Comprende versione **ebook**

Wayne W. **Daniel** • Chad L. **Cross**

Biostatistica

Concetti di base per l'analisi statistica
delle scienze dell'area medico-sanitaria

III edizione

EdiSES

TERZA EDIZIONE

BIOSTATISTICA

Concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria

WAYNE W. DANIEL, PH.D.

Professor Emeritus
Georgia State University

CHAD L. CROSS, PH.D., PSTAT®

Statistician
Office of Informatics and Analytics
Veterans Health Administration

Associate Graduate Faculty
University of Nevada, Las Vegas

EdiSES

Titolo originale:
WAYNE W. DANIEL, CHAD L. CROSS
BIOSTATISTICS A Foundation for Analysis in the Health Sciences

BIOSTATISTICA *Concetti di base per l'analisi statistica delle scienze dell'area medico-sanitaria* – III edizione

9 8 7 6 5 4 3 2 1 0
2024 2023 2022 2021 2020 2019

Le cifre sulla destra indicano il numero e l'anno dell'ultima ristampa effettuata

Fotocomposizione:
Grafic&Design di *Ettore Menna* – Napoli

Stampato presso la:
Tipolitografia Sograte S.r.l.
Zona Ind. Regnano – Città di Castello (PG)

per conto della
EdiSES S.r.l. – Piazza Dante, 89 – Napoli

http://**www.edises.it** e-mail: **info@edises.it**

ISBN 978-88-3319-041-9

AUTORI

***Hanno collaborato all'edizione italiana*:**

ALESSANDRA CARTOCCI
Università degli Studi di Siena

PAOLA DALMASSO
Università degli Studi di Torino

MARTA DI NICOLA
Università degli Studi di Chieti-Pescara

LUIGI FERRANTE
Università Politecnica delle Marche

MONICA FERRARONI
Università degli Studi di Milano

VINCENZO GUARDABASSO
Università degli Studi di Catania

DOMENICA MATRANGA
Università degli Studi di Palermo

STEFANIA ROSSI
Università degli Studi di Siena

PAOLA TELLAROLI
Università degli Studi di Padova

PAOLO TREROTOLI
Università degli Studi di Bari

MARIA ELISABETTA ZANOLIN
Università degli Studi di Verona

Revisione a cura di:

ANNA CHIARA FRIGO - *Università degli Studi di Padova*

EGLE PERISSINOTTO - *Università degli Studi di Padova*

PREFAZIONE ALL'EDIZIONE ITALIANA

La III edizione del testo "Biostatistica" è stata scritta per gli studenti e gli specializzandi di vari corsi di laurea di indirizzo medico-biologico, compresi quelli in Scienze MM.FF.NN., Professioni sanitarie e per tutti quelli che applicano la statistica in campo biosanitario. Questo testo prevede conoscenze basilari di matematica, di algebra e di calcolo.

All'interno del testo sono stati svolti esercizi ed esempi per spiegare concetti di base che utilizzano grandi insiemi di dati, consultabili all'indirizzo www.wiley.com/college/daniel. Inoltre, a questo sito ci si può riferire ogni qualvolta vengono citati "file di Excel" e per eseguire direttamente calcoli per un'analisi statistica al computer. Nella versione italiana del libro, gli esempi che nel testo americano erano trattati con Minitab, SAS o SPSS, sono stati svolti con R e in particolare con l'interfaccia grafica R Commander. La scelta è ricaduta su questo programma perché è gratuito e facilmente scaricabile da tutti. L'utilizzo di software gratuiti è legato al desiderio di agevolare l'approccio alla statistica, incoraggiando la pratica mediante lo svolgimento di numerosi esercizi.

Desidero ringraziare la Dott.ssa Tellaroli per il contribuito alla traduzione in italiano del libro e allo svolgimento degli esempi R. Un ringraziamento va ai colleghi che hanno curato molto puntualmente la revisione dei capitoli. Alla Dott.ssa Cavestri, alla Dott.ssa Favorito e al Dottor Solenne di EdiSES va tutta la mia ammirazione per la pazienza e la dedizione che hanno mostrato nel tenere le fila di questo progetto.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

PRESENTAZIONE DELLA TERZA EDIZIONE

L'XI edizione di *Biostatistica* è stata predisposta con l'obiettivo di richiamare un pubblico vasto. Le edizioni precedenti del libro sono state utilizzate dagli Autori e dai loro colleghi in diversi contesti. Agli studenti universitari iscritti a corsi di Laurea in Scienze Biologiche, Professioni sanitarie e Matematica questa edizione fornisce un'introduzione ai concetti della statistica applicata. Come le precedenti, questa edizione è strutturata per soddisfare le esigenze dei neolaureati in vari ambiti, come le Scienze infermieristiche, le Scienze applicate e la Sanità pubblica e che sono alla ricerca di una solida preparazione nei metodi quantitativi. Per i professionisti che già lavorano in campo medico questa edizione rappresenta un utile testo di riferimento.

