

Hubert M. Blalock jr.

Statistica per la ricerca sociale



Fondamenti e metodi per l'analisi empirica nelle scienze sociali a.a. 2012-2013
Corso di laurea in **Teorie, culture e tecniche per il Servizio sociale**
Università degli Studi di Macerata

IL MULINO

Correlazione e regressione (I)

In questo capitolo e in quello immediatamente successivo prenderemo in considerazione le relazioni tra due scale a intervalli. Nel capitolo XIX, in cui affronteremo la correlazione multipla e la correlazione parziale, estenderemo la trattazione al caso in cui si esaminino contemporaneamente tre o più variabili a scala a intervalli. Per il momento, peraltro, ci limiteremo al caso in cui si abbiano due misurazioni al livello di scala a intervalli per ogni soggetto: ad esempio, si può avere, per ognuno degli adulti maschi di una determinata comunità, il grado di istruzione espresso in numero di anni di scuola e il reddito annuo, oppure per ogni città la percentuale di addetti all'industria e il tasso di crescita della popolazione.

In problemi di questo tipo, spesso, non siamo interessati solo nel verificare se la relazione esiste o meno e ad avere una misura della «strettezza» della relazione, ma anche a descrivere la *natura* della relazione tra le due variabili, cosicché si possa predire il valore di una variabile conoscendo l'altra. Per esempio, possiamo voler prevedere il reddito futuro di una persona dal numero di anni per cui ha frequentato la scuola o il tasso di crescita di una città dalla percentuale dei suoi cittadini addetti all'industria. Quando vogliamo soprattutto sapere, a scopo esplorativo, *quali* variabili sono connesse a una determinata variabile, ci interesseranno soprattutto le misure della «strettezza» della relazione, quali, ad esempio, i coefficienti di correlazione; d'altro lato, una volta trovate quali sono le variabili connesse in modo significativo, è molto probabile che sia necessario ricorrere all'analisi della regressione per poter predire con esattezza una variabile utilizzando l'altra.

Il lettore dovrebbe a questo punto conoscere bene i *tests* di significatività e le misure di associazione; purtuttavia, è bene iniziare la nostra trattazione studiando i particolari problemi

connessi alla cosiddetta «predizione». Ciò perché il concetto di regressione precede logicamente quello di correlazione ed è più importante di esso sotto il profilo teorico; gli argomenti che verremo via via sviluppando lo proveranno. Dopo aver parlato della predizione, passeremo a trattare delle misure della strettezza della relazione. Nel capitolo XVIII, che costituisce una continuazione del presente, esamineremo diversi *tests* di significatività, nonché il coefficiente di cograduazione che può essere usato per verificare l'esistenza o meno di una relazione tra due scale ordinali.

17.1. Regressione lineare e minimi quadrati

In certo qual modo lo scopo ultimo di tutte le scienze è la predizione: ciò non significa naturalmente che sia di secondaria importanza il «capiere» o interpretare in termini di causa ed effetto perché due o più variabili sono connesse tra di loro. Forse è più corretto dire che questa «comprensione» è lo scopo ultimo e che possiamo predire in modo sempre più corretto solo nella misura in cui capiamo sempre meglio il perché della connessione. Forse, se non vi fossero limiti alla nostra comprensione, potremmo effettuare predizioni senza sbagliare mai. Per esempio, se conosciamo *sia* le leggi del moto planetario, *che* il campo gravitazionale all'interno del sistema solare *e* la posizione e la velocità di Venere in un determinato momento, potremo predirne il moto futuro. Ma, anche se prescindiamo da questa posizione rigidamente deterministica, resta il fatto che ogni scienza ha come scopo la predizione.

In sociologia e nelle altre scienze sociali le predizioni sono, spesso per mera necessità, espresse in termini molto vaghi. Generalmente, questo avviene perché non è possibile raggiungere il livello di misurazione della scala a intervalli. Così, possiamo predire che, man mano che aumenta lo status di un soggetto all'interno di un gruppo, egli si adegnerà in misura sempre maggiore alle norme del gruppo stesso. Un'affermazione del genere non implica che esista una relazione causale unilaterale; dice semplicemente che esiste una relazione positiva tra status e conformismo. Usando per analogia un termine matematico, il cui uso non sarebbe in questo caso del tutto giustificato, diciamo che lo status è una *funzione* del conformismo o che il

conformismo è funzione dello status, evitando di toccare l'argomento della causalità. Si noti, tuttavia, che in tale espressione non si cerca di descrivere *la natura* della relazione, se non in quanto ci si limita ad affermare che essa è positiva. A meno che tutte e due le variabili non siano al livello di scala a intervalli è molto difficile poter dire di più.

Ma supponiamo di trovarci di fronte a due scale a intervalli: diviene allora possibile descrivere con maggiore accuratezza come una variabile varia col variare dell'altra. Per esempio, possiamo essere in grado di dire che ogni anno di istruzione fa sì che il reddito aumenti di mille dollari. In questo caso, ci troveremo di fronte a una relazione molto semplice, cioè a una relazione lineare, o a linea retta. Ma la maggior parte delle relazioni non sono così semplici, anche se, come meglio vedremo in seguito, è spesso possibile avvicinarsi molto alla natura reale della relazione, pur postulando la sua linearità. Il modo più elegante, e nel tempo stesso più semplice, di descrivere la relazione che esiste tra due o più variabili consiste nel servirsi di un'equazione matematica. Probabilmente lo studente conosce le leggi fisiche che regolano la relazione tra pressione, temperatura e volume dei gas ($PV/T=k$), oppure che indicano che relazione vi è tra l'accelerazione di un corpo che cade, la distanza percorsa e il tempo impiegato a cadere. Ciascuna di queste equazioni matematiche può essere espressa anche mediante un tipo di curva geometrica. Fortunatamente, in sociologia abbiamo generalmente a che fare con equazioni molto semplici e con le curve più semplici possibili (linee rette). Man mano che aggiungiamo nuove variabili, diviene sempre più difficile rappresentare la relazione con curve geometriche, perché non abbiamo a disposizione tutte le dimensioni necessarie, ma per il momento possiamo trascurare questo aspetto del problema.

Supponiamo che si abbia una variabile dipendente Y che debba essere predetta a partire da una variabile indipendente X . In qualche problema sarà chiaro che X precede Y nel tempo: così, per esempio, normalmente la scuola precede il guadagno di un reddito. In questi casi è possibile un'interpretazione del genere, ma dobbiamo essere molto cauti ed evitare di dire sia che esiste una relazione causale, sia che X è l'unica variabile che esercita influenza su Y . Se la direzione del rapporto di causalità è ambigua o se si pensa che ciascuna delle due variabili sia causa dell'altra, si può fornire una *spiegazione* teorica del rapporto

solo facendo ricorso a modelli che si basano su sistemi di equazioni che non possono essere trattati in questo libro¹. Se il nostro fine è semplicemente quello di mettere a punto stime o quello di predire «a breve» $\hat{Y}$ da X , non vi sono ambiguità di questo tipo, ma bisogna sempre tener presente che nella statistica non vi è nulla che possa impedirci di effettuare operazioni matematiche che dal punto di vista teorico non hanno alcun significato. In questo capitolo ed in quelli che seguono, daremo per scontato che la variabile Y , scelta come variabile dipendente in senso matematico, è anche variabile dipendente dal punto di vista della causalità, così che l'interpretazione teorica dei risultati non presenta difficoltà.

Abbiamo già visto che, nel caso X e Y siano tra di loro indipendenti, non ci è possibile predire Y partendo da X , o meglio che il conoscere X non migliora le nostre capacità predittive in merito a Y . Se le variabili non sono indipendenti, invece, il conoscere X ci aiuta a predire Y , e tanto più stretto è il rapporto di dipendenza tra le variabili, tanto più accurata è la predizione. In seguito vedremo come misurare la forza della relazione per mezzo dei coefficienti di correlazione; per ora, però, concentriamoci sul *come* possiamo predire Y partendo da X . Per esempio, possiamo voler prevedere quale sarà il reddito futuro di un soggetto, sapendo che ha un diploma di scuola media inferiore, cioè, complessivamente, ha frequentato la scuola media per tre anni. Se non conoscessimo il grado di istruzione, il modo migliore a nostra disposizione per la valutazione del reddito consisterebbe nel fare riferimento al reddito medio di tutti i maschi adulti (ovviamente nel caso che non vi sia inflazione), ma, dal momento che conosciamo il grado di istruzione, dovremmo essere in grado di effettuare una predizione più accurata.

L'equazione di regressione. Il problema può essere formulato in questi termini. Immaginiamo che per ogni dato valore della variabile indipendente X (istruzione) abbiamo una distribuzione di Y (redditi). In altre parole, per ogni diverso livello di istruzione esisterà nell'universo una certa distribuzione di redditi: non

tutte le persone che hanno un diploma di scuola media avranno lo stesso reddito, ma i loro redditi si distribuiranno attorno a una media. Vi sarà una di queste distribuzioni per i diplomati delle scuole elementari, un'altra per i diplomati della scuola media superiore, una terza per i laureati, ecc. Ciascuna diversa distribuzione di redditi (corrispondenti a un valore fisso di X) avrà una media e potremo individuare il valore di questa media con un punto utilizzando il sistema degli assi ortogonali che già conosciamo. La linea o luogo che unisce tutte queste medie di Y corrispondenti a un valore fisso di X , viene indicata come la *curva di regressione di Y su X* . La figura 17.1 rappresenta un'equazione di regressione.

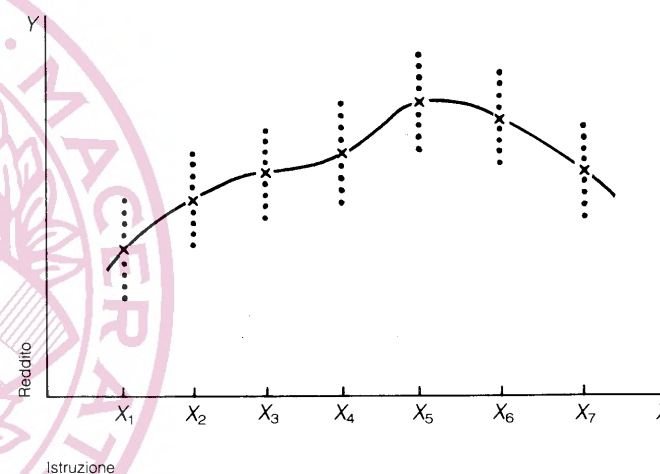


FIG. 17.1. Forma generale della regressione di Y su X ovvero luogo delle medie dei valori di Y corrispondenti a determinati valori di X

Le equazioni di regressione sono le «leggi» di una scienza. In qualche caso la variabilità dei valori attorno all'equazione di regressione è minima; in questi casi sono possibili predizioni molto accurate, ed eventuali deviazioni dalla «legge» sono spesso considerate come frutto di errori di misurazione o effetto di altre variabili minori che non è possibile «controllare» completamente. La «legge» viene, quindi, formulata come se esistesse una relazione perfetta tra X e Y . Idealmente, tutti i punti dovrebbero cadere esattamente sulla curva e la relazione potrebbe essere considerata come una funzione matematica perfetta nella quale

¹ Si vedano C. Christ, *Econometric Models and Methods*, New York, John Wiley and Sons, 1966, p. III; J. Johnston, *Econometric Methods*, New York, McGraw-Hill, 1972², trad. it. *Econometria*, Milano, Angeli, 1978³; N. K. Namboodiri, L. F. Carter, H. M. Blalock, *Applied Multivariate Analysis and Experimental Designs*, New York, McGraw-Hill, 1975, capitoli IV, X e XI.

vi è un solo Y per ogni X . Nelle scienze sociali non possiamo raggiungere una tale perfezione e neppure avvicinarci ad essa. Normalmente, ci aspettiamo una notevole variabilità attorno all'equazione di regressione e preferiamo fare riferimento alle medie e alle varianze delle distribuzioni di Y per ogni X . Ciononostante, in linea di principio, il procedimento è lo stesso in tutte le scienze, anche se le «leggi» delle scienze sociali non possono raggiungere né approssimare la precisione delle leggi della fisica.

