

5.1. Associazione

I primi capitoli concernevano l'analisi statistica di osservazioni su una singola variabile. In alcuni problemi i dati erano stati divisi in due gruppi e la dicotomia avrebbe potuto essere considerata come definizione di una seconda variabile; tuttavia i problemi a due campioni sono esempi piuttosto artificiosi di relazioni tra due variabili.

In questo capitolo si esaminerà più in generale l'associazione tra due variabili quantitative, limitatamente alle situazioni in cui l'andamento generale del fenomeno è lineare. Il termine lineare significa che al variare di una variabile, l'altra variabile segue, *in media*, un andamento che può essere rappresentato approssimativamente da una linea retta. Situazioni più complesse saranno discusse nei capitoli IX e X.

La tecnica grafica di base per il caso di due variabili è il *diagramma di dispersione* ed è buona norma rappresentare i dati in questa forma prima di affrontare qualunque analisi numerica. In Fig. 5 ne vien dato un esempio. In generale i dati si riferiscono a un certo numero di *individui*, ognuno dei quali fornisce informazioni su due variabili. Nel diagramma di dispersione a ogni variabile è associato uno dei due assi coordinati e a ogni individuo corrisponde un punto le cui coordinate sono i valori osservati delle variabili. In Fig. 5.1 gli individui sono città e le variabili sono il tasso di mortalità infantile e un indice opportuno di sovraffollamento.

Il diagramma di dispersione fornisce una descrizione globale della distribuzione di ogni variabile e delle relazioni tra variabili. L'ulteriore analisi statistica risponde a diversi scopi. Innanzitutto dà una misura numerica di alcune fondamentali caratteristiche della relazione, così come media e deviazione standard danno misure concise delle più importanti caratteristiche della distribuzione di una singola variabile. In secondo luogo permette al ricercatore che lo desidera di prevedere il valore di una variabile quando sia noto il valore di un'altra. Di regola ciò è impossibile con assoluta certezza ma si può dire qualcosa sul valore medio e sulla variabilità del valore previsto. Per esempio dalla Fig. 5.1 risulta con una discreta approssimazione che

una città con 0,6 persone per stanza aveva nel 1961 una mortalità infantile media di circa 20 per mille nati vivi, estendendosi l'intervallo di variazione da 14 a 26.

In terzo luogo questo gruppo di tecniche permette di stabilire se la tendenza che risulta dai dati è o non è significativa. Per esempio in base ai dati di Fig. 5.1 si può con sicurezza affermare che la mortalità infantile cresce in

