STATISTICA PER LE SCIENZE SOCIALI

UNIVERSITATIS · MACERATENSIS · SIGILLVM

il Mulino

GEORGE W. BOHRNSTEDT
DAVID KNOKE

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

Regressione e correlazione bivariata

Capitolo 6

Le tavole di contingenza discusse nel capitolo precedente mettono in luce la covariazione fra coppie di variabili discrete esistente in un campione di osservazioni. In questo capitolo, invece, esamineremo le relazioni fra coppie di variabili continue. Le procedure di regressione e correlazione bivariata assumono che la *forma* della relazione fra Y e X sia *lineare* e che la variabile dipendente sia distribuita normalmente a ogni livello della variabile indipendente. Tuttavia, anche quando questi assunti vengono violati la regressione e la correlazione risultano, in genere, **robuste**, cioè solo raramente inducono il ricercatore a trarre conclusioni errate dai risultati ottenuti.

1. Diagrammi di dispersione e rette di regressione

Per rappresentare graficamente la relazione fra coppie di variabili continue bisogna utilizzare tecniche diverse da quelle descritte nei capitoli precedenti. Il **diagramma di dispersione** rappresenta un utile strumento per visualizzare il modo in cui due variabili continue covariano. Per costruire un diagramma di dispersione bisogna innanzitutto tracciare un piano cartesiano. I valori della variabile indipendente (X) vengono collocati sull'asse orizzontale, mentre i valori della variabile dipendente (Y) vengono posti sull'asse verticale. Quindi, in corrispondenza delle coordinate X, Y appropriate, si traccia nel grafico, sotto forma di punto, la posizione di tutte le osservazioni disponibili. L'insieme dei punti così tracciati mostra il modo in cui le due variabili covariano. La figura 6.1 illustra quattro forme «tipiche» di relazioni bivariate.

A titolo di esempio, supponiamo ora di ipotizzare che negli Stati Uniti il reddito personale sia correlato con il livello di istruzione dei cittadini:

a.a. 2012-2013
Fondamenti e metodi per l'analisi empirica nelle scienze sociali

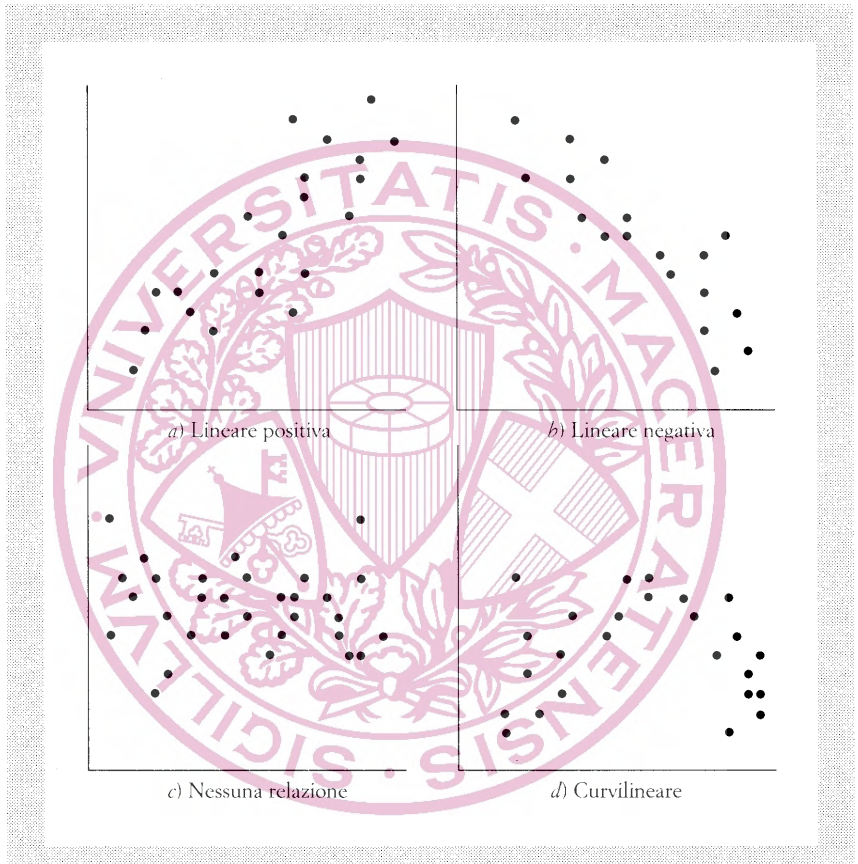

FIG. 6.1. Diagrammi di dispersione raffiguranti quattro tipi ideali di relazioni bivariate.

H_1: Maggiore è la percentuale dei cittadini di uno stato che possiedono una laurea, maggiore è il reddito annuale pro capite di quello stato.

La figura 6.2 illustra graficamente la relazione fra il reddito annuale pro capite registrato nei 51 stati americani (nel 1985) e la percentuale di cittadini con almeno 16 anni di istruzione formale (nel 1980). La «nuvola» dei punti, che si muove dall'angolo in basso a sinistra verso l'angolo in alto a destra, mette chiaramente in luce l'esistenza di una covariazione positiva: i valori bassi di una variabile tendono a essere associati ai valori bassi dell'altra, così come tendono a essere associati i valori più elevati di entrambe le variabili. La figura, dunque, fornisce un sostegno visivo alla nostra ipotesi.

