Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 24

Sommario

Sommario

Il materiale coperto è contenuto nel capitolo 10 del testo di Agresti e Finlay (2012)

- Tipi di indagine
- Associazione e causalità
- Controllo statistico
- Relazioni spurie
- Variabili soppressorie
- Variabili intervenienti
- Cause multiple
- Interazioni tra variabili

Tipi fondamentali di indagine

Tipi fondamentali di indagine

Si distinguono due tipi fondamentali di indagine

1. gli esperimenti
2. gli studi osservazionali.

Studi sperimentali

- Negli esperimenti il fenomeno oggetto di studio è sotto il controllo del ricercatore. Si studia l'effetto di uno o più trattamenti sulle unità sperimentali.
- Nel caso più semplice, i soggetti vengono assegnati casualmente dal ricercatore al gruppo di trattamento o al gruppo di controllo. L'obiettivo è comparare gli effetti del trattamento sulla risposta.
- La randomizzazione consente di produrre dei gruppi in cui tutte le altre variabili, anche quelle non misurate, hanno approssimativamente la stessa distribuzione e quindi sono comparabili.
- Le differenze sistematiche nelle risposte sono attribuite all'effetto del trattamento.

Illustrazione

Illustrazione. Nell'esperimento sull'efficacia di un farmaco, per esempio, i pazienti vengono assegnati in maniera casuale al gruppo di trattamento o al gruppo di controllo.

Né i pazienti né i ricercatori sanno a quale gruppo appartengono i pazienti (esperimento *double-blind*).

I gruppi sono in tutto confrontabili tranne per la modalità del trattamento, e quindi eventuali differenze nella risposta non possono essere dovute ad altro che al trattamento.

Studi osservazionali

- Negli studi osservazionali il fenomeno oggetto di studio esiste in natura e non è controllabile dallo sperimentatore.
- I soggetti si auto-assegnano al gruppo di controllo o al trattamento: il ricercatore deve osservare passivamente ciò che accade.

Studi osservazionali

Illustrazione. studio sugli effetti del fumo.

- Trattamento: fumo.
- Risposta: sviluppo malattie repiratorie.
- Auto-selezione degli individui al trattamento: fumatori vs non fumatori (controllo).

Limiti degli studi osservazionali

- Negli studi osservazionali la mancanza di controllo sull'assegnazione dei trattamenti fa sì che i gruppi non siano mai totalmente comparabili e quindi eventuali differenze nella risposta sono attribuibili
  - al fumo e/o
  - ad altri fattori non controllati (fattori ereditari, età, sesso).
- Anche se con metodi statistici è possibile misurare le differenze nella risposta aggiustate per eventuali non comparabilità nei gruppi trattati, agli studi osservazionali non possono essere assegnate interpretazioni di natura causale.

Associazione e causalità

Associazione e causalità

L'associazione tra due o più variabili non prova l'esistenza di un rapporto di tipo causale.

- Tra due variabili può esserci associazione e non causalità.
- Il concetto di causalità è difficile e controverso.

Associazione e causalità

- La nozione di causalità riveste un ruolo centrale nella ricerca scientifica. Tutti sono familiari con il concetto di "nesso causale", almeno in un senso informale.
  - Es. essere esposti ad un virus *causa* l'influenza.
- Non tutti sanno però come un possibile nesso causale possa essere sottoposto a verifica in maniera rigorosa.

Associazione e causalità

Tale verifica è particolarmente problematica in psicologia dove le relazioni tra le variabili hanno una natura probabilistica piuttosto che deterministica.

- Es. cosa causa la delinquenza giovanile?
  - Essere poveri?
  - Provenire da famiglie con un solo parente?
  - La mancanza di educazione morale e religiosa?
  - Fattori genetici?
  - Una combinazione di questi e altri fattori?

Associazione e causalità

Affinché possa essere considerata di natura causale, una relazione tra due variabili deve soddisfare tre criteri.

1. Esiste "covariazione" tra la variabile dipendente e la variabile indipendente.
2. Esiste una sequenza temporale tra le variabili. Il cambiamento nella presunta variabile indipendente deve precedere il cambiamento nella presunta variabile dipendente.
3. Assenza di spiegazioni alternative. Es. l'associazione fra le due variabili potrebbe essere il prodotto di una causa comune ad entrambe (relazione spuria).