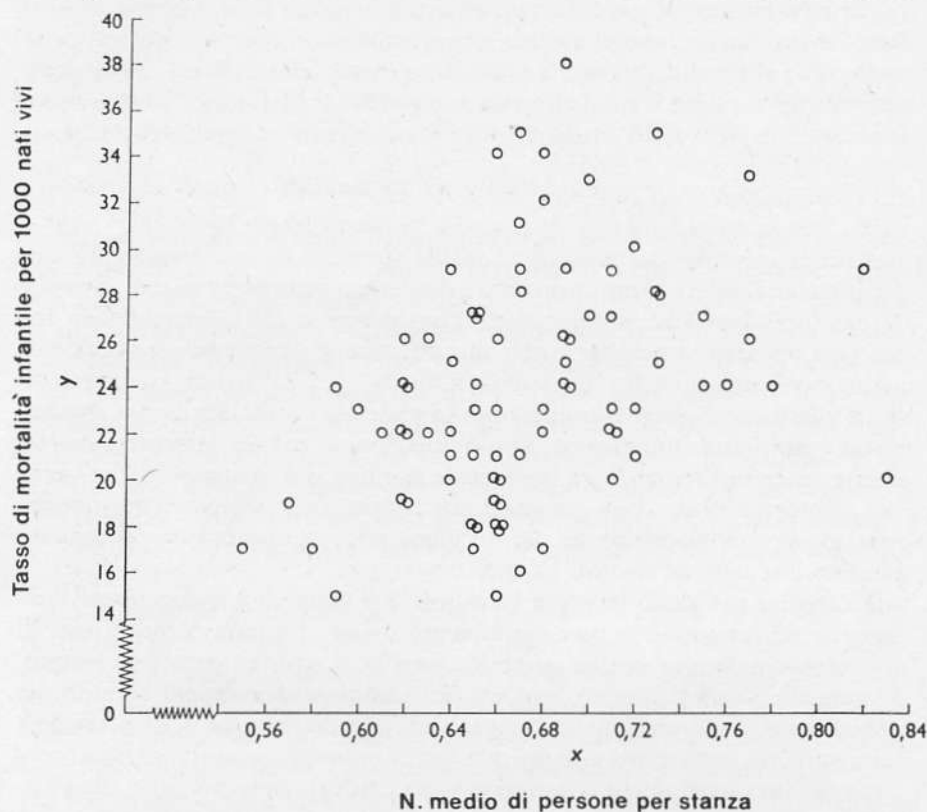


Fig. 5.1. Diagramma di dispersione relativo al numero medio di persone per stanza e alla mortalità infantile per 1000 nati vivi in 83 capoluoghi di contea dell'Inghilterra e del Galles nel 1961.

media al crescere dell'indice di sovraffollamento o non si tratta piuttosto di una tendenza apparente dovuta interamente al caso?

Un'ulteriore applicazione riguarda la possibilità di correggere le misure di una variabile tenendo conto degli effetti di un'altra variabile. In uno studio sui volumi espiratori forzati (FEV) dei lavoratori della industria del cadmio esposti per un certo numero di anni ai vapori di cadmio è stato istituito un confronto con altri lavoratori non esposti. Il valore medio di FEV del primo

gruppo era minore di quello del secondo gruppo. Il secondo gruppo erano più vecchie, tendendo a diminuire con l'età. La differenza delle medie di FEV è dovuta alla differenza delle medie di età. Per rispondere alla domanda sulla relazione tra FEV e età, bisogna distinguere la causa.

Bisogna distinguere la causa. Due variabili sono correlate se la conoscenza del valore di una di esse permette di conoscere l'altra. Vi è ad esempio una correlazione tra i divorzi in Gran Bretagna e il consumo di tabacco. Se il tabacco sia una causa di divorzio, le coppie divorziate si separano a causa di un processo che implica un processo di divorzio.

Un'ulteriore distinzione è quella tra correlazione e causalità. Due variabili possono essere correlate se sono scelte a caso, o se sono correlate per una situazione in cui i due fattori sono correlati dal ricercatore. Un esempio è quello dello studio di una popolazione scolastica su individui. In questo caso non è possibile scegliere a caso, ma caso tuttavia può essere una popolazione scelta dalla stessa area e dalla stessa situazione. Lo studio di un tipo di relazione tra due variabili è un diagramma di dispersione. La distribuzione della distribuzione è data nel paragrafo 5.3.

5.2. Regressione lineare

Supponiamo di disporre di due variabili x e y su un grande numero di individui, in cui y varia mediamente con x . Se y è una variabile casuale, il valore medio di y varii in funzione di x , si chiama regressione lineare. Se x è noto, si chiama regressione lineare. La regressione è indicata con la lettera b . Le tesi riguardanti il fatto che la regressione delle altezze e valori

gruppo era minore di quello del secondo; tuttavia i lavoratori del primo gruppo erano più vecchi dei lavoratori del secondo gruppo e si sa che FEV tende a diminuire con l'età. Si pone pertanto il problema di determinare se la differenza delle medie di FEV può essere spiegata unicamente dalla differenza di età. Per rispondere a tale quesito bisogna analizzare dettagliatamente la relazione tra FEV e età; il metodo è descritto nel paragrafo 9.5.

Bisogna distinguere accuratamente tra *concetto di associazione* e *concetto di causa*. Due variabili sono associate se la distribuzione dell'una dipende dalla conoscenza del valore dell'altra; ciò non significa che una variabile sia *causa* dell'altra. Vi è ad esempio una forte associazione tra numero assoluto di divorzi in Gran Bretagna e quantità di tabacco ivi importato (gli "individui" sono nel diagramma di dispersione i singoli anni). Non ne consegue però che il tabacco sia una causa perturbatrice dell'equilibrio coniugale o che le coppie divorziate si consolino col tabacco. In altri termini l'associazione non implica un processo causale.