L'espressione di una relazione ipotizzata in forma statistica richiede un sistema di notazione specifico che sia in grado di denotare sia le variabili che le osservazioni. Per cominciare, in un'analisi di regressione la variabile dipendente è indicata con il simbolo Y, mentre la varia-

Page 2 of 18

a.a. 2012-2013
Fondamenti e metodi per l'analisi empirica nelle scienze sociali
Page 3 of 18

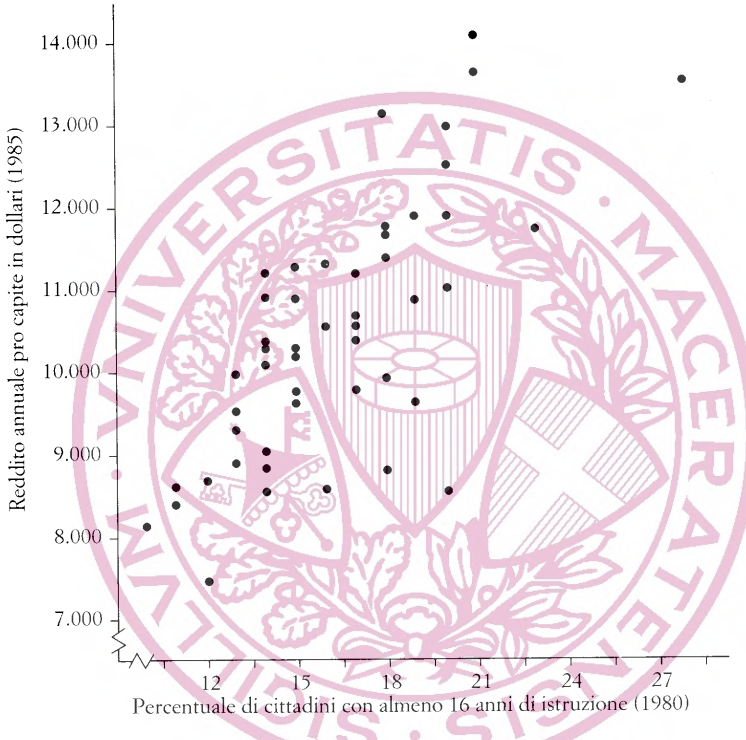

FIG. 6.2. Diagramma di dispersione raffigurante la relazione fra il reddito annuale pro capite registrato negli stati degli Usa (1985) e la percentuale di cittadini di quegli stati con almeno 16 anni di istruzione formale (1980).

bile indipendente (a volte chiamata **variabile esplicativa**) è indicata con il simbolo X. Per indicare un'osservazione individuale specifica – come una persona, una città o uno stato – si utilizza un indice deponente che, semplicemente, rappresenta il numero dell'osservazione stessa (da 1 a N, dove N denota la numerosità del campione). Dunque, in generale X_i indica il valore assunto dalla variabile X in corrispondenza dell'osservazione i. Sostituendo l'indice deponente generale con un numero specifico si indica il valore di X corrispondente a una specifica osservazione. Nell'esempio precedente, X_{37} denota il valore assunto dalla variabile «Percentuale di laureati» nel 37° stato della lista (Oregon), pari a 17,9%. Analogamente, Y_{37} indica il reddito annuale pro capite dell'Oregon, pari a 9.925 dollari.

Quando stimiamo la **relazione lineare** fra una variabile dipendente (Y) e una variabile indipendente (X), «regrediamo» Y rispetto a X, producendo una **relazione lineare bivariata** o, più semplicemente, una

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

regressione bivariata. In termini algebrici, la forma generale di una equazione lineare è scritta come segue:

$$Y = a + bX$$

Come si può vedere, il valore di Y (l'ordinata) equivale alla somma di una costante a – che corrisponde al punto in cui la retta «intercetta» o incrocia l'asse verticale – più il prodotto del gradiente (o inclinazione) b moltiplicato per il valore di X. L'intercetta della retta a corrisponde al valore di Y quando $X = 0$. L'inclinazione della retta b, invece, esprime la quantità di cambiamento prodotta in Y quando il valore di X varia di una unità.

Tipicamente, i ricercatori sociali assumono che due variabili sono correlate in modo lineare, ad eccezione dei casi in cui hanno forti ragioni per credere fin dall'inizio che una forma non lineare, ad esempio una funzione esponenziale o logaritmica, rappresenti la relazione in modo più accurato. Una relazione lineare costituisce la forma più elementare e, quindi, una ragionevole prima approssimazione. Poiché la *parsimonia* (intesa come semplicità) rappresenta uno degli obiettivi principali della scienza, cominciare l'analisi cercando di determinare se una funzione lineare descrive in modo soddisfacente la relazione fra due variabili è una scelta sensata. Solo dopo avere (eventualmente) stabilito in modo evidente che una linea retta *non* rappresenta una data relazione in modo adeguato si può passare a considerare forme funzionali più complesse. Il capitolo 9 illustra diverse forme non lineari, fra cui la regressione logistica.

Se il reddito annuale pro capite dei vari stati avesse una relazione perfetta con la percentuale di laureati, tutte le osservazioni ricadrebbero su una linea retta. Chiaramente, il diagramma di dispersione illustrato nella figura 6.2 rappresenta una relazione che è lontana dall'essere perfettamente lineare; non così lontana, però, da apparire puramente casuale. La figura, inoltre, *non* suggerisce che una funzione curvilineare offrirebbe una descrizione migliore della relazione osservata. Pertanto, possiamo procedere all'illustrazione di tecniche che permettono di stabilire in quale misura una linea retta approssima la covariazione osservata fra due variabili continue.