Nella figura 17.1 abbiamo indicato la natura in generale delle equazioni di regressione in termini di linea passante per le medie dei valori di Y corrispondenti ai valori dati di X . Dovremo ora semplificare notevolmente i termini del problema per renderlo più facile sotto il profilo statistico. Benché il concetto di regressione sia di applicazione generale, la maggior parte della statistica conosciuta si basa soltanto sul modello più semplice di essa. In particolare, diamo per scontato per ora 1) che la forma dell'equazione di regressione sia lineare, 2) che le distribuzioni dei valori di Y per ogni valore dato di X siano di forma normale, e 3) che le varianze delle distribuzioni di Y per ogni valore di X siano uguali. Esamineremo ora questi requisiti uno dopo l'altro, dedicando però un'attenzione maggiore al primo.

Se la regressione di Y su X è lineare, ovvero se la relazione è a linea retta, possiamo scrivere un'equazione come segue:

$$Y = \alpha + \beta X \quad (17.1)$$

in cui sia α che β sono costanti.

L'equazione (17.1) implica che il rapporto tra X e Y sia esatto, ma presto introdurremo nella equazione un termine di errore o «residuo». Un altro modo di scrivere la stessa equazione è $E(Y/X) = \alpha + \beta X$, in cui la simbologia $E(X/Y)$ sottolinea che ci stiamo occupando del valore atteso di Y che dipende da X . L'inclinazione della linea di regressione è data da β dal momento che questa costante indica di quante unità cambia Y per una determinata variazione unitaria in relazione a un determinato X . Il fatto che la relazione sia lineare, significa che, per ogni determinata variazione di X , diciamo cinque unità, si verifica sempre la stessa variazione in Y (cioè cinque volte β), qualunque sia la posizione sull'asse delle X . (Si veda la figura 17.2). È facile vedere che se $\beta = 1$ e se le unità di X e di Y sono rappresentate

da distanze uguali sui rispettivi assi, la retta di regressione verrà ad essere inclinata di 45° rispetto all'asse delle X ; un β maggiore dell'unità indicherà un'inclinazione maggiore e, tanto più sarà inclinata la retta, tanto maggiore sarà la variazione di Y per ogni determinata variazione di X . Allo stesso modo, se β è minore dell'unità, ma maggiore di zero, occorrerà una variazione maggiore di X per produrre una variazione di Y . Nel caso limite in cui la retta sia perfettamente orizzontale, β diviene zero e ogni variazione di X non apporterà nessuna variazione di Y . In altre parole, se $\beta = 0$, non esiste relazione lineare tra X e Y : il conoscere X non aiuta a predire Y sulla base di un modello lineare². Se β è negativo, sappiamo che la relazione tra le due variabili è negativa: coll'aumentare di X , Y diminuisce.

Una linea retta può essere completamente determinata se si conoscono due punti di essa, o un solo punto e l'inclinazione della retta. Pertanto esiste una e una sola retta con equazione $Y = \alpha + \beta X$, dato naturalmente per scontato che α e β siano quantità costanti anche se generiche. Se sono dati sia α che β , possiamo tracciare la retta prendendo in considerazione semplicemente due punti di essa. Sappiamo che quando $X = 0$, $Y = \alpha$; pertanto il punto $(0, \alpha)$ giace sulla retta; inoltre, quando $Y = 0$ abbiamo $0 = \alpha + \beta X$, cioè $X = -\alpha/\beta$. Questo punto $(-\alpha/\beta, 0)$ è naturalmente il punto in cui la retta incrocia l'asse delle X . Se non è comodo individuare questi due punti, se ne possono individuare altri con lo stesso metodo³.

Requisiti relativi alla X e residuo. Sinora non ci siamo occupati esplicitamente del fatto che, dal momento che i punti sono dispersi attorno alla curva di regressione, dovremmo rappresentare il valore effettivo di Y relativo ad ogni caso mediante una equazione che contiene un termine di disturbo cioè un residuo specifico per ogni caso. Se indichiamo con Y_i e X_i i valori relativi all' i -esimo caso, possiamo rappresentare la relazione (lineare) come segue

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

dove ε_i rappresenta il residuo di cui dobbiamo ora occuparci. Questo termine può essere visto come risultante contemporanea-

² Come meglio vedremo in seguito, se le due variabili sono indipendenti, β sarà sempre zero, ma, se β è zero, le variabili possono anche non essere indipendenti.

³ Per un esempio si veda *infra*, pp. 494-495.

mente da errori nella misurazione di Y (ma non di X) e da tutte quelle cause di Y che non sono state esplicitamente introdotte nell'equazione. Se per la maggior parte queste cause trascurate hanno effetti di minor conto quando sono prese una per una e se agiscono quasi indipendentemente l'una dall'altra, sarà ragionevole dare per scontato che il valore atteso del residuo, $E(\varepsilon_i)$, sia zero e che gli ε_i si distribuiscano normalmente. Cosa importantissima, il residuo sarà indipendente, statisticamente, da X . Si può dimostrare che nell'usare il metodo dei minimi quadrati per stimare i coefficienti di regressione α e β è necessario dare per scontato che $E(\varepsilon)=0$ e che X_i ed ε non sono correlati. Il requisito della distribuzione normale e quello della omoscedasticità, cioè che σ_i^2 sia costante per tutti livelli di X , saranno necessari per i *tests* di significatività e per la determinazione degli intervalli di fiducia.

Il requisito veramente determinante alla base della analisi della regressione è quello relativo alla reciproca indipendenza di X e del residuo. Nelle applicazioni sperimentali spesso siamo in grado di scegliere livelli costanti di X (come, ad esempio, quando manteniamo la temperatura costante ad intervalli di 50 gradi). In casi del genere, dal momento che il livello della X è tenuto sotto controllo e, presumibilmente, non manipolato in modo che vari sistematicamente al variare del residuo, non dobbiamo in genere preoccuparci dell'esistenza del requisito, ma a pensarci bene, risulterà che in molti esperimenti non è detto che le cose vadano effettivamente così, perché, nel manipolare X , è possibile che vengano alterati anche altri fattori lasciati fuori dalla equazione e pertanto inglobati nel residuo. Nella ricerca non sperimentale normalmente X e Y sono osservati e non manipolati e pertanto sia X che Y sono variabili casuali o *stocastiche* che hanno una distribuzione di probabilità. A volte la distribuzione di X sarà approssimativamente normale, anche se questo non è necessario per l'analisi della regressione. Ma è *essenziale* che stabiliamo qualcosa in merito alla distribuzione congiunta di X e del residuo ε_i . Se fin dall'inizio abbiamo qualche elemento per indicare una distribuzione particolare, questo può bastare; ma, sfortunatamente, quasi sempre nella pratica non abbiamo informazioni del genere. Nella grandissima maggioranza dei casi diamo per scontato che X_i ed ε_i siano statisticamente indipendenti, il che risulta giustificato se le cause trascurate di Y sono *i*) numerose, poco importanti quando sono

prese una per una e non correlate tra di loro oppure *ii*) non correlate ad X quando uno o due fattori esclusi predominano. Se non si è disposti a dare per scontato tutto questo in un caso particolare, si dovrà cercare di identificare i più importanti fattori di disturbo trascurati per introdurli esplicitamente nell'equazione come ulteriori variabili. Nel capitolo XIX parleremo della regressione multipla nella quale sono appunto inclusi ulteriori fattori causali.

Uno dei vantaggi offerti dalla teoria statistica della regressione consiste nel fatto che essa è abbastanza sviluppata da specificare in termini espliciti quali sono i requisiti che debbono essere dati per scontati a proposito del residuo. Dovrebbe risultare ovvio che ciò che abbiamo detto a proposito del comportamento delle variabili trascurate si applica anche a *tutti* i procedimenti dei quali ci siamo occupati sinora. Per esempio, se si riscontra una differenza statisticamente significativa tra medie o proporzioni e se si vuole attribuire un ruolo causale alla variabile indipendente (per esempio, il sesso) in questo rapporto, bisognerà anche essere disposti a dare per scontato che le variabili trascurate non sono connesse in modo sistematico alla scala nominale dicotomizzata (ad esempio, il sesso). Non si può aggirare il problema costituito dalla mancanza di requisiti cambiando semplicemente il tipo di analisi e sperando che con questo il problema scompaia.

Usiamo le lettere greche perché per il momento abbiamo a che fare con tutto l'universo. In un'equazione del genere sia α che β possono essere interpretati in termini geometrici. Se poniamo X uguale a zero, vediamo subito che $Y=\alpha$; pertanto α rappresenta il punto in cui la linea di regressione incrocia l'asse delle Y (cioè in cui $X=0$).

