Analisi bivariata con variabili quantitative

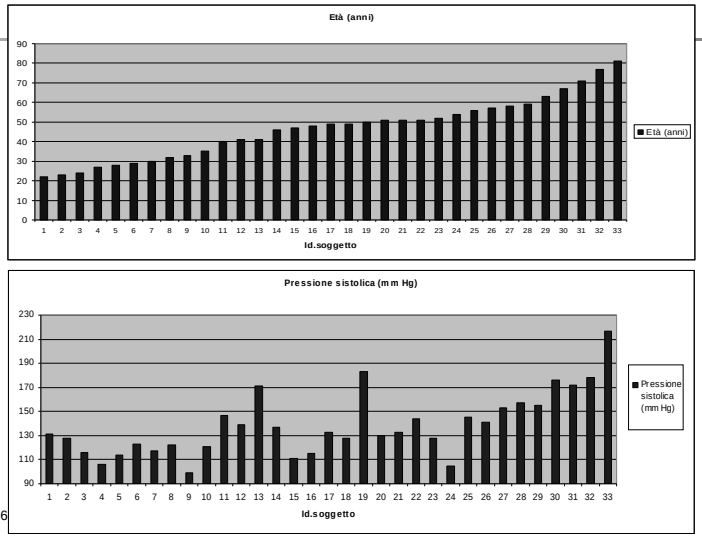
Regressione lineare
Correlazione lineare

LA REGRESSIONE LINEARE

- In un campione di 33 donne, di età compresa tra 22 e 81 anni, è stata misurata la pressione sistolica (in mm di mercurio).
- È noto dalla letteratura medica che, all'aumentare dell'età, tende ad aumentare anche la pressione sistolica.
- I dati di cui disponiamo sono i seguenti; riguardano, da una parte (variabile X) l'età, espressa in anni compiuti, delle 33 donne; dall'altra la pressione sistolica (variabile Y), misurata in mm Hg, delle stesse 33 donne. Li vediamo organizzati nella tabella che segue.

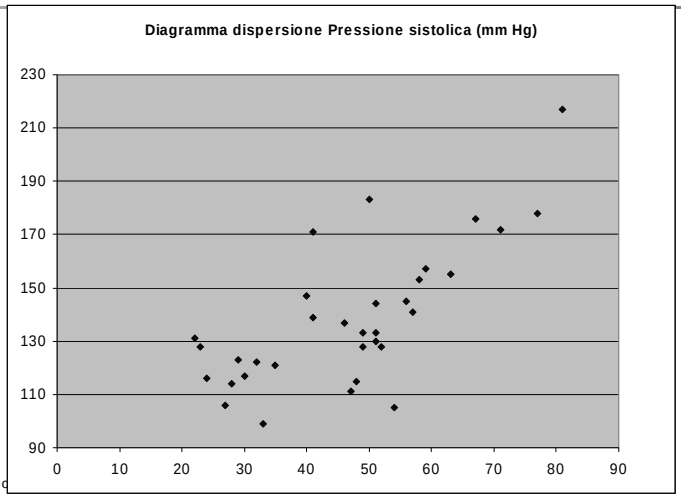
Identificativo soggetto	Età (anni)	Pressione sistolica (mm Hg)
1	22	131
2	23	128
3	24	116
4	27	106
5	28	114
6	29	123
7	30	117
8	32	122
9	33	99
10	35	121
11	40	147
12	41	139
13	41	171
14	46	137
15	47	111
16	48	115
17	49	133
18	49	128
19	50	183
20	51	130
21	51	133
22	51	144
23	52	128
24	54	105
25	56	145
26	57	141
27	58	153
28	59	157
29	63	155
30	67	176
31	71	172
32	77	178
33	81	217

Confronto tra il grafico della 'X' e il grafico della 'Y'



- Dal semplice confronto dei due grafici (a barre), ci accorgiamo che **esiste una relazione** tra i due fenomeni: anche se non possiamo ancora dire in quale misura, si può vedere chiaramente come, al crescere dell'età, tende ad aumentare, anche se con una spiccata variabilità individuale, anche la pressione sistolica.
- Ancora più indicativo, in proposito, è il terzo grafico che vediamo nella prossima diapositiva, nel quale i punti rappresentano le **coppie ordinate** (X,Y) di dati.
- Un grafico di questo tipo è detto DIAGRAMMA DI DISPERSIONE (o ***scatter plot***).

Diagramma di dispersione tra X e Y



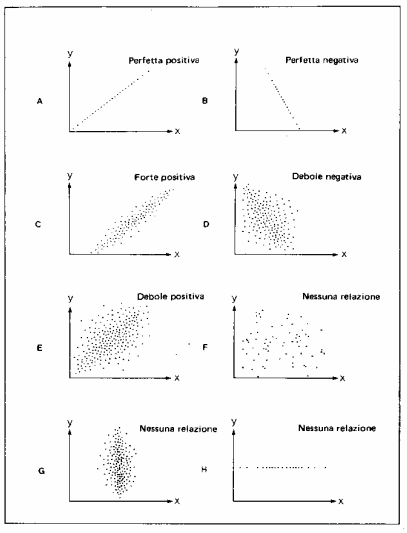
- È quindi evidente che tra i due fenomeni considerati esiste una RELAZIONE, in questo caso POSITIVA: al crescere dell'età, si riscontra una tendenza all'aumento della pressione sistolica.

- La metodologia statistica ci offre un metodo per tradurre in termini quantitativi la presenza di questa "relazione" tra fenomeni: gli strumenti sono, con i differenti significati che vedremo nel seguito, il COEFFICIENTE DI REGRESSIONE ed il COEFFICIENTE DI CORRELAZIONE.

