Lezione IX:

Regressione e Correlazione

Cattedra di Biostatistica – Dipartimento di Scienze Biomediche, Università degli Studi "G. d'Annunzio" di Chieti – Pescara

Prof. Enzo Ballone

RELAZIONE TRA DUE VARIABILI QUANTITATIVE

- Quando si considerano due o più caratteri (variabili) si possono esaminare anche il tipo e l'intensità delle relazioni che sussistono tra loro.
- Nel caso in cui per ogni individuo si rilevino congiuntamente due variabili quantitative, è possibile verificare se esse variano simultaneamente e quale relazione "matematica" sussista tra queste variabili.

RELAZIONE TRA DUE VARIABILI QUANTITATIVE

- Si ricorre all'analisi della regressione e a quella della correlazione:
- **analisi della regressione**: per sviluppare un modello statistico che possa essere usato per prevedere i valori di una variabile, detta dipendente o più raramente predetta ed individuata come l'effetto, sulla base dei valori dell'altra variabile, detta indipendente o esplicativa, individuata come la causa.
- **analisi della correlazione**: per misurare l'intensità dell'associazione tra due variabili quantitative, di norma non legate direttamente da causa-effetto, facilmente mediate da almeno una terza variabile, ma che comunque variano congiuntamente.

RELAZIONE TRA DUE VARIABILI QUANTITATIVE

- Quando per ciascuna unità di un campione o di una popolazione si rilevano due caratteristiche, si ha una **distribuzione doppia** e i dati possono essere riportati in forma tabellare:

unità	carattere X	carattere Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
…	…	…
n	x_n	y_n

- Se il numero di dati è ridotto, la distribuzione doppia può riguardare una tabella che riporta tutte le variabili relative ad ogni unità o individuo misurato.

RELAZIONE TRA DUE VARIABILI QUANTITATIVE

- Se il numero di dati è grande, si ricorre ad una sintesi tabellare chiamata distribuzione doppia di frequenze in cui si suddividono, *eventualmente*, le unità del collettivo in classi Xi e Yj per i due caratteri e si contano le unità che hanno contestualmente entrambe le modalità (n_{ij}):

	Y_1	Y_2	…	Y_j	…	Y_k	Tot.
X_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1k}	n_{1*}
X_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2k}	n_{2*}
…	…	…	…	…	…	…	…
X_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ik}	n_{i*}
…	…	…	…	…	…	…	…
X_h	n_{h1}	n_{h2}	…	n_{hj}	…	n_{hk}	n_{h*}
Tot.	N_{*1}	n_{*2}	…	n_{*j}	…	n_{*k}	n

RELAZIONE TRA DUE VARIABILI QUANTITATIVE

- I totali delle righe e delle colonne rappresentano due distribuzioni semplici e sono dette **distribuzioni marginali** della distribuzione doppia.
- Le frequenze riportate in una colonna o in una riga sono dette **distribuzioni parziali** della distribuzione doppia.
- Una distribuzione doppia può essere rappresentata graficamente con :
- **diagrammi di dispersione** : si riportano le singole coppie di misure osservate considerando ogni coppia della distribuzione come coordinate cartesiane di un punto del piano; si ottiene in tal modo una **nuvola di punti**, che descrive in modo visivo la relazione tra le due variabili

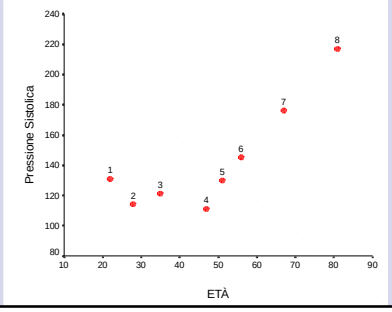
Esempio 1:

- In Tab. sono riportati i valori assunti dai due caratteri quantitativi età (ETA') e pressione sistolica (PAS) misurati in un campione di 8 soggetti:

soggetto	ETA' (anni)	PAS (mm Hg)
1	22	131
2	28	114
3	35	121
4	47	111
5	51	130
6	56	145
7	67	176
8	81	217