Associazione e causalità

Diciamo che X *causa* Y se:

- esiste una relazione tra X e Y;
- la relazione è asimmetrica (una variazione di X causa una variazione di Y, ma non viceversa);
- una variazione di X ha come risultato una variazione di Y quali che siano le influenze di altri fattori.

Associazione e causalità

- Se tutti questi tre criteri risultano soddisfatti, allora si conclude che le evidenze empiriche vanno a sostegno dell'esistenza di un nesso causale.
- Se uno o più dei criteri precedenti non sono soddisfatti, allora si conclude che non vi è nesso causale.

Associazione e causalità

- Non è mai possibile provare che una prima variabile sia la causa di una seconda variabile, dato che i nessi causali sono inferiti dallo sperimentatore e mai osservati direttamente in quanto tali.
- L'unica cosa che il ricercatore può fare è falsificare le ipotesi di natura causale.
- Un'ipotesi causale può essere dimostrata falsa se le evidenze empiriche non soddisfano almeno uno dei requisiti elencati in precedenza.

Associazione e causalità

Associazione e causalità

In psicologia è difficile stabilire con certezza relazioni causali per i seguenti motivi:

- le teorie psicologiche non sono sufficientemente precise da consentire di identificare i nessi causali;
- possono esserci tante possibili cause e le teorie psicologiche non riescono ad isolare le cause corrette;
- le sequenze temporali non sono chiaramente definite.

Sequenza temporale

Sequenza temporale

Sequenza temporale

Sequenza temporale

La *causa* deve precedere l'*effetto*.

- Nella ricerca sperimentale la sequenza temporale è sotto il controllo dello sperimentatore: somministrazione di un farmaco o placebo -> decorso della malattia.
- Il successo o meno del trattamento è osservato *dopo* che il trattamento viene somministrato, dunque la sequenza temporale è certa.

Sequenza temporale

Sequenza temporale

- Precedenza logica: l'età e il genere sono antecedenti al reddito, per esempio, e quindi devono essere considerati come cause piuttosto che come effetti.
- Nell'esempio della tavola 10.1 sullo *scouting*, S, e la delinquenza giovanile, D, sono possibili sia $S \rightarrow D$ che $D \rightarrow S$.

Eliminazione di spiegazioni alternative
Evidenze aneddotiche

Eliminazione di spiegazioni alternative

Eliminazione di spiegazioni alternative

Eliminazione di spiegazioni alternative
Evidenze aneddotiche

I requisiti dell'associazione statistica e della sequenza temporale risultano soddisfatti in molti casi in cui è assente una relazione causale.

- Vi potrebbe essere, infatti, una terza variabile Z che è al contempo causa di X e di Y, per cui se Z viene controllata, l'associazione tra X e Y scompare, ossia Y risulta statisticamente indipendente da X e la presunta relazione causale tra X e Y si rivela spuria.
- L'associazione tra due variabili può essere mediata da una terza variabile che interviene tra le due.
- L'associazione compare per errori sistematici nello studio (es. errori di campionamento).

Evidenze aneddotiche

Le evidenze aneddotiche non sono sufficienti per provare che non vi è una relazione causale tra due variabili (es. fumo e cancro).

- L'associazione tra due variabili, infatti, non deve essere deterministica per essere causale.
- Per provare che non vi è relazione causale è necessario stabilire che viene violato il criterio dell'associazione statistica, o quello della sequenza temporale, o che vi sono spiegazioni alternative.

Controllo statistico

Controllo statistico

- Per affermare che X causa Y è necessario stabilire che l'associazione tra X e Y viene preservata quando l'influenza di altre variabili è rimossa.
- Si dice che una variabile è *controllata* quando la sua influenza su altre variabile viene eliminata.
- In un esperimento, una variabile può essere controllata mantenendo costante il suo valore mentre si studia la relazione tra altre variabili.
  - Es. in un esperimento di laboratorio sugli effetti di un farmaco sui ratti il ricercatore può controllare l'età o la dieta dei ratti (tutti i ratti hanno la stessa età o dieta).