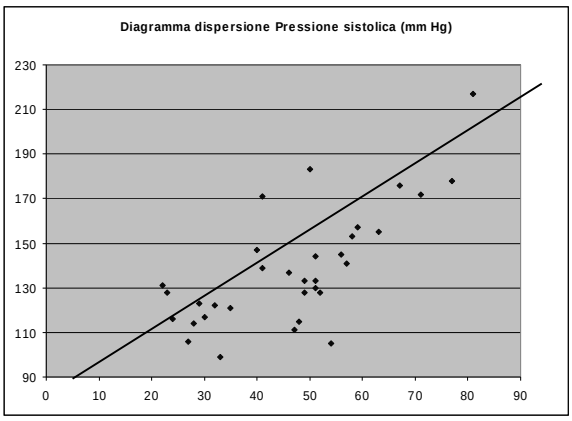
- Il primo passo che dobbiamo fare è quello di identificare un **MODELLO INTERPRETATIVO**, che ci permetta di LEGGERE LA RELAZIONE ESISTENTE TRA I DATI nel miglior modo possibile.

- Il modello di gran lunga più utilizzato (anche se ne esistono altri) è il cosiddetto MODELLO LINEARE; tale modello consiste sostanzialmente nel tracciare una RETTA che "INTERPOLI" i dati (secondo un certo CRITERIO, che vedremo successivamente).

Possibili configurazioni dei dati in relazione al “modello lineare”



Nel grafico che vediamo qui, abbiamo tracciato (per ora, in modo approssimativo) la retta che interpola (cioè “passa attraverso”) i punti sul piano cartesiano.



La retta di regressione e il coefficiente di regressione

- Il modello lineare si esprime con l'EQUAZIONE della retta interpolante, che possiamo scrivere nel seguente modo :

- $y = a + bx$

La retta di regressione e il coefficiente di regressione

- Questa equazione ha due PARAMETRI, che sono i seguenti:
- *a* è l' "intercetta", cioè il punto di intersezione tra la retta e l'asse delle Y;
- *b*, che è il parametro più importante, non è altro che il COEFFICIENTE ANGOLARE della retta interpolante.

- La retta che interpola i dati è detta **RETTA DI REGRESSIONE**, e il parametro *b* è detto **COEFFICIENTE DI REGRESSIONE**.

IL METODO DEI MINIMI QUADRATI

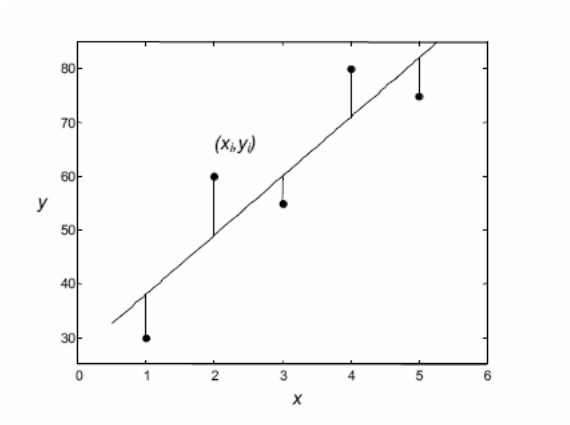
- Ma in che modo, a partire dai dati a nostra disposizione, possiamo ricavare la retta che "approssima meglio" i dati ?
- Il criterio comunemente adottato, e che si è rivelato il migliore sulla base delle sue proprietà matematiche (sulle quali, in questa sede, non ci dilunghiamo), è il cosiddetto **METODO DEI MINIMI QUADRATI** :
- tale metodo consiste nella MINIMIZZAZIONE della somma delle distanze al quadrato tra i punti (teorici) idealmente giacenti sulla retta e i punti empirici corrispondenti ai dati.

- Il metodo dei minimi quadrati consiste, in pratica, nel porre la seguente condizione di minimo:

$$\sum_{i=1}^{n}\left(y_i^* - y_i\right)^2 = \min$$

- Dove :
- y_i^* è l'ordinata del "punto teorico" situato sulla retta di regressione;
- y_i è l'ordinata del punto "empirico" corrispondente, cioè di uno dei nostri dati.

In termini grafici:



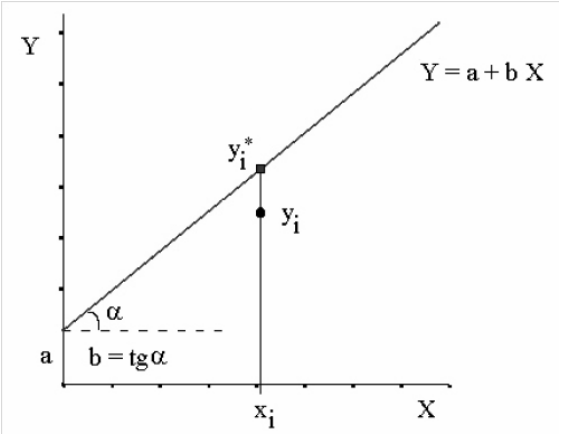
- Dall'applicazione del criterio dei minimi quadrati, tralasciando qui i dettagli tecnici, si ottiene la seguente formula:

$$b_{Y|X} = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2}$$

- In statistica, il coefficiente di regressione, *b*, è una delle grandezze più studiate e più importanti.
- Esso esprime, **NELLA STESSA UNITA' DI MISURA DELLA VARIABILE Y**, quante unità "servono" in media (in più o in meno: il coefficiente di regressione può essere POSITIVO, come nel nostro esempio, oppure NEGATIVO), per ogni corrispondente unità "di incremento" della variabile X.

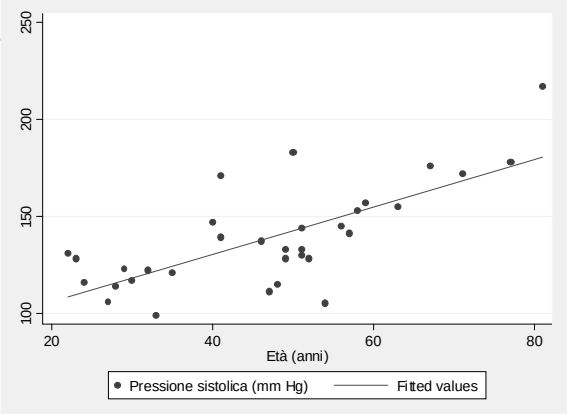
In termini geometrici

Il coefficiente angolare, b, è pari alla tangente trigonometrica dell'angolo α, formato dall'intersezione tra la retta di regressione e una retta parallela all'asse delle X.