La varietà di argomenti fornita da questo testo e le centinaia di esercizi pratici inclusi permettono ai docenti una grande flessibilità nella progettazione di corsi strutturati su più livelli. A tal fine, offriamo i seguenti suggerimenti relativi ai contenuti che riteniamo essere più utili in aula.

Come le precedenti edizioni di questo libro, anche questa richiede alcuni prerequisiti matematici oltre ad una solida competenza algebrica. Abbiamo dato enfasi all'apprendimento pratico e intuitivo dei principi piuttosto che dei concetti astratti sottesi ad alcuni metodi, che necessitano di maggiori competenze matematiche. Perciò abbiamo deciso di proporre problemi ed esempi presi direttamente dalla Letteratura biomedica, invece che utilizzare esempi costruiti *ad hoc*. Crediamo che questo renda il testo più interessante per gli studenti e che fornisca un aiuto più concreto ai professionisti della sanità che faranno riferimento al testo nello svolgimento del loro lavoro.

Per la maggior parte degli esempi e delle tecniche statistiche riportate in questo volume sarà descritto l'uso di software. L'esperienza ci ha portati a decidere di includere in questa edizione esempi trattati con R. Siamo infatti convinti che l'utilizzo di questo strumento arricchisca la presentazione dei materiali e dia agli studenti l'opportunità di comprendere le varie tecniche utilizzate dagli statistici professionisti.

MODIFICHE E AGGIORNAMENTI IN QUESTA EDIZIONE

La maggior parte dei capitoli include correzioni e chiarimenti che migliorano il materiale presentato e lo rendono più facilmente comprensibile ed accessibile. In questa edizione sono state apportate alcune modifiche specifiche e alcuni miglioramenti che crediamo siano contributi preziosi, ringraziamo quindi i revisori delle precedenti edizioni per i loro commenti e suggerimenti a riguardo.

I cambiamenti specifici in questa edizione includono: il testo aggiuntivo sull'installazione e l'utilizzo del programma di R nel Capitolo 1, gli esempi svolti con il programma R presenti nei vari capitoli, il testo aggiuntivo riguardante le misure di dispersione nel Capitolo 2, il testo aggiuntivo nel Capitolo 6, una nuova introduzione ai modelli lineari nel Capitolo 8 che unisce i concetti della regressione e dell'ANOVA nei Capitoli 8-11, l'aggiunta dell'ANOVA a due fattori per misure ripetute nel Capitolo 8, una discussione sulle similitudini fra l'ANOVA e la regressione nel Capitolo 11, un nuovo testo approfondito e corredato di esempi sul test della bontà di adattamento del modello di regressione logistica e la regressione di Poisson nel Capitolo 11, il test di McNemar nel Capitolo 12.

Il cambiamento più importante di questa edizione è il nuovo Capitolo 14 sull'Analisi della Sopravvivenza. Questo capitolo è dovuto alle richieste dei revisori e all'esperienza degli Autori in termini di crescente utilizzo di questi metodi nella ricerca applicata. In questo nuovo capitolo abbiamo incluso alcuni dei materiali contenuti nel Capitolo 12 delle precedenti edizioni, e aggiunto materiali ed esempi dettagliati. Forniamo un'introduzione al concetto di censura, alle stime di Kaplan-Meier, ai metodi per il confronto delle curve di sopravvivenza e al modello di regressione di Cox per rischi proporzionali. Avendo tutto questo nuovo materiale, abbiamo deciso di spostare i contenuti delle statistiche demografiche in un nuovo Capitolo 15.