Abbiamo detto in precedenza che dobbiamo supporre che, per ogni valore di X , valori di Y si distribuiscano normalmente attorno alla media; sarà anche opportuno supporre, che, per gli X stocastici per ogni valore di Y , i valori di X si distribuiscano normalmente. Diciamo che la distribuzione congiunta di X e Y è una *distribuzione normale bivariata* e intendiamo con questo che abbiamo due variabili ciascuna delle quali è distribuita normalmente attorno all'altra. Una distribuzione normale bivariata può essere rappresentata da una superficie a tre dimensioni come quella raffigurata alla figura 17.3. L'altezza della superficie a ogni determinato punto (X, Y) è proporzionale al numero di casi

che si situano a quel punto. È quindi necessario un diagramma a tre dimensioni per rappresentare la distribuzione congiunta di X e Y , proprio come era necessario un diagramma a due dimensioni per rappresentare la distribuzione di frequenza del solo X . L'esatta forma della distribuzione, che assomiglia molto a un cappello a pan di zucchero, dipenderà da quanto forte è la relazione tra le due variabili. Se le due variabili sono ambedue espresse in termini di unità di scarto quadratico medio, tanto più è stretta la relazione, tanto più stretto è il cappello. Nel caso estremo, in cui Y è predetto esattamente da X , e pertanto tutti i punti giacciono esattamente sulla retta di regressione, lo scarto quadratico medio delle Y per ogni determinato X è zero, e pertanto è zero anche lo spessore del cappello. D'altro lato, nel caso che non vi sia relazione tra X e Y il cappello viene ad avere una base quasi esattamente circolare. Ogni piano perpendicolare al piano XY taglierà la superficie secondo una curva normale e ogni piano parallelo al piano XY la taglierà formando una ellissi.

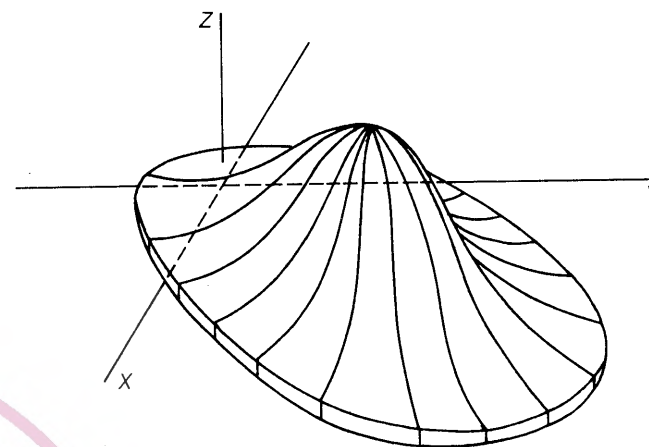


FIG. 17.3. La distribuzione normale bivariata (riprodotto, per gentile concessione, da A. M. Mood, *Introduction to the Theory of Statistics*, New York, McGraw-Hill Book Company, Inc., 1950, figura 41, p. 165)

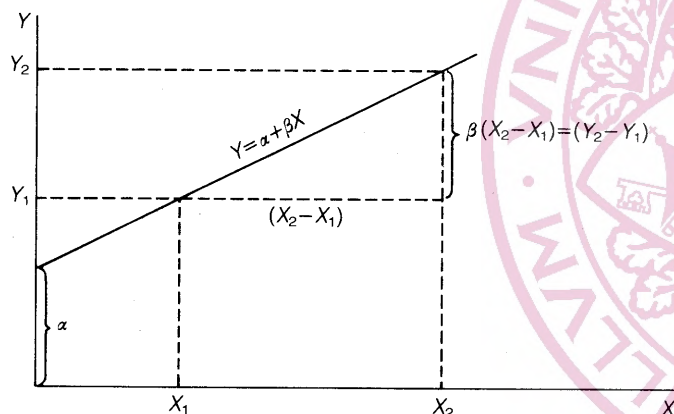


FIG. 17.2. Equazione di regressione lineare, con interpretazioni geometriche di α e di β

La distribuzione normale bivariata ha la proprietà che la regressione di Y su X è lineare; pertanto, se abbiamo una distribuzione normale bivariata, sappiamo che nel caso che tracciamo una linea attraverso le medie dei valori di Y corrispondenti a un determinato valore di X , otteniamo una retta. Naturalmente, non è sempre vero il contrario, non è detto, cioè, che

se la regressione è lineare la distribuzione congiunta delle due variabili sia una normale bivariata.

Per i *tests* di significatività, dovremo anche dare per scontato che gli scarti quadratici medi delle Y siano uguali per ogni X , indipendentemente dal valore che X assume; avremo modo di parlarne a proposito della correlazione, dal momento che la correlazione è sostanzialmente una misura della dispersione attorno alla retta di regressione. Per il momento è sufficiente affermare che, se la distribuzione congiunta è una bivariata normale, gli scarti quadrati delle Y per ciascuna delle X saranno effettivamente identici. Questa proprietà di avere le varianze uguali è nota come *omoscedasticità* ed è analoga a quella che abbiamo assunto come requisito a proposito dell'analisi della varianza e cioè che $\sigma_1 = \sigma_2 = \dots = \sigma_k$.

Metodo dei minimi quadrati; ipotesi lineare. Il modello di regressione che abbiamo preso in considerazione sinora è piuttosto semplice, ma, sfortunatamente, nella sua forma teorica, non presenta una grande utilità. Avviene, infatti, molto di rado che si abbia a disposizione un numero di casi sufficiente per esaminare la distribuzione dei valori di Y per diversi valori consecutivi di

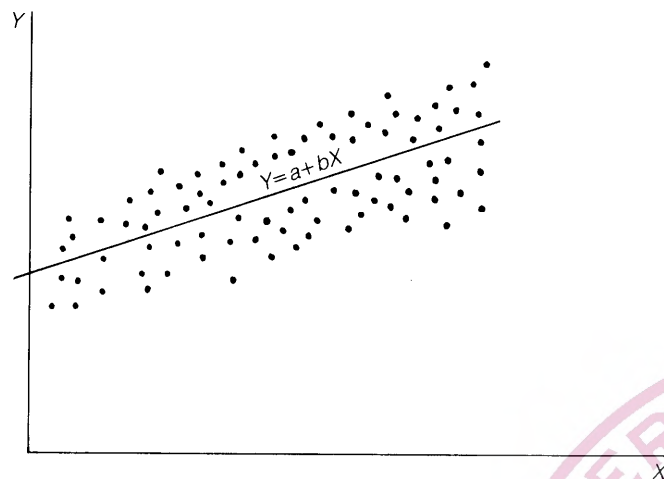


FIG. 17.4. Diagramma a nube di punti e retta individuata col metodo dei minimi quadrati

X ; in generale si riscontra che esistono relativamente pochi casi in cui gli X sono identici o quasi identici. Se riportiamo i punti immagine dei casi osservati su di un piano cartesiano X, Y nel modo ormai consueto, otteniamo una serie dispersa di punti del genere di quella illustrata alla figura 17.4. Quando effettuiamo una distribuzione dei punti con tale sistema, otteniamo ciò che generalmente viene denominato un *diagramma a nube di punti*, o *diagramma di dispersione*. È bene prendere l'abitudine di tracciare un diagramma a nube di punti in ogni caso prima di procedere con l'analisi; in molti casi l'osservare il diagramma sarà sufficiente a indicare se vale la pena di procedere. Per esempio, se i punti appaiono dispersi più o meno a caso per tutto il piano, è chiaro che tra le due variabili non vi è nessuna relazione, o che comunque la relazione è molto debole.

Dopo aver proiettato i vari valori su di un diagramma, si può voler tracciare un tipo di curva che sia quella che meglio si adatta ai punti; questo si può fare tracciando «ad occhio» una curva (nel nostro caso una retta), ma esistono metodi molto più scientifici per farlo. Uno di questi metodi è il metodo dei minimi quadrati del quale ci occuperemo in questo paragrafo. Il nostro scopo è ora, però, diverso da quello che avevamo in precedenza, quando parlavamo dell'analisi della regressione. Allora tracciava-

mo una linea attraverso le medie delle distribuzioni degli Y : ora vogliamo avvicinarci il più possibile ai punti usando una curva qualsiasi che più delle altre si adatti ai punti stessi.

Per potere usare il metodo dei minimi quadrati dobbiamo stabilire a priori che tipo di curva vogliamo usare. Nel caso dell'analisi della regressione, la forma della curva viene ad essere effettivamente determinata dalla linea passante attraverso le medie delle distribuzioni degli Y , se sono disponibili tutti i dati relativi all'universo. Ancora una volta ci serviremo della curva più semplice, la linea retta, come curva dei minimi quadrati. Ciò significa che dovremo adattare ai dati la linea retta più appropriata, determinata secondo il metodo dei minimi quadrati, ottenendo un'equazione del tipo

$$Y = a + bX \quad (17.2)$$

Risulterà che i valori di a e b ottenuti con questo metodo sono le stime senza errore sistematico più efficienti dei parametri dell'universo α e β , se l'equazione di regressione è in effetti una retta e diamo per scontato *a*) la casualità del campionamento, *b*) che $E(\varepsilon_i) = 0$ e *c*) che X_i ed ε_i sono statisticamente indipendenti.

Il metodo dei minimi quadrati consiste nel trovare quell'unica retta che ha la proprietà di ridurre al minimo la somma dei quadrati degli scarti tra i valori teorici sulla linea stessa e i valori empirici di Y . Pertanto, se misuriamo gli scarti dei punti sulle linee verticali tracciate da ciascuno dei punti stessi alla retta dei minimi quadrati, eleviamo al quadrato tali scarti e li sommiamo, la somma risultante sarà minore di quella ottenibile mediante qualsiasi altra linea retta. (Si veda la figura 17.5). Si noti che abbiamo parlato di scarti in verticale e non di distanze sulla perpendicolare o di scarti in orizzontale; è, naturalmente, possibile tracciare una linea che renda minime le distanze sulle perpendicolari alla retta (il cosiddetto metodo dei minimi quadrati ortogonali), ma le equazioni che si ottengono con quest'ultimo metodo sono molto meno convenienti. Se si usassero gli scarti in orizzontale, la linea dei minimi quadrati calcolata in tal modo potrebbe servire per stimare la regressione di X su Y . È facile vedere come la retta che si ottiene minimizzando la somma dei quadrati degli scarti in verticale, non renda necessariamente

minima anche la somma dei quadrati degli scarti calcolati in orizzontale. Possiamo, così, ottenere due diverse rette dei minimi quadrati e solo nel caso in cui tutti i punti giacciono esattamente sulla stessa retta le due rette coincideranno. Rendendo minima la somma dei quadrati delle distanze verticali, si trova la retta che ha la proprietà di rendere uguale a zero la somma degli scarti positivi e negativi e di rendere minimo lo scarto quadratico medio dei punti rispetto alla retta stessa. Riprenderemo in seguito e con maggiori particolari il concetto di scarto quadratico medio delle Y rispetto alla retta.