- Ad esempio, se si studiasse la relazione (che è positiva) tra peso (Y) e altezza (X), il coefficiente di regressione ci direbbe quanti grammi in più di peso si hanno (in media) per ogni centimetro in più di altezza.
- Nel nostro esempio, il coefficiente *b* ci dirà quanti **mm di mercurio** di pressione sistolica IN PIU' si avranno per ogni **incremento unitario** di età (**anno**).

La <u>vera</u> retta di regressione



EQUAZIONE DELLA RETTA DI REGRESSIONE:

y = 1,222*x + 81,517

- Tornando al nostro esempio, il coefficiente di regressione relativo ai nostri dati è risultato essere pari a +**1,22**: accade, cioè, che, in media, ad ogni anno di età in più, si hanno circa **1,22** millimetri di mercurio in più di pressione sistolica.
- Questo risultato, se ci pensate bene, non è di poco conto, visto che l'ipertensione è un fattore di rischio che svolge un ruolo fondamentale nelle malattie cardiovascolari. Il fatto che sia così fortemente legato all'età, fa sì che esso si sommi ad **altri** Fattori Di Rischio fortemente correlati all'età, come, ad esempio, l'inattività fisica, l'eccesso ponderale, ecc.

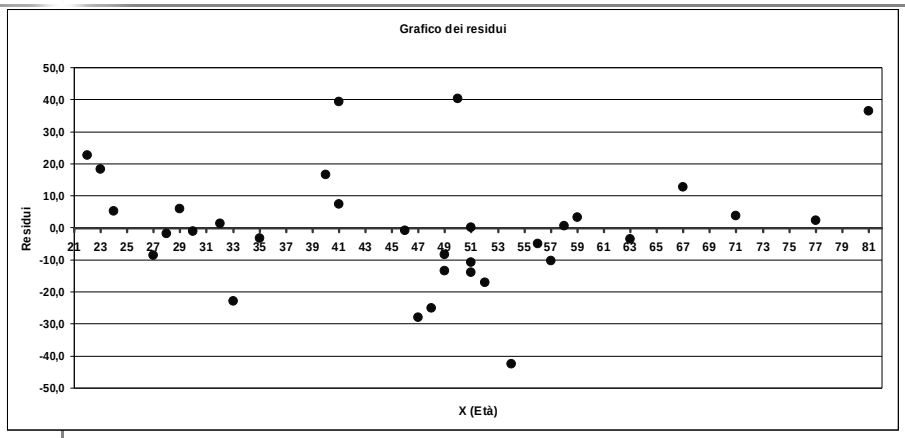
Analisi dei residui

- La quantità
- $y_i - y_i^*$
- è detta **RESIDUO** del valore *i*-esimo della Y rispetto al modello di regressione.
- L'analisi dei residui – e del relativo grafico – è molto utile perché permette di analizzare in modo efficace due cose:
- l'**adattamento** più o meno buono del modello rispetto ai dati;
- 2) l'eventuale presenza nel *dataset* di DATI ANOMALI (*outliers*).

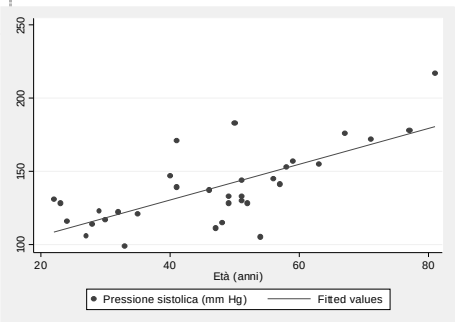
Con i dati del nostro esempio, abbiamo questi <u>residui</u>:

Identificativo soggetto	Età (anni)	Residui
1	22	22,6
2	23	18,4
3	24	5,2
4	27	-8,5
5	28	-1,7
6	29	6,0
7	30	-1,2
8	32	1,4
9	33	-22,8
10	35	-3,3
11	40	16,6
12	41	7,4
13	41	39,4
14	46	-0,7
15	47	-28,0
16	48	-25,2
17	49	-8,4
18	49	-13,4
19	50	40,4
20	51	-13,8
21	51	-10,8
22	51	0,2
23	52	-17,1
24	54	-42,5
25	56	-4,9
26	57	-10,2
27	58	0,6
28	59	3,4
29	63	-3,5
30	67	12,6
31	71	3,7
32	77	2,4
33	81	36,5

In termini grafici:

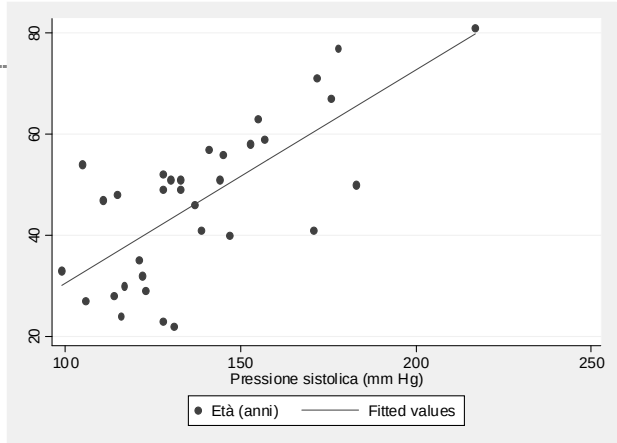


La correlazione lineare



Torniamo al diagramma di dispersione relativo al nostro esempio.
La retta di regressione che abbiamo tracciato ha X come **VARIABILE INDIPENDENTE** (l'età in anni compiuti) e Y come VARIABILE DIPENDENTE (la pressione sistolica).

