

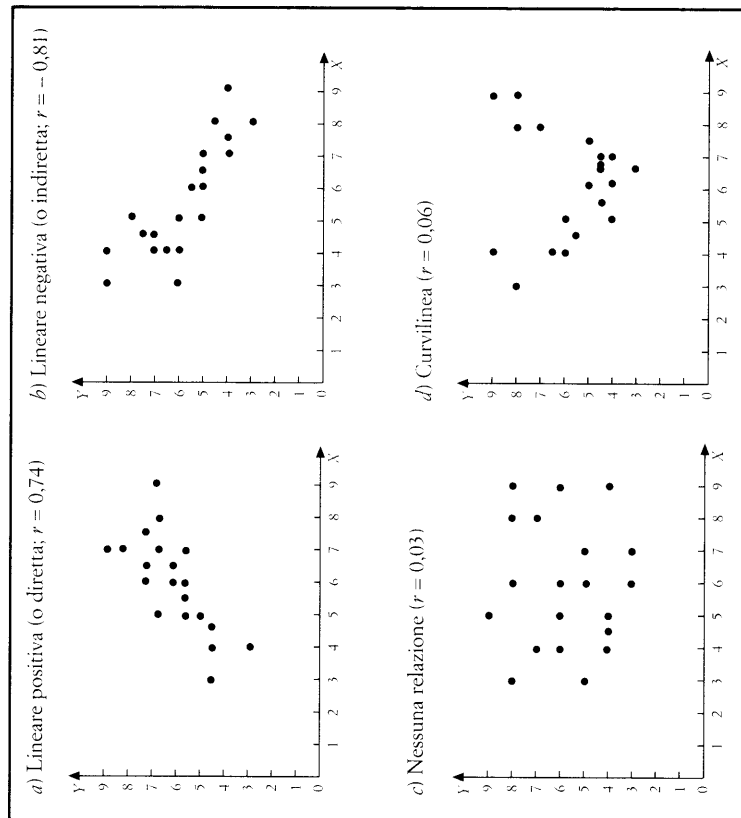


fig. 3.3. Diagrammi di dispersioni raffiguranti quattro tipi di relazioni fra due variabili.

tuaioni base ne dobbiamo aggiungere una quarta, quella della relazione curvilinea, rappresentata in figura 3.3d: abbiamo studenti con voto alto in italiano che hanno voto alto anche in matematica (come nel caso della relazione positiva), però abbiamo anche studenti con voto basso in italiano che hanno voto alto in matematica (come nella relazione negativa).

Le relazioni curvilinee sono più difficili da studiare e su di esse non ci soffermeremo in questa nostra introduzione. Ci concentreremo sulla relazione lineare anche perché essa rappresenta la forma più semplice di relazione fra variabili: è la prima ipotesi che il ricercatore fa ed è naturale che una trattazione semplificata come la nostra ad essa si limiti.

Tuttavia valga questo avvertimento: il fatto di trovare che fra due varia-

bili non c'è relazione sulla base degli strumenti che stiamo qui per presentare, sta a significare assenza di relazione *lineare*, il che non esclude la presenza di relazioni curvilinee (in particolare la fig. 3.3d presenta un coefficiente di correlazione lineare $r = 0,06$, non dissimile da $r = 0,03$ della fig. 3.3c dove effettivamente siamo in assenza di relazione).

7.2. Retta di regressione

Abbiamo detto che quando c'è una relazione (lineare) i punti del diagramma di dispersione tendono ad allinearsi lungo una retta. Data una nuvola di punti, possiamo allora tracciare una retta, che chiamiamo **retta interpolante**, la quale meglio di tutte le altre si avvicini ai punti e sintetizzi (naturalmente con una certa semplificazione) la nuvola. Noi sappiamo che in generale una retta è esprimibile mediante la seguente equazione:

$$Y = a + bX$$

dove a è l'*intercetta* della retta sull'asse delle Y (cioè l'ordinata della retta quando l'ascissa è 0) e b è l'*inclinazione* della retta (cioè la variazione dell'ordinata quando l'ascissa varia di un'unità). Per alcuni esempi si veda la figura 3.4.