Un'ulteriore distinzione si pone tra situazioni in cui entrambe le variabili possono essere considerate come variabili casuali, quando cioè gli individui sono scelti a caso, o almeno senza alcun riferimento al valore delle variabili, e situazioni in cui i valori di una variabile sono deliberatamente prefissati dal ricercatore. Un esempio della prima situazione potrebbe essere rappresentato dallo studio della relazione tra altezza e pressione arteriosa nell'età scolare su individui dello stesso sesso e di un determinato gruppo di età. In questo caso non è stato possibile selezionare il campione interamente a caso tuttavia può essere considerato approssimativamente rappresentativo di una popolazione scolare dello stesso sesso e della stessa età proveniente dalla stessa area e dallo stesso tipo di scuola. Un esempio della seconda situazione è lo studio della crescita infantile in un dato intervallo di età. Il tipo di relazione tra età e altezza così come può venir rappresentata da un diagramma di dispersione dipende molto dall'intervallo di età prescelto e dalla distribuzione delle età in tale intervallo. Il problema sarà ripreso nel paragrafo 5.3.

5.2. Regressione lineare

Supponiamo di disporre di osservazioni riguardanti le variabili aleatorie x e y su un grande numero di individui e che si voglia conoscere il modo in cui y varia mediamente al variare di x . Se è appropriato pensare a y come a una variabile casuale per ogni valore di x , ci si può chiedere come il valore medio di y vari in funzione di x . La distribuzione di probabilità di y , quando sia noto x , si chiama distribuzione *condizionale* o *condizionata* e l'attesa condizionata è indicata con $E(y/x)$. A questo livello del discorso non si fanno ipotesi riguardanti il fatto che x sia o non sia una variabile casuale. Nello studio delle altezze e valori di pressione arteriosa in soggetti scelti a caso entrambe

le variabili sono casuali; se invece x e y fossero rispettivamente l'altezza e il peso di bambini scelti in base all'età solo la y sarebbe una variabile casuale.

L'attesa condizionata $E(y/x)$ dipende in generale da x ; questa dipendenza è indicata col nome di *funzione di regressione* di y su x . La Fig. 5.2 ne riporta due esempi; la regressione di Fig. 5.2 (b) differisce in due modi da quella di Fig. 5.2 (a): la curva di 5.2 (b) è una retta — la retta di regressione di y su x ; inoltre in 5.2 (b) la variabilità di y per un dato valore di x è costante mentre in 5.2 (a) la variabilità decresce al crescere di x ; si dice che la regressione in (b) è *omoscedastica* mentre quella in (a) è *eteroscedastica*.

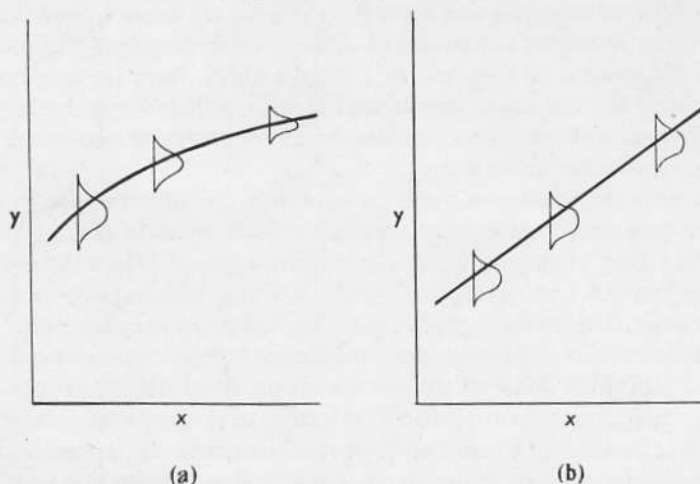


Fig. 5.2. Due curve di regressione di y su x : (a) non lineare e eteroscedastica; (b) lineare e omoscedastica. Le distribuzioni riportate sono quelle relative ai valori di y per determinati valori di x .