Anche se, dal punto di vista del fenomeno analizzato, la cosa non ha alcun senso, si potrebbe ricavare anche la retta che ha Y come VARIABILE INDIPENDENTE e la X come VARIABILE DIPENDENTE.
Si otterrebbe allora il risultato rappresentato nel diagramma di dispersione che segue:



La correlazione lineare

- Si otterrebbe, cioè, un coefficiente di regressione pari a **+0,422** ; un risultato che (per quanto sia del tutto privo di significato dal punto di vista empirico) è **DIVERSO** da quello visto in precedenza;
- Questo accade perché l'analisi della regressione **HA UNA DIREZIONE**: si prende in esame ciò che accade a una variabile (dipendente) IN CORRISPONDENZA delle variazioni dell'altra variabile (indipendente). Come abbiamo visto, il coefficiente di regressione è espresso NELLA STESSA UNITA' DI MISURA della variabile dipendente, Y.

Il coefficiente di correlazione

- Potrebbe rivelarsi utile, invece, disporre di una misura statistica della relazione tra X e Y che NON VADA LETTA IN UNA PRECISA DIREZIONE, ma che esprima semplicemente il COVARIARE o il CONTROVARIARE delle due variabili, senza che sia necessaria leggere la misura "rispetto a" una certa variabile.
- Questo compito è svolto dal COEFFICIENTE DI CORRELAZIONE (***r***); il coefficiente di correlazione ha le seguenti caratteristiche:
- è un NUMERO PURO, e non è quindi espresso in ALCUNA UNITA' DI MISURA, né dell'una, né dell'altra variabile.
- è un numero che **VARIA TRA –1 E +1**:

Il coefficiente di correlazione

- **r = -1** CORRELAZIONE NEGATIVA PERFETTA
- **r = 0** ASSENZA TOTALE DI CORRELAZIONE
- **r = 1** CORRELAZIONE POSITIVA PERFETTA.

- Va sottolineato che il coefficiente di correlazione si riferisce, sempre e comunque, a una relazione di tipo **LINEARE** tra i dati; se, per caso, la relazione tra X e Y NON FOSSE LINEARE ma, ad esempio, parabolica, il coefficiente di correlazione (lineare) NON TERREBBE CONTO in alcun modo dell'esistenza di questa relazione.

Richiami da una lezione precedente

Ricordate la deviazione standard (o scarto quadratico medio) di X?

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}}$$

Il quadrato della deviazione standard è detto VARIANZA di X [Var(X)]

$$Var(X) = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n}$$

La varianza, moltiplicata per n, è detta DEVIANZA di X [Dev(X)]

$$Dev(X) = n \cdot Var(X) = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

Come si ottiene la formula del coefficiente di correlazione?

- Il coefficiente di correlazione non è altro che la MEDIA GEOMETRICA dei due COEFFICIENTI DI REGRESSIONE (l'uno, relativo ad Y rispetto ad X; l'altro, relativo ad X rispetto ad Y) calcolati sui nostri dati:

$$r = \sqrt{b_{Y|X} \cdot b_{X|Y}} =$$

$$= \sqrt{\frac{Cov(X,Y)}{Var(X)} \frac{Cov(X,Y)}{Var(Y)}} =$$

$$= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} =$$

$$= \frac{\sum \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum \left(x_i - \overline{x}\right)^2 \sum \left(y_i - \overline{y}\right)^2}}$$

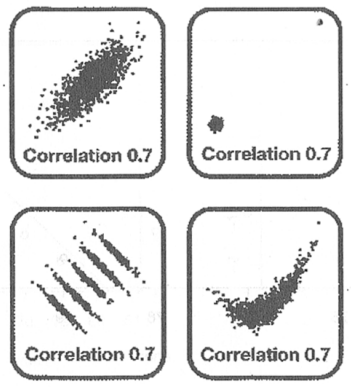
Il coefficiente di correlazione

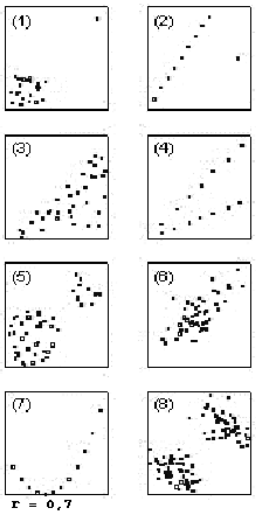
- Il coefficiente di correlazione permette di dare conto in modo immediato, con un semplice numero, della **relazione lineare** esistente tra le due variabili.
- Così, sui dati del nostro esempio, si calcola un coefficiente di correlazione pari a:
- **r = 0,7178**: si tratta di un valore MOLTO ELEVATO.

ATTENZIONE !

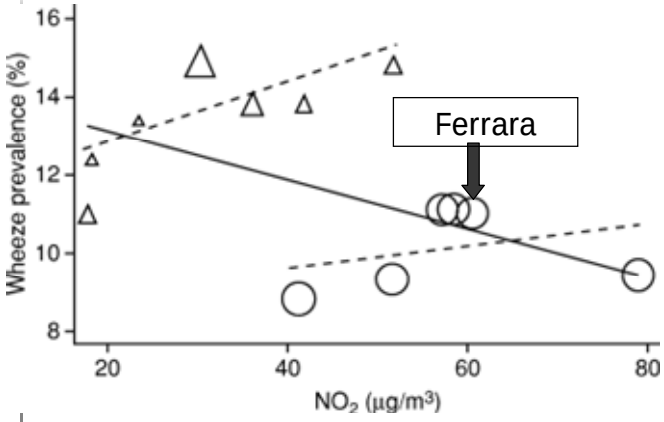
- Spesso, il solo calcolo del coefficiente di correlazione può trarre in inganno; è sempre opportuno osservare bene, prima, il diagramma di dispersione. Spesso, infatti, un valore di *r* elevato (0,7, come nel nostro esempio) può nascondere la presenza di dati anomali, oppure la presenza di una relazione non lineare ma, ad esempio, parabolica, oppure la presenza, nei dati, di due o più sottogruppi nei quali la tendenza dei dati è di direzione opposta ma che, una volta aggregati, creano una **relazione spuria** opposta a quella "realmente" calcolata.
- Vediamo tutti questi "casi limite" nelle due figure che seguono:

ATTENZIONE: il coefficiente di correlazione, per tutti questi diagrammi di dispersione, è <u>sempre</u> 0,7.