RINGRAZIAMENTI

Molti revisori, studenti e docenti hanno contribuito a migliorare questo testo con discussioni, revisioni attente e domande puntuali. In particolare, vorremmo ringraziare:

- Dr. Sheniz Moonie, University of Nevada, Las Vegas
- Dr. Guogen Shan, University of Nevada, Las Vegas
- Dr. Gian Jhangri, University of Alberta
- Dr. Tina Cunningham, Eastern Virginal Medical School
- Dr. Shakhawat Hossain, University of Winnipeg
- Dr. Milind Phadnis, University of Kansas Medical Center
- Dr. David Anderson, Xavier University of Louisiana
- Dr. Derek Webb, Bemidji State University
- Dr. Keiji Oda, Loma Linda University
- Dr. David Zietler, Grand Valley State University
- Dr. Genady Grabarnik, St. John's University

• Dr. Al Bartolucci, University of Alabama at Birmingham
• Dr. Hwanseok Choi, University of Southern Mississippi
• Dr. Mark Kelley, University of Pittsburgh at Bradford
• Dr. Wan Tang, Tulane University
• Dr. Phil Gona, University of Massachusetts, Boston
• Dr. Jill Smith, University of California, Riverside
• Dr. Ronnie Brown, University of Baltimore
• Dr. Apoorv Goel, Indiana University-Purdue University Indianapolis
• Dr. Daniel Yorgov, Indiana University-Purdue University Fort Wayne

Dobbiamo ringraziare altre tre persone per i loro importanti contributi al testo. Il Dott. John P. Holcomb della Cleveland State University ha aggiornato molti esempi ed esercizi che troverete nel testo. Il Dott. Edward Danial della Morgan State University ha eseguito una revisione esaustiva della nona edizione, e i suoi preziosi commenti rappresentano un apporto sostanziale al libro. Il Dott. Jodi B. A. McKibben della Uniformed Services University of the Health Sciences ha fornito una revisione dettagliata della decima edizione del libro.

Dobbiamo ringraziare anche i Professori Geoffrey Churchill e Brian Schott della Georgia State University che hanno programmato i codici per generare alcune delle tabelle in Appendice, e il Professor Lillian Lin, che ha letto e commentato il materiale sulla regressione logistica nelle edizioni precedenti del libro. Inoltre, il Dott. James T. Wassell ha fornito assistenza con alcuni dei metodi di analisi della sopravvivenza presentati nelle precedenti edizioni.

Siamo grati ai molti ricercatori nel campo delle scienze mediche che pubblicano i loro risultati e rendono disponibili i loro dati, fornendo materiali preziosi per la pratica degli studenti di biostatistica.

Note finali

Sono eternamente grato di aver avuto l'opportunità di lavorare con il Dr. Wayne Daniel su diverse edizioni di questo testo. Sono stato invitato da Wayne a lavorare con lui in vari incarichi a partire dall'ottava edizione. Da allora, ho avuto il piacere di conoscere Wayne e di apprezzare i suoi alti standard e le sue aspettative. Sfortunatamente Wayne non ha potuto partecipare a questa edizione. Sono onorato che mi abbia affidato di portare avanti la sua eredità.

Chad L. Cross
Las Vegas, Nevada

Materiale di supporto per i docenti

I docenti che utilizzano il testo a scopo didattico possono scaricare dal sito **www.edises.it**, previa registrazione all'area docenti, le immagini del libro in formato Power Point.

INDICE GENERALE

10.5 USO DELL'EQUAZIONE DI REGRESSIONE MULTIPLA

Come abbiamo visto nel capitolo precedente, per ottenere un valore calcolato di Y, cioè $\hat{y}$, quando è fissato un valore di X, possiamo usare l'equazione di regressione. Similmente, possiamo usare l'equazione di regressione multipla per ottenere un valore di $\hat{y}$, quando vengono fissati i valori di due o più variabili X presenti nell'equazione.

Come nel caso della regressione lineare semplice, nella regressione multipla possiamo interpretare il valore $\hat{y}$ in uno dei seguenti modi. Il primo si riferisce all'interpretazione di $\hat{y}$ come una stima della media delle sottopopolazioni dei valori di Y che si assume esistano per delle particolari combinazioni di valori di X_i. In base a questa interpretazione, $\hat{y}$ è chiamata *stima* e l'equazione, quando è usata per questo scopo, è chiamata l'*equazione di stima*. La seconda interpretazione di $\hat{y}$ è che corrisponde al valore più probabile che assume Y per dei valori dati di X_i. In questo caso $\hat{y}$ è chiamato il *valore predetto* di Y e l'equazione è chiamata *equazione di previsione*. In entrambi i casi, quando l'assunzione di normalità del Paragrafo 10.2 è verificata, si possono costruire degli intervalli intorno al valore $\hat{y}$. Quando $\hat{y}$ è interpretato come una stima della media della popolazione, l'intervallo è chiamato *intervallo di confidenza*, mentre quando $\hat{y}$ è interpretato come un valore previsto di Y, l'intervallo è chiamato *intervallo di previsione*. Adesso vediamo come sono costruiti questi intervalli.