La situazione rappresentata in Fig. 5.2 (b) è importante non solo per la sua semplicità ma anche perchè regressioni lineari e omoscedastiche sono frequenti in ogni lavoro scientifico. Nella discussione che segue si introdurrà un'ulteriore semplificazione assumendo che la distribuzione di y dato x sia normale.

Il modello può allora essere descritto dicendo che per un dato valore di x , y segue una distribuzione normale con media

$$E(y/x) = \alpha + \beta x$$

(l'equazione generale di una retta) e varianza σ^2 (costante). Un insieme qualunque di dati consiste di n coppie di osservazioni indicate con (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) dove ogni y_i è un'osservazione indipendente dalle altre e proveniente dalla distribuzione $N(\alpha + \beta x_i, \sigma^2)$. Come si possono stimare i parametri α , β e σ^2 che caratterizzano il modello?

Considerazioni te
col metodo dei *min*.
mente quelle quan
 $\Sigma(y_i - \bar{y})^2$ dove $\bar{y}$

Anche dal punto d
disegnare una linea
gramma di dispersio
distanze dei punti c
parallela all'asse y (
Analiticamente si

Una dimostrazione
noti che per ogni insi
 $\Sigma(p_i - \bar{p})^2$ è $P = \bar{p}$.

come si può vedere e
destra della (5.4) è
 $P = \bar{p}$. Consideriamo
 b il minimo di S si ha

(dal precedente risulta

come in (5.2). La line

da cui consegue che

Si deve ora trovare

$$S = \Sigma(y_i - \bar{y} - b(x_i - \bar{x}))^2 \\ = \Sigma(x_i - \bar{x})^2 (b - \frac{\Sigma y_i (x_i - \bar{x})}{\Sigma (x_i - \bar{x})^2})^2$$

* Le stime di α e β sec
(vedi paragrafo 4.11); in
minimo.

Considerazioni teoriche* portano alla seguente regola: α e β sono stimati col metodo dei *minimi quadrati*; le stime corrispondenti, a e b , sono esattamente quelle quantità che minimizzano la somma dei quadrati *residua*, $\Sigma(y_i - Y_i)^2$ dove Y_i è dato dall'equazione di regressione stimata

$$Y_i = a + bx_i \quad (5.1)$$

Anche dal punto di vista intuitivo il proposito è interessante: si tratta di disegnare una linea di regressione che passi in mezzo agli n punti del diagramma di dispersione in modo da minimizzare la somma dei quadrati delle distanze dei punti dalla retta, essendo queste ultime misurate in direzione parallela all'asse y (Fig. 5.3).

Analiticamente si dimostra che a e b sono dati dalle seguenti formule

$$a = \bar{y} - b\bar{x} \quad (5.2)$$

$$b = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \quad (5.3)$$

Una dimostrazione che non fa ricorso all'analisi è la seguente. Innanzitutto si noti che per ogni insieme di n numeri $p_1, p_2, \dots, p_n$ la quantità P che minimizza $\Sigma(p_i - P)^2$ è $P = \bar{p}$. Per dimostrarlo si osservi che per ogni valore di p si ha

$$\Sigma(p_i - P)^2 = \Sigma(p_i - \bar{p})^2 + n(\bar{p} - P)^2 \quad (5.4)$$

come si può vedere espandendo $(p_i - P)^2$ in $((p_i - \bar{p}) + (\bar{p} - P))^2$. Il membro di destra della (5.4) è minimo quando il secondo termine vale zero cioè quando $P = \bar{p}$. Consideriamo ora $S = \Sigma(y_i - Y_i)^2 = \Sigma(y_i - a - bx_i)^2$. Per ogni valore di b il minimo di S si ha quando

$$a = \Sigma(y_i - bx_i)/n$$

(dal precedente risultato con $p_i = y_i - bx_i$) cioè quando

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

come in (5.2). La linea di regressione può perciò essere così scritta

$$Y_i = a + bx_i = \bar{y} - b\bar{x} + bx_i = \bar{y} + b(x_i - \bar{x}) \quad (5.5)$$

da cui consegue che essa passa per il punto $(\bar{x}, \bar{y})$.

