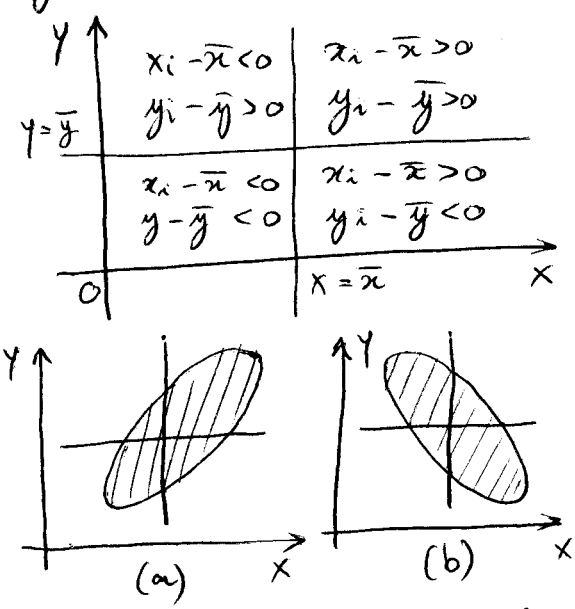
CORRELAZIONE E REGRESSIONE

LA DIPENDENZA TRA DUE CARATTERI DI TIPO QUANTITATIVO VIENE CHIAMATA CORRELAZIONE, PER DISTINGUERLA DALLA DIPENDENZA TRA DUE CARATTERI QUALITATIVI, CHIAMATA CONNESSIONE.

SIANO X ED Y DUE VARIABILI STATISTICHE DI MEDIE $\bar{x}$ E $\bar{y}$, RILEVATE SU UN COLLETTIVO DI N UNITÀ STATISTICHE. SIANO $x_1, x_2, \dots, x_N$ I VALORI OSSERVATI DI X E $y_1, y_2, \dots y_N$ I VALORI OSSERVATI DI Y. SI CHIAMA COVARIANZA DI X E Y, E SI INDICA CON σ_{xy}, IL VALORE COSÌ DEFINITO:

$$\sigma_{xy} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{N}$$

SI OSSERVI COME QUESTA DEFINIZIONE RICORDI QUELLA DI BARICENTRO DI UN SISTEMA DI CORPI. IN PRATICA SI TROVA IL "BARICENTRO" DI UNA DISTRIBUZIONE DI VALORI STATISTICI. SE LA COVARIANZA È POSITIVA, VUOL DIRE CHE LA MAGGIOR PARTE DEI PRODOTTI $(x_i - \bar{x})(y_i - \bar{y})$ SONO POSITIVI, E QUINDI LA MAGGIOR PARTE DEI PUNTI DI COORDINATE (x_i, y_j) DEVE CADERE ALL'INTERNO DEL QUADRANTE DEL PIANO $X\text{o}Y$ A DESTRA DI $y = \bar{y}$ E SOPRA $x = \bar{x}$, COME SI VEDE NELLO SCHEMA QUI A DESTRA, OPPURE IN QUELLO A SINISTRA DI $y = \bar{y}$ E SOTTO $x = \bar{x}$, PER CUI LA DISTRIBUZIONE DEI VALORI È LA (a). SE INVECE LA COVARIANZA È NEGATIVA, LA MAGGIORANZA DEI PRODOTTI $(x_i - \bar{x})(y_i - \bar{y})$ È NEGATIVA, E QUINDI LA MAGGIOR PARTE DEI PUNTI DI COORDINATE (x_i, y_j) DEVE CADERE INTERNAMENTE AGLI ALTRI DUE QUADRANTI, PER CUI LA DISTRIBUZIONE DEI VALORI È SIMILE ALLA (b). ORA, LA (a) È UNA DISTRIBUZIONE DI TIPO LINEARE CRESCENTE, LA (b) È UNA DISTRIBUZIONE DI TIPO LINEARE DECRESCENTE. SE INVECE LA COVARIANZA È NULLA, I PUNTI SONO SPARPAGLIATI SENZA ALCUNA REGOLARITÀ, OPPURE SONO DISPOSTI SECONDO RELAZIONI MOLTO LONTANE DA QUELLE LINEARI, COME AD ESEMPIO UNA RELAZIONE QUADRATICA.