Data una nuvola di punti, il problema è dunque quello di determinare i parametri a e b della retta che meglio di tutte le altre è in grado di sintetizzarla. Senza entrare in merito alle dimostrazioni, possiamo dire che la statistica ci permette di determinare a e b a partire dai valori di X e Y sui casi, utilizzando il cosiddetto «metodo dei minimi quadrati», che consiste nel trovare quella retta che ha la proprietà di ridurre al minimo la somma dei quadrati degli scarti fra i valori teorici di Y (sulla retta) ed i valori di Y osservati (nei dati). Le formule sono le seguenti:

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$a = \bar{Y} - b\bar{X}$$

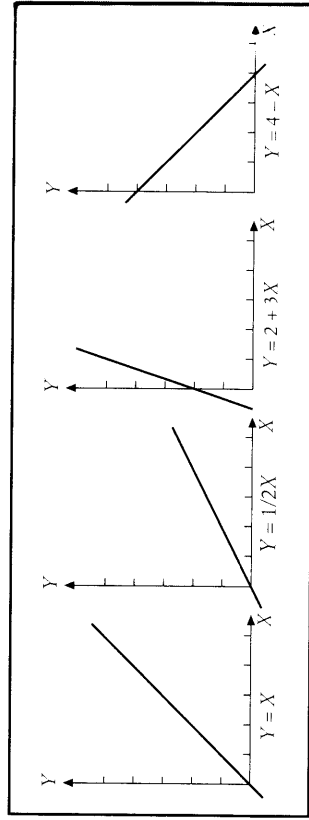


fig. 3.4. Alcuni esempi di rette e loro equazioni.

Chiamiamo la retta definita dall'equazione $Y = a + bX$, dove a e b siano calcolati attraverso le formule appena date, *retta di regressione*³¹; b è il *coefficiente di regressione*, mentre a è l'*intercetta* della retta (o la *costante*³² dell'equazione della retta).

Soffermiamoci un attimo sulla formula di b . Il numeratore è la somma dei prodotti degli scarti di Y e di X dalle rispettive medie, la quale, quando è divisa per il numero di casi N , ci dà la **covarianza**, che indichiamo con S_{YX} :

$$S_{YX} = \frac{\sum (Y_i - \bar{Y})(X_i - \bar{X})}{N}$$

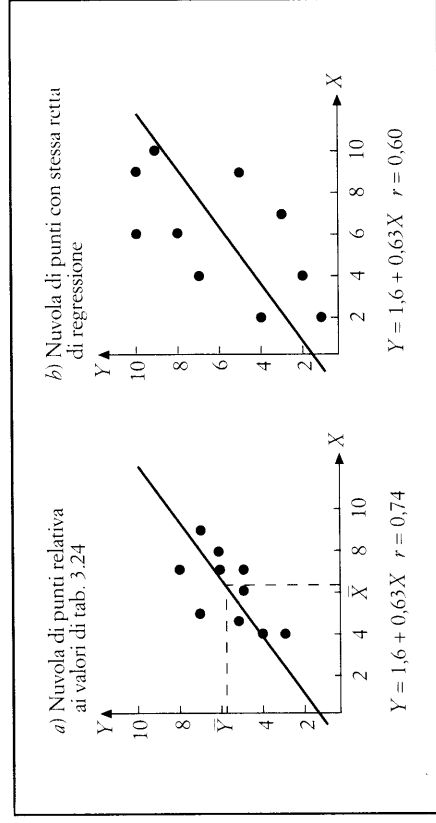
Il denominatore della formula di b è la somma degli scarti al quadrato dei valori di X dalla sua media; dividendoli per N abbiamo la varianza di X :

$$S_X^2 = \frac{\sum (X_i - \bar{X})^2}{N}$$

Dividendo la covarianza di YX per la varianza di X i due N al denominatore si annullano, ottenendo la formula sopra presentata, che quindi può essere scritta anche in questa forma:

$$b = \frac{\text{covarianza fra } X \text{ e } Y}{\text{varianza di } X} = \frac{S_{YX}}{S_X^2}$$

In tabella 3.26 abbiamo riportato i calcoli dei parametri a e b per il no-

fig. 3.5. Diagrammi di dispersione aventi stessa retta di regressione ma differente r .