Esempio 1:

- Diagramma di Dispersione (a Scatter)



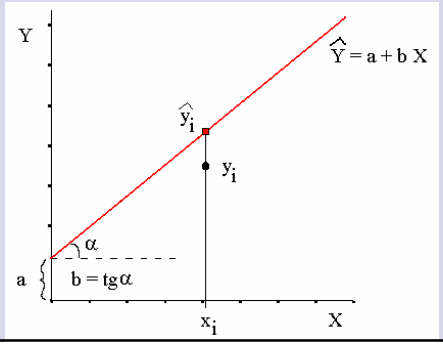
Domande:

- di quanto varia la pressione sistolica all'aumentare dell'età ?
- la relazione tra le due variabili è tendenzialmente lineare?

Regressione lineare semplice

- La “forma” di relazione matematica più semplice tra due variabili è la regressione lineare semplice, rappresentata dalla retta di regressione:
- $\hat{Y} = a + b \times X$
- dove :
- $\hat{Y}$ valore stimato di Y attraverso il modello regressivo
- X valore empirico di X
- a intercetta della retta di regressione
- b coefficiente di regressione (= coeff. angolare della retta)

Regressione lineare semplice



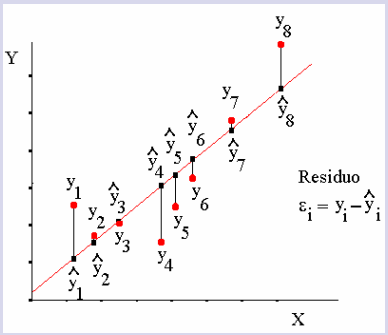
Regressione lineare semplice

- Per stimare la retta che meglio approssima la distribuzione dei punti, si può partire considerando che ogni punto osservato Yi si discosta dalla retta di una certa quantità ei detta errore o **residuo**

$$Y_i = a + b \times X_i + \varepsilon_i$$

- Ogni valore ε_i può essere positivo o negativo:
  - positivo quando il punto Y sperimentale è sopra la retta
  - negativo quando il punto Y sperimentale è sotto la retta

Regressione lineare semplice



Metodo dei minimi quadrati

- la retta migliore per rappresentare la distribuzione dei punti è quella che **minimizza** la somma:
- $\sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
- Secondo il principio dei minimi quadrati si stimano matematicamente a e b:
- b = $\frac{CODEV(X,Y)}{DEV(X)}$ e a = $\bar{y} - b\cdot\bar{x}$
- dove:
- CODEV(X,Y) = Codevianza di X e Y = $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- DEV(X) = Devianza di X = $\sum_{i=1}^{n}(x_i - \bar{x})^2$
- DEV(Y) = Devianza di Y = $\sum_{i=1}^{n}(y_i - \bar{y})^2$

Metodo dei minimi quadrati

- Esempio 1:

n°	ETA' (X)	PAS (Y)	$X-\bar{x}$	$Y-\bar{y}$	$(X-\bar{x})^2$	$(Y-\bar{y})^2$	$(X-\bar{x})(Y-\bar{y})$
1	22	131	-26.4	-12.1	696.96	146.41	+319.44
2	28	114	-20.4	-29.1	416.16	846.81	+593.64
3	35	121	-13.4	-22.1	179.56	488.41	+296.14
4	47	111	-1.4	-32.1	1.96	1030.41	+44.94
5	51	130	+2.6	-13.1	6.76	172.61	-43.06
6	56	145	+7.6	+1.9	57.76	3.61	+14.44
7	67	176	+18.6	+32.9	345.96	1082.41	+611.94
8	81	217	+32.6	+73.9	1062.76	5461.21	+2409.14
	$\bar{x}$=48.4	$\bar{y}$=143.1	0	0	DEV(X) 2767.88	DEV(Y) 9230.88	CODEV(X,Y) 4255.62