LA COVARIANZA PUÒ ESSERE CALCOLATA ANCHE CON LA FORMULA $\sigma_{xy} = \frac{\sum_{i=1}^{N} x_i y_i}{N} - \bar{x}\bar{y}$

UNA VOLTA APPURATA UNA CORRELAZIONE TRA DUE VARIABILI STATISTICHE X E Y, OCCORRE CAPIRE IL GRADO DI TALE CORRELAZIONE, COME SI FA CON IL TEST DEL CHI QUADRO; L'IDEA È QUELLA DI METTERE A RAPPORTO LA COVARIANZA CON IL SUO VALORE MASSIMO. SE σ_x E σ_y SONO LE DEVIAZIONI STANDARD DI X ED Y, LA COVARIANZA PUÒ ASSUMERE VALORI COMPRESI IN QUESTO INTERVALLO:

$$-\sigma_x \sigma_y \leq \sigma_{xy} \leq +\sigma_x \sigma_y$$

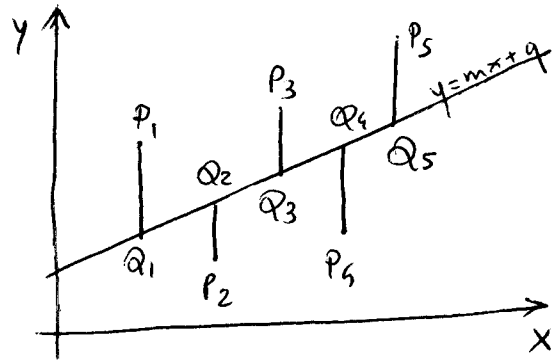
SI CHIAMA ALLORA COEFFICIENTE DI CORRELAZIONE LINEARE DI DUE VARIABILI X E Y IL COEFFICIENTE COSÌ DEFINITO: (→)

(→)

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

RISULTA $-1 \leq \rho \leq 1$. IL SEGNO DEL COEFFICIENTE DI CORRELAZIONE LINEARE È LO STESSO DELLA COVARIANZA; $\rho > 0$ INDICA UNA RELAZIONE LINEARE CRESCENTE, MENTRE $\rho < 0$ NE INDICA UNA DECRESCENTE. L'INDICE DI CORRELAZIONE LINEARE È UGUALE A ± 1 SE E SOLO SE LA CORRELAZIONE È PERFETTAMENTE LINEARE; TANTO PIÙ ρ SI AVVICINA A ± 1, TANTO PIÙ LA CORRELAZIONE È LINEARE. SE $\rho = 0$, TRA X E Y NON VI È ALCUNA CORRELAZIONE LINEARE. COME PERÒ DETERMINARE L'EFFETTIVA FUNZIONE LINEARE CHE MEGLIO INTERPRETA TALE LEGAME?

CONSIDERIAMO UNA FUNZIONE LINEARE DI EQUAZIONE $y = mx + q$ E, PER OGNI PUNTO $P(x_i, y_i)$ CHE RAPPRESENTA I DATI, CONSIDERIAMO IL CORRISPONDENTE PUNTO $Q_i(x_i, y'_i)$ DI ASCISSA x_i APPARTENENTE ALLA RETTA CHE COSTITUISCE IL GRAFICO DI $y = mx + q$, A FIANCO SI VEDE LA COSTRUZIONE DEI SEGMENTI P_1Q_1, P_2Q_2, ... P_iQ_i ..., P_nQ_n CON $N = 5$. IN GENERALE $\overline{P_iQ_i} = |y_i - y'_i|$ ELEVIAMO AL QUADRATO TALI LUNGHEZZE E LE SOMMIAMO: $\sum_{i=1}^{N} (y_i - y'_i)^2$



QUESTA SOMMA ESPRIME CON UN UNICO NUMERO UNA STIMA DELLA DISTANZA FRA I DATI y_i OSSERVATI E I VALORI TEORICI y'_i CALCOLATI SUL GRAFICO DELLA RETTA. COME FUNZIONE LINEARE CHE MEGLIO APPROSSIMA I DATI, CONVIENE SCEGLIERE QUELLA PER CUI QUESTA SOMMA RISULTA MINIMA. TALE RETTA PRENDE IL NOME DI RETTA DI REGRESSIONE.