L'intervallo di confidenza per la media di una sottopopolazione dei valori di Y per dei particolari valori delle X_i Abbiamo visto che l'intervallo di confidenza al $100(1 - \alpha)\%$ per un parametro può essere costruito seguendo la procedura che prevede di sommare e sottrarre dallo stimatore una quantità uguale al coefficiente di attendibilità, che corrisponde a $1 - \alpha$, moltiplicato per l'errore standard della stima. Inoltre, abbiamo visto che, nella regressione multipla, lo stimatore è dato da:

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_{1j} + \hat{\beta}_2 x_{2j} + \cdots + \hat{\beta}_k x_{kj} \tag{10.5.1}$$

Se denominiamo l'errore standard dello stimatore come $s_{\hat{y}}$, l'intervallo di confidenza al $100(1 - \alpha)\%$ per la media di Y, specificato X_i è

$$\hat{y}_j \pm t_{(1-\alpha/2),\, n-k-1}\, s_{\hat{y}j} \tag{10.5.2}$$

L'intervallo di previsione per un valore particolare di Y dati particolari valori delle X_i Quando si interpreta $\hat{y}$ come valore Y più probabile in base a determinati valori delle X_i, possiamo costruire un intervallo di previsione nello stesso modo in cui è stato costruito l'intervallo di confidenza. L'unica differenza tra i due è l'errore standard. L'errore standard della previsione è leggermente più grande dell'errore standard della stima e ciò rende l'intervallo di previsione più ampio dell'intervallo di confidenza.

Se indichiamo l'errore standard della previsione con $s'_{\hat{y}}$, l'intervallo di previsione al $100(1 - \alpha)\%$ è

$$\hat{y}_j \pm t_{(1-\alpha/2),\, n-k-1}\, s'_{\hat{y}j} \tag{10.5.3}$$

I calcoli di $s_{\hat{y}_j}$ e $s'_{\hat{y}_j}$ nel caso della regressione multipla sono complicati e non verranno trattati in questo testo. Il lettore che vuole sapere come vengono calcolate queste statistiche può consultare il libro di Anderson e Bancroft (3), altri riferimenti elencati alla fine di questo capitolo e del Capitolo 9 e delle precedenti edizioni di questo testo. Nell'esempio seguente si illustra come utilizzare R Commander per ottenere intervalli di confidenza per la media di Y e gli intervalli di previsione per un particolare valore di Y.

ESEMPIO 10.5.1

Riferiamoci nuovamente all'Esempio 10.3.1. Innanzitutto, vogliamo costruire l'intervallo di confidenza al 95% del valore medio della CDA (Y) in una popolazione di soggetti sessantottenni (X_1) che hanno frequentato 12 anni di scuola (X_2). Quindi, supponiamo di avere un soggetto di 68 anni che ha frequentato la scuola per 12 anni. Quale sarà il valore previsto di CDA di questo soggetto?

Soluzione: La stima puntuale del valore medio della CDA è:

$$\hat{y} = 5.49407 - 0.18412(68) + 0.61078(12) = 0.303$$

Anche la previsione puntuale, che è la stessa della stima puntuale ottenuta prima, è

$$\hat{y} = 5.49407 - 0.18412(68) + 0.61078(12) = 0.303$$

Per ottenere un intervallo di confidenza e l'intervallo di previsione per i parametri per cui abbiamo appena calcolato una stima puntuale e una previsione puntuale, scriviamo nella finestra Script di R in R Commander le seguenti istruzioni:

```
new <- data.frame(AGE = 68, EDLEVEL=12)
pred.w.clim <- predict(RegModel.2, new, interval="confidence", level = .95)
pred.w.plim <- predict(RegModel.2, new, interval="prediction", level = .95)
pred.w.clim
pred.w.plim
```

ottenendo l'output:

```
> pred.w.clim
        fit       lwr      upr
1 0.3029173 -1.038481 1.644315

> pred.w.plim
        fit       lwr      upr
1 0.3029173 -6.093481 6.699316
```

Interpretiamo questi intervalli nel solito modo. Prendiamo in esame per primo l'intervallo di confidenza. Siamo confidenti al 95% che l'intervallo da –1.038 a 1.644 includa la media della sottopopolazione dei valori Y per la combinazione specificata dei valori X_i, poiché questo parametro sarebbe incluso in circa il 95% degli intervalli che possono essere costruiti nel modo illustrato.

Ora consideriamo il soggetto che ha 68 anni e 12 anni di istruzione. Siamo confidenti al 95% che questo soggetto avrebbe ottenuto un punteggio CDA tra –6.093 e 6.699. Il fatto che l'intervallo di previsione sia più ampio dell'intervallo di confidenza non dovrebbe sorprenderci. Dopotutto, è più facile stimare la risposta media di quanto non sia stimare un'osservazione individuale.