stro esempio sui voti scolastici. Sulla base di questi valori possiamo tracciare la retta di regressione. In figura 3.5a abbiamo riportato i 10 casi sui quali abbiamo effettuato i calcoli (sui valori di X e Y di tab. 3.26), ed abbiamo tracciato la retta di regressione. Dato che la sua equazione è data da $Y = 1,6 + 0,63X$, e poiché bastano due punti per poter tracciare una retta, basta sostituire due qualsiasi valori di X nell'equazione appena scritta, trovare i due corrispondenti valori di Y e per i due punti così individuati tracciare la retta. In genere si assumono come punti di passaggio della retta la sua intercetta sull'asse Y (nel nostro caso $X = 0$ e $Y = 1,6$) ed il punto definito dalle medie delle due variabili (nel nostro caso $X = 6,2$ e $Y = 5,5$). Infatti una delle proprietà della retta di regressione è che passa per il punto definito dalle medie delle due variabili.

Naturalmente la retta di regressione non dà una rappresentazione perfetta della nuvola dei punti. Essa ci permette di predire Y noto X nell'ipotesi di una relazione lineare perfetta, nel caso cioè che tutti i punti siano sulla retta:

$$\hat{Y} = a + bX$$

Ma i punti non sono mai perfettamente allineati su una retta. L'equazione che ci dà il valore reale di Y è allora la seguente:

$$Y = a + bX + e$$

dove e è l'**errore** o **residuo**, cioè lo scarto fra valore predetto dall'equazione di regressione e valore effettivamente osservato nei dati:

$$Y - \hat{Y} = e$$

Si noti la differente notazione delle due equazioni: $\hat{Y}$ (con l'accento circonflesso) è il valore di Y *stimato* (predetto) dall'equazione di regressione; Y (senza accento circonflesso) è invece il valore osservato.

Possiamo chiederci qual è il significato da attribuire al coefficiente di regressione b , al di là del suo significato geometrico di inclinazione della retta di regressione. Esso può essere interpretato come la variazione che fa registrare la variabile Y al variare di un'unità di misura della variabile X . Se l'unità d'analisi è costituita dalla regione, se X è il tasso di immigrazione della regione e Y il tasso di criminalità, un'equazione $Y = 1,3 + 0,42 X$ ci dice che all'aumentare di una unità del tasso di immigrazione aumenta di 0,42 unità il tasso di criminalità.

Un ultimo problema riguarda la *direzione del legame* fra X e Y . Come abbiamo già accennato, nel modello della regressione Y è la variabile dipendente e X è la variabile indipendente: noi *prevediamo* Y a partire da X e non viceversa. Nel nostro esempio prevediamo il voto di matematica sulla base del voto in italiano. Se invertissimo la collocazione delle due variabili ($Y =$ voto in italiano e $X =$ voto in matematica) otterremmo dei valori di a e b differenti (in particolare otterremmo $b = 0,86$ e $a = 1,4$). Nel nostro esempio sui voti in italiano e in matematica non esiste di fatto una direzione causale (le due variabili sono correlate, ma non ha senso dire che è la prima ad influenzare la seconda o viceversa); tuttavia nella grande maggioranza delle situazioni esiste chiaramente un nesso causale e quindi una direzione della relazione. Naturalmente la decisione su quale sia la variabile indipendente e quale la dipendente è una decisione che il ricercatore compie in tutta autonomia, e la statistica non può aiutarlo. Va detto che in ogni caso, anche quando come nel nostro esempio sui voti non esiste una direzione causale, per applicare la tecnica della regressione il ricercatore deve preventivamente decidere che una variabile è indipendente e l'altra è dipendente (anche se si tratta di una scelta forzata).

7.3. Coefficiente di correlazione

La retta di regressione ci dice qual è la *forma* della relazione fra X e Y : ci dice per esempio che assume una configurazione lineare, sintetizzabile in una retta avente una certa intercetta ed una certa inclinazione. Non ci dice però nulla sulla *forza* (o strettezza, o intensità) di questa relazione. In figura 3.5b abbiamo riportato accanto al diagramma di dispersione sul quale abbiamo calcolato la retta di regressione un secondo diagramma, con una nuvola ancora di 10 punti ed avente la stessa retta di regressione. Ma la seconda nuvola ha i punti assai più dispersi rispetto alla retta di regressione. Pur essendo le due nuvole di punti esprimibili dalla stessa retta ($Y = 1,6 + 0,63X$), appare evidente che nel primo caso le due variabili sono fortemente correlate fra di loro, mentre nel secondo caso la loro relazione è più debole. Si capisce anche che la forza della relazione è collegabile al modo con il quale i punti si distribuiscono attorno alla retta di regressione: se essi sono addensati attorno alla retta, allora la relazione è forte; se sono piuttosto dispersi attorno alla relazione è debole.