Un esempio con dati reali.
Relazione tra biossido di azoto (inquinamento atmosferico) e asma: l'effetto dell'inquinamento SI COMBINA con quello dell'ozono dovuto al clima mediterraneo (temperature più elevate), dando luogo a valori elevati della prevalenza di asma anche in presenza di valori bassi di NO_2. [Fonte: Studio ISAYA]



△ Città dell'area climatica mediterranea **Media NO_2: 31**

○ Città dell'area climatica subconti-nentale **Media NO_2: 58**

La bontà dell'adattamento lineare: l'indice di determinazione lineare

- Finora abbiamo visto due metodi (regressione e correlazione) che ci permettono di misurare l'intensità della relazione tra X e Y;
- Siamo partiti, però, dal presupposto che la relazione tra i dati fosse **lineare.** Abbiamo, pertanto, ipotizzato che tale relazione esistesse e fosse di forma lineare (e non quadratica, cubica, ecc.);
- La misura che vediamo ora serve a dare una valutazione della "bontà" dell'**adattamento** (*to fit*) lineare dei dati empirici al modello (retta di regressione) da noi utilizzato.

La bontà dell'adattamento lineare: l'indice di determinazione lineare

- Prima di tutto, sono necessarie due definizioni. La quantità:

$$Dev(Y)_{regr} = \sum_i \left(y_i^* - \bar{y}\right)^2$$

- è detta ***devianza di regressione*** della Y, ed esprime la variazione tra i punti teorici situati sulla retta di regressione ed il ***valore medio*** della variabile Y.

La bontà dell'adattamento lineare: l'indice di determinazione lineare

- Invece, la quantità:

$$Dev(Y)_{disp} = \sum_i \left(y_i - y_i^*\right)^2$$

- è detta ***devianza di dispersione*** della Y, ed esprime la variazione tra i punti teorici situati sulla retta ed i rispettivi punti empirici.

La bontà dell'adattamento lineare: l'indice di determinazione lineare

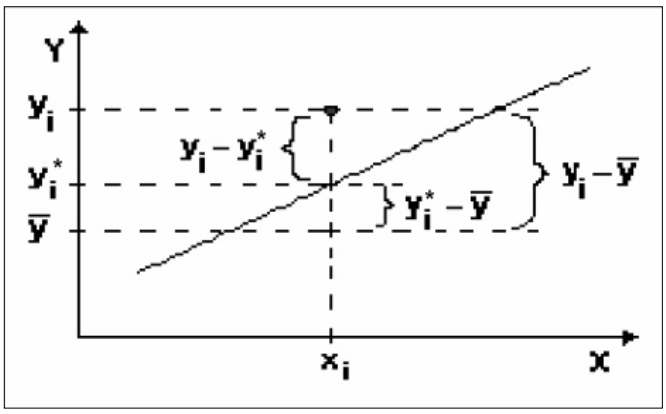
- Si può dimostrare che vale la relazione :

$$Dev(Y) = Dev(Y)_{regr} + Dev(Y)_{disp}$$

- ovvero che la **devianza totale (complessiva)** della Y è scomponibile additivamente tra le due "devianze" sopra descritte, la prima delle quali (la devianza di regressione) sarà **tanto più grande quanto maggiore è la parte della variabilità complessiva della Y che viene *assorbita* dal modello lineare, cioè dalla relazione lineare con la X**.

- Infatti, nel caso limite in cui i punti empirici coincidessero *tutti* con i punti teorici giacenti sulla retta, la devianza di dispersione si annullerebbe.

In termini grafici:



Il coefficiente di determinazione lineare (R^2)

- Pertanto, il rapporto:

$$R^2 = \frac{Dev(Y)_{regr}}{Dev(Y)} = 1 - \frac{Dev(Y)_{disp}}{Dev(Y)}$$

- È un indicatore della bontà dell'adattamento dei dati (Y) alla retta, cioè alla **relazione lineare** con la variabile indipendente X.

Proprietà di R^2

- È un rapporto di composizione, e pertanto varia tra 0 e 1.
- Si ha:
- **$R^2 = 0$** RELAZIONE LINEARE INESISTENTE;
- **$R^2 = 1$** RELAZIONE LINEARE PERFETTA (tutti i punti empirici giacciono sulla retta di regressione).

Proprietà di R^2

- Si può dimostrare che il coefficiente di determinazione lineare R^2 è pari al quadrato di ***r***, coefficiente di correlazione.
- Questa proprietà fa sì che non sia necessario calcolare R^2, ma si possa ricavare direttamente da ***r***.

Con i dati del nostro esempio

- $R^2 = (r)^2 =$
- $=(0,7178...)^2 =$ **0,5153...**

- Si tratta di un valore discreto, ma non molto elevato:
- In sostanza, questo significa che **solo il 51,5%** della variabilità totale della Y è attribuibile alla relazione lineare con la X; la rimanente parte deve essere attribuita ad altri fattori esplicativi.

Un esempio (in ambito epidemiologico)

In questo articolo, viene ipotizzata una forte relazione esistente tra consumo di sale, consumo di nitrati e tassi di mortalità per tumore dello stomaco.