Il punto di partenza è un'**equazione predittiva** nella quale il valore assunto dalla variabile dipendente in corrispondenza di ciascuna osservazione i è una funzione lineare esatta del corrispondente valore della variabile indipendente:

$$\hat{Y}_i = a + b_{YX} X_i$$

Il deponente di b indica l'ordine della regressione, con la variabile indipendente che segue quella dipendente. L'accento circonflesso posto sopra Y_i indica che si tratta di un valore predetto (o «atteso»), che

Page 4 of 18

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

può essere uguale o diverso rispetto al corrispondente valore osservato (Y_i senza circonflesso). I veri dati analizzati dai ricercatori sociali non seguono mai relazioni lineari perfette, come abbiamo già visto esaminando la figura 6.2. Di conseguenza, dobbiamo tenere conto delle deviazioni dalla predizione lineare formulando un **modello di regressione lineare**:

$$Y_i = a + b_{YX}X_i + e_i$$

L'assenza del circonflesso sul simbolo della variabile dipendente indica che si tratta del valore osservato di Y, anziché di quello atteso. L'**errore** e_i rappresenta la porzione del valore assunto da Y in corrispondenza dell'osservazione i che *non* è predetta dalla sua relazione lineare con X. In altri termini, e_i misura la discrepanza fra i valori osservati di Y e quelli attesi in base all'equazione di regressione lineare. Nell'ambito delle N osservazioni, alcuni errori di predizione saranno positivi (cioè maggiori di zero), altri saranno negativi (cioè minori di zero), e altri ancora potranno essere nulli (cioè esattamente uguali a zero). Esattamente come le deviazioni dalla media di una distribuzione, la somma degli errori è sempre uguale a zero perché i valori positivi e negativi si controbilanciano esattamente.

L'errore di regressione è anche detto **residuo**, perché rappresenta la quantità che rimane dopo avere sottratto l'equazione predittiva dal modello di regressione lineare:

$$Y_i - \hat{Y}_i = [a + b_{YX}X_i + e_i] - [a + b_{YX}X_i] = e_i$$

Uno dei compiti principali dell'analisi di regressione è quello di stimare i valori dei due coefficienti di regressione utilizzando i dati osservati. Le stime di a e b_{YX} devono minimizzare i residui, cioè fare sì che gli errori di predizione prodotti da quella equazione siano minori di quelli prodotti da qualsiasi altra relazione lineare.

A un diagramma di dispersione è possibile sovrapporre una **retta di regressione**. Questa retta possiede una proprietà molto utile: assumendo una data relazione lineare, per ogni valore di X_i predice un preciso valore di $\hat{Y}_i$. Dunque, la posizione delle osservazioni «reali» nel grafico cartesiano può essere confrontata con la posizione da esse occupata lungo la retta di regressione, al fine di determinare l'entità degli errori di predizione. Nel prossimo paragrafo presenteremo un metodo per stimare i due parametri del modello di regressione lineare utilizzando i dati campionari, per calcolare la retta di regressione e per determinare la vicinanza fra i valori predetti e quelli osservati.

Page 5 of 18

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

2. La stima dell'equazione di regressione lineare

2.1. Il criterio dei minimi quadrati

Per stimare un'equazione di regressione bivariata si utilizzano tutte e N le osservazioni disponibili. Gli stimatori dei due coefficienti (a e b_{YX}) devono soddisfare il criterio dei *minimi quadrati*. Poiché, come abbiamo già detto, la somma dei residui e_i è sempre uguale a zero, elevandoli al quadrato si eliminano i segni negativi, cosicché la somma degli errori al quadrato è sempre maggiore di zero. L'obiettivo del criterio dei minimi quadrati è il seguente: sommando le differenze al quadrato fra ogni valore osservato Y_i e il corrispondente valore $\hat{Y}_i$ predetto dall'equazione di regressione prescelta, si deve ottenere una quantità minore di quella che si otterrebbe utilizzando qualsiasi altra equazione di regressione lineare. In altri termini,

$$\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{N} e_i^2$$

deve essere minimo. Gli stimatori che soddisfano questa proprietà producono stime dei **minimi quadrati comuni** (OLS, dall'inglese *Ordinary Least Squares*) di a e b_{YX}.

Lo stimatore OLS del **coefficiente di regressione bivariata** b_{YX} è calcolato come segue:

$$b_{YX} = \frac{\sum(Y_i - \overline{Y})(X_i - \overline{X})}{\sum(X_i - \overline{X})^2}$$

Il numeratore della formula corrisponde alla somma dei prodotti delle deviazioni di Y e X dalle rispettive medie. Quando questa somma è divisa per $N - 1$ si ottiene la **covarianza**, indicata con il simbolo s_{YX}. Formalmente:

$$s_{YX} = \frac{\sum(Y_i - \overline{Y})(X_i - \overline{X})}{N-1} =$$

$$= \frac{N\sum X_i Y_i - \sum X_i \sum Y_i}{N(N-1)}$$

Il denominatore della formula di b_{YX}, a sua volta, è la somma delle deviazioni al quadrato di X dalla sua media. Come abbiamo visto nel capitolo 3, dividendo questa somma per $N - 1$ si ottiene la varianza campionaria:

$$s_X^2 = \frac{\sum(X_i - \overline{X})^2}{N-1}$$

Page 6 of 18

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

Poiché $N - 1$ appare sia nel denominatore della covarianza che in quello della varianza campionaria, quando si calcola il rapporto fra queste due misure tale termine viene eliminato. Dunque, lo stimatore OLS del coefficiente di regressione bivariata può essere espresso anche nel modo seguente:

$$b_{YX} = \frac{s_{YX}}{s_X^2}$$

Sebbene le due formule di b_{YX} presentate sopra siano concettualmente accurate, entrambe sono difficili da calcolare senza un computer. La seguente **formula computazionale di *b***, che non richiede il calcolo di alcuna deviazione, fornisce identici risultati numerici è può essere risolta utilizzando una semplice calcolatrice tascabile:

$$b_{YX} = \frac{N\sum Y_i X_i - \sum Y_i \sum X_i}{N\sum X_i^2 - (\sum X_i)^2}$$