Metodo dei minimi quadrati

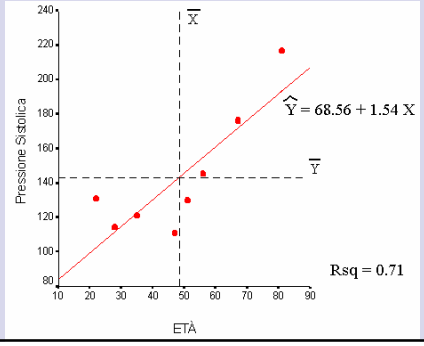
- Si ottiene:
- coeff. di regressione

$$b = \frac{4255.62}{2767.88} = 1.54$$

- intercetta

$$a = 143.1 - 1.54*48.4 = 68.56$$

Metodo dei minimi quadrati



Metodo dei minimi quadrati

- Supposto “accettabile” il modello regressivo lineare, affrontiamo le seguenti domande:
- di quanto aumenta *mediamente* la pressione sistolica all’aumentare di un anno di eta ?
- che valore ha la pressione alla nascita ?
- Interpretando i valori dei coefficienti della retta di regressione si può dire:
- l’aumento medio della pressione è di circa **b=1.5 mmHg** per l’aumento di un anno di età.

Metodo dei minimi quadrati

- **Il coeff. di regressione esprime di quanti varia mediamente la variabile dipendente al variare di una unità della variabile indipendente.**
- alla nascita il valore della pressione *sarebbe* (!) di a**=68.56 mmHg**, ***ma*** questa è una indicazione teorica perché non è possibile stimare il valore della pressione arteriosa per età fuori del range considerato (2281 aa).
- **L'intercetta è quel valore che assume la variabile dipendente quando quella indipendente è uguale a 0.**

Metodo dei minimi quadrati

- Esempio 2:
  - X = Consumo pro-capite di tabacco per sigarette (kg/anno),
  - Y = Quoziente di mortalità per tumore maligno della laringe, trachea, bronchi e polmoni (x 100.000 abitanti)

Anni	X	Y
1985	0.281	5.05
1986	0.417	5.07
1987	0.485	5.81
1988	0.604	6.50
1989	0.648	7.16
1990	0.657	8.38
1991	0.660	8.14
1992	0.719	8.05
1993	0.761	8.56
1994	0.790	9.00

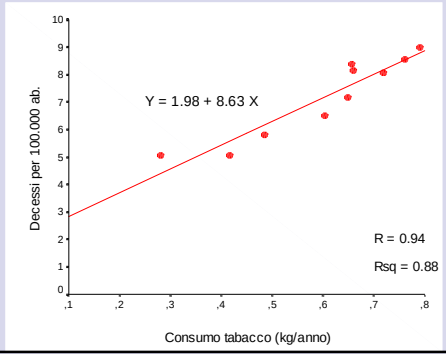
Metodo dei minimi quadrati

b = 8.63 a = 1.98

Y = 1.98 + 8.63 X

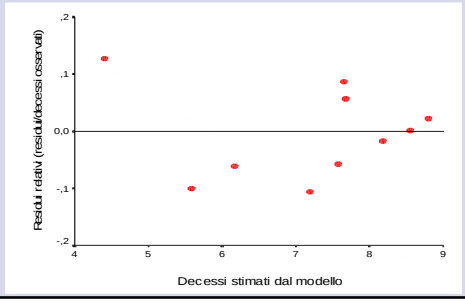
- Qualora il consumo annuo di tabacco pro-capite aumenti di 1 kg si avrà, mediamente, un aumento di circa 9/100.000 della mortalità nella popolazione analizzata.