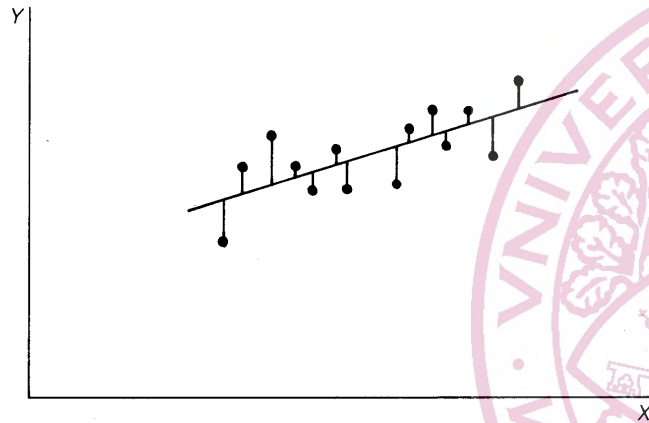


FIG. 17.5. Equazione che minimizza la somma dei quadrati degli scarti sulla verticale per la stima della regressione di Y su X

Per ottenere la retta con il metodo dei minimi quadrati dobbiamo, dunque, calcolare i fattori a e b che permetteranno di tracciare la retta avente la proprietà desiderata. Questo tipo di problema può essere facilmente risolto mediante il calcolo infinitesimale che dà le seguenti formule di calcolo per a e b ⁴:

⁴ Per gli studenti che hanno qualche nozione di calcolo infinitesimale possiamo delineare come si svolge la derivazione. Partiamo dalla equazione $Y_i = a + bX_i + e_i$, dove e_i è un termine residuo che può essere utilizzato per la stima di ε_i , il residuo della equazione di regressione. Vogliamo rendere minima la somma dei quadrati di questi residui, e cioè la quantità $\sum e_i^2 = \sum (Y_i - a - bX_i)^2$, rispetto alle due quantità a e b che

$$a = \frac{\sum_{i=1}^N Y_i - b \sum_{i=1}^N X_i}{N} = \bar{Y} - b\bar{X} \quad (17.3)$$

e

$$b = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2} = \frac{\sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i^2} \quad (17.4)$$

in cui $x_i = X_i - \bar{X}$ e $y_i = Y_i - \bar{Y}$. Si noti che in queste formule a e b sono le incognite, mentre gli altri valori sono derivati direttamente dai dati. Una volta che è stato calcolato b si può ricavare facilmente a utilizzando la prima delle due formule. Possiamo, quindi, concentrare la nostra attenzione sul calcolo di b .

Il numeratore di b è composto di $\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})$, che è la cosiddetta «codevianza» di X e Y . Questa quantità è analoga alla somma dei quadrati (che abbiamo adoperato per l'analisi della varianza) con la differenza che invece di elevare al quadrato $(X - \bar{X})$ o $(Y - \bar{Y})$ li moltiplichiamo l'uno per l'altro, ottenendo una misura di come X e Y varino assieme; di qui il nome codevianza. Se dividiamo la codevianza per N otteniamo ciò che, per analogia, viene detta la *covarianza*. Si vedrà immediatamente che il termine b è dato dal rapporto tra la covarianza e la varianza di X .

Se esaminiamo più da vicino la codevianza di X e Y , vediamo che, contrariamente a quanto avviene per la somma dei quadrati, essa può essere sia positiva che negativa. Se tra X e Y esiste una relazione positiva, a valori elevati di X tenderanno a corrispondere valori elevati di Y ; pertanto se $X > \bar{X}$, sarà anche vero che $Y > \bar{Y}$, e il prodotto di $(X - \bar{X})$ per $(Y - \bar{Y})$ sarà positivo come la somma di tutti i prodotti. Se invece X e Y sono tra di loro in relazione negativa, normalmente avremo che se X è maggiore di $\bar{X}$, Y è minore di $\bar{Y}$, e il prodotto delle due

sono assunte come incognite. Assumiamo derivate parziali rispetto ad a e b , le poniamo uguali a zero e risolviamo le due equazioni risultanti (chiamate *equazioni normali*) per a e b . Lo stesso procedimento può essere generalizzato al caso della regressione multipla.

differenze sarà negativo. Se tra X e Y non vi è relazione, circa la metà dei prodotti sarà negativa, mentre l'altra metà sarà positiva, dal momento che X e Y varieranno indipendentemente l'uno dall'altro: in questo caso b sarà zero, o comunque molto vicino a zero. Pertanto, tanto più forte è la relazione tra X e Y , tanto maggiore è il valore numerico della codevarianza. Come meglio vedremo tra poco, la codevarianza appare anche al numeratore dell'espressione per il calcolo del coefficiente di correlazione, la misura della forza della relazione. Per ottenere b , si divide la codevarianza per la somma dei quadrati relativa a X e si ottiene così la stima dell'inclinazione dell'equazione di regressione.

Per il calcolo della codevarianza è generalmente conveniente usare una formula di calcolo, direttamente analoga a quella che già abbiamo usato per il calcolo della somma dei quadrati e che si ottiene allo stesso modo. La formula di calcolo per b diviene dunque⁵:

$$b = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \quad (17.5)$$

Nell'equazione (17.5) sia il numeratore che il denominatore sono stati moltiplicati per N per ridurre gli errori di arrotondamento derivanti dalla divisione e per facilitare i calcoli mediante calcolatrice⁶.

Problema. Poniamo di avere i dati della tabella 17.1, dove è indicata per una serie di città degli Stati Uniti, la percentuale di negri (X) e la differenza tra il reddito mediano dei bianchi e dei negri (Y) riportata come misura di discriminazione economica.

Da questi dati possiamo calcolare cinque somme che, unitamente a N , cioè al numero dei casi, sono tutto ciò di cui abbiamo bisogno per risolvere i problemi di regressione e di correlazione. Eccetto una, queste somme ci servono per calcolare a e b . I calcoli possono essere riassunti così:

$$\begin{array}{ll} N=13 & \sum Y=8.557 \\ \sum X=62,88 & \sum Y^2=6.192.505 \\ \sum X^2=432,2768 & \sum XY=43.943,32 \end{array}$$

⁵ Si veda l'esercizio 5) al termine del capitolo.

⁶ In questa formula e in quelle che seguono gli indici sono stati tralasciati, dal momento che occorre effettuare la somma per tutti gli N casi.

TAB. 17.1. Dati per un problema di correlazione

Percentuale di negri (X)	Differenza di reddito (Y) (in dollari)	Percentuale di negri (X)	Differenza di reddito (Y) (in dollari)
2,13	809	4,62	859
2,52	763	5,19	228
11,86	612	6,43	897
2,55	492	6,70	867
2,87	679	1,53	513
4,23	635	1,87	335
		10,38	868

Fonte: Censimento USA, 1950.

L'unica quantità che sinora non abbiamo ancora adoperato è $\sum XY$. Ponendo i valori così ottenuti nella formula per a e b , otteniamo:

$$\begin{aligned} b &= \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2} \\ &= \frac{13(43.943,32) - (62,88)(8.557)}{13(432,2768) - (62,88)^2} = \frac{33.199,0}{1.665,7} = 19,931 \end{aligned}$$

e:

$$\begin{aligned} a &= \frac{\sum Y - b\sum X}{N} \\ &= \frac{8.557 - (19,931)(62,88)}{13} = 561,83 \end{aligned}$$

L'equazione lineare che otteniamo è pertanto:

$$Y_p = a + bX = 561,83 + 19,931X$$

in cui abbiamo usato Y_p per indicare che il valore di Y è stato predetto mediante un'equazione dei minimi quadrati; come già abbiamo detto, a e b ottenuti con questo metodo sono le stime

senza errore sistematico più efficienti di α e β , i veri coefficienti della regressione, se il residuo ε_i nella equazione $Y = \alpha + \beta X_i + \varepsilon_i$ ha un valore atteso di zero e non è connesso ad X e se ci troviamo effettivamente di fronte ad un campione casuale estratto dall'universo del quale ci stiamo occupando.

La retta ottenuta col metodo dei minimi quadrati è la miglior stima della vera regressione, se la regressione stessa è lineare. L'equazione ottenuta col metodo dei minimi quadrati ha inoltre la proprietà di passare attraverso il punto $(\bar{X}, \bar{Y})$ che costituisce la media sia di X che di Y . Questo può essere dimostrato mediante l'equazione (17.3); infatti se

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

è vero anche che

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

il che significa che questi valori di X e di Y sono tra quelli che corrispondono all'equazione e pertanto il punto di coordinate $\bar{X}$, $\bar{Y}$ giace sulla retta.

Nel problema che stiamo esaminando, se conosciamo il valore di X (percentuale di negri tra gli abitanti) per una città qualsiasi, la stima più accurata che possiamo effettuare del valore di Y è data dal valore di Y che otteniamo dall'equazione dei minimi quadrati a partire dal valore dato di X . Dal momento che i valori delle Y sono differenze in dollari tra il reddito mediano dei bianchi e dei negri, vediamo che a un aumento di 1 nella percentuale di negri tra gli abitanti corrisponde un aumento di 19,93 dollari nella differenza tra reddito mediano dei negri e dei bianchi. Un diagramma di dispersione e la relativa retta ottenuta col metodo dei minimi quadrati per questo problema sono stati tracciati nella figura 17.6. Per illustrare l'uso dell'equazione, possiamo pensare a questo esempio: se sapessimo che in una città gli abitanti negri costituiscono l'8 per cento del totale, potremmo stimare la differenza tra la mediana del reddito dei bianchi e quello dei negri così:

$$Y_p = a + b(8) = 561,83 + (19,931)(8) = 721,28 \text{ dollari}$$

Dalla figura 17.6 si può vedere che graficamente avremmo potuto

ottenere un risultato approssimativamente uguale. Incidentalmente, si noti che ponendo $X=8$ e risolvendo per Y abbiamo individuato un secondo punto della retta che ci può servire per tracciare la retta stessa attraverso il diagramma di dispersione.

17.2. Correlazione

D'ora in avanti, supponiamo che X sia una variabile stocastica e quindi al di fuori del controllo del ricercatore. Non soltanto vogliamo conoscere la *forma* o la natura della relazione tra X e Y , cosicché possiamo predire una variabile a partire dall'altra, ma dobbiamo anche sapere qual è la «forza» o «strettezza» della relazione: infatti, se la relazione è molto debole, non ha senso utilizzare la variabile X per predire Y . Spesso i ricercatori vogliono sapere innanzitutto *quali* sono le variabili più strettamente connesse a una data variabile dipendente tra una vasta gamma di variabili disponibili: negli studi a carattere esplorativo di questo genere l'analisi della regressione assume solo un'importanza

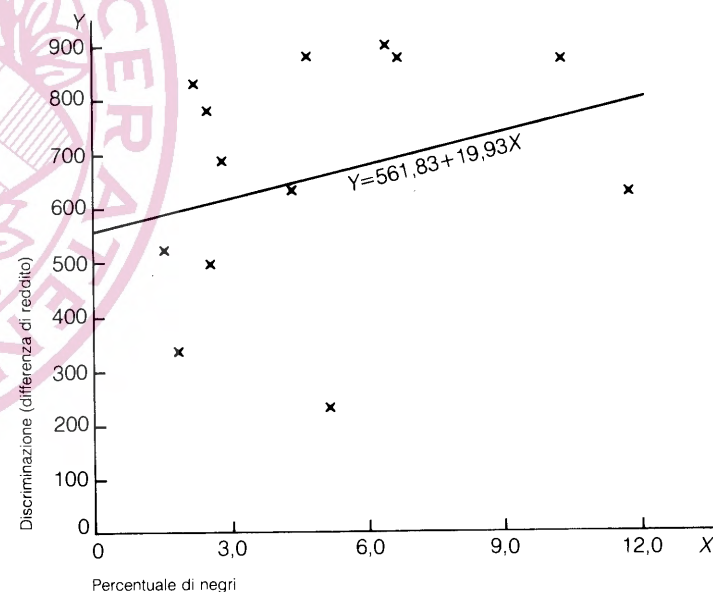


Fig. 17.6. Diagramma di dispersione e retta dei minimi quadrati per i dati della tabella 17.1

tanza secondaria. Man mano che le scienze sociali progrediscono e le variabili più importanti vengono identificate, si può sempre più concentrare l'attenzione sui problemi connessi colla predizione. Secondo alcuni statistici, sinora si è dedicata troppa attenzione alla correlazione e troppo poco all'analisi della regressione; se questo corrisponda a verità, dipende ovviamente dal livello raggiunto dalla scienza contro la quale si appunta la critica.