■

ESERCIZI

Per ognuno degli esercizi seguenti calcolare il valore di y e costruire (a) l'intervallo di confidenza al 95% e (b) gli intervalli di previsione al 95% per dei valori specifici di X_i.

10.5.1 Riferendosi ai dati dell'Esercizio 10.3.1, sia $x_{1j} = 95$ e $x_{2j} = 35$.

10.5.2 Riferendosi ai dati dell'Esercizio 10.3.2, sia $x_{1j} = 50$, $x_{2j} = 20$ e $x_{3j} = 22$.

10.5.3 Riferendosi ai dati dell'Esercizio 10.3.3, sia $x_{1j} = 5$ e $x_{2j} = 6$.

10.5.4 Riferendosi ai dati dell'Esercizio 10.3.4, sia $x_{1j} = 1$ e $x_{2j} = 2$.

10.5.5 Riferendosi ai dati dell'Esercizio 10.3.5, sia $x_{1j} = 90$ e $x_{2j} = 80$.

10.5.6 Riferendosi ai dati dell'Esercizio 10.3.6, sia $x_{1j} = 50$, $x_{2j} = 95.0$, $x_{3j} = 2.00$, $x_{4j} = 6.00$, $x_{5j} = 75$ e $x_{6j} = 70$.

10.6 MODELLO DI CORRELAZIONE MULTIPLA

Nel Capitolo precedente abbiamo sottolineato che, mentre l'analisi della regressione riguarda la forma della relazione fra le variabili, l'obiettivo dell'analisi della correlazione è conoscere a fondo la forza della relazione. Questo è vero anche nel caso multivariato e, in questo paragrafo, studiamo i metodi per misurare la forza della relazione fra molte variabili. Prima, comunque, definiamo il modello e le assunzioni sulle quali si basa l'analisi.

L'equazione del modello Possiamo scrivere il modello di correlazione come

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_k x_{kj} + \epsilon_j \tag{10.6.1}$$

dove y_i è un valore specifico proveniente dalla popolazione dei valori della variabile Y, i β sono i coefficienti di regressione definiti nel Paragrafo 10.2, gli x_{ij} sono particolari valori (noti) delle variabili casuali X_i. Il modello è simile al modello di regressione multipla, ma con un'importante differenza. Nel modello di regressione multipla dato dall'Equazione 10.2.1, le X_i sono variabili non casuali, mentre nel modello di correlazione multipla le X_i sono variabili casuali. In altri termini, nel modello di correlazione c'è una distribuzione congiunta di Y e delle X_i, che chiamiamo *distribuzione multivariata*. In tale modello, le variabili non sono più pensate come indipendenti o dipendenti, poiché logicamente esse sono interscambiabili, per cui ciascuna delle X_i può assumere il ruolo di Y.

Di solito, da una popolazione di interesse vengono estratti dei campioni casuali di unità di associazione da cui si ottengono dei valori delle X_i e della Y.

Con i metodi descritti nel Paragrafo 10.3, adattiamo un piano o un iperpiano dei minimi quadrati ai dati campionari utilizzando le equazioni risultanti, come fatto in precedenza. Se possiamo assumere che la distribuzione della popolazione dalla quale abbiamo estratto il campione è normale, cioè assumiamo che la distribuzione congiunta di Y e delle X_i è una *distribuzione normale multivariata*, possiamo applicare le procedure inferenziali. In più, si possono calcolare le misure campionarie del grado della relazione fra le variabili e, sotto l'assunzione che esso provenga da una popolazione normale multivariata, i parametri corrispondenti possono essere stimati attraverso gli intervalli di confidenza e, inoltre, si possono eseguire i test per la verifica delle ipotesi. In particolare, possiamo calcolare una stima del *coefficiente di correlazione multipla* che misura la dipendenza fra Y e le X_i. Questa è una ovvia estensione del concetto della correlazione fra due variabili che abbiamo trattato nel Capitolo 9. Possiamo calcolare anche i *coefficienti di correlazione parziale* che misurano l'intensità della relazione tra ogni coppia di variabili, dopo aver eliminato l'influenza di tutte le altre variabili.

Il coefficiente di correlazione multipla Come primo passo per analizzare la relazione tra le variabili, consideriamo il coefficiente di correlazione multipla.