Metodo dei minimi quadrati



Metodo dei minimi quadrati

- Bontà del modello: ANALISI DEI RESIDUI



Metodo dei minimi quadrati

- Esempio 3:
  - Età e Statura di 10 ragazzi

Ragazzo	X = Età (anni)	Y = Statura (cm)
1	6	115
2	6	120
3	7	122
4	8	130
5	8	128
6	9	134
7	10	136
8	10	140
9	11	145
10	12	151

Metodo dei minimi quadrati

- **b = 5.47** : un aumento di 1 anno di età comporta in media un aumento di circa 5.5 cm di altezza.
- OSSERVAZIONE GENERALE
- Si può studiare anche la dipendenza (sempre ***in media***) della variabile X dalla Y; in tal caso si ottiene la retta di regressione di **Y su X** con coefficienti:
- b' = $\frac{CODEV(X,Y)}{DEV(Y)}$ e a = $\bar{x} - b' \cdot \bar{y}$,

Metodo dei minimi quadrati

- Esempio 4:
  - Studio della relazione tra Capacità Vitale CV (= volume massimo di aria che è possibile contenere nei polmoni dopo un'inspirazione profonda) di un campione di fumatori rispetto al numero di sigarette fumate giornalmente dagli stessi.

Metodo dei minimi quadrati

Esempio 4

Sogg.	N° Sigarette (X)	CV (l aria) (Y)	$(X-\bar{x})^2$	$(Y-\bar{y})^2$	$(X-\bar{x})(Y-\bar{y})$
1	2	6.5	78.77	3.80	-17.31
2	4	6.5	47.27	3.80	-13.41
3	5	6.0	34.52	2.10	-8.52
4	6	5.9	23.77	1.82	-6.58
5	7	5.5	15.02	0.90	-3.68
6	8	5.5	8.27	0.90	-2.73
7	9	5.0	3.52	0.20	-0.84
8	10	4.0	0.77	0.30	0.48
9	11	4.0	0.02	0.30	-0.07
10	12	4.4	1.27	0.02	-0.17
11	13	4.1	4.52	0.20	-0.96
12	14	3.5	9.77	1.10	-3.28
13	15	3.4	17.02	1.32	-4.74
14	16	3.2	26.27	1.82	-6.92
15	20	2.8	83.27	3.06	-15.97
16	22	2.5	123.77	4.20	-22.81
	$\bar{x}$=10.87	$\bar{y}$=4.55	DEV(X)	DEV(Y)	CODEV(X.Y)
			477.75	25.88	- 107.51

Metodo dei minimi quadrati

- Esempio 4
  - **b = - 0.225** : ogni sigaretta in più fumata comporta **in media** una diminuzione di capacità vitale pari a 0.225 l.
  - **a = 6.99** : valore medio di CV per non fumatori.

Metodo dei minimi quadrati

- **Valore predittivo dell'analisi della regressione**
  - La semplice rappresentazione grafica dei valori osservati e della retta di regressione fornisce alcune indicazioni importanti per l'interpretazione delle relazioni esistenti tra le due variabili.
  - Il valore del coefficiente angolare indica quanto aumenta in media la variabile dipendente Y all'aumento di una unità della variabile indipendente X.

Metodo dei minimi quadrati

- Se si cambia la scala della variabile indipendente o predittiva X (per esempio l'altezza misurata in mm o in m e non più in cm) lasciando invariata quella della variabile dipendente o predetta Y, muta proporzionalmente anche il valore del coefficiente angolare b.

Metodo dei minimi quadrati

- Nell'analisi della regressione:
  - è frequente, specialmente negli utilizzi predittivi, il ricorso al tempo come variabile indipendente;
  - viene spesso dimenticato che qualsiasi previsione o stima di Y derivata dalla retta è valida solo entro il campo di variazione della variabile indipendente X;
  - non è dimostrato che la relazione esistente tra le due variabili sia dello stesso tipo anche per valori minori o maggiori di quelli sperimentali rilevati.