Lo stimatore OLS dell'**intercetta** a è dato semplicemente dalla seguente formula:

$$a = \overline{Y} - b_{YX}\overline{X}$$

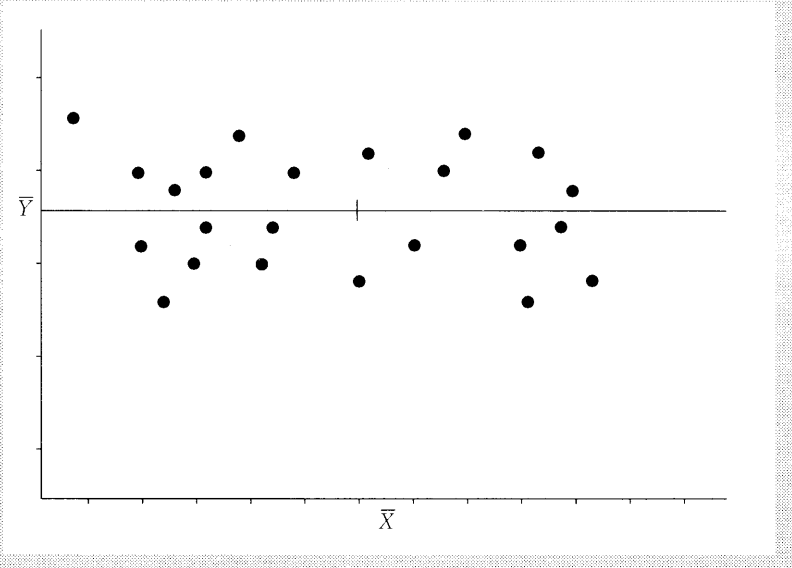

FIG. 6.3. Retta di regressione quando $b_{YX} = 0$.

Page 7 of 18

a.a. 2012-2013
Fondamenti e metodi per l'analisi empirica nelle scienze sociali
Page 8 of 18

TAB. 6.1. *Calcolo delle medie, delle varianze, del coefficiente di correlazione e dei coefficienti di regressione relativi alla relazione fra reddito annuale pro capite (Y) e percentuale di cittadini con almeno 16 anni di istruzione*

Stato	Y	X	X^1	YX	Y^2
Alabama	8.681	12,2	148,84	105.908,2	75.359.761
Alaska	13.650	21,1	445,21	288.015,0	186.322.500
Arizona	10.561	17,4	302,76	183.761,4	111.534.721
Arkansas	8.389	10,8	116,64	90.601,2	70.375.321
California	11.885	19,6	384,16	232.946,0	141.253.225
Colorado	11.713	23,0	529,00	269.399,0	137.194.369
Connecticut	14.090	20,7	428,49	291.663,0	198.528.100
Delaware	11.375	17,5	306,25	199.062,5	129.390.625
Florida	11.271	14,9	222,01	167.937,9	127.035.441
Georgia	10.191	14,6	213,16	148.788,6	103.856.481
Hawaii	11.003	20,3	412,09	223.360,9	121.066.009
Idaho	8.567	15,8	249,64	135.358,6	73.393.489
Illinois	11.302	16,2	262,44	183.092,4	127.735.204
Indiana	9.978	12,5	156,25	124.725,0	99.560.484
Iowa	10.096	13,9	193,21	140.334,4	101.929.216
Kansas	10.684	17,0	289,00	181.628,0	114.147.856
Kentucky	8.614	11,1	123,21	95.615,4	74.200.996
Lousiana	8.836	13,9	193,21	122.820,4	78.074.896
Maine	9.042	14,4	207,36	130.204,8	81.757.764
Maryland	12.967	20,4	416,16	264.526,8	168.143.089
Massachusetts	12.510	20,0	400,00	250.200,0	156.500.100
Michigan	10.902	14,3	204,49	155.898,6	118.853.604
Minnesota	11.186	17,4	302,76	194.636,4	125.126.596
Mississippi	7.483	12,3	151,29	92.040,9	55.995.289
Missouri	10.283	13,9	193,21	142.933,7	105.740.089
Montana	8.781	17,5	306,25	153.667,5	77.105.961
Nebraska	10.546	15,5	240,25	163.463,0	111.218.116
Nevada	11.200	14,4	207,36	161.280,0	125.440.000
New Hampshire	11.659	18,2	331,24	212.193,8	135.932.281
New Jersey	13.129	18,3	334,89	240.260,7	172.370.641
New Mexico	8.814	17,6	309,76	155.126,4	77.686.596
New York	11.765	17,9	320,41	210.593,5	138.415.225
North Carolina	9.517	13,2	174,24	125.624,4	90.573.289

Questa formula garantisce che la coppia di coordinate $(\overline{X}, \overline{Y})$ ricada sempre sulla retta di regressione, indipendentemente dai valori specifici di a e b_{YX}.

Una equazione di regressione lineare stima una **media condizionata** di Y, cioè un valore atteso della variabile dipendente $(\hat{Y}_i)$ per ogni specifico valore della variabile indipendente (X_i). Se non esiste alcuna relazione fra Y e X, allora il coefficiente b_{YX} risulta uguale a zero. Come ci mostra la figura 6.3, in una situazione di questo tipo la retta di regressione corre parallela all'asse orizzontale, all'esatta distanza di $\overline{Y}$ unità dall'origine. Inoltre, tutti i valori attesi risultano uguali ad a, in quanto $\hat{Y}_i = a + 0X_i$ implica che $\hat{Y}_i = a$ e, quindi, $\hat{Y}_i = \overline{Y}$. Quando l'inclinazione della retta di regressione è uguale a zero, la conoscenza dei valori assunti da X non ci permette di predire un valore di Y diverso dalla sua media $\overline{Y}$. Quando, invece, fra Y e X esiste effettivamente una relazione di tipo lineare, la media condizionata di Y dipen-

a.a. 2012-2013
Fondamenti e metodi per l'analisi empirica nelle scienze sociali
Page 9 of 18

TAB. 6.1. (*segue*)