Il coefficiente di correlazione r del quale ci occuperemo in questo paragrafo è stato usato per la prima volta da Karl Pearson e viene spesso denominato coefficiente di correlazione «prodotto-momento» [presso gli statistici italiani, coefficiente di correlazione del Bravais. *N.d.T.*] per distinguerlo da altre misure di associazione. Questo coefficiente misura la variabilità dei dati attorno alla equazione *lineare* dei minimi quadrati. Esiste un coefficiente ρ (ro) relativo all'universo che misura in che modo la vera equazione di regressione si adatta ai dati e l' r ne è una stima ottenuta misurando gli scarti dalla retta ottenuta col metodo dei minimi quadrati.

Dal momento che l'equazione di regressione rappresenta il luogo delle medie delle distribuzioni di Y per ogni dato valore di X , sarebbe anche possibile misurare la variabilità dei dati attorno a questa linea calcolando lo scarto quadratico medio attorno ad essa⁷. La maggior parte di coloro che si dedicano alla ricerca applicata nei vari campi, si sono ormai abituati all'uso del coefficiente di correlazione e però è improbabile che il coefficiente di correlazione sia abbandonato nel prossimo futuro. Esso ha il vantaggio di poter essere interpretato facilmente e ha la caratteristica di variare da $-1,0$ a $+1,0$, il che è molto apprezzato dai pratici. Come meglio vedremo in seguito, la relazione tra il coefficiente di correlazione e lo scarto quadratico medio rispetto alla retta dei minimi quadrati è molto semplice e di questo fatto potremo servirci per interpretare r .

Abbiamo già detto come il limite superiore di r sia $+1,0$: se tutti i punti giacciono esattamente sulla retta, r sarà $+1,0$, o $-1,0$, a seconda che la relazione sia positiva o negativa. Se i punti sono sparsi a caso, r sarà zero. Tanto più la retta si adatta

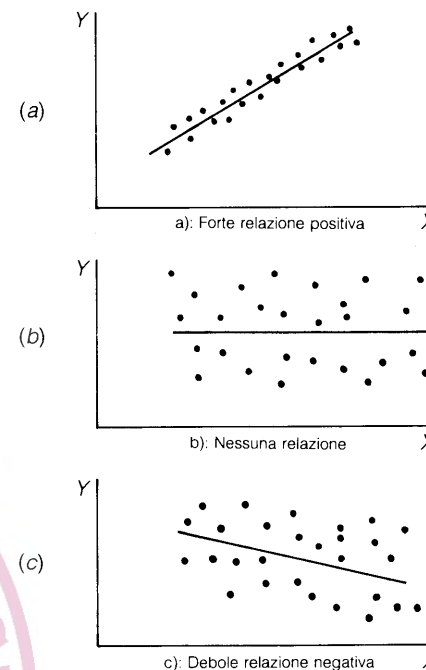


FIG. 17.7. Diagrammi di dispersione relativi a casi in cui esistono differenze nella direzione e nella forza della relazione tra X e Y

ai punti, tanto maggiore è l' r ; questo fatto è illustrato nella figura 17.7.

Si noti, però, che r è una misura di relazione *lineare*, dal momento che esso misura in qual modo la retta ottenuta col metodo dei minimi quadrati si adatta ai punti: non bisogna cadere nell'errore di pensare che, per il solo fatto che r è uguale a zero, non vi sia una relazione tra le due variabili. Se non vi è relazione, r sarà approssimativamente zero e i punti saranno dispersi a caso, ma vi possono essere dei casi in cui vi è una relazione curvilinea perfetta, eppure r è zero, il che indica solo che non esiste una *linea retta* che si adatti ai dati. Ciò è esattamente quanto si verifica nel caso illustrato dalla figura 17.8. Pertanto, ove il ricercatore trovi una correlazione uguale a zero, dovrà essere molto cauto nel dedurre che tra le due variabili non vi è relazione alcuna. In linea di massima, basterà osservare il diagramma di dispersione per rendersi conto se è vero che non

⁷ In seguito parleremo più a lungo di questa misura. Per ora basterà notare che essa è costituita da una estensione del concetto di scarto quadratico medio, con la differenza che la media di Y non è più considerata costante, ma funzione di X .

esiste correlazione, o se è vero, invece, che esiste una relazione ma che essa è abbastanza curvilinea da giustificare un coefficiente di correlazione pari a zero. Nella maggior parte dei problemi affrontati dalla sociologia, le relazioni possono essere rappresentate con sufficiente approssimazione mediante una linea retta, ma ciò non significa che non si debba stare in guardia contro eventuali eccezioni.

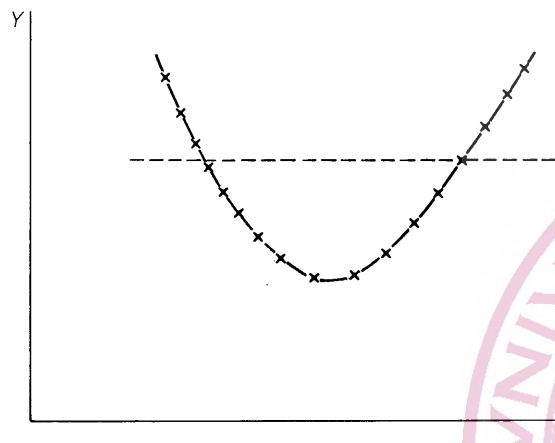


FIG. 17.8. Diagramma di dispersione relativo a un caso di relazione curvilinea perfetta in cui $r=0$

Sinora non abbiamo ancora dato una definizione del coefficiente di correlazione, ma possiamo farlo agevolmente a partire dalla formula:

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{[\sum (X - \bar{X})^2][\sum (Y - \bar{Y})^2]}} = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} \quad (17.6)$$

In parole, il coefficiente di correlazione è il rapporto tra la covarianza e la radice quadrata del prodotto della devianza di X per la devianza di Y . Dividendo il numeratore e il denominatore per N e scrivendo questa quantità sotto radice come N^2 , possiamo vedere come r possa anche essere definito come il rapporto tra la covarianza e il prodotto degli scarti quadratici medi di X e Y . La covarianza misura la variazione congiunta di X e di Y , ma la sua dimensione dipende dalla variabilità totale esistente nelle

due variabili. Dal momento che il valore della covarianza può essere notevolmente superiore all'unità, non è conveniente servirsi direttamente di essa come misura di associazione; conviene invece standardizzare la covarianza dividendola per il prodotto dei due scarti quadratici medi e ottenendo così una misura che varia da 1,0 a $-1,0$.

Abbiamo già visto che la covarianza è zero nel caso in cui tra X e Y non vi sia relazione alcuna, ma è relativamente facile dimostrare che il limite superiore di r corrisponde all'unità. Partiamo dal caso in cui b è positivo e tutti i punti giacciono esattamente sulla retta di regressione. Per ogni Y , possiamo quindi scrivere $Y = a + bX$. Dal momento che anche il punto $(\bar{X}, \bar{Y})$ giace sulla retta, abbiamo $\bar{Y} = a + b\bar{X}$. Pertanto, per tutti i punti sulla retta:

$$Y - \bar{Y} = (a + bX) - (a + b\bar{X}) = b(X - \bar{X})$$

da cui deriva:

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

e:

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

Esaminando il numeratore e il denominatore della formula per il coefficiente di correlazione si vedrà che, in questo caso, r è uguale a $+1,0$. Allo stesso modo si può dimostrare che se tutti i punti giacciono sulla retta e questa ha una inclinazione negativa, r risulterà pari a $-1,0$.

È necessario sottolineare anche la relazione che esiste tra il coefficiente di correlazione e l'inclinazione delle due rette di regressione ottenibili dai dati. Se indichiamo con b_{yx} l'inclinazione dell'equazione dei minimi quadrati che stima la regressione di Y su X e indichiamo con b_{xy} l'inclinazione dell'equazione che stima la regressione di X su Y , abbiamo, per simmetria, che

$$b_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (Y - \bar{Y})^2}$$

e quindi:

$$X = a_{xy} + b_{xy}Y$$

Pertanto r ha lo stesso numeratore dei due b e, se i due b sono zero, si può dedurre che anche r è zero e viceversa. Questo potrebbe portare a ritenere che la forza della relazione è proporzionale all'inclinazione della retta di regressione, ma ciò è vero solo nel caso in cui il denominatore rimanga costante. Come meglio vedremo in seguito, i valori numerici di b dipendono dalle dimensioni dell'unità di misura usata.

Il valore di r è stato standardizzato così che è indipendente dalle dimensioni relative degli scarti quadratici medi di X e di Y . Se così non fosse, saremmo in un grosso guaio, perché ci troveremmo di fronte a una misura che, nel caso ragionissimo in termini di dollari, sarebbe diversa da quella che avremmo nel caso ragionissimo in termini di centesimi di dollaro. Dalle formule che abbiamo dato per r e per b , si può vedere come r^2 possa essere espresso in termini di b . Così:

$$r^2 = b_{yx}b_{xy} = \frac{[\sum xy]^2}{\sum x^2 \sum y^2} \quad (17.7)$$

Lo studente dovrebbe a questo punto essere in grado di provare che quando r è 1,0 (o $-1,0$), $b_{yx} = 1/b_{xy}$, e che ciò significa che le due equazioni dei minimi quadrati coincidono. In linea di massima, man mano che r si avvicina a zero, l'angolo formato dalle due rette aumenta sino a che, quando $r=0$, le due rette sono perpendicolari o ortogonali.