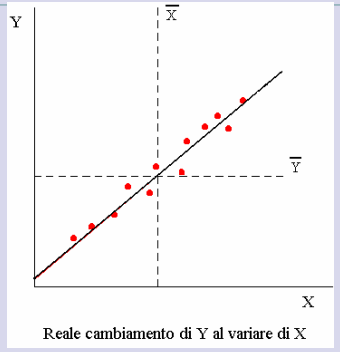
Metodo dei minimi quadrati

- **Significatività della retta di regressione**
  - Con il metodo dei minimi quadrati è sempre possibile ottenere la retta che meglio si adatta ai dati rilevati, indipendentemente dalla dispersione dei punti intorno alla retta.
  - Tuttavia il semplice calcolo della retta non è sufficienti ai fini dell'analisi statistica.

Metodo dei minimi quadrati

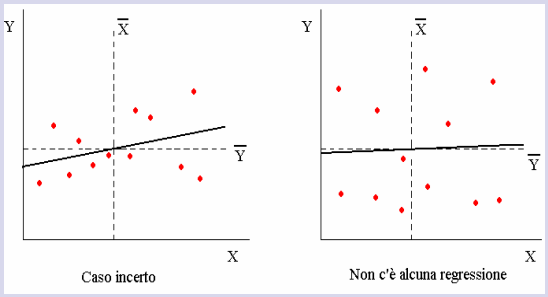
- La retta potrebbe indicare:
  - **una relazione reale** tra le due variabili, se il valore di b è alto e la dispersione dei punti intorno alla retta è ridotta;
  - **relazione casuale o non significativa**, quando la dispersione dei punti intorno alla retta è approssimativamente uguale a quella intorno alla media.

Metodo dei minimi quadrati



Reale cambiamento di Y al variare di X

Metodo dei minimi quadrati



Caso incerto

Non c'è alcuna regressione

Metodo dei minimi quadrati

- Il coefficiente angolare b della retta di regressione, che determina la quantità di variazione di Y per ogni unità aggiuntiva di X, è calcolato *da osservazioni sperimentali*.
- Ciò che tuttavia interessa al ricercatore è la relazione esistente nella popolazione, e sebbene il valore di b sia differente da zero, non è detto che nella popolazione al variare di X si abbia una variazione di Y.
- La significatività del coefficiente di regressione nella popolazione (b) può essere saggiata mediante la verifica dell'ipotesi nulla:

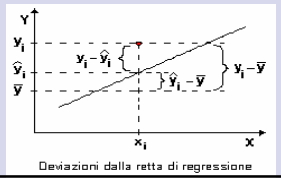
$$H_0 : \beta = 0.$$

Metodo dei minimi quadrati

- Accettando H_0 si assume che il valore reale del coefficiente angolare sia **β** = 0, dunque al variare di X, Y resta costante e uguale al valore dell'intercetta a, pertanto non esiste alcun legame tra X e Y.
- Rifiutando H_0. si accetta l'ipotesi alternativa H_1: **β ≠ 0**. dunque al variare di X si ha una corrispondente variazione sistematica di Y.
- Un metodo per la verifica della significatività della retta calcolata è il **test F di Fisher-Snedecor**, che si basa sulla scomposizione delle devianze.

Metodo dei minimi quadrati

- La somma dei quadrati delle distanze tra i tre punti y_i, $\hat{y}_i$ e $\bar{y}$ definiscono le tre devianze: devianza totale, devianza della regressione o devianza dovuta alla regressione, devianza d'errore o devianza residua:



Deviazioni dalla retta di regressione

Metodo dei minimi quadrati

- Devianza Totale = $\sum_{i=1}^{n}(y_i - \bar{y})^2$
- Devianza di Regressione = $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- Devianza Residua = $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- **Devianza Totale = Devianza di Regress. + Devianza Res.**