Si deve ora trovare il valore di b che minimizza

$$S = \Sigma(y_i - \bar{y} - b(x_i - \bar{x}))^2$$

$$= \Sigma(x - \bar{x})^2 \left(b - \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2} \right)^2 + \Sigma(y - \bar{y})^2 - \frac{(\Sigma(x - \bar{x})(y - \bar{y}))^2}{\Sigma(x - \bar{x})^2} \quad (5.6)$$

* Le stime di α e β secondo i "minimi quadrati" sono anche stime di massima verosimiglianza (vedi paragrafo 4.11); inoltre tra tutte le stime non viziate sono quelle che hanno errore standard minimo.

come si vede facilmente sviluppando il membro di destra; (l'indice i è stato eliminato per semplicità). Il secondo e terzo termine della (5.6) non dipendono da b ; il primo termine è una somma di quadrati e non può essere negativo; tuttavia esso vale zero quando

$$b = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2}$$

che è pertanto il valore richiesto di b , come in (5.3).

Va notato che il valore minimo di S è dato dagli ultimi due termini della (5.6) e cioè

$$\sum(y - \bar{y})^2 - \frac{(\sum(x - \bar{x})(y - \bar{y}))^2}{\sum(x - \bar{x})^2} \quad (5.7)$$

Infine si può dimostrare che una stima non viziata di σ^2 è

$$s_0^2 = \frac{\sum(y - Y)^2}{n - 2} \quad (5.8)$$

essendo la somma dei quadrati residua $\sum(y - Y)^2$ ottenuta dalla (5.7). Il divisore $n - 2$ rappresenta i gradi di libertà residui e s_0^2 è la *varianza residua*.

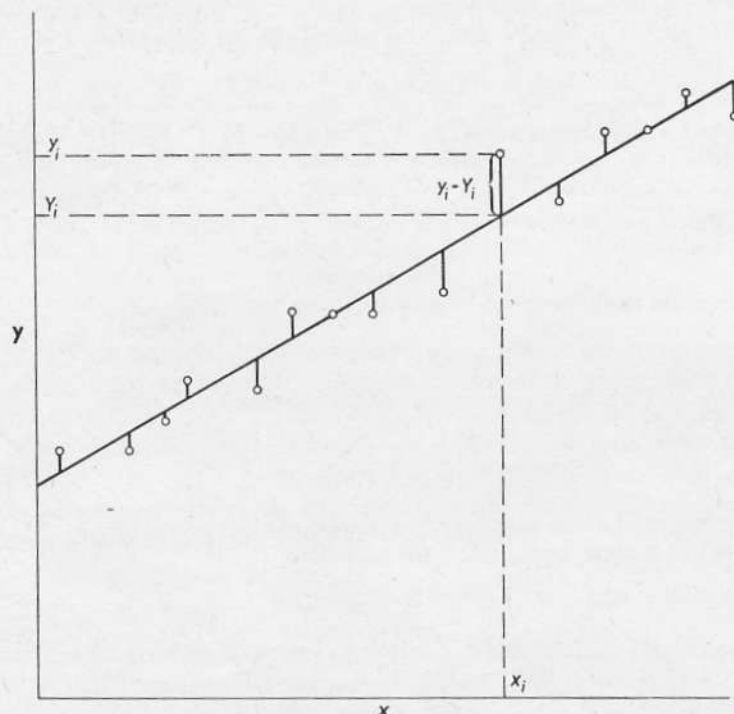


Fig. 5.3. Curva di regressione lineare calcolata col metodo dei minimi quadrati dimostrante gli scarti tra valori osservati y_i e valori Y_i calcolati mediante l'equazione di regressione.

