

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	,905	,602		1,502	,207
	STRESS	,619	,106	,946	5,814	,004

a. Dependent Variable: INCUBI

- E' fondamentale verificare l'ipotesi nulla che il coefficiente di regressione sia uguale a zero.
  - A questo scopo si divide *b* per il sue errore standard, e si ottiene una *t*
- **Esercizio: calcolare la *t* usando β**

Intervallo di confidenza

- È utile comprendere l'intervallo di confidenza del coefficiente *b*
- Informa sull'ambito di variazione (probabilistico) del coefficiente *b* nella popolazione

$$b \pm t_{\alpha/2}\sigma_b \, ; \beta \pm t_{\alpha/2}\sigma_\beta$$

- Ci da un'idea quindi di ciò che potremmo trovare in successivi campionamenti e ricerche

- Dalla tabella precedente:

$$b = .62; \sigma_b = .11; t_{(gdl=4).025} = 2.78$$

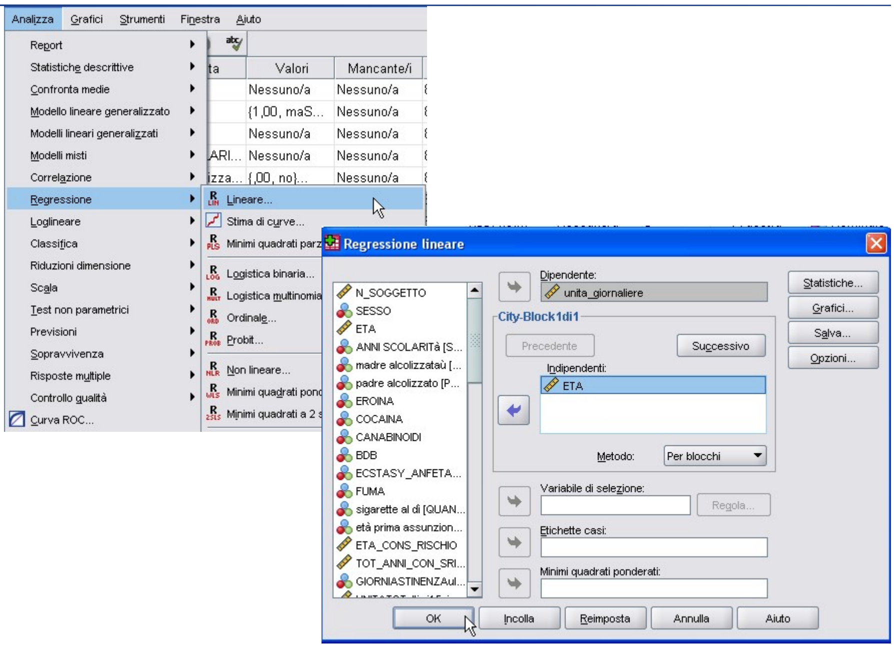
$$.31 \le b \le .93$$

Interpretazione

L'interpretazione si basa su:

- Significatività della regressione
- Forza della spiegazione (R^2).
- Direzione e grandezza del coefficiente di regressione.
  - Significatività del coefficiente di regressione (statistica *t*)
  - Il coefficiente di regressione può essere non standardizzato (cambiamento in y in seguito a cambiamento unitario in x)
  - standardizzato (cambiamento in y in deviazioni standard di y, al cambiamento di una deviazione standard di x)
  - Il C.I. del coefficiente fornisce suggerimenti sulla replicabilità dell'effetto

Con SPSS



Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima
1	,053ª	,003	-,006	10,11782

a. Predittori: (Costante), ETA

Anova[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	31,037	1	31,037	,303	,583ª
	Residuo	11055,982	108	102,370		
	Totale	11087,019	109			

a. Predittori: (Costante), ETA

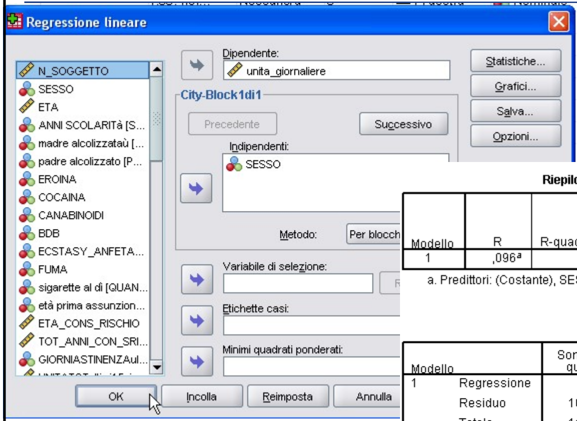
b. Variabile dipendente: unita_giornaliere

Coefficienti[a]

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.
		B	Deviazione standard Errore	Beta		
1	(Costante)	15,802	4,521		3,495	,001
	ETA	-,055	,100	-,053	-,551	,583

a. Variabile dipendente: unita_giornaliere

Il sesso predice i cicchetti?



Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima
1	,096ª	,009	,000	10,08508

a. Predittori: (Costante), SESSO

Anova[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	102,470	1	102,470	1,007	,318ª
	Residuo	10984,548	108	101,709		
	Totale	11087,019	109			

a. Predittori: (Costante), SESSO

b. Variabile dipendente: unita_giornaliere

Report

unita_giornaliere

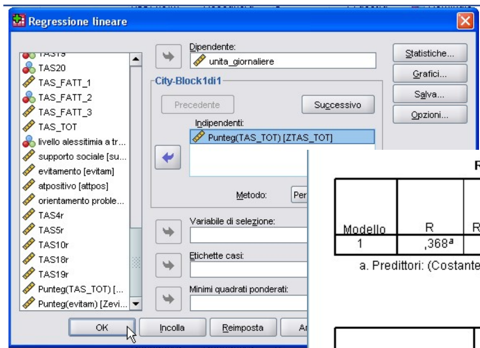
SESSO	Media	N	Deviazione std.
maSCHIO	12,8333	84	9,49213
FEMMINA	15,1051	26	11,84264
Totale	13,3703	110	10,08542

Coefficienti[a]

Modello		Coefficienti non standardizzati B	Coefficienti non standardizzati Deviazione standard Errore	Coefficienti standardizzati Beta	t	Sig.
1	(Costante)	12,833	1,100		11,663	,000
	SESSO	2,272	2,263	,096	1,004	,318

a. Variabile dipendente: unita_giornaliere

Alessetimia predice i cicchetti?



Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima
1	,368ª	,135	,127	9,42266

a. Predittori: (Costante), Punteg(TAS_TOT)

Anova[b]

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	1498,075	1	1498,075	16,873	,000ª
	Residuo	9588,944	108	88,787		
	Totale	11087,019	109			

a. Predittori: (Costante), Punteg(TAS_TOT)

b. Variabile dipendente: unita_giornaliere

Quasi 4 birre in più per ogni deviazione standard di variazione in alessitimia

Coefficienti[a]

Modello		Coefficienti non standardizzati B	Coefficienti non standardizzati Deviazione standard Errore	Coefficienti standardizzati Beta	t	Sig.
1	(Costante)	13,370	,898		14,882	,000
	Punteg(TAS_TOT)	3,707	,903	,368	4,108	,000

a. Variabile dipendente: unita_giornaliere