Possiamo anche dire che il ricercatore è interessato alla forza della relazione prima ancora che alla sua forma. Infatti in caso di relazione debole, non ha molto senso andare a vederne la forma, andare cioè a calcolare i parametri della retta di regressione: ha poco senso prevedere Y a partire da X se poi questa previsione contiene un ampio margine di errore. In genere quindi nella ricerca sociale prima si vanno a vedere quali sono le variabili indipendenti più correlate con la dipendente, e poi si procede ad esplorare meglio la forma della retta di regressione per quelle variabili che risultano più correlate.

Per misurare la forza di una relazione fra due variabili cardinali si utilizza il coefficiente di correlazione r (detto anche, dal nome dello statista che l'ha utilizzato la prima volta, **coefficiente di correlazione di Pearson**, oppure coefficiente di correlazione «prodotto-momento», oppure ancora coefficiente di correlazione del Bravais³³).

La formula del coefficiente di correlazione è la seguente:

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

l'autore decide di utilizzare tutti e cinque gli indicatori per il suo indice. Altre volte la matrice di correlazione mette in evidenza misure molto differenziate, correlazioni elevate fra alcune variabili e virtuale assenza di correlazione fra altre, e sulla base della sua ispezione il ricercatore può impostare la successiva analisi.

Dopo aver tessuti gli elogi di questa misura della forza della correlazione, avanziamo ora due importanti note di cautela. Innanzitutto, anche se è implicito in quanto siamo finora venuti dicendo, va ricordato che r è un coefficiente di correlazione *lineare*; se fra le variabili c'è una relazione che però non è lineare, r ci dice che non c'è relazione. Si torni indietro alla figura 3.3d: di fronte ad una netta relazione curvilinea fra X e Y , il valore di r è praticamente pari a zero. In questi casi sarà l'ispezione del diagramma di dispersione ad avvertirci dell'esistenza di una relazione curvilinea.

Il secondo problema è quello della sensibilità di r ai valori estremi. In figura 3.6 abbiamo riportato due esempi ancora riferiti ai voti scolastici. Nel primo esempio i punti formano una nuvola dalla quale appare evidentemente assenza di relazione; tuttavia l'esistenza di un caso deviante (uno studente con 10 in matematica e 10 in italiano) «attrae» verso di sé la retta di regressione, dando forma ad una situazione di fatto informe: infatti il coefficiente di correlazione con tutti i punti, compreso quello estremo, è $r = 0,60$; ma se si toglie il caso deviante esso diventa $r = 0,04$, segnalando quindi l'assenza di relazione. Simile è il secondo esempio (fig. 3.6b): questa volta la correlazione fra i punti ci sarebbe, ma il caso anomalo di uno studente con 10 in italiano e 4 in matematica fa deviare verso di sé la retta di regressione creando una situazione di scarso allineamento dei punti attorno alla retta (senza il caso deviante $r = 0,92$, che scende a $r = 0,21$ quando anche il caso anomalo è presente). Si noti che i casi devianti alterano sia i valori di r sia quelli dei parametri a e b . In questi casi sarebbe opportuno condurre l'analisi escludendo i casi devianti.

Per entrambe le situazioni presentate, quella della relazione curvilinea e quella dei casi devianti, l'unico modo per evitare di incorrere in errori grossolani è quello di non accontentarsi dei valori dei coefficienti di correlazione, ma di ispezionare attentamente anche i diagrammi di dispersione.