Il coefficiente di correlazione multipla è la radice quadrata del coefficiente di determinazione multiplo e, di conseguenza, il suo valore può essere calcolato facendo la radice quadrata dell'Equazione 10.4.2, cioè:

$$R_{y.12\ldots k} = \sqrt{R^2_{y.12\ldots k}} = \sqrt{\frac{\sum(\hat{y}_j - \bar{y})^2}{\sum(y_j - \bar{y})^2}} = \sqrt{\frac{SSR}{SST}} \tag{10.6.2}$$

Per illustrare i concetti e le tecniche dell'analisi della correlazione multipla, consideriamo un esempio.

ESEMPIO 10.6.1

Wang *et al*. (A-4), usando femori di cadaveri umani di soggetti di età compresa fra 16 e 19 anni, hanno studiato la resistenza dell'osso e misurato il reticolo di collagene all'interno dell'osso. Due variabili che misurano il reticolo di collagene sono la porosità (P, espressa in percentuale) e la forza di tensione (S). La misura di resistenza dell'osso (W, in Newton) è la forza richiesta per fratturare l'osso. I 29 femori utilizzati in questo studio erano privi di patologie ossee. Vogliamo analizzare la natura e la forza della relazione tra le tre variabili. Le misure sono riportate nella tabella seguente.

TABELLA 10.6.1 Resistenza dell'osso e proprietà del reticolo di collagene di 29 femori

W	P	S
193.6	6.24	30.1
137.5	8.03	22.2
145.4	11.62	25.7
117.0	7.68	28.9
105.4	10.72	27.3
99.9	9.28	33.4
74.0	6.23	26.4
74.4	8.67	17.2
112.8	6.91	15.9
125.4	7.51	12.2
126.5	10.01	30.0
115.9	8.70	24.0
98.8	5.87	22.6
94.3	7.96	18.2
99.9	12.27	11.5
83.3	7.33	23.9
72.8	11.17	11.2
83.5	6.03	15.6
59.0	7.90	10.6

(Continua)

W	P	S
87.2	8.27	24.7
84.4	11.05	25.6
78.1	7.61	18.4
51.9	6.21	13.5
57.1	7.24	12.2
54.7	8.11	14.8
78.6	10.05	8.9
53.7	8.79	14.9
96.0	10.40	10.3
89.0	11.72	15.4

Fonte: Xiaodu Wang, Ph.D. Riprodotto per gent. conc. dell'autore.

Soluzione: Utilizziamo R Commander per eseguire l'analisi dei nostri dati. Il lettore interessato a ricavare le formule e le procedure aritmetiche impiegate può consultare i testi indicati alla fine di questo capitolo e del Capitolo 9, nonché nella precedente edizione di questo testo. Per ottenere un'equazione di previsione dei minimi quadrati e il coefficiente di correlazione multipla come desiderato in questa parte dell'analisi, possiamo allora usare la procedura di regressione multipla di R Commander. Quando facciamo questo per i valori campionari di Y, X_1 e X_2, salvati nelle colonne, rispettivamente, dalla 1 alla 3, otteniamo un output come in Figura 10.6.1.

L'equazione dei minimi quadrati è allora

$$\hat{y} = 35.61 + 1.451x_{1j} + 2.3960x_{2j}$$

```
Call:
lm(formula = W ~ P + S, data = Esel061)

Residuals:
    Min      1Q  Median      3Q     Max
-33.907 -19.594  -0.517  10.159  76.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.6138    29.1296   1.223  0.23245
P             1.4509     2.7632   0.525  0.60397
S             2.3960     0.7301   3.282  0.00294 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.42 on 26 degrees of freedom
Multiple R-squared:  0.2942, Adjusted R-squared:  0.2399
F-statistic: 5.419 on 2 and 26 DF,  p-value: 0.01078
```

FIGURA 10.6.1 Output della procedura di regressione lineare multipla con R Commander per i dati della Tabella 10.6.1.

Questa equazione può essere usata per ottenere stime e valori predetti e può essere valutata con i metodi discussi nel Paragrafo 10.4.