Stato	Y	X	X^2	YX	Y^2
North Dakota	9.635	14,8	219,04	142.598,0	92.833.225
Ohio	10.371	13,7	187,69	142.082,7	107.557.641
Oklahoma	9.754	15,1	228,01	147.285,4	95.140.516
Oregon	9.925	17,9	320,41	177.657,5	98.505.625
Pennsylvania	10.288	13,6	184,96	139.916,8	105.842.944
Rhode Island	10.892	15,4	237,16	167.736,8	118.635.664
South Carolina	8.890	13,4	179,56	119.126,0	79.032.100
South Dakota	8.553	14,0	196,00	119.742,0	73.153.809
Tennessee	9.290	12,6	158,76	117.054,0	86.304.100
Texas	10.373	16,9	285,61	175.303,7	107.599.129
Utah	8.535	19,9	396,01	169.846,5	72.846.225
Vermont	9.619	19,0	361,00	182.761,0	92.525.161
Virginia	11.894	19,1	364,81	227.175,4	141.467.236
Washington	10.866	19,0	361,00	206.454,0	118.069.956
Washington D.C.	13.530	27,5	756,25	372.075,0	183.060.900
West Virginia	8.141	10,4	108,16	84.666,4	66.275.881
Wisconsin	10.298	14,8	219,04	152.410,4	106.048.804
Wyoming	9.782	17,2	295,84	168.250,4	95.687.524
Somme (Σ)	531.306	832,1	14.134,55	8.849.810	5.652.405.000

$$\overline{Y} = 531.306/51 = 10.417,8$$

$$\overline{X} = 832,1/51 = 16,3$$

$$s_Y^2 = \frac{(51)(5.652.405.000) - (531.306)^2}{(51)(50)} = 2.347.682,10$$

$$s_X^2 = \frac{(51)(14.134,55) - (832,1)^2}{(51)(50)} = 11,165$$

$$s_{XY} = \frac{(51)(8.849.810) - (531.306)(832,1)}{(51)(50)} = 3.623,8$$

de dallo specifico valore di X_i. Inoltre, tutte le medie condizionate ricadono sulla retta determinata dall'intercetta e dal coefficiente di regressione stimati.

2.2. Un'applicazione della regressione lineare

Tornando all'esempio della relazione fra reddito annuale pro capite e percentuale di laureati negli Stati Uniti, la tabella 6.1 contiene tutte le informazioni necessarie per calcolare b_{YX} e a, assumendo il reddito come variabile dipendente. Innanzitutto, b_{YX} assume il seguente valore:

$$b_{YX} = \frac{(51)(8.849.810) - (531.306)(832,1)}{(51)(14.134,6) - 692.390,4} = 324,5$$

a.a. 2012-2013
Fondamenti e metodi per l'analisi empirica nelle scienze sociali
Page 10 of 18

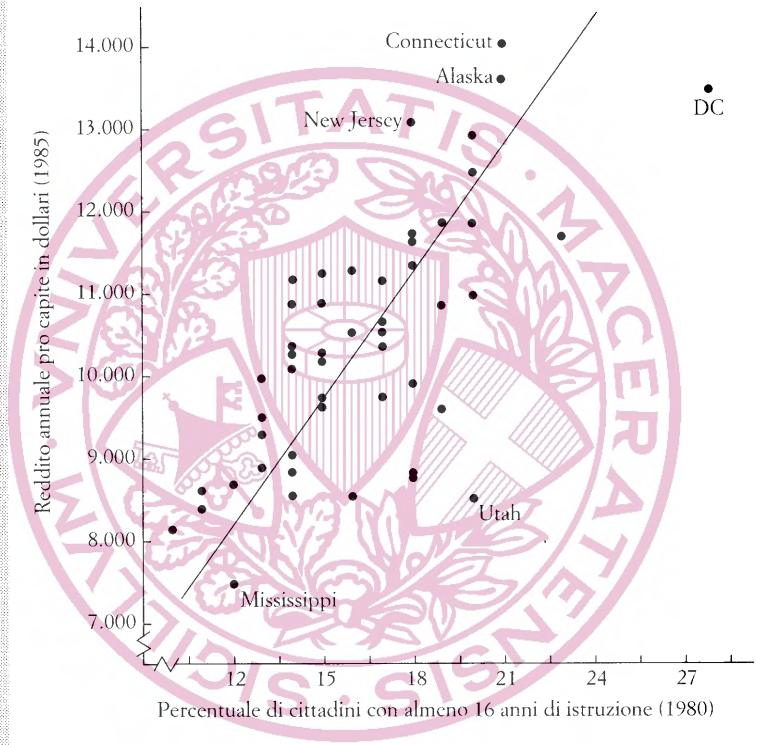

FIG. 6.4. Relazione fra reddito annuale pro capite registrato negli stati degli Usa (1985) e percentuale di cittadini di quegli stati con almeno 16 anni di istruzione formale (1980): diagramma di dispersione e retta di regressione.

In alternativa, $b_{YX} = s_{XY}/s_X^2 = 3.623{,}8/11{,}165 = 324{,}6$ (la leggera differenza fra i due valori è dovuta agli errori di arrotondamento).

Come abbiamo già detto, il coefficiente di regressione bivariata esprime la misura in cui Y varia ogni volta che X varia di una unità. La nostra stima del parametro b_{YX}, quindi, ci dice che a ogni aumento dell'1% nella quota dei laureati di uno stato è associato un aumento (atteso) del reddito annuale pro capite pari a 324,6 dollari.