DATE DUE VARIABILI X E Y, DI VALORI MEDI RISPETTIVAMENTE $\bar{x}$ E $\bar{y}$, LA RETTA DI REGRESSIONE CHE ESPRIME Y IN FUNZIONE DI X È LA RETTA CHE PASSA PER IL PUNTO DI COORDINATE $(\bar{x}; \bar{y})$ E CHE HA COME COEFFICIENTE ANGOLARE IL COSIDDETTO COEFFICIENTE DI REGRESSIONE, DEFINITO DA $m = \frac{\sigma_{xy}}{\sigma_x^2}$. L'EQUAZIONE È ALLORA $y - \bar{y} = m(x - \bar{x})$. LA RETTA DI REGRESSIONE È DETTA ANCHE RETTA DEI MINIMI QUADRATI.

VEDIAMO ORA UN ESEMPIO PRATICO DI TUTTO QUESTO. SU CINQUE PAZIENTI È STATA RILEVATA L'ETÀ (X) E IL TASSO DI COLESTEROLO NEL SANGUE:

ETÀ	25	33	40	55	72
COLESTEROLO	125	130	137	146	170

DETERMINIAMO ANZITUTTO IL COEFFICIENTE DI CORRELAZIONE LINEARE, RICAVANDOLO DALLA COVARIANZA σ_{xy} E DALLE DEVIAZIONI STANDARD σ_x DELL'ETÀ E σ_y DEL TASSO DI COLESTEROLO. IN MODO DA AGEVOLARE I CALCOLI, CHE SONO TANTI E COMPLESSI, CONVIENE ORGANIZZARE I DATI IN UNA TABELLA COME QUESTA:

(→)

(→)

x_i	y_i	$x_i y_i$	x_i^2	y_i^2
25	125	3125	625	15625
33	130	4290	1089	16900
40	137	5480	1600	18769
55	146	8030	3025	21316
72	170	12240	5184	28900
$\sum_i x_i = 225$	$\sum_i y_i = 708$	$\sum_i x_i y_i = 33165$	$\sum_i x_i^2 = 11523$	$\sum_i y_i^2 = 101510$

A QUESTO PUNTO OTTERREMO:

$$\bar{x} = \frac{\sum_i x_i}{5} = 45$$

$$\bar{y} = \frac{\sum_i y_i}{5} = 141{,}6$$

$$\sigma_x^2 = \frac{\sum_i x_i^2}{5} - \bar{x}^2 = \frac{11523}{5} - 45^2 = 279{,}6$$

$$\sigma_y^2 = \frac{\sum_i y_i^2}{5} - \bar{y}^2 = \frac{101510}{5} - 141{,}6^2 = 251{,}44$$

$$\sigma_{xy} = \frac{\sum_i x_i y_i}{5} - \bar{x}\bar{y} = \frac{33165}{5} - 45 \cdot 141{,}6 = 261$$

CONCLUDIAMO PERCIÒ CHE IL COEFFICIENTE DI CORRELAZIONE LINEARE NEL NOSTRO CASO VALE:

$$\rho = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \frac{261}{\sqrt{279{,}6} \cdot \sqrt{251{,}44}} = 0{,}984$$

PERCIÒ LA RELAZIONE SI PUÒ CONSIDERARE LINEARE AL 98,4%. CERTAMENTE LA RETTA DI REGRESSIONE È UN MODELLO CHE INTERPRETA MOLTO BENE LA CORRELAZIONE TRA X E Y. RICORDIAMO CHE ρ VIENE CHIAMATO ANCHE INDICE DI BRAVAIS - PEARSON, ESSENDO STATO INTRODOTTO DAL FISICO MATEMATICO AUGUSTE BRAVAIS (1811 - 1863) E KARL PEARSON (1857 - 1936). SCRIVIAMO ALLORA NEL NOSTRO CASO L'EQUAZIONE DELLA RETTA DI REGRESSIONE. IL SUO COEFFICIENTE ANGOLARE È DATO DALLA FORMULA SOPRA RIPORTATA:

$$m = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{261}{279{,}6} = 0{,}933$$

L'EQUAZIONE DELLA RETTA DI REGRESSIONE PERCIÒ È:

$$y - 141{,}6 = 0{,}933(x - 45)$$

CIOÈ:

$$y = 0{,}933x + 99{,}594$$

AD ESEMPIO, SE IN ESSA INSERISCO $x = 55$, OTTENGO $y = 150{,}909$, UN VALORE MOLTO VICINO ALL'$y = 146$ OSSERVATO (LO SCARTO È APPENA DEL 3%)