International Journal of Epidemiology
© International Epidemiological Association 1996
Vol. 25, No. 3
Printed In Great

Dietary Salt, Nitrate and Stomach Cancer Mortality in 24 Countries

J V JOOSSENS,* M J HILL,** P ELLIOTT,† R STAMLER,‡ J STAMLER,‡ E LESAFFRE,* A DYER,‡ R NICHOLS† AND H KESTELOOT* ON BEHALF OF EUROPEAN CANCER PREVENTION (ECP) AND THE INTERSALT COOPERATIVE RESEARCH GROUP

Joossens J V (Department of Epidemiology, Leuven University, Kapucijnenvoer 33, B-3000 Leuven, Belgium), Hill M J, Elliott P, Stamler R, Stamler J, Lesaffre E, Dyer A, Nichols R and Kesteloot H on behalf of European Cancer Prevention (ECP) and the INTERSALT Cooperative Research Group. Dietary salt, nitrate and stomach cancer mortality in 24 countries. *International Journal of Epidemiology* 1996; **25**: 494–504.

Background. High salt and nitrate intake are considered as risk factors for stomach cancer, but little is known about possible interactions. This ecological study examines the respective importance of both factors for stomach cancer mortality at the population level using data obtained under standardized conditions and with biochemical analyses performed in the same laboratories.

Method. Randomly selected 24-hour urine samples from 39 populations, sampled from 24 countries (N = 5756 people for sodium, 3303 for nitrate) were obtained from the INTERSALT study. Median sodium and nitrate levels were age- and sex-standardized between ages 20–49 years and averaged per country. Ecological correlation–regression analyses were done in relation to national stomach cancer mortality rates.

Results. The Pearson correlation of stomach cancer mortality with sodium for the 24 countries was: 0.70 in men and 0.74 in women (both $P < 0.001$), and with nitrate: 0.63 ($P = 0.001$) in men and 0.56 ($P < 0.005$) in women. In multiple regression of stomach cancer mortality, using sodium and nitrate as independent variables, the adjusted R^2 was 0.61 in men and 0.54 in women (both $P < 0.001$). Addition of the interaction term (sodium × nitrate) to the previous model increased the adjusted R^2 to 0.77 in men, and to 0.63 in women. The analysis of this model showed that the importance of nitrate as risk factor for stomach cancer mortality increased markedly with higher sodium levels. However, the relationship of stomach cancer mortality with sodium was always stronger than with nitrate.

Conclusions. Salt intake, measured as 24-hour urine sodium excretion, is likely the rate-limiting factor of stomach cancer mortality at the population level.

Keywords: stomach cancer mortality, atrophic gastritis, 24-hour urine sodium, 24-hour urine nitrate, *Helicobacter pylori*, fruits and vegetables

Consumo di sale e mortalità nei maschi

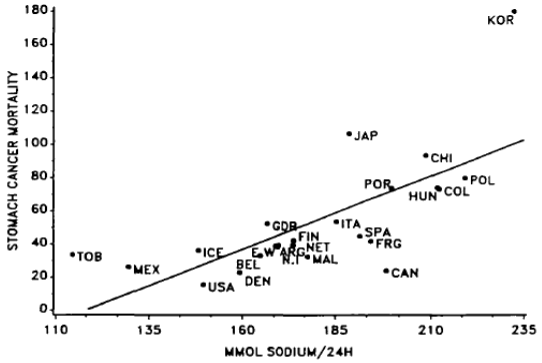


FIGURE 1 *Linear regression between stomach cancer mortality per 100 000/year, age-adjusted between 45–74 years and mmol Na/24-hours, r = 0.70, P < 0.001 in men, n = 24*
ARG = Argentina; BEL = Belgium; CAN = Canada; CHI = P.R. of China; COL = Colombia; DEN = Denmark; E.W = England and Wales; FIN = Finland; FRG = Fed. Rep. of Germany; GDR = German Dem. Rep.; HUN = Hungary; ICE = Iceland; ITA = Italy; JAP = Japan; KOR = South Korea; MAL = Malta; MEX = Mexico; NET = the Netherlands; N.I = Northern Ireland; POL = Poland; POR = Portugal; SPA = Spain; TOB = Trinidad and Tobago; USA = United States.

Consumo di sale e mortalità nelle femmine

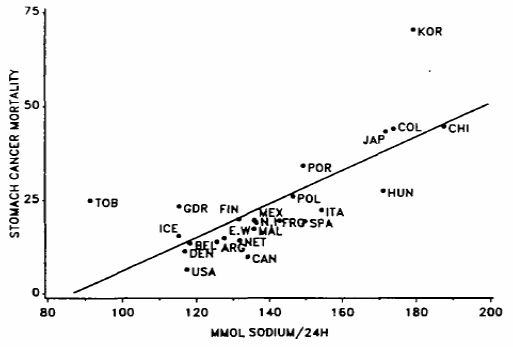


FIGURE 2 *Linear regression between stomach cancer mortality per 100 000/year, age-adjusted between 45–74 years and mmol Na/24-hours, r = 0.74, P < 0.001 in women. n = 24*

ARG = Argentina; BEL = Belgium; CAN = Canada; CHI = P.R. of China; COL = Colombia; DEN = Denmark; E.W = England and Wales; FIN = Finland; FRG = Fed. Rep. of Germany; GDR = German Dem. Rep.; HUN = Hungary; ICE = Iceland; ITA = Italy; JAP = Japan; KOR = South Korea; MAL = Malta; MEX = Mexico; NET = the Netherlands; N.I = Northern Ireland; POL = Poland; POR = Portugal; SPA = Spain; TOB = Trinidad and Tobago; USA = United States.

Un altro esempio, tratto da un articolo (piuttosto discutibile)

Intelligence 38 (2010) 93–100

Contents lists available at ScienceDirect

Intelligence

In Italy, north–south differences in IQ predict differences in income, education, infant mortality, stature, and literacy

Richard Lynn

University of Ulster, Coleraine, Northern Ireland, United Kingdom

ARTICLE INFO

Article history:
Received 13 January 2009
Received in revised form 27 July 2009
Accepted 27 July 2009
Available online 19 August 2009

Keywords:
IQ
Income
Infant mortality
Stature
Education
Italy

ABSTRACT

Regional differences in IQ are presented for 12 regions of Italy showing that IQs are highest in the north and lowest in the south. Regional IQs obtained in 2006 are highly correlated with average incomes at $r = 0.937$, and with stature, infant mortality, literacy and education. The lower IQ in southern Italy may be attributable to genetic admixture with populations from the Near East and North Africa.