Come vediamo nella Figura 10.6.1, l'output di regressione multipla fornisce anche il coefficiente di determinazione multiplo, che nel presente esempio è

$$R^2_{y.12} = 0.294$$

Il coefficiente di correlazione multipla, pertanto, è

$$R_{y.12} = \sqrt{0.294} = 0.542$$

L'interpretazione di $R_{y.12}$

Interpretiamo $R_{y.12}$ come una misura della correlazione fra le variabili forza richiesta per la frattura, porosità e forza del reticolo di collagene nel campione di 29 femori di soggetti di età compresa fra 16 e 19 anni. Se i dati costituiscono un campione casuale della popolazione di persone di questo tipo, possiamo considerare $R_{y.12}$ come una stima di $\rho_{y.12}$, cioè il vero coefficiente di correlazione multipla nella popolazione. Possiamo interpretare $R_{y.12}$ anche come il coefficiente di correlazione semplice fra y_j e $\hat{y}$ cioè, rispettivamente, fra i valori osservati e i valori calcolati della variabile dipendente. Se il coefficiente di correlazione è pari a 1, significa che vi è una perfetta corrispondenza fra i valori di Y osservati e calcolati, mentre quando il coefficiente è pari a 0, significa che, fra i valori osservati e calcolati di Y, non esiste una relazione lineare. Il coefficiente di correlazione multipla è sempre di segno positivo.

Possiamo saggiare l'ipotesi nulla che $\rho_{y.12\ldots k} = 0$ calcolando

$$F = \frac{R^2_{y.12\ldots k}}{1 - R^2_{y.12\ldots k}} \cdot \frac{n - k - 1}{k} \tag{10.6.3}$$

Il valore numerico ottenuto dall'Equazione 10.6.3 viene confrontato con il valore tabulato di F con k e $n - k - 1$ gradi di libertà. Il lettore ricorderà che tale test è identico al test H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ descritto nel Paragrafo 10.4.

Per il nostro esempio, testiamo l'ipotesi nulla che $\rho_{y.12} = 0$ contro l'ipotesi alternativa che $\rho_{y.12} \neq 0$. Calcoliamo

$$F = \frac{0.294}{1 - 0.294} \cdot \frac{29 - 2 - 1}{2} = 5.41$$

Poiché 5.41 è maggiore di 4.27, con $p < 0.025$, possiamo rifiutare l'ipotesi nulla a un livello di significatività uguale a 0.025 e concludere che la forza necessaria per fratturare l'osso è correlata con la porosità e la misura di forza della rete di collagene nella popolazione campionata.

Il valore calcolato di F, per testare l'ipotesi H_0 nulla che il coefficiente di correlazione multipla nella popolazione sia uguale a zero, è indicato nella Figura 10.6.1 ed è pari a 5.42. I due valori calcolati di F differiscono per effetto di differenze nell'arrotondamento dei calcoli intermedi.

■

La correlazione parziale Il ricercatore può avere bisogno di una misura per la forza della relazione lineare tra due variabili eliminando l'effetto delle restanti variabili. Tale misura si chiama *coefficiente di correlazione parziale*. Ad esempio, il coefficiente di correlazione parziale $r_{y.12}$ è una misura della correlazione fra Y e X_1 controllando per l'effetto di X_2.

I coefficienti di correlazione parziale possono essere calcolati dai *coefficienti di correlazione semplice*. I coefficienti di correlazione semplice misurano la correlazione fra due variabili quando non viene fatto alcuno sforzo per controllare le altre variabili. In altri termini, essi sono i coefficienti per ogni coppia di variabili che possono essere ottenuti con i metodi della correlazione semplice trattati nel Capitolo 9.

Supponiamo di avere tre variabili, Y, X_1 e X_2. Il coefficiente di correlazione parziale campionario, che misura la correlazione fra Y e X_1 controllando per X_2 viene scritto $r_{y1.2}$. Nei pedici il simbolo alla destra del punto indica la variabile per la quale si controlla, mentre i due simboli alla sinistra indicano quali variabili vengono correlate. Nel caso di tre variabili, si possono calcolare due coefficienti di correlazione parziale campionaria che possiamo calcolare e cioè $r_{y2.1}$ e $r_{12.y}$.

Il coefficiente di determinazione parziale Il quadrato del coefficiente di correlazione parziale è chiamato coefficiente di determinazione parziale. Esso fornisce un'informazione utile circa le interrelazioni fra le variabili. Consideriamo ad esempio $r_{y1.2}$. Il suo quadrato, $r^2_{y1.2}$, ci dice quanta parte della variabilità residua di Y è spiegata da X_1 dopo che X_2 ha spiegato la massima variabilità possibile di Y.