Metodo dei minimi quadrati

- Dal rapporto della devianza dovuta alla regressione e quella residua con i rispettivi **gradi di libertà** (1 ed n-1 gdl rispettivamente) si stimano la varianza dovuta alla regressione e la varianza residua.
- Il rapporto:

$$\frac{\text{Varianza di Regressione}}{\text{VarianzaResidua}}$$

- determina il **valore del test F con 1 e n-2 gdl ($F_{(1.n-2)}$).**

Metodo dei minimi quadrati

- Se l'F calcolato è inferiore a quello tabulato per la probabilità prefissata e i gdl corrispondenti. si accetta l'ipotesi nulla H_0 (non esiste regressione lineare statisticamente significativa)
- Se l'F calcolato supera quello tabulato si rifiuta l'H_0 e si accetta H_1 (la regressione lineare tra le due variabili è significativa)

Metodo dei minimi quadrati

- Se $\beta = 0$, la varianza dovuta alla regressione e quella residua sono stime indipendenti e non viziate della variabilità dei dati
- Se $\beta \neq 0$, la varianza residua è una stima non viziata della variabilità dei dati, mentre la varianza dovuta alla regressione è stima di una grandezza maggiore della varianza residua.
- Di conseguenza. il rapporto tra le due varianze è da ritenersi utile alla verifica dell'ipotesi $\beta = 0$.

Metodo dei minimi quadrati

- Rifiutare H_0 :
  - non significa che non esiste relazione tra le due variabili. ma solamente che non esiste una relazione di tipo lineare
  - significa che potrebbe esistere una relazione di tipo differente, come quella curvilinea di secondo grado o di grado superiore

Metodo dei minimi quadrati

- La trasformazione di uno o di entrambi gli assi è spesso sufficiente per ricondurre una relazione di tipo curvilineo a quella lineare:
  - la crescita esponenziale di una popolazione nel tempo. generata da tassi costanti. diviene lineare con la trasformazione logaritmica del tempo, di norma riportato sull'asse delle ascisse
  - la relazione curvilinea tra lunghezza e peso di individui della stessa specie diviene lineare con la trasformazione mediante radice cubica del peso. correlato linearmente al volume
  - l'analisi statistica permette qualsiasi tipo di trasformazione che determini una relazione lineare tra due variabili

Metodo dei minimi quadrati

- Esempio 1
  - Con le misure delle caratteristiche ETA' e PAS rilevate sugli 8 individui è stata determinata le retta di regressione .

$$\hat{PAS} = 68.56 + 1.54 \cdot ETA'$$

  - Supposto il campione estratto dalla popolazione oggetto di studio *significativo*, con le tecniche dell'inferenza statistica occorre verificare:
    - se la retta può essere assunta come rappresentativa di un rapporto lineare tre le due variabili;
    - se è corretto affermare che, nella popolazione di riferimento, ad una variazione di età corrisponde un cambiamento lineare della pressione sistolica;
    - se, mediante il test F, $\beta = 0$ (ip. H_0) oppure $\beta \neq 0$ (ip. H_1).

Metodo dei minimi quadrati

- Si calcola la seguente tabella:

	Devianza	gdl	Varianza
Regressione	6543.1	1	6543.1
Residua	2687.8	6	447.9
Totale	9230.9	7	

- $F_{(1.6)} = \frac{6543.1}{447.9} = 14.61$

Metodo dei minimi quadrati

- il **valore critico** riportato nelle tavole di F per 1 e 6 gdl e per un livello di significatività =0.01 è pari a 13.75;
- il **valore calcolato** di F è **superiore** a quello critico;
- per p<0.01 **si rifiuta H_0**: si può supporre un rapporto lineare tra le variazioni di età e pressione sistolica.
- La stima della significatività della retta o verifica dell'esistenza di una relazione lineare tra le variabili può essere condotta anche con il **test t di Student**, con risultati equivalenti al test F.