Rimane da affrontare ancora un problema: quello dell'interpretazione del coefficiente di correlazione. Abbiamo detto che il coefficiente di correla-

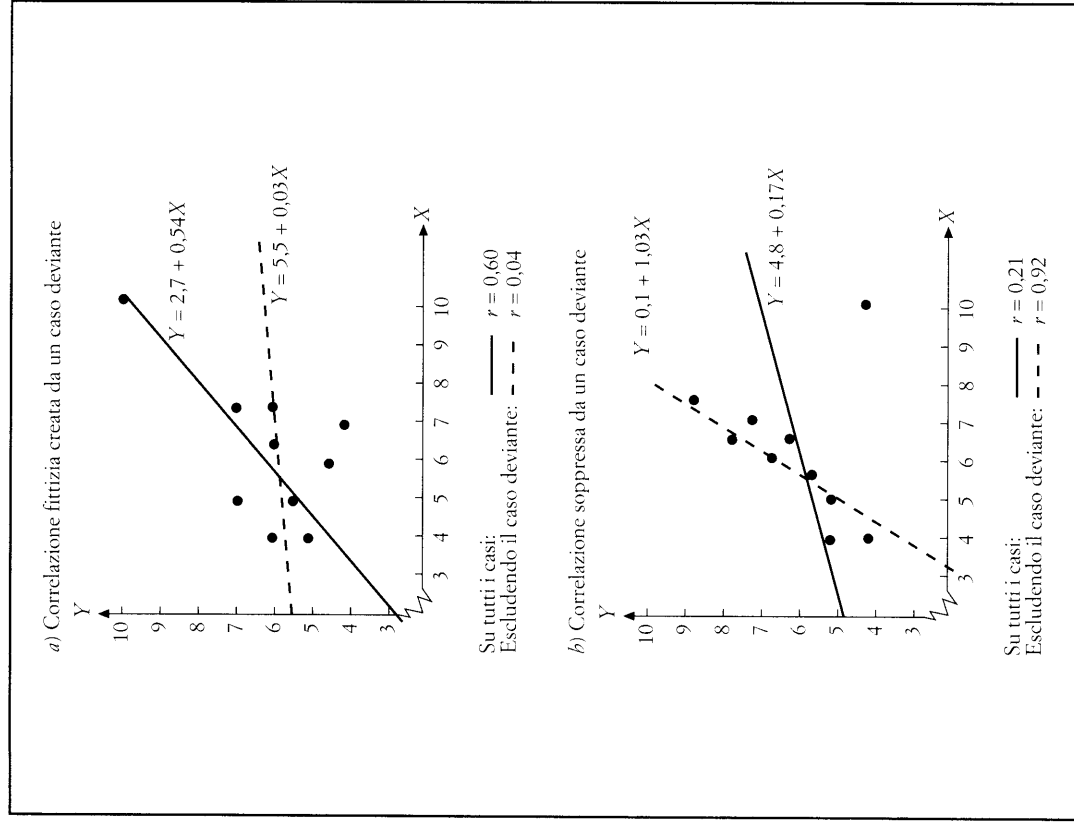


fig. 3.6. Effetto di casi devianti (outliers) sulla retta di regressione.

zione è una misura di quanto i punti siano prossimi alla retta di regressione. Vediamo di chiarire meglio questo concetto e di dargli un'interpretazione non solo geometrica. Come già abbiamo anticipato, se tutti i punti giacesse- ro esattamente sulla retta di regressione, allora i valori predetti dalla retta di regressione $\hat{Y}_i$ coinciderebbero con i valori Y_i osservati. Questo non succe- de, in generale, per cui abbiamo per ogni valore di Y uno scarto $\hat{Y}_i - Y_i$ che è diverso da zero e che chiamiamo *errore* o *residuo* e indichiamo con la lette- ra e (quindi $e_i = \hat{Y}_i - Y_i$).

Seguendo un procedimento assai simile a quello già presentato per l'analisi della varianza, possiamo scrivere lo scarto di un singolo valore dalla media generale nel modo seguente, sottraendo ed aggiungendo il corrispon- dente valore $\hat{Y}$ predetto dall'equazione di regressione:

$$(Y_i - \bar{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

Abbiamo così scomposto lo scarto di ogni singolo valore osservato dalla sua media in due componenti: a) lo scarto del valore osservato dal valore predetto dall'equazione di regressione $(Y_i - \hat{Y}_i)$, scarto che corrisponde all'errore e_i ; b) lo scarto del valore predetto dalla retta di regressione dalla me- dia $(\hat{Y}_i - \bar{Y})$. In figura 3.7 diamo una rappresentazione grafica di questa scomposizione. Si può dimostrare che elevando al quadrato entrambi i membri di questa eguaglianza e sommando su tutti i casi si ottiene:

$$\begin{aligned} \sum (Y_i - \bar{Y})^2 &= \sum (Y_i - \hat{Y}_i)^2 + \sum (\hat{Y}_i - \bar{Y})^2 \\ SQ_{totale} &= SQ_{non\ spiegata} + SQ_{spiegata} \end{aligned}$$