Poiché $\overline{X} = 16{,}3\%$, $\overline{Y} = 10.417{,}8$ dollari e $b_{YX} = 324{,}6$, possiamo facilmente calcolare l'intercetta: $a = 10.417{,}8 - (324{,}6)(16{,}3) = 5.126{,}8$. Pertanto, l'equazione di regressione bivariata completa è la seguente:

$$\hat{Y}_i = 5.126{,}8 + 324{,}6X_i$$

La figura 6.4 mostra la retta di regressione stimata sovrapposta al diagramma di dispersione corrispondente. Se la scala dell'asse orizzon-

a.a. 2012-2013
Fondamenti e metodi per l'analisi empirica nelle scienze sociali
Page 11 of 18

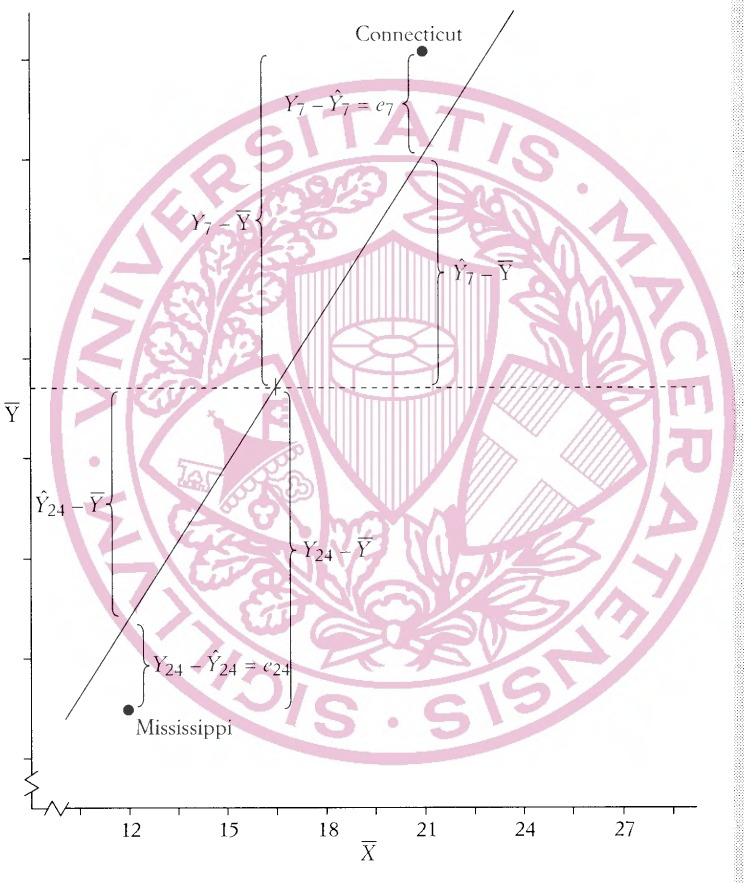

FIG. 6.5. Retta di regressione e residui.

tale fosse sufficientemente estesa a sinistra, la retta incrocerebbe l'asse verticale in corrispondenza del valore 5.126,8, equivalente a $X = 0$. Sebbene la retta risulti abbastanza vicina ad alcune delle 51 osservazioni, nessun caso ricade esattamente su di essa.

La figura 6.5 illustra il concetto di residui, o *errori di predizione*. Come abbiamo già accennato, un residuo è semplicemente la differenza fra un valore osservato e il corrispondente valore predetto dall'equazione di regressione, cioè $e_i = Y_i - \hat{Y}_i$. La figura 6.5 illustra due errori di predizione, uno positivo (corrispondente all'osservazione 7, cioè al Connecticut) e uno negativo (corrispondente all'osservazione 24, cioè al Mississippi). Come vedremo nel prossimo paragrafo, i residui forniscono un mezzo per misurare la forza della relazione lineare fra due variabili continue.

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

3. *R*-quadrato e correlazione

3.1. La scomposizione della somma dei quadrati

Un modo per determinare la forza della covariazione fra due variabili consiste nel misurare il grado complessivo di vicinanza fra i valori osservati e la retta di regressione stimata. Se tutte le osservazioni ricadessero esattamente sulla retta di regressione, gli N valori attesi $\hat{Y}_i$ sarebbero privi di qualsiasi errore (cioè, $\hat{Y}_i - Y_i = 0$). Come abbiamo già osservato, tuttavia, quando si lavora con dati reali la predizione perfetta è pressoché impossibile. In ogni regressione è sempre presente una certa quantità – più o meno grande – di errore. Il nostro obiettivo realistico consiste nel determinare i contributi relativi della predizione e dell'errore alla variazione totale osservata nella variabile dipendente. Poiché la variazione in Y è dovuta sia all'effetto esercitato da X, sia all'errore casuale, possiamo suddividere la somma dei quadrati totale in due componenti: una sistematica e una casuale.

Come già sappiamo, una deviazione corrisponde alla differenza fra una data osservazione Y_i e la media della distribuzione corrispondente $\overline{Y}$. Se a una deviazione sottraiamo e aggiungiamo il valore predetto da una data equazione di regressione ($\hat{Y}_i$), otteniamo la seguente identità:

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

Ogni osservazione, dunque, è formata da due componenti:

1. $Y_i - \hat{Y}_i$ esprime la discrepanza fra un'osservazione e il suo valore atteso; tale discrepanza corrisponde all'errore e_i;
2. $\hat{Y}_i - \overline{Y}$ rappresenta la porzione di un valore osservato che è attribuibile alla relazione lineare fra Y e X.

Elevando al quadrato entrambi i lati di questa identità e sommando lungo tutte le osservazioni si ottengono la **somma dei quadrati della regressione** e la **somma dei quadrati dell'errore**. Riordinando questi termini abbiamo:

$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$$

Il termine posto alla sinistra del segno uguale è chiamato **somma dei quadrati totale** e corrisponde al numeratore della varianza. I due termini a destra del segno uguale, invece, rappresentano rispettivamente la somma dei quadrati della regressione e la somma dei quadrati dell'errore. Utilizzando una notazione semplificata, l'identità può essere espressa come segue:

$$SQ_{TOTALE} = SQ_{REGRESSIONE} + SQ_{ERRORE}$$

Page 12 of 18

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

Questa equazione mostra che, assumendo l'esistenza di una relazione lineare fra X e Y, tutta la variazione osservata nella variabile dipendente (SQ_{TOTALE}) può essere suddivisa in una porzione attribuibile alla sua relazione con la variabile indipendente ($SQ_{REGRESSIONE}$) e in un'altra porzione attribuibile agli errori di predizione (SQ_{ERRORE}).