Metodo dei minimi quadrati

- **Il test t** è :
  - fondato su calcoli didatticamente meno evidenti di quelli del test F. ma offre il vantaggio di poter essere applicato sia in test unilaterali (β>0 ? oppure β<0 ?) che in test bilaterali (β≠0?);
  - basato sul rapporto tra il valore del coefficiente di regressione b (che rappresenta la risposta media di Y ai diversi valori di X entro il suo intervallo di variazione) ed il suo errore standard SE(b):

Metodo dei minimi quadrati

- SE(b) = $\sqrt{\frac{\text{Varianza Residua}}{\text{DEV(X)}}}$
- $t_{(n-2)} = \frac{b-\beta}{SE(b)}$
- dove β è il valore atteso e i gdl sono n-2.

	Coefficiente	Errore Standard	t	Significatività
Constante	68.748	20.850	3.297	.016
ETA	1.538	.402	3.822	.009

- Osservazione: **t**(n-2) = $\sqrt{F_{(1,n-2)}}$.

Regressione

- **COEFFICIENTE DI DETERMINAZIONE**
  - Per una regressione lineare semplice, ma più in generale per qualsiasi regressione da quella curvilinea a quella lineare multipla, il coefficiente di determinazione **r^2** è la proporzione di variazione totale della variabile dipendente spiegata da quella indipendente:

$$r^2 = \frac{\text{Devianza di Regressione}}{\text{Devianza Totale}}$$

Regressione

- Espresso a volte in percentuale ed indicato in alcuni testi con **R^2** o **Rsq**, serve per misurare "quanto" della variabile dipendente Y sia predetto dalla variabile indipendente X e, quindi, per valutare la bontà dell'equazione di regressione ai fini della previsione sui valori della Y.
- E' una misura che ha scopi descrittivi dei dati raccolti. Non è legata ad inferenze statistiche, ma a scopi pratici, specifici dell'uso della regressione come metodo per prevedere Y conoscendo X.

Regressione

- Il suo valore, compreso tra 0 e 1, è tanto più elevato quanto più la retta passa vicino ai punti, fino a raggiungere 1 (o 100%) quando tutti i punti sperimentali sono collocati esattamente sulla retta e quindi ogni Y_i può essere predetto con precisione totale dal corrispondente valore di X_i
- Nell'esempio con le 8 osservazioni di età e pressione. il valore del coefficiente di determinazione è:

$$r^2 = \frac{6543.1}{9230.9} = 0{,}71$$

Regressione

- Ciò significa che, noto il valore dell'età, quello della pressione è stimato mediante attraverso la retta di regressione con una approssimazione di circa il 71%.
- Il restante $1-r^2=29\%$ è determinato dalla variabilità individuale di scostamento dalla retta ed indica la parte di variabilità della variabile risposta imputabile eventualmente ad altri fattori diversi dall'età.
- La valutazione del valore di r^2 è in stretto rapporto con la disciplina oggetto di studio. Si può ritenere in alcuni ambiti che il modello lineare abbia un ***buon fitting*** con i valori sperimentali se $\mathbf{r^2 > 0.6}$, ma va detto anche che nelle scienze sociali spesso si reputa alto un valore uguale a 0.30 mentre i fisici stimano basso un valore pari a 0.98.

Correlazione Lineare Semplice

- Una misura della bontà del modello lineare può essere ottenuta studiando l'***interdipendenza*** tra due caratteri statistiche quantitativi X e Y.
- Uno degli indici molto noto per una tale misura è il ***Coefficiente di Correlazione Lineare r***:

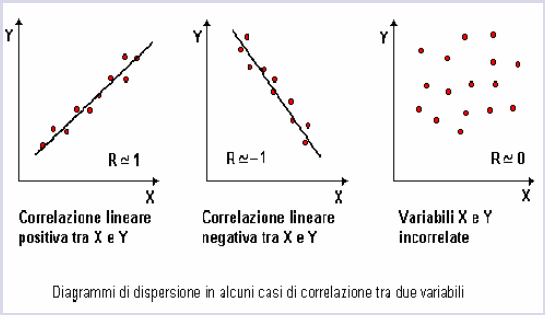
$$r = \frac{\mathrm{CODEV}(X,Y)}{\sqrt{\mathrm{DEV}(X)\cdot\mathrm{DEV}(Y)}}$$

- Tale quantità, indicata anche con **R**, varia tra -1 e 1.