Infine possiamo presentare una formula di calcolo per r che utilizza le cinque somme che abbiamo calcolato per la determinazione di a e b ⁸. La formula è:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}} \quad (17.8)$$

Il numeratore naturalmente è già stato calcolato e così anche gran parte del denominatore. La correlazione tra percentuali di negri e la differenza tra redditi mediani delle città prese in

considerazione nell'esempio numerico presentato poc'anzi è quindi:

$$\begin{aligned} r &= \frac{13(43.943,32) - (62,88)(8.557)}{\sqrt{[13(432,2768) - (62,88)^2][13(6.192.505) - (8.557)^2]}} \\ &= \frac{33.199}{110.120} = 0,301 \end{aligned}$$

Si noti che se si aggiungono o si sottraggono gli stessi valori sia da X che da Y il coefficiente di correlazione non cambia; allo stesso modo si può cambiare la «scala» delle due variabili dividendole o moltiplicandole per una qualsiasi costante senza che r cambi. Ciò equivale a dire che la correlazione tra reddito e istruzione è la stessa sia che il reddito venga misurato in dollari, sia che venga misurato in centesimi. Ma, se è vero che il coefficiente di correlazione non cambia se vengono effettuate queste trasformazioni dei dati, lo stesso non può dirsi dell'equazione dei minimi quadrati: l'aggiungere o il sottrarre valori cambierà il valore numerico di a e il cambiare la «scala» porterà a un cambiamento dell'inclinazione della retta. Per esempio, se ogni valore di X è diviso per 10, lasciando invariato Y , ne risulterà un b moltiplicato per 10 rispetto a quello originario. Sarebbe bene che il lettore si rendesse conto di questo e cercasse di spiegarne il motivo, esaminando le formule di r , a e b . Molte delle cose che siamo venuti dicendo possono essere usate per semplificare i calcoli. Per esempio, se X comprende o un numero molto grande o un decimale molto piccolo, può essere conveniente cambiare la «scala» della variabile per ridurre il rischio di errori di calcolo; oppure, se la variabile X è composta di valori quali 1.207, 1.409, 1.948, ecc., sarà consigliabile sottrarre 1.000 da ognuno di questi valori. Certi tipi di abbreviazioni usate per i calcoli possono essere adoperate solo se tutti i valori sono positivi e, pertanto, nel calcolare r sarà opportuno per potersene avvalere aggiungere a ciascun valore un numero leggermente superiore al valore negativo più grande che figura tra i dati.

Prima di procedere oltre, possiamo sottolineare un'altra cosa in merito al coefficiente di correlazione: dal momento che esso utilizza sia le varianze che le covarianze, il suo valore è molto sensibile a valori elevati in una delle due variabili. Inoltre, la

⁸ Se si eccettuano i casi in cui potrebbe sorgere confusione, continueremo a usare d'ora innanzi b al posto di b_{yx} .

grandezza di r dipende dalla variabilità generale della variabile indipendente. La figura 17.9 illustra questi fatti. Nella figura 17.9 *a* l'effetto della presenza di uno o due valori molto elevati fa sì che risulti una correlazione abbastanza elevata quando negli altri casi non vi è quasi correlazione, mentre nella figura 17.9 *b*, si verifica il contrario in quanto una relazione lineare abbastanza forte è sciupata da pochi casi estremi non allineati col resto dei punti. In quest'ultimo caso siamo probabilmente di fronte a un caso di relazione non lineare. In un diagramma di dispersione questi casi permettono di capire che cosa succede in ogni particolare circostanza. Vediamo che cosa si può fare se si verifica una delle due situazioni descritte.

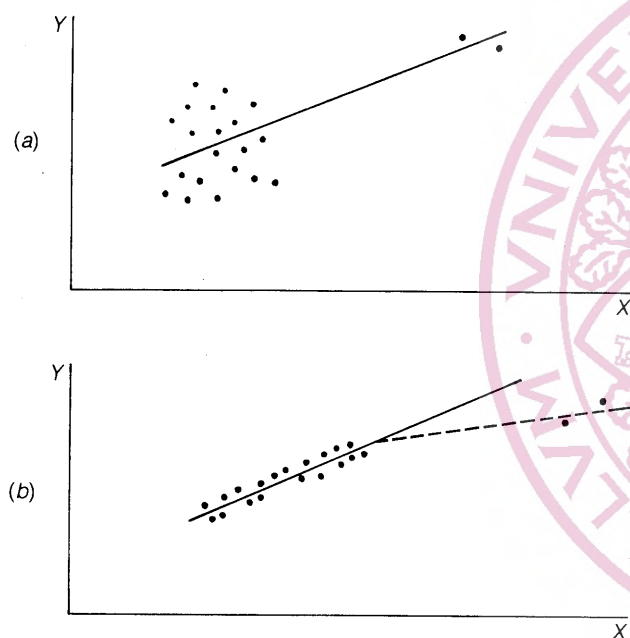


FIG. 17.9. Diagrammi di dispersione che illustrano il possibile effetto di casi estremi rispetto alla distribuzione di X

La figura 17.9 illustra molto bene l'osservazione fatta in precedenza, che, cioè, la grandezza del coefficiente di correlazione dipende dall'estensione del campo di variazione delle due variabili; se i casi molto elevati fossero stati più numerosi, la

distribuzione risultante avrebbe potuto essere come nella figura 17.10. La correlazione su tutta la distribuzione avrebbe potuto, cioè, essere alta, ma essere molto vicina a zero entro ogni possibile campo di variazione limitato di X . In effetti, ciò è chiaro indice del fatto che la variabilità di X è insufficiente a controbattere l'effetto di disturbo delle variabili incontrollate che possano agire sulla relazione; infatti, in pratica, X è tenuto quasi costante. Pertanto, se un diagramma di dispersione risulta simile a quello illustrato nella figura 17.9 *a*, sarà bene cercare di estendere la variabilità di X includendo nella distribuzione un numero maggiore di casi estremi.

Se non è materialmente possibile trovare altri casi estremi, o se l'interesse del ricercatore è diretto principalmente ai casi non estremi, sarà bene escludere del tutto dallo studio i pochi casi estremi rispetto ai quali si dispone di dati. Poniamo, per esempio, che X sia la grandezza della città e la città di New York risulti compresa nel campione: è evidente che, a meno che non sia possibile trovare altre città di dimensioni comparabili (il che chiaramente non può avvenire perché non esistono negli Stati Uniti città del genere), sarà necessario limitare l'analisi alle città con meno di 500.000 abitanti. In qualche caso sarà opportuno calcolare due r , uno includendo i valori estremi e uno escludendoli. È chiaro, comunque, che la soluzione adottata dipenderà dal tipo di problema e dagli interessi particolari del ricercatore. Tuttavia, è bene essere sempre consci del fatto che uno o due valori estremi possono avere un effetto esagerato sulla dimensione di r e, se questo si verifica, occorre tenerne conto. Il campo di variazione delle variabili dovrebbe sempre essere aggiunto al coefficiente di correlazione; ancora una volta si dimostra che l'adoperare una sola misura sintetica, per buona che essa sia rispetto alle altre misure, può indurre a errori di interpretazione.

Se i dati risultano invece come quelli della figura 17.9 *b*, possiamo aspettarci una relazione non lineare. Anche in questo caso, sarebbe bene ottenere altri valori estremi, o, se questi sono solo uno o due, escluderli addirittura dall'analisi. Casi di questo genere illustrano il fatto che, entro un campo di variazione limitato, una relazione può essere approssimativamente lineare, ma che, quando il campo di variazione viene esteso, il modello lineare può essere del tutto inadatto. È bene, pertanto, evitare di generalizzare più di quanto lo permettano i dati a disposizione; una conclusione prudente del tipo «entro i limiti di... e di...»

la relazione appare approssimativamente lineare» sarà la più appropriata.

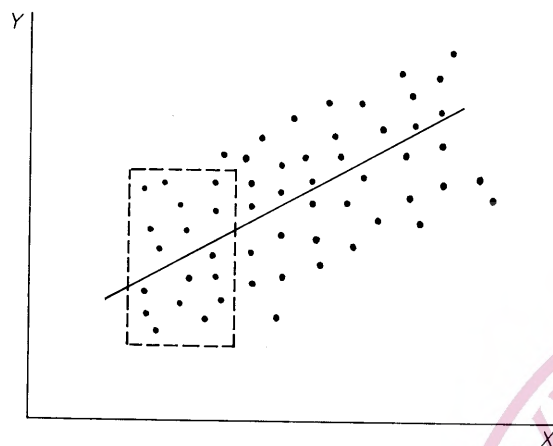


FIG. 17.10. Diagramma di dispersione che illustra come sia possibile che non esista una relazione se il campo di variazione di X è limitato, mentre esiste una relazione positiva su tutta la distribuzione

Un confronto tra correlazioni e coefficienti di inclinazione delle rette. Quanto abbiamo appena detto a proposito della sensibilità dei coefficienti di correlazione alle differenze della devianza di X rispetto alla dispersione prodotta dai fattori estranei mette in luce uno dei problemi fondamentali di *tutte* le misure di associazione. Il nostro scopo primario dovrebbe essere il cercare di capire che tipo di «legge» collega X e Y , se la relazione tra di esse è o meno lineare e, nel caso sia lineare, quale è l'inclinazione della retta. Nel confrontare i risultati di due diversi studi o di diversi sotto-campioni in uno stesso studio, dobbiamo riconoscere che è perfettamente possibile riscontrare differenze notevoli tra i coefficienti di correlazione anche se sono in atto le stesse «leggi», misurate dalla inclinazione delle rette. E cioè, gli r possono essere diversi anche se i coefficienti di inclinazione non lo sono e ciò può essere dovuto solo a differenze nella devianza della variabile indipendente X o a differenze nel livello a cui i fattori estranei che producono variazioni casuali nella Y sono stati portati sotto controllo. Come vedremo quando parleremo dell'analisi della covarianza, quando effettuiamo un *test* per l'interazione, sottoponiamo a verifica l'esistenza di una differen-

za tra i coefficienti di inclinazione, non l'esistenza di una differenza tra le correlazioni. Nel prossimo capitolo tratteremo brevemente dei *tests* per le differenze tra coefficienti di correlazione, ma il lettore deve essere avvertito sin d'ora che tali differenze, se in effetti esistono, si prestano a gravi errori di interpretazione.