Il calcolo dei coefficienti di correlazione parziale Per le tre variabili possono essere calcolati i seguenti coefficienti di correlazione semplice:

r_{y1}, coefficiente di correlazione semplice fra Y e X_1
r_{y2}, coefficiente di correlazione semplice fra Y e X_2
r_{12}, coefficiente di correlazione semplice fra X_1 e X_2

La procedura di correlazione di R Commander può essere usata per calcolare i coefficienti di correlazione semplice come mostrato nella Figura 10.6.2. Dall'output della Figura 10.6.2 si vede che $r_{12} = -0.08$, $r_{y1} = 0.043$ e $r_{y2} = 0.536$. Per ottenere anche i valori della significatività statistica p, selezionare l'opzione "p-values a coppie".

```
 Pearson correlations:
        P       S      W
P  1.0000 -0.0802 0.0433
S -0.0802  1.0000 0.5355
W  0.0433  0.5355 1.0000

 Number of observations: 29

 Pairwise two-sided p-values:
  P      S      W
P        0.6794 0.8235
S 0.6794        0.0028
W 0.8235 0.0028

 Adjusted p-values (Holm's method)
  P      S      W
P        1.0000 1.0000
S 1.0000        0.0083
W 1.0000 0.0083
```

Original menu → Statistics → Summaries → Correlation matrix…

FIGURA 10.6.2 Output della procedura di regressione lineare multipla con R Commander per i dati della Tabella 10.6.1.

I coefficienti di correlazione parziale campionaria, che si possono calcolare nel caso di tre variabili, sono:

1. il coefficiente di correlazione parziale fra Y e X_1 al netto dell'effetto di X_2:

$$r_{y1.2} = \left(r_{y1} - r_{y2}r_{12}\right) \Big/ \sqrt{\left(1-r_{y2}^2\right)\left(1-r_{12}^2\right)} \tag{10.6.4}$$

2. il coefficiente di correlazione parziale fra Y e X_2 al netto dell'effetto di X_1:

$$r_{y2.1} = \left(r_{y2} - r_{y1}r_{12}\right) \Big/ \sqrt{\left(1-r_{y1}^2\right)\left(1-r_{12}^2\right)} \tag{10.6.5}$$

3. il coefficiente di correlazione parziale fra X_1 e X_2 al netto dell'effetto di Y:

$$r_{12.y} = \left(r_{12} - r_{y1}r_{y2}\right) \Big/ \sqrt{\left(1-r_{y1}^2\right)\left(1-r_{y2}^2\right)} \tag{10.6.6}$$

ESEMPIO 10.6.2

Per illustrare il calcolo dei coefficienti di correlazione parziale campionaria riferiamoci all'Esempio 10.6.1 e calcoliamo il coefficiente di correlazione parziale fra le variabili: quantità di forza per la frattura ($W = Y$), porosità ($P = X_1$) e la forza del reticolo di collagene ($S = X_2$).

Soluzione: Invece di calcolare i coefficienti di correlazione parziale dai coefficienti di correlazione semplice con le Equazioni dalla 10.6.4 alla 10.6.6, utilizziamo R Commander per ottenerli.

La procedura di R Commander per calcolare i coefficienti di correlazione parziale è uguale a quella per calcolare i coefficienti di correlazione, basterà soltanto scegliere sulla finestra come tipo di correlazione "Parziale". La procedura per i dati della Tabella 10.6.1 è mostrata nella Figura 10.6.3 insieme all'output che mostra che $r_{y1.2} = 0.102$, $r_{12.y} = -0.122$ e $r_{y2.1} = 0.541$.

■

Il test di ipotesi per i coefficienti di correlazione parziale Possiamo testare l'ipotesi nulla che tutti i coefficienti di correlazione parziale della popolazione siano pari a 0 attraverso il test t. Per esempio per verificare H_0: $\rho_{y1.2\ldots k} = 0$ calcoliamo

$$t = r_{y1.2\ldots k}\sqrt{\frac{n-k-1}{1-r_{y1.2\ldots k}^2}} \tag{10.6.7}$$

che si distribuisce come una t di Student con $n - k - 1$ gradi di libertà.

Illustriamo la procedura per il nostro esempio testando H_0: $\rho_{y1.2} = 0$, contro l'alternativa H_A: $\rho_{y1.2} \neq 0$. Il valore di t calcolato è

$$t = 0.102\sqrt{\frac{29-2-1}{1-(0.102)^2}} = 0.523$$

Wayne W. **Daniel** • Chad L. **Cross**

Biostatistica

Accedi all'**ebook** e ai **contenuti digitali** › **Espandi** le tue risorse › con un libro che **non pesa** e si **adatta** alle dimensioni del tuo **lettore**

All'interno del volume il **codice personale** e le istruzioni per accedere alla versione **ebook** del testo e agli ulteriori servizi. L'accesso alle risorse digitali è **gratuito** ma limitato a **18 mesi dalla attivazione del servizio**.

EdiSES
www.edises.it

ISBN 978-88-3319-041-9

€ 49,00