L'autore ha utilizzato alcune variabili…

Table 1
Descriptive statistics for IQs and related variables for Italian regions.

Region	Reading	Math	Science	Mean education	IQ	Stature 1855	Stature 1910	Stature 1927	Stature 1980	Per cap income 1970	Per cap income 2003	Infant mortality 1955–7	Infant mortality 1999–02	Literacy 1880	Years educ 1951	Years educ 1971	Years educ 2001	Latitude
Friuli-Venezia	519	513	534	522	103	165.3	167.4	171.7	178.0	8985	20,750	38.3	2.50	45.9	5.2	5.7	9.0	46.0
Trentino	508	508	521	512	101			169.2	177.1	10,930	23,079	44.9	3.47	45.9	5.1	5.7	8.9	46.0
Veneto	511	510	524	515	101	164.8	167.1	169.0	177.0	9223	20,338	36.7	3.17	–	4.6	5.3	8.8	45.5
Tuscany	–	–	–	–	–	163.7	166.2	169.8	175.8	10,022	19,666	35.2	3.24	38.1	4.4	5.2	8.6	43.5
Lombardy	491	487	499	492	100	162.5	166.2	168.1	175.2	11,693	22,639	45.4	3.61	63.0	5.2	–	–	45.0
Piedmont	506	492	508	502	100	162.1	167.0	168.9	175.3	10,964	20,519	–	3.86	67.8	5.1	5.5	8.6	45.0
Liguria	483	473	488	481	97	163.2	167.7	169.7	175.1	9517	20,000	40.8	4.05	55.5	5.1	5.9	9.0	44.5
Emilia Romana	496	494	510	500	100	163.4	166.6	169.2	175.4	10,058	22,439	36.2	3.73	52.2	4.6	5.2	8.7	44.5
Umbria	–	–	–	–	–	161.9	164.6	167.2	175.8	7815	17,070	39.8	3.76	28.6	4.1	4.9	8.7	43.0
Lazio	–	–	–	–	–	161.8	164.6	167.6	175.5	10,317	20,207	–	–	41.8	4.8	5.8	9.4	41.5
Abruzzi Basilicata	446	443	451	447	92	159.6	162.8	164.2	174.0	6814	15,480	68.1	4.56	18.1	3.8	4.6	8.5	41.0
Campania	438	436	442	439	90	160.2	162.9	164.9	173.1	6481	11,862	62.2	5.21	24.6	3.6	4.7	8.2	40.5
Puglia Arulia	440	435	447	441	91	159.7	163.1	164.3	173.3	6313	12,030	70.4	5.88	20.0	3.4	4.5	8.0	40.0
Sardinia	435	429	449	438	90	158.5	160.6	162.1	171.6	8054	13,722	53.6	4.10	19.1	3.4	4.6	8.2	40.0
Calabria	–	–	–	–	–	158.3	162.2	163.3	172.4	6128	11,595	117.5	5.54	14.6	3.5	4.5	8.0	39.0
Sicily	424	423	433	427	89	160.2	163.3	164.7	172.7	6525	12,488	57.0	6.62	19.1	3.5	4.5	8.0	37.0

E su queste ha poi calcolato i coefficienti di correlazione.

Table 2
Correlation matrix for variables in Table 1.

Measure	Reading	Math	Science	Mean educ	IQ	Stature 1855	Stature 1910	Stature 1927	Stature 1980	Income 1970	Income 2003	Infant mortality 1955–57	Infant mortality 1999–02	Literacy 1880	Years educ 1951	Years educ 1971	Years educ 2001
Math	0.993																
Science	0.993	0.994															
Mean educ	0.997	0.998	0.998														
IQ	0.993	0.991	0.990	0.993													
Height 1855	0.918	0.938	0.927	0.929	0.918												
Height 1910	0.906	0.897	0.877	0.894	0.902	0.929											
Height 1927	0.926	0.918	0.911	0.919	0.925	0.964	0.965										
Height 1980	0.936	0.953	0.936	0.944	0.933	0.939	0.868	0.919									
Income 1970	0.729	0.677	0.678	0.694	0.736	0.691	0.843	0.782	0.604								
Income 2003	0.914	0.914	0.908	0.913	0.937	0.815	0.841	0.842	0.821	0.934							
Inf Mrt 1955–7	−0.844	−0.834	−0.860	−0.847	−0.841	−0.775	−0.671	−0.716	−0.661	−0.684	−0.718						
Inf Mrt 1999–02	−0.867	−0.861	−0.883	−0.873	−0.861	−0.774	−0.666	−0.758	−0.807	−0.745	−0.826	0.670					
Literacy 1880	0.863	0.829	0.820	0.838	0.861	0.748	0.875	0.807	0.666	0.902	0.876	−0.661	−0.642				
Years educ 1951	0.922	0.899	0.893	0.905	0.929	0.820	0.901	0.894	0.830	0.889	0.936	−0.631	−0.765	0.924			
Years educ 1971	0.883	0.855	0.860	0.866	0.871	0.782	0.831	0.858	0.788	0.864	0.877	−0.642	−0.750	0.863	0.965		
Years educ 2001	0.886	0.880	0.878	0.882	0.886	0.721	0.685	0.761	0.796	0.774	0.850	−0.716	−0.868	0.689	0.862	0.908	
Latitude	0.970	0.961	0.956	0.964	0.963	0.863	0.874	0.870	0.875	0.798	0.899	−0.676	−0.897	0.842	0.888	0.802	0.726