Correlazione Lineare Semplice

- Un valore di **r** vicino a 1 indica una associazione stretta o molto stretta tra le due variabili; si parla in tal caso di *correlazione lineare positiva* tra X e Y: all'aumentare di una variabile aumenta anche l'altra.
- Un valore di **r** vicino a –1 denota un'alta o molto alta *correlazione lineare negativa (discordanza)* tra X e Y: all'aumentare di una di esse l'altra diminuisce.
- Un valore di **r** = 0 o prossimo a 0 indica *indifferenza* (*indipendenza)* tra le variabili.

Correlazione Lineare Semplice



Diagrammi di dispersione in alcuni casi di correlazione tra due variabili

Correlazione Lineare Semplice

- L'analisi della correlazione misura **solo** il grado di associazione spazio-temporale di due fenomeni; il coefficiente **r** è semplicemente una misura dell'intensità dell'associazione tra due variabili.
- Nell'es. 1, utilizzando i calcoli della Tabella costruita a pag. 7, si ha:

$$r = \frac{+4255.62}{\sqrt{2767.88 \cdot 9230.88}} = +0.842$$

- e si registra, quindi, un apprezzabile grado di correlazione lineare positiva tra l'età e la pressione sistolica *per i dati presi in esame.*

Correlazione Lineare Semplice

- Valori di r intorno all'80% o superiori possono, in teoria, far ritenere buona l'associazione lineare: ma va tenuto conto dell'ambito disciplinare e della numerosità dei dati.
- Il coefficiente di correlazione può essere calcolato come media geometrica dei coefficienti di regressione lineare di Y su X (b) e di X su Y (b'):

$$r = \pm\sqrt{b \cdot b'} \quad .$$

- Inoltre il valore r2 è proprio il coefficiente di determinazione.

Correlazione Lineare Semplice

- Per quanto attiene l'esempio n.4 relativo al n° di sigarette fumate (X) e la capacità vitale (Y), il valore di r è

$$r = \frac{-107.51}{\sqrt{477.75 \cdot 25.88}} = -0.967,$$

- dunque c'è correlazione lineare negativa tra i due caratteri presi in esame.
- Inoltre:

$$r2 = 93\%$$

- che è la parte di variazione totale della CV spiegata dal modello regressivo.

Correlazione Lineare Semplice

- Un valore basso o nullo di **r** non deve essere interpretato come assenza di una qualsiasi forma di relazione tra le due variabili:
  - è assente solo una relazione di tipo lineare,
  - tra le due variabili possono esistere relazioni di tipo non lineare.

Cenni sulla regressione multipla

- Esempio 5
  - Dati rilevati su 8 soggetti:

Soggetto	Sesso	Età	PAS	PAD	Fumo
1	1	22	131	70	5
2	0	28	114	75	8
3	1	35	121	80	30
4	0	47	111	75	20
5	0	51	130	70	15
6	1	56	145	80	0
7	1	67	176	85	25
8	1	81	217	90	10

Cenni sulla regressione multipla

- Modelli regressivi lineari:

PAD = 65 + 0.28 Età $r = 0.78$ $r^2 = 0.61$
$t = 3.067$ $p = 0.022$

PAS = 69 + 1.53 Età $r = 0.84$ $r2 = 0.71$
$t = 3.82$ $p = 0.009$

PAS = 75 + 1.55 Età – 0.54 Fumo $t_{Età} = 3.67$ $p = 0.014$
$t_{Fumo} = -0.66$ $p = 0.536$

PAD = 54 + 0.13 PAS + 0.16 Fumo + 0.07 Età

- la t non è significativa per nessuna delle variabili.