Le grandezze a e b sono molto usate in particolare nella regressione della linea di

L'espressione $a + b(x - \bar{x})$ di x e di y dalle rispettive medie viata analoga alla (5.3) simile a quello svolto

$$\sum(x_i - \bar{x})(y_i - \bar{y})$$

Espressioni equivalenti. Bisogna osservare che a deve essere una grandezza media può essere negativa

Esempio 5.1. La Tabella 5.1, l'incremento di peso del peso alla nascita. L'associazione non è strettamente lineare. In base a

$n =$

$$\begin{aligned} \sum x^2 &= 409880 \\ (\sum x)^2/n &= 399618, \\ \sum(x - \bar{x})^2 &= 10262, \end{aligned}$$

Per prevedere l'incremento di peso conoscere la regressione

L'equazione della linea di regressione

Le grandezze a e b sono i *coefficienti di regressione*; questa denominazione è molto usata in particolare per il coefficiente angolare b , cioè per la inclinazione della linea di regressione.

L'espressione al numeratore della (5.3) è la somma dei prodotti degli scarti di x e di y dalle rispettive medie. Per il calcolo è più utile una formula abbreviata analoga alla (1.3) per il calcolo della devianza. Con un ragionamento simile a quello svolto per derivare la (1.3) si trova che

$$\begin{aligned}\Sigma(x_i - \bar{x})(y_i - \bar{y}) &= \Sigma(x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y}) \\ &= \Sigma x_i y_i - \bar{y} \Sigma x_i - \bar{x} \Sigma y_i + n \bar{x} \bar{y} \\ &= \Sigma x_i y_i - \frac{2(\Sigma x_i)(\Sigma y_i)}{n} + \frac{(\Sigma x_i)(\Sigma y_i)}{n} \\ &= \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}\end{aligned}\quad (5.9)$$

Espressioni equivalenti sono $\Sigma x_i y_i - \bar{x} \Sigma y_i$, $\Sigma x_i y_i - \bar{y} \Sigma x_i$ e $\Sigma x_i y_i - n \bar{x} \bar{y}$. Bisogna osservare che mentre una somma di quadrati di scarti dalla media deve essere una grandezza non negativa, una somma di prodotti di scarti dalla media può essere negativa (in tal caso per la (5.3) anche b sarà negativo).

Esempio 5.1. La Tab. 5.1 dà per 32 bambini i valori di x , il peso alla nascita, e di y , l'incremento di peso tra il 70° e il 100° giorno di vita espresso come percentuale del peso alla nascita. La Fig. 5.4 riporta un diagramma di dispersione che suggerisce un'associazione negativa tra le due variabili. La tendenza sembra ragionevolmente lineare. In base alla Tab. 5.1 il calcolo procede così

$$\begin{aligned}n &= 32 & \Sigma x &= 3576 & \Sigma y &= 2281 \\ & & \bar{x} &= 3576/32 & \bar{y} &= 2281/32 \\ & & &= 111,75 & &= 71,28 \\ \Sigma x^2 &= 409880 & \Sigma xy &= 246032 & \Sigma y^2 &= 179761 \\ (\Sigma x)^2/n &= 399618,00 & (\Sigma x)(\Sigma y)/n &= 254901,75 & (\Sigma y)^2/n &= 162592,53 \\ \Sigma(x - \bar{x})^2 &= 10262,00 & \Sigma(x - \bar{x})(y - \bar{y}) &= -8869,75 & \Sigma(y - \bar{y})^2 &= 17168,47\end{aligned}$$

Per prevedere l'incremento di peso in base al peso alla nascita è necessario conoscere la regressione di y su x

$$b = -\frac{8869,75}{10262,00} = -0,8643$$

L'equazione della linea di regressione è per la (5.5)

$$\begin{aligned}Y &= 71,28 - 0,8643(x - 111,75) \\ &= 167,87 - 0,8643x\end{aligned}$$

Tabella 5.1. *Pesi alla nascita di 32 neonati e incremento di peso tra il 70esimo e il 100esimo giorno dalla nascita espresso come percentuale del peso iniziale.*