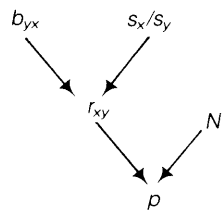
Può essere utile concepire un coefficiente di correlazione r_{xy} come funzione di due tipi di variabili, l'inclinazione della retta b_{yx} ed un fattore s_x/s_y che implica il rapporto tra i due scarti quadratici medi che riguardano il campione o sotto-campione del quale ci si sta occupando. Così

$$r_{xy} = b_{yx}(s_x/s_y)$$

Naturalmente il valore numerico di b_{yx} è determinato, non solo dalla legge che connette X a Y , ma anche dalla scelta della unità di misura fatta dal ricercatore. Anche il fattore s_x/s_y è funzione di tali unità che naturalmente sono note prima rispetto ai dati dell'universo o del campione; ma il rapporto s_x/s_y sarà specifico a ciascun campione (lo stesso vale per σ_x/σ_y per ciascun universo) e sarà usato per produrre la misura standardizzata r_{xy} . Il coefficiente di correlazione presenta il vantaggio di essere standardizzato così da non dipendere dalla scelta delle unità di misura, ma, sfortunatamente, deve essere standardizzato in termini di qualcosa, e questo qualcosa risulta essere una quantità che non resta costante al variare dei campioni o degli universi. Questo deve essere capito molto bene e quindi i coefficienti di inclinazione non standardizzati debbono *sempre* essere riportati a fianco dei coefficienti di correlazione così che non possano esservi delle incomprensioni.

In termini leggermente diversi, possiamo riconoscere l'esistenza di una gerarchia di fini o scopi scientifici nella inferenza statistica e nella stima. In primo luogo sottoponiamo a verifica la significatività per decidere se esiste un rapporto che non può essere attribuito alla sola azione del caso; e da questo punto di vista abbiamo già sottolineato che il livello di probabilità o di significatività è funzione della forza della relazione e delle dimensioni del campione. Se abbiamo un campione di dimensioni molto grandi, possiamo ottenere un livello di probabilità bassissimo anche con una relazione tanto debole da essere in pratica priva di significato. Ma dopo aver trovato almeno una relazione

di forza moderata, abbiamo dinnanzi un compito ancora più importante e cioè quello di stimare la natura di tale relazione misurata dal coefficiente di correlazione in caso di linearità. Quando i coefficienti di correlazione sono moderatamente alti, invece di confrontarli direttamente, noi stimiamo l'inclinazione delle rispettive rette e confrontiamo tali stime nel nostro *test* per l'interazione. Il processo può essere rappresentato da un diagramma quale il seguente in cui la direzione delle frecce rappresenta il «fluire della causalità» (per esempio, le probabilità subiscono l'influenza della forza della relazione e delle dimensioni del campione) che spesso si svolge in direzione opposta a quella seguita dalle varie fasi del procedimento di analisi statistica. Il diagramma indica che p è funzione di due variabili, una delle quali (le dimensioni del campione) non presenta in sé interesse, e che il coefficiente di correlazione r_{xy} è anch'esso funzione di due fattori l'uno dei quali (s_x/s_y) non ci interessa. Il nostro scopo è quello di far risalire l'analisi verso l'alto del diagramma sino a giungere alla stima dei coefficienti di regressione invece di restare al livello della probabilità o a quello delle affermazioni relative ai coefficienti di correlazione.



Si può dimostrare che ogni qualvolta si ha a che fare con le misure di associazione al livello delle scale ordinali, questa distinzione tra inclinazione delle rette di regressione e misure di associazione viene meno. Ma nel caso delle dicotomie si può dimostrare che, se si segue la regola di sistemare la variabile indipendente al vertice della tabella, e quindi con i dati segnati in verticale, calcolando le proporzioni o le percentuali dall'alto verso il basso ed effettuando i confronti orizzontalmente, le differenze di proporzioni che ne risultano possono essere considerate un caso speciale di inclinazione della retta b_{yx} mentre ϕ diviene un caso particolare di r_{xy} . Se le proporzioni sono calcolate nell'altra direzione, la differenza delle proporzioni diviene un

caso particolare di b_{xy} , cosicché ne possiamo trarre ulteriore conforto per la regoletta empirica a suo tempo consigliata. Questi risultati possono essere ottenuti assegnando valori di 0 e di 1 sia alla X che alla Y ed usando poi le formule fondamentali per il calcolo di r_{xy} e di b_{yc} (si veda l'esercizio 5) alla fine del capitolo).

Come si interpreta il coefficiente di correlazione. Per poter interpretare r nei casi in cui il suo valore non sia 0 o 1,0, dobbiamo tornare al concetto di variabilità attorno all'equazione di regressione. Abbiamo definito la varianza attorno alla media di Y come:

$$\sigma_y^2 = \frac{\sum (Y - \mu_y)^2}{M}$$

in cui M rappresenta le dimensioni dell'universo (in contrapposito con le dimensioni del campione N) e in cui usiamo gli indici per mettere in chiaro il fatto che dobbiamo distinguere tra le due variabili. Pertanto, il normale concetto di varianza implica scarti da un valore medio *fisso*, la media generale. Ma possiamo anche ottenere la media degli Y che corrispondono a un determinato X e diamo per scontato che questi valori varino col variare di X in modo da produrre il fenomeno della regressione lineare. Possiamo, quindi, rendere più generale il concetto di media, ottenendo una specie di media *sub-conditione* di Y per un determinato valore di X che possiamo indicare col simbolo $\mu_{y|x}$ o con $E(Y/X)$.

Estendendo allo stesso modo anche il concetto di varianza possiamo ottenere una misura della variabilità attorno all'equazione di regressione:

$$\sigma_{y|x}^2 = \frac{\sum (Y - \mu_{y|x})^2}{M} \quad (17.9)$$

in cui il simbolo $\sigma_{y|x}^2$ viene usato per sottolineare il fatto che la grandezza della variabilità attorno all'equazione di regressione, esattamente come il valore della media di cui abbiamo parlato in precedenza, varia col variare di X . In altre parole, per ogni valore di X abbiamo una media degli Y e una varianza attorno a tale media. Non è necessario che la variabilità attorno

alla media sia la stessa per tutti gli X , per quanto noi diamo per scontato che si dia omoscedasticità, cioè che le varianze siano uguali⁹.

Abbiamo ora due misure delle variabilità di Y . La prima misura, la variabilità attorno a quel valore di Y , la media generale μ_y , che costituirebbe la stima migliore di Y se non si conoscesse il valore di X . In altre parole, se ci viene chiesto di predire Y senza conoscere il valore di X , il modo più sicuro di rispondere è costituito da μ_y , o nel caso che si abbiano a disposizione solo i dati di un campione, $\bar{Y}$. Se si conosce X , invece, si predirà quel valore di Y che giace sulla retta di regressione in corrispondenza del valore di X conosciuto. Eccezion fatta per il caso in cui non vi sia relazione alcuna tra X e Y , il conoscere X dovrebbe facilitare la predizione di Y ; se la relazione fosse perfetta la predizione potrebbe essere esatta, in quanto tutti i punti verrebbero a giacere esattamente sulla retta. In linea di massima, una predizione esatta è fuori discussione, ma se assumiamo che la distribuzione di Y sia normale e che lo scarto quadratico medio $\sigma_{y|x}$ sia costante, possiamo sapere in ogni momento qual è il rischio di commettere errori, nonché qual è la grandezza degli errori stessi. Inoltre, cosa molto più importante ai nostri fini, possiamo mettere a confronto i due scarti quadratici medi e ottenere una misura di quanto migliora la predizione per il fatto di conoscere X , utilizzando la tecnica, già ben conosciuta, dell'analisi della varianza.

Nell'analisi della varianza si divideva la devianza totale, ovvero la somma generale dei quadrati, in parti spiegate e parti non spiegate. Useremo di nuovo la stessa tecnica, ottenendo, quasi fossero un sottoprodotto, i valori di $\sigma_{y|x}^2$ e r^2 . Potremo così dare un significato al coefficiente di correlazione. Anzitutto, dobbiamo trasformare gli scarti di ciascun Y da $\bar{Y}$ in somme di due diverse quantità, cioè $(Y - Y_p) + (Y_p - \bar{Y})$ (si veda un'illustrazione grafica di questo alla figura 17.11). La prima di queste quantità, cioè $Y - Y_p$, rappresenta lo scarto del valore di Y dalla retta dei minimi quadrati e indica la quantità di errore che si commette quando si usa Y_p per predire Y . La seconda espressione rappresenta lo scarto tra la retta dei minimi quadrati (per un X determinato) e il valore $\bar{Y}$. Nella maggior parte dei casi

quest'ultima quantità corrisponde alla riduzione nell'errore che si ottiene conoscendo Y_p . Se ora eleviamo al quadrato i due termini dell'equazione e sommiamo per tutti i casi, otteniamo:

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_p)^2 + 2 \sum (Y - Y_p)(Y_p - \bar{Y}) + \sum (Y_p - \bar{Y})^2$$

Fortunatamente il termine centrale si elide e restiamo con

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_p)^2 + \sum (Y_p - \bar{Y})^2 \quad (17.10)$$

$$SQ \text{ generale} = SQ \text{ non spiegata} + SQ \text{ spiegata}$$

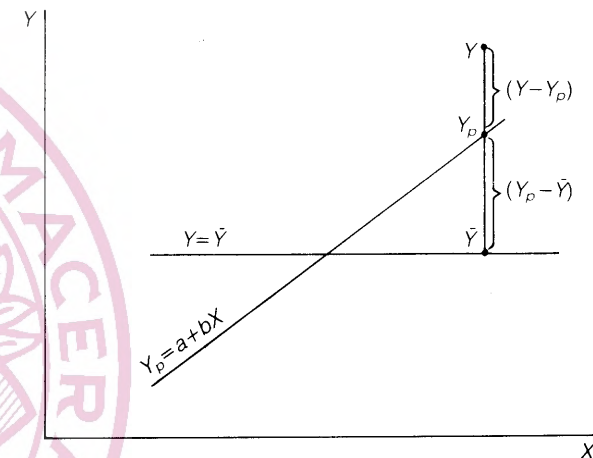


FIG. 17.11. Rappresentazione geometrica del fatto che gli scarti dalla media $\bar{Y}$ corrispondono alla somma degli scarti dalla retta dei minimi quadrati cogli scarti della retta dei minimi quadrati da $\bar{Y}$

La prima quantità posta alla destra del segno di uguale nell'equazione rappresenta la somma dei quadrati degli scarti dei valori reali di Y dalla retta dei minimi quadrati; questa è la quantità di devianza che non viene spiegata, in quanto si tratta dell'errore che si commette predicendo Y mediante la retta dei minimi quadrati. Ciò che resta indica quanto abbiamo guadagnato in esattezza usando Y_p al posto di $\bar{Y}$ e può essere, quindi, chiamata la somma dei quadrati «spiegata», senza naturalmente far riferimento a una spiegazione di indole causale, ma solo all'associazione tra le due variabili. Vediamo ora più da vicino le due quantità.