Chiamiamo il primo membro di quest'eguaglianza *somma dei quadrati (o devianza) totale*, pari al numeratore della varianza della variabile dipen- dente Y . Essa risulta scomposta in due componenti: *somma dei quadrati (de- vianza) spiegata* dalla regressione e *somma dei quadrati (devianza) non spie- gata*. In questo modo abbiamo scomposto la variazione della variabile indi- pendente Y in una parte predetta (o «spiegata») dalla retta di regressione ed una parte attribuibile agli errori di predizione.

Ora si può dimostrare che il quadrato del coefficiente di correlazione,

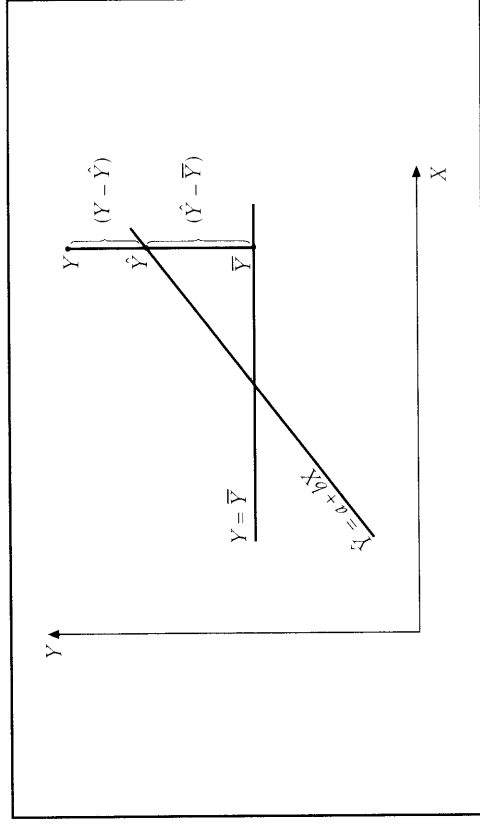


fig. 3.7. Retta di regressione, valore osservato, valore stimato ed errore (residuo) della stima.

detto anche *coefficiente di determinazione*, è pari alla proporzione di variazio- ne totale della variabile dipendente che è spiegata dalla retta di regressione:

$$\text{Coefficiente di determinazione: } r^2 = \frac{SQ_{spiegata}}{SQ_{totale}}$$

Quindi il quadrato del coefficiente di correlazione può essere interpre- tato (in modo del tutto simile a quanto avevamo visto per η^2 nell'analisi della varianza) come quella *parte di variazione della variabile dipendente che è spiegata dalla variazione della variabile indipendente* (potremmo anche dire che è «determinata» dalla variazione della indipendente, da cui il nome di coefficiente di determinazione)³⁴. Se per esempio fra X e Y troviamo che $r = 0,60$, ciò significa che il 36% della variazione di Y è spiegata da X , cioè è attribuibile alla variazione di X . Se il ricercatore se la sente di dare un'inter- pretazione in termini causali, può dire che il 36% della variazione di Y è causato da variazioni di X . Se troviamo una correlazione di $r = 0,60$ fra titolo di studio e reddito, possiamo dire che il 36% delle differenze di reddito fra gli intervistati è attribuibile a differenze di titolo di studio.

Quindi, una correlazione di 0,50 fra due variabili non va interpretata

come la metà di una correlazione perfetta. In questo caso solo il 25% della variazione della variabile dipendente è spiegata dalla variabile indipendente. Per arrivare ad una spiegazione del 50% della variabilità della variabile dipendente occorre arrivare a coefficienti di correlazione attorno a 0,70. Ciò significa che in generale il ricercatore considererà degne di interesse solo le correlazioni con r almeno attorno a 0,30, quando cioè almeno il 10% della variabilità della variabile dipendente è spiegata dalla indipendente.