Peso alla nascita	Aumento di peso tra il 70° e il 100° giorno come % di x
72	68
112	63
111	66
107	72
119	52
92	75
126	76
80	118
81	120
84	114
115	29
118	42
128	48
128	50
123	69
116	59
125	27
126	60
122	71
126	88
127	63
86	88
142	53
132	50
87	111
123	59
113	76
106	72
103	90
118	68
114	93
94	91

Per disegnare tale linea basta conoscere le coordinate di due punti su di essa ma è più prudente calcolarne tre. Ponendo $x = 80, 100$ e 140 come convenienti numeri compresi nell'intervallo di variazione di x (vedi Fig. 5.4) si ha

x	Y
80	98,73
100	81,44
140	46,87

In Fig. 5.4 è riportata la linea di regressione individuata.

In situazioni in cui si desidera considerare la regressione di y su x , per un dato y ,

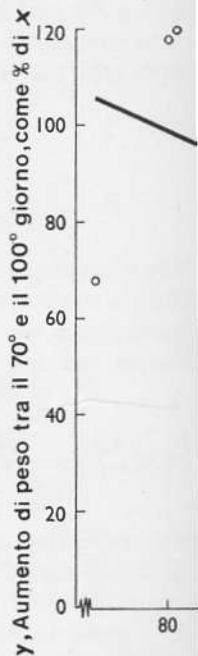


Fig. 5.4. Diagramma di dispersione dei dati relativi al peso alla nascita e all'incremento di peso tra il 70° e 100° giorno espresso come percentuale del peso iniziale.

La linea di regressione di y su x è quella già trovata scambiando i ruoli di x e y . La sua equazione è utile scrivere l'equazione

con $b_{y \cdot x}$ dato dalla (5.10) e $a_{y \cdot x}$ dato dalla (5.11).

con

In Fig. 5.4 è riportata anche la linea di regressione che passa per i punti così individuati.

In situazioni in cui sia x che y sono variabili casuali può essere utile considerare la regressione di x su y ; tale regressione indica come il valor medio di x , per un dato y , varii in funzione di y .

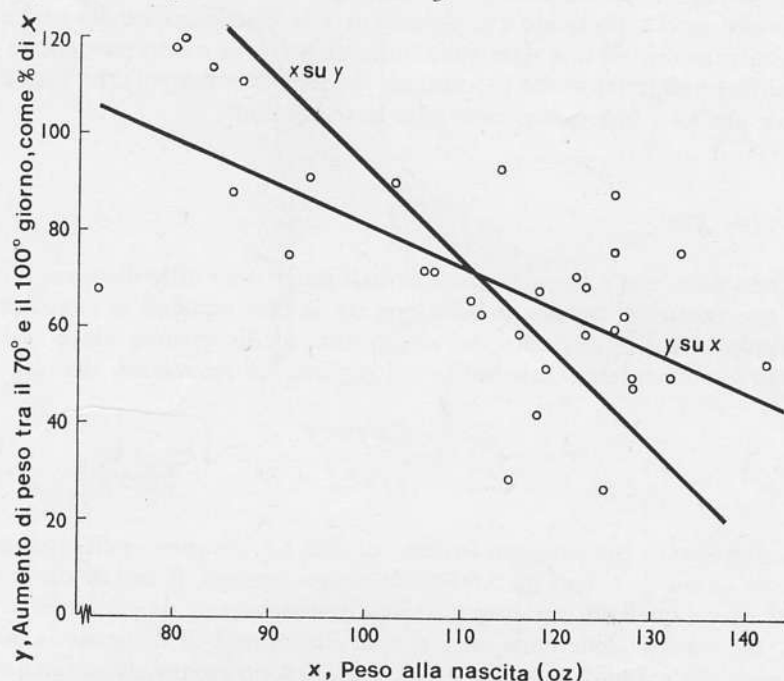


Fig. 5.4. Diagramma di dispersione per peso alla nascita x e incremento di peso tra il 70° e 100° giorno espresso come percentuale di x in 32 neonati; sono riportate anche le due linee di regressione.