Se dividiamo la somma dei quadrati non spiegata per il

⁹ Tale requisito di omoscedasticità è probabile venga meno ogni qual volta la variabile dipendente Y è una dicotomia o una proporzione; ne parleremo in breve al capitolo XX.

numero dei casi, otteniamo la varianza del campione $s_{y|x}^2$ attorno alla retta dei minimi quadrati. Pertanto:

$$s_{y|x}^2 = \frac{\sum (Y - Y_p)^2}{N} \quad (17.11)$$

Se vogliamo ottenere una stima senza errore sistematico della varianza dell'universo attorno alla retta dei minimi quadrati, dobbiamo dividere non per N , ma piuttosto per il numero appropriato di gradi di libertà. In questo caso, abbiamo perso due gradi di libertà nel calcolare a e b come stime rispettivamente di α e di β ; pertanto se vogliamo stimare $\sigma_{y|x}^2$ useremo

$$\hat{\sigma}_{y|x}^2 = \frac{\sum (Y - Y_p)^2}{N - 2} \quad (17.12)$$

Così la somma dei quadrati non spiegata può essere facilmente trasformata in una stima della varianza attorno alla retta dei minimi quadrati. È abbastanza facile vedere come il procedimento sia stato analogo a quello seguito nell'analisi della varianza: il concetto di variabilità attorno all'equazione dei minimi quadrati è stato semplicemente messo al posto del concetto di variabilità all'interno delle classi di X .

Passiamo ora alla somma dei quadrati spiegata $(Y_p - \bar{Y})^2$: si può agevolmente dimostrare che questa quantità è equivalente a $r^2 [\sum (Y - \bar{Y})^2]$ cioè a $r^2 \sum y^2$. Dal momento che $Y_p = a + bX$ e $\bar{Y} = a + b\bar{X}$, abbiamo:

$$(Y_p - \bar{Y}) = b(X - \bar{X})$$

Pertanto:

$$\begin{aligned} \sum (Y_p - \bar{Y})^2 &= b^2 \sum (X - \bar{X})^2 = b^2 \sum x^2 \\ &= \frac{(\sum xy)^2}{(\sum x^2)^2} (\sum x^2) = \frac{(\sum xy)^2}{\sum x^2} \\ &= \frac{(\sum xy)^2}{\sum x^2 \sum y^2} (\sum y^2) = r^2 \sum y^2 \end{aligned}$$

$$= r^2 \sum (Y - \bar{Y})^2$$

Abbiamo dunque dimostrato che

$$r^2 = \frac{\sum (Y_p - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{SQ \text{ spiegata}}{SQ \text{ generale}}$$

che può anche essere espressa in un formato analogo a quello usato per le misure di RPE e cioè

$$r^2 = \frac{SQ \text{ generale} - SQ \text{ non spiegata}}{SQ \text{ generale}}$$

dove le somme dei quadrati equivalgono agli errori di predizione. Allo stesso modo è possibile dimostrare che r^2 rappresenta anche il rapporto tra la devianza spiegata di X e quella totale dello stesso X . Il quadrato del coefficiente di correlazione può essere, pertanto, interpretato come espressione della porzione della devianza totale di una variabile che è spiegata dalla variabilità dell'altra. La quantità di $\sqrt{1 - r^2}$, a volte indicata come *coefficiente di alienazione*, rappresenta la radice quadrata di quella parte della somma generale dei quadrati che non è spiegata dalla variabile indipendente.

Occorre notare che non esiste un'interpretazione diretta e facile per r ; in linea di massima, anzi, è facile essere indotti in errore dal valore numerico di r dal momento che questo, eccezion fatta per i casi in cui $r=0$ o $1,0$, sarà di valore maggiore del corrispondente r^2 . Può, quindi, sembrare che un r di $0,5$ corrisponda alla metà di una correlazione perfetta, mentre invece esso indica che nel caso specifico una variabile spiega solo il 25 per cento, cioè un quarto, della devianza dell'altra variabile. Una correlazione di $0,7$ indica che leggermente meno della metà della devianza in una variabile è spiegata dall'altra. Vediamo, inoltre, che correlazioni di $0,3$ o minori di $0,3$ significano che solo una minima parte della devianza è spiegata. La tabella 17.2 indica che relazione vi è tra le quantità che abbiamo menzionato.

Dal momento che $1 - r^2$ rappresenta la devianza che non viene spiegata abbiamo:

$$(1 - r^2) [\sum (Y - \bar{Y})^2] = \sum (Y - Y_p)^2$$

e pertanto:

$$(1-r^2) \frac{\sum (Y-\bar{Y})^2}{N} = \frac{\sum (Y-Y_p)^2}{N}$$

ovvero:

$$(1-r^2)s_y^2 = s_{y|x}^2$$

Ne deriva che:

$$s_{y|x} = \sqrt{1-r^2} s_y$$

Il risultato ci indica in che misura possiamo ridurre lo scarto quadratico medio se conosciamo X . (Si veda l'ultima colonna della tabella 17.2). Quando $r=0$, i due scarti quadratici medi sono uguali; la cosa è perfettamente ovvia, quando si pensi che in questo caso particolare la retta dei minimi quadrati verrebbe ad essere una retta orizzontale indicata dall'equazione $Y=\bar{Y}$. Quando r^2 è uguale a uno, $s_{y|x}$ è naturalmente zero dal momento che tutti i punti giacciono sulla retta dei minimi quadrati.

TAB. 17.2. Relazione esistente tra r , r^2 , $1-r^2$ e $\sqrt{1-r^2}$

r	r^2	$1-r^2$	$\sqrt{1-r^2}$
0,90	0,81	0,19	0,44
0,80	0,64	0,36	0,60
0,70	0,49	0,51	0,71
0,60	0,36	0,64	0,80
0,50	0,25	0,75	0,87
0,40	0,16	0,84	0,92
0,30	0,09	0,91	0,95
0,20	0,04	0,96	0,98
0,10	0,01	0,99	0,995

Esaminando la tabella 17.2 vediamo che, solo quando r è molto grande, la grandezza dello scarto quadratico medio viene

molto ridotta. Se r è 0,80, lo scarto quadratico medio attorno alla retta dei minimi quadrati è di 0,60 rispetto allo scarto quadratico medio normale, ma se r è 0,40, il miglioramento nella predizione di Y da X è ben poco.

Città	Indice di integrazione	Indice di eterogeneità	Indice di mobilità
Rochester	19,0	20,6	15,0
Syracuse	17,0	15,6	20,2
Worcester	16,4	22,1	13,6
Erie	16,2	14,0	14,8
Milwaukee	15,8	17,4	17,6
Bridgeport	15,3	27,9	17,5
Buffalo	15,2	22,3	14,7
Dayton	14,3	23,7	23,8
Reading	14,2	10,6	19,4
Des Moines	14,1	12,7	31,9
Cleveland	14,0	39,7	18,6
Denver	13,9	13,0	34,5
Peoria	13,8	10,7	35,1
Wichita	13,6	11,9	42,7
Trenton	13,0	32,5	15,8
Grand Rapids	12,8	15,7	24,2
Toledo	12,7	19,2	21,6
San Diego	12,5	15,9	49,8
Baltimore	12,0	45,8	12,1
South Bend	11,8	17,9	27,4
Akron	11,3	20,4	22,1
Detroit	11,1	38,3	19,5
Tacoma	10,9	17,8	31,2
Flint	9,8	19,3	32,2
Spokane	9,6	12,3	38,9
Seattle	9,0	23,9	34,2
Indianapolis	8,8	29,2	23,1
Columbus	8,0	27,4	25,0
Portland (Oregon)	7,2	16,4	35,8

Fonte: R. C. Angell, *The Moral Integration of American Cities*, in «American Journal of Sociology», LVII (1951), 2 (riprodotta per gentile concessione dell'autore e dell'editore. Copyright 1951 by the University of Chicago).

Glossario

Coefficiente di alienazione
 Coefficiente di correlazione
 Covarianza
 Distribuzione normale bivariata
 Equazione dei minimi quadrati
 Inclinazione della retta di regressione
 Regressione di Y su X

Esercizi

1) Partendo dai semplici numeri:

$X:$ 1 2 3 4 5
 $Y:$ 2 0 2 4 6

che sono in relazione perfetta secondo la formula $Y = -4 + 2X$, si verifichi che $r=1,0$; $b=2$ e $a=-4$.

2) I dati presentati nella tabella 17.2 si riferiscono a 29 città degli Stati Uniti con un numero di abitanti superiore a 100.000 (non sono comprese città del sud) e sono tratti dallo studio di Angell sulla «integrazione morale» nelle città americane. L'indice di integrazione morale è costruito utilizzando il tasso di criminalità oltre al tasso di spese assistenziali. L'eterogeneità di ciascuna città è stata misurata sulla base del numero di cittadini non di razza bianca o nati all'estero in essa residenti. Come seconda variabile indipendente è stato introdotto un indice di mobilità basato sul numero di persone che in media vengono a risiedere nella città, o se ne vanno dalla stessa, ogni anno.

a) Si tracci un diagramma di dispersione relativo al rapporto tra integrazione morale ed eterogeneità.

b) Per le stesse variabili si calcolino r , a e b e si tracci, attraverso il diagramma di dispersione, la retta dei minimi quadrati (l'integrazione morale è l' Y) [Risp. $r=-0,156$; $a=13,9$; $b=-0,049$]

c) Qual è la grandezza relativa dello scarto quadratico medio attorno alla retta dei minimi quadrati in rapporto allo scarto quadratico medio attorno a Y ?

3) Per risolvere gli esercizi del capitolo XIX vi sarà bisogno delle correlazioni tra l'integrazione morale e la mobilità e tra la

eterogeneità e la mobilità. Si calcolino ambedue questi r [Risp. $r=-0,456$; $r=-0,513$]

*4) Si verifichi l'esattezza della formula di calcolo (17.5) relativa a b e della formula di calcolo (17.8) relativa ad r .

5) Assegnando valori di 0 e 1 sia ad X che ad Y ed usando poi le formule definitorie di b ed r , si dimostri che Φ è un caso particolare di r e che una differenza di proporzioni è un caso particolare di coefficiente di inclinazione.

Bibliografia

Croxton F. E. e D. J. Cowden, *Applied General Statistics*, Englewood Cliffs, N. J., Prentice-Hall, Inc., 1955, capitoli XIX-XX.

Hagood M. J. e D. O. Price, *Statistics for Sociologists*, cit., cap. XXIII.

McNemar Q., *Psychological Statistics*, cit., capitoli VIII, IX e X.

Wallis W. A. e H. V. Roberts, *Statistics: A New Approach*, cit., cap. XVII.

Barberi B., *Nozioni di calcolo statistico*, cit., capitoli IX-X.

Boldrini M., *Statistica. Teoria e metodi*, cit., cap. XV.

D'Addario, *Rappresentazione analitica delle funzioni statistiche*, in *Lezioni di metodologia statistica per ricercatori*, cit., vol. VI, pp. 3-58.

Gini C., *Corso di statistica*, cit., cap. IX (3).

Parenti G., *Metodologia statistica*, cit., cap. VII.

Salvemini T., *Regressione e correlazione*, Torino, Boringhieri, 1959.