3.2. Il coefficiente di determinazione

Quanti errori di predizione siamo in grado di evitare utilizzando un modello di regressione? Se non disponessimo di alcuna informazione sulla relazione fra X e Y, per ogni X_i la nostra migliore predizione di Y_i sarebbe semplicemente la media campionaria $\overline{Y}$. In questo caso la somma degli errori al quadrato $\sum(Y_i - \hat{Y}_i)^2$ si ridurrebbe a $\sum(Y_i - \overline{Y})^2$, in quanto la media prenderebbe il posto di tutti i valori attesi. Ciò significa che la somma dei quadrati dell'errore risulterebbe identica alla somma dei quadrati totale e, quindi, la somma dei quadrati della regressione sarebbe uguale a zero. Nella misura in cui l'informazione su X_i ci aiuta a migliorare la nostra predizione di Y_i, gli errori di predizione si riducono.

Poiché la somma dei quadrati totale dipende dalla dimensione del campione, per confrontare distribuzioni diverse è necessario standardizzare i valori assunti dalle sue componenti. La somma dei quadrati della regressione e la somma dei quadrati dell'errore contengono l'informazione necessaria per costruire una misura standardizzata di *riduzione proporzionale dell'errore* (RPE) applicabile all'analisi di regressione. Come abbiamo già visto nel capitolo precedente (cfr. cap. 5, par. 3), tutte le misure RPE possiedono la seguente forma generale:

$$\text{Statistica RPE} = \frac{\text{Errore senza la regola di decisione} - \text{Errore con la regola di decisione}}{\text{Errore senza la regola di decisione}}$$

Se l'applicazione di una regola di decisione produce lo stesso errore che si ottiene applicando *nessuna* regola di decisione, allora la statistica RPE assume valore zero; in questo caso, la regola di decisione non aiuta – neppure minimamente – a ridurre l'errore di predizione. Se, invece, l'applicazione della regola di decisione elimina *completamente* l'errore, allora la statistica RPE assume un valore pari a 1.

La regressione lineare può essere vista come una regola di decisione che, per ogni osservazione del campione, ci dice quale valore $\hat{Y}_i$ dobbiamo attenderci in media *dato* il valore X_i osservato. Se esiste una relazione lineare fra Y e X, la regressione produrrà un errore di predizione complessivo inferiore a quello che si avrebbe utilizzando solo la media di Y come regola di predizione.

Page 13 of 18

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

La statistica RPE applicabile alla regressione è detta **coefficiente di determinazione** (R^2_{YX}), in quanto indica la proporzione della variazione totale di Y che è «determinata» dalla sua relazione lineare con X. La formula di R-quadrato è la seguente

$$R^2_{YX} = \frac{\sum(Y_i - \bar{Y})^2 - \sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

ovvero

$$R^2_{YX} = \frac{SQ_{TOTALE} - SQ_{ERRORE}}{SQ_{TOTALE}} = 1 - \frac{SQ_{ERRORE}}{SQ_{TOTALE}}$$

Dunque, R-quadrato equivale a 1 meno il quoziente del rapporto fra la somma dei quadrati dell'errore e la somma dei quadrati totale. Il coefficiente di determinazione può assumere valori compresi fra 0 e 1. Quando si verifica una predizione perfetta (e quindi $SQ_{ERRORE} = 0$), allora $R^2_{YX} = 1$; quando, invece, tutta la variazione osservata è dovuta all'errore (e quindi $SQ_{ERRORE} = SQ_{TOTALE}$), allora $R^2_{YX} = 0$. Poiché la somma dei quadrati della regressione è uguale alla differenza fra la somma dei quadrati totale e la somma dei quadrati dell'errore, R-quadrato può essere calcolato anche nel seguente modo:

$$R^2_{YX} = \frac{SQ_{REGRESSIONE}}{SQ_{TOTALE}}$$

Chiaramente, il coefficiente di determinazione equivale alla proporzione della somma dei quadrati totale attribuibile al modello di regressione stimato.

La formula computazionale del coefficiente di determinazione bivariato si basa sul rapporto fra la covarianza al quadrato e il prodotto delle varianze di entrambe le variabili implicate:

$$R^2_{YX} = \frac{s^2_{YX}}{s^2_X s^2_Y}$$

Applicando questa formula all'esempio sulla relazione fra reddito annuale pro capite e percentuale di laureati negli Usa (cfr. tab. 6.1), si ottiene:

Page 14 of 18

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

$$R^2_{YX} = \frac{(3.623,8)^2}{(11,165)(2.347.682,10)} = 0,501$$

Dunque, il 50,1% della variazione nei redditi annuali pro capite osservata negli Stati Uniti è «spiegata» (in senso statistico) dalla percentuale dei cittadini di ciascuno stato che sono laureati. Per gli standard delle scienze sociali, un'associazione bivariata di questa ampiezza riveste una indubbia rilevanza sostanziale.

La quantità $1 - R^2_{YX}$ rappresenta la proporzione di varianza della variabile Y che *non può* essere attribuita alla sua relazione lineare con X. Per questo, essa è chiamata **coefficiente di non determinazione**. Nel nostro esempio questo coefficiente assume un valore pari a $1 - 0,501 = 0,499$. Ciò significa che la relazione lineare fra Y e X non è in grado di spiegare il 49,9% della varianza di Y. Nel capitolo 8, dedicato alla regressione multipla, vedremo che la proporzione di varianza non spiegata può essere ridotta aggiungendo altre variabili indipendenti al modello di regressione.