La linea di regressione di x su y può calcolarsi con formule analoghe a quelle già trovate scambiando tra loro x e y . Per evitare confusioni tra le due linee è utile scrivere l'equazione di regressione di y su x così

$$Y = \bar{y} + b_{y \cdot x}(x - \bar{x})$$

con $b_{y \cdot x}$ dato dalla (5.3). La linea di regressione di x su y è allora data dall'equazione

$$X = \bar{x} + b_{x \cdot y}(y - \bar{y})$$

con

$$b_{x \cdot y} = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(y - \bar{y})^2}$$

Dalla Fig. 5.4 si vede che le due linee sono in generale differenti; entrambe passano per il punto $(\bar{x}, \bar{y})$, che è perciò il loro punto d'intersezione. Nell'esempio 5.1 interessava principalmente la regressione di y su x poichè è naturale studiare la variazione del peso in funzione del peso alla nascita e analizzare la distribuzione delle variazioni di peso per ogni particolare valore di peso alla nascita piuttosto che determinare la distribuzione dei pesi alla nascita in funzione di una data variazione di peso (in modo particolare se quest'ultima è espressa come percentuale del peso alla nascita). In altre circostanze possono interessare entrambe le regressioni.

5.3. Correlazione

Quando sia x che y sono variabili casuali può essere utile disporre di un indice che misuri di quanto la relazione tra le due variabili si avvicina al caso ideale in cui ogni punto del diagramma di dispersione giace esattamente su una linea retta. Tale indice è il *coefficiente di correlazione* definito da

$$r = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 \Sigma(y - \bar{y})^2}} = \sqrt{b_{y \cdot x} \cdot b_{x \cdot y}} \quad (5.10)$$

Si può dimostrare che per ogni insieme di dati r è compreso nell'intervallo che va da -1 a $+1$. La Fig. 5.5 riporta cinque insiemi di dati in cui la variabilità di x e quella di y non cambiano sostanzialmente da un insieme all'altro. Le sensibili differenze tra i cinque diagrammi di dispersione sono sintetizzate dai valori di r . Le Figg. 5.5 (a) e (e) sono esempi di perfetta correlazione in cui il valore di y è esattamente determinato come funzione lineare di x ; in altri termini i punti giacciono esattamente su una linea retta; se le variabili crescono in modo concorde, come in (a), allora $r = +1$; se invece una decresce al crescere dell'altra, come in (e), allora $r = -1$. In entrambi questi casi le due linee di regressione coincidono. La Fig. 5.5 (c) rappresenta la situazione opposta in cui $r = 0$; i due coefficienti di regressione $b_{y \cdot x}$ e $b_{x \cdot y}$ valgono zero e le linee di regressione sono tra loro perpendicolari. In (b) e in (d) sono rappresentate situazioni intermedie in cui si ha rispettivamente $0 < r < 1$ (b) e $-1 < r < 0$ (d); in questi casi le rette di regressione formano un angolo ma sono orientate sostanzialmente nella stessa direzione.

Dalla formula abbreviata per la somma dei quadrati e dei prodotti si deduce una formula alternativa per il coefficiente di correlazione la quale è più conveniente per il calcolo

$$r = \frac{\Sigma xy - (\Sigma x)(\Sigma y)/n}{\sqrt{(\Sigma x^2 - (\Sigma x)^2/n)(\Sigma y^2 - (\Sigma y)^2/n)}} \quad (5.11)$$

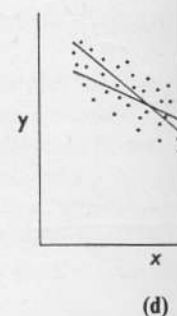
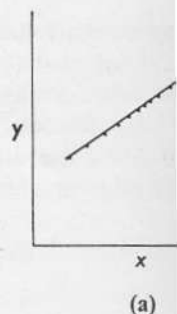


Fig. 5.5. Cinque diagrammi di dispersione con coefficiente di correlazione diverso.

Per la (5.7) la somma dei quadrati di y su x è

$$\Sigma(y - \bar{y})^2 (1 - r^2)$$

che suggerisce un'...