Va tuttavia detto che nelle scienze sociali è difficile trovare correlazioni elevate (per cui anche valori di r attorno a 0,30 sono considerati sostanzialmente rilevanti, analogamente a quanto avevamo già detto per l' r^2 nell'analisi della varianza). I fenomeni umani sono molto complessi ed è difficile trovare che una variabile «è spiegata» in proporzione elevata da un'altra sola variabile. Infatti ciò che non è spiegato – e che viene indifferentemente chiamato errore o residuo – andrebbe più correttamente collocato sotto la categoria del «residuo» che sotto quella dell'«errore», in quanto va interpretato come variazione di Y dovuta a variabili diverse da X , piuttosto che come errore di osservazione o di campionamento; ed è assolutamente plausibile che in una realtà che di necessità è multivariata, esigua sia la capacità esplicativa di un modello bivariato. Si apre a questo punto l'ampissimo capitolo dell'**analisi multivariata**.

Anche per il coefficiente di correlazione r così come per quello di regressione b esistono dei test di significatività che permettono di stabilire se i valori trovati sono sufficientemente elevati da poter falsificare l'ipotesi nulla. Essi tuttavia non sono molto utilizzati. Per questo motivo e per esigenze di concisione non li riportiamo in questo manuale e rimandiamo il lettore interessato ad un qualsiasi testo di statistica.

7.4. La correlazione ecologica

Nel paragrafo 12 del precedente capitolo abbiamo introdotto l'argomento dei dati aggregati, dei dati cioè per i quali l'unità d'analisi è un aggregato di individui, fra i quali i più importanti sono quelli in cui l'aggregato è un territorio (comune, provincia, regione, contea, ecc.), per la buona ragio-

ne che di questo genere sono i dati delle statistiche ufficiali. Poiché questi dati sono poco costosi in quanto già raccolti dallo stato, per il motivo che sono disponibili anche per fenomeni per i quali sarebbe impossibile raccogliere dati individuali (si pensi solo alle ricerche sul passato), per il fatto che riguardano l'intera popolazione invece che più riduttivi campioni, essi rappresentano una base informativa di fondamentale importanza per la ricerca sociale. Le variabili di questi dati sono per la maggior parte variabili cardinali, come abbiamo già spiegato nel capitolo precedente. È quindi naturale che lo studio della relazione fra due variabili di questo genere (per esempio il tasso di industrializzazione ed il reddito pro capite) avvenga utilizzando gli strumenti tipici delle variabili cardinali, e cioè la regressione e la correlazione. È d'uso, fra gli studiosi, chiamare «correlazione ecologica» la correlazione fra due variabili aggregate (dove evidentemente l'attributo «ecologico»³⁵ fa riferimento non alla difesa del territorio, quanto piuttosto alla dimensione territoriale del fenomeno studiato). In questo modo contrapponiamo *correlazione ecologica*, dove l'unità d'analisi è un aggregato di individui (non necessariamente un aggregato territoriale), a *correlazione individuale* dove l'unità d'analisi è l'individuo (per esempio dati derivanti da questionari).

Nella maggioranza dei casi, l'interrogativo del ricercatore sociale è riferito agli individui: per esempio, come votano gli operai? Per rispondere a questa domanda egli dovrà intervistare un campione di operai, con grande dispendio di energie e di costi. Ma poiché egli dispone, per esempio per tutti i comuni della nazione, delle percentuali di voti ai partiti e della percentuale di operai residenti, non potrebbe forse egli dalla correlazione ecologica (sui comuni) fra percentuale di operai e percentuale di voti al partito A, al partito B, ecc. risalire alla correlazione individuale (sugli individui) fra partito votato e classe sociale?

In linea di massima la risposta a questa domanda è negativa. In un noto articolo del 1950, Robinson dimostrò che la correlazione ecologica è sempre più elevata di quella individuale, con differenze talvolta stupefacenti che capovolgono i risultati: per esempio trovò che la correlazione ecologica fra la percentuale di neri e di analfabeti era negli Stati Uniti $r = 0,946$ quando le unità d'analisi erano le nove divisioni geografiche del Bureau of Census, scendeva (pur rimanendo altissima) a 0,773 quando i dati erano a livello di stato, mentre diventava trascurabile, crollando allo 0,203, quando il dato era