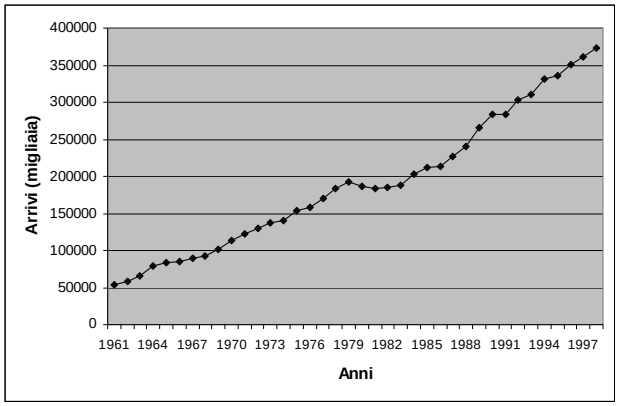
Fare previsioni con le serie storiche: un esempio

Serie storica degli arrivi turistici nel complesso dei paesi europei, dal 1960 al 1998

ANNO	ARRIVI (MILIONI)	ANNO	ARRIVI (MILIONI)
1961	53,99	1980	186,11
1962	58,78	1981	184,00
1963	66,16	1982	185,13
1964	78,60	1983	187,95
1965	83,73	1984	202,28
1966	85,50	1985	212,11
1967	90,00	1986	214,14
1968	92,50	1987	227,46
1969	101,00	1988	240,90
1970	113,00	1989	266,28
1971	122,00	1990	282,88
1972	130,00	1991	283,03
1973	137,00	1992	303,01
1974	140,00	1993	310,79
1975	153,86	1994	331,48
1976	157,82	1995	335,60
1977	170,57	1996	350,26
1978	183,58	1997	361,51
1979	193,00	1998	372,52

In termini grafici

Se osserviamo il grafico, possiamo facilmente notare che si tratta di un andamento **quasi perfettamente lineare**.



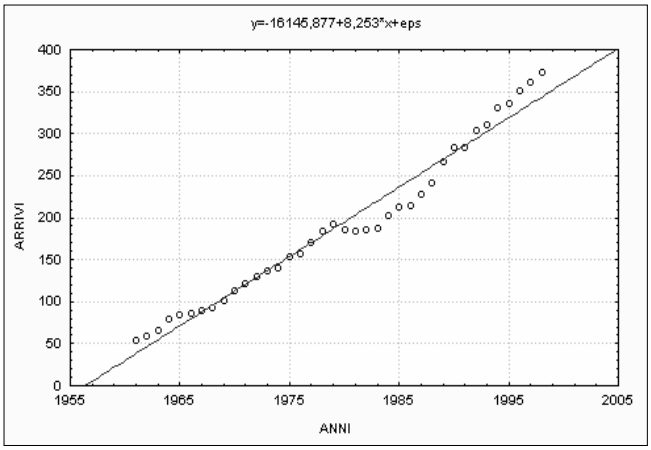
Uso della regressione lineare a scopo <u>previsivo</u>

- Pertanto, per la serie storica appena vista, anziché utilizzare il metodo decompositivo (che abbiamo già visto), possiamo applicare a questi dati un **metodo analitico**.
- Possiamo, cioè, adattare ai dati una retta di regressione lineare, e cercare così di **prevedere** l'andamento futuro della serie storica.
- Infatti, l'analisi di regressione può avere anche una finalità **<u>previsiva</u>** (*estra-polazione*), oltre a quella **<u>interpretativa</u>** vista sopra.

- Per fare questo, utilizzeremo gli anni di calendario (1960 …. 1998) nella loro qualità di "variabile indipendente" X, mentre gli arrivi turistici saranno la nostra Y.

- Stimando come di consueto i parametri della retta con il metodo dei minimi quadrati, si ottiene la retta:
- y = -16145,877+8,253x
- Pertanto, **b = 8,253**
- Questo significa che, in media, durante il periodo considerato, si è avuto, per ogni anno di calendario, un incremento pari a **8,253 milioni di arrivi** in più.

Ecco la retta di regressione stimata



- Sottolineiamo l'adattamento molto buono dei dati alla retta interpolante: infatti il coefficiente di determinazione lineare è pari al **97,05%**; appena il 3% della variabilità della Y non è spiegato dalla relazione lineare con la X.

La previsione

- Supponiamo di voler prevedere il dato che si sarebbe registrato nell'anno 2001.
- Come possiamo fare? È semplice: basta sostituire "2001" (cioè un certo valore della X) nella nostra equazione della retta (adesso, ha un ruolo importante anche l'intercetta):
- L'equazione è $y = f(X) = -16145{,}877 + 8{,}253x$
- Pertanto:
- Y = f(2001) = -16145,877+8,253**2001* =
- = **368,376**

La previsione

- Pertanto, supponendo che anche per gli anni successivi si mantenga la tendenza lineare all'aumento degli arrivi turistici che si è registrata negli anni dal 1960 al 1998, prevediamo che nel 2001 ci saranno:
- Circa **368,4 milioni** di arrivi turistici nei paesi europei.

Oppure…

- Se, invece, riteniamo più realistico considerare l'andamento solo degli anni dal 1980 al 1998, che fanno registrare un incremento annuo più marcato, dobbiamo usare l'equazione:
- $y = f(x) = -22765{,}47+11{,}579x$
- La previsione per il 2001 diventerà, con la nuova equazione:
- $y = f(2001) = -22765{,}47+11{,}579*2001 =$
- $= \mathbf{404{,}109}$

Utilizzando sempre la seconda equazione, possiamo proseguire con gli anni successivi …

	Anno	Arrivi
Dato reale	1995	**335,60**
"	1996	**350,26**
"	1997	**361,51**
"	1998	**372,52**
Previsione	*1999*	***380,95***
"	*2000*	***392,53***
"	*2001*	***404,11***
"	*2002*	***415,69***
	…	**…**
Previsione	*2005*	***450,42***