3.3. Il coefficiente di correlazione

La radice quadrata di R^2_{YX}, che riassume la relazione lineare fra due variabili continue, è chiamata **coefficiente di correlazione** di Pearson. La sua formula è la seguente:

$$r_{YX} = \sqrt{R^2_{YX}} =$$

$$= \frac{s_{XY}}{s_Y s_X}$$

L'utilità del coefficiente di correlazione consiste nel mostrare la direzione della relazione fra X e Y. Mentre R^2_{YX} non ci dice nulla sul fatto che le due variabili siano direttamente o inversamente correlate, il segno positivo o negativo che accompagna r_{YX} indica la direzione della covariazione. Questo segno deve essere uguale al segno del coefficiente di regressione b_{YX}. Nel nostro esempio, la correlazione fra reddito e istruzione a livello statale risulta pari a +0,708. A differenza di R^2_{YX}, r_{YX} può assumere valori che variano fra –1, corrispondente a una relazione inversa perfetta, e +1, corrispondente a una relazione diretta perfetta. Se $r_{YX} = 0$, allora Y e X sono indipendenti.

Dalla definizione di R-quadrato segue che il coefficiente di correlazione è *simmetrico*, cioè $r_{YX} = r_{XY}$:

Page 15 of 18

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

$$r_{YX} = \frac{s_{YX}}{s_Y s_X} =$$

$$= \frac{\sum (Y_i - \overline{Y})(X_i - \overline{X})/(N-1)}{s_Y s_X} =$$

$$= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})/(N-1)}{s_X s_Y} =$$

$$= \frac{s_{XY}}{s_X s_Y} =$$

$$= r_{XY}$$

Sebbene il coefficiente di correlazione sia simmetrico, è importante notare che il coefficiente bivariato r_{YX} che si ottiene «regredendo» Y rispetto a X *non è uguale* al coefficiente bivariato b_{YX} che si ottiene «regredendo» X rispetto a Y. Nel paragrafo 3.5 vedremo la relazione fra questi due coefficienti e R-quadrato.

3.4. Correlazione fra valori Z

Se si standardizzano due variabili Y e X (trasformandole in Z_Y e Z_X) e si correlano fra loro si ottengono risultati interessanti. Come già sappiamo, una distribuzione di valori Z ha sempre varianza e deviazione standard pari a 1. Poiché $r_{YX} = s_{YX}/(s_Y s_X)$, la formula della correlazione fra due variabili standardizzate è la seguente:

$$r_{Z_Y Z_X} = \frac{s_{Z_Y Z_X}}{s_{Z_Y} s_{Z_X}} = \frac{s_{Z_Y Z_X}}{(1)(1)} = s_{Z_Y Z_X}$$

Dunque, la correlazione fra due variabili standardizzate equivale alla loro covarianza. Inoltre, poiché

$$\overline{Z}_Y = \overline{Z}_X = 0$$

e

$$s_{Z_Y Z_X} = \frac{\sum (Z_{Y_i} - \overline{Z}_Y)(Z_{X_i} - \overline{Z}_X)}{N-1}$$

Il coefficiente di correlazione può essere espresso interamente in termini di variabili standardizzate:

Page 16 of 18

a.a. 2012-2013

Fondamenti e metodi per l'analisi empirica nelle scienze sociali

$$r_{Z_Y Z_X} = s_{Z_Y Z_X} = \frac{\sum Z_Y Z_X}{N-1}$$

Infine, dato che

$$Z_{Y_i} = \frac{(Y_i - \overline{Y})}{s_Y}$$

e

$$Z_{X_i} = \frac{(X_i - \overline{X})}{s_X}$$

ne deriva che

$$r_{Z_Y Z_X} = \frac{\sum Z_Y Z_X}{N-1} =$$

$$= \frac{\sum [(Y_i - \overline{Y})/s_Y][(X_i - \overline{X})/s_X]}{N-1} =$$

$$= \frac{\sum (Y_i - \overline{Y})(X_i - \overline{X})/(N-1)}{s_Y s_X} =$$

$$= \frac{s_{YX}}{s_Y s_X}$$

$$= r_{YX}$$

Dunque, la correlazione fra due variabili standardizzate Z_Y e Z_X è esattamente uguale alla correlazione fra le due variabili originali Y e X. Ciò significa che il coefficiente di correlazione rimane invariato tanto se le variabili sono misurate nella loro unità originale, quanto se vengono trasformate in valori Z.

3.5. Relazione fra regressione e correlazione

La relazione fra un coefficiente di regressione e il corrispondente coefficiente di correlazione bivariata è funzione delle deviazioni standard di entrambe le variabili implicate. Innanzitutto, ricordiamo che b_{YX} può essere espresso come rapporto fra la covarianza di Y e X e la varianza di X: $b_{YX} = s_{YX}/s_X^2$. Riordinando questi termini al fine di isolare la covarianza, si ottiene: $s_{YX} = b_{YX} s_X^2$. Se nell'equazione della cor-

Page 17 of 18

relazione sostituiamo s_{YX} con $b_{YX}s_X^2$, otteniamo: $r_{YX} = (b_{YX}s_X^2)/s_Y s_X$. Pertanto, un coefficiente di correlazione può essere espresso in termini di un coefficiente di regressione bivariata:

$$r_{YX} = b_{YX}\frac{s_X}{s_Y}$$

A sua volta, il coefficiente di regressione può essere espresso in termini del coefficiente di correlazione:

$$b_{YX} = r_{YX}\frac{s_Y}{s_X}$$

Dunque, conoscendo i valori di b_{YX}, s_Y e s_X è possibile calcolare r_{YX}. In alternativa, conoscendo i valori di r_{YX}, s_Y e s_X si può calcolare b_{YX}. Tornando ancora una volta al nostro esempio, si ha:

$$r_{YX} = b_{YX}\frac{s_X}{s_Y} = 324,6\left(\frac{3,34}{1.532,21}\right) = 0,7076$$

e

$$b_{YX} = r_{YX}\frac{s_Y}{s_X} = 0,7076\left(\frac{1.532,21}{3,34}\right) = 324,6$$

a.a. 2012-2013
Fondamenti e metodi per l'analisi empirica nelle scienze sociali
Page 18 of 18