Controllo statistico

- Nelle ricerche osservazionali le variabili che si vogliono controllare (es. l'intelligenza) non possono essere tenute costanti.
- Possiamo però approssimare il tipo di controllo che si ottiene negli esperiementi ragruppando insieme le osservazioni con valori simili rispetto alla variabile che si vuole controllare.
  - Es. per controllare l'effetto del grado di istruzione si possono ragruppare in gruppi gli individui che fanno parte del campione esaminato (licenza media, superiore, laurea, dottorato).
- Questa forma di controllo si dice *controllo statistico* anziché sperimentale.

Illustrazione

- Agresti e Finlay considerano un campione di dati che contiene, tra l'altro, i punteggi di un test di abilità matematica e l'altezza degli studenti (figura 10.1).
- Tra abilità matematica e altezza c'è una correlazione pari a 0.81: studenti più alti hanno punteggi più alti di abilità matematica.
- L'altezza influenza l'abilità matematica, oppure ci sono delle spiegazioni alternative?

Illustrazione

- Una variabile che certamente influenza entrambe le variabili "altezza" e "abilità matematica" è l'età. All'aumentare dell'età entrambe queste variabili aumentano.
- Possiamo eliminare l'effetto dell'età dall'associazione tra altezza e ablilità matematica controllando statisticamente questa variabile.
- Questo risultato si ottiene esaminando separatamente la relazione tra altezza e ablilità matematica in gruppi d'età omogenei.

Illustrazione

- Se questo viene fatto, all'interno di ciascun gruppo d'età il coefficiente di correlazione diventa prossimo a zero.
- L'altezza non è dunque una causa dell'abilità matematica perché l'associazione tra altezza e ablilità matematica scompare quando l'età viene tenuta costante.

Variabile di controllo

- Mantenendo costante una *variabile di controllo* possiamo eliminare la porzione dell'associazione tra X e Y che è dovuta alla sua variazione.
- I grafici che rappresentano la relazione tra due variabili mantenendo costante il valore di una variabile di controllo vengono detti *parziali*.
  - Nell'esempio precedente abbiamo descritto la relazione tra due variabili quantitative (abilità matematica e altezza) nel caso in cui una terza variabile (età) è mantenuta costante.
  - E' anche possibile descrivere l'associazione tra due variabili qualitative controllando statisticamente l'effetto di una terza variabile.

Illustrazione

- Nella tabella 10.1 viene riportata un'ipotetica relazione bivariata tra *scouting* e comportamenti delinquenziali.

		Comportamenti delinquenziali	
		sì	no
scout	sì	36 (.09)	364 (.91)
	no	60 (.15)	340 (.85)

- Si noti che la proporzione di comportamenti delinquenziali è più alta tra coloro che non si dedicano allo scouting.

Illustrazione

Nella tabella seguente, la relazione tra scouting e comportamenti delinquenziali viene esaminata controllando la variabile "Partecipazione ad attività religiose".

		Partecipazione ad attività religiose					
		bassa		media		alta	
Delinq.		sì	no	sì	no	sì	no
scout	sì	10 (.20)	40 (.80)	18 (.12)	132 (.88)	8 (.04)	192 (.96)
	no	40 (.20)	160 (.80)	18 (.12)	132 (.88)	2 (.04)	48 (.96)

In ciascuna tavola parziale, la proporzione di comportamenti delinquenziali è la stessa tra scout e non scout.

Illustrazione

- Quando la partecipazione ad attività religiose è controllata, la relazione tra scouting e comportamenti delinquenziali scompare.
- Possiamo dunque concludere che non vi è un nesso causale tra scouting e comportamenti delinquenziali.

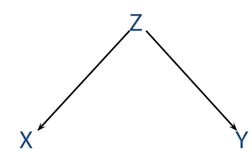
Relazione spuria

Relazione spuria

- Un'associazione tra due varibili X e Y è detta *spuria* se dipende da una terza variabile Z e scompare quando tale variabile viene controllata.
- Se l'associazione tra X e Y dipende dall'effetto di una terza variabile Z, allora non può essere interpretata come evidenza di un nesso causale tra X e Y.
  - Es. relazione spuria tra altezza e abilità matematica;
  - Es. relazione spuria tra scouting e comportamenti delinquenziali.

Relazione spuria

Rappresentazione grafica di una relazione spuria tra X e Y. L'associazione tra X e Y scompare quando la variabile Z viene controllata.



Relazione spuria

- In pratica, anche se la relazione tra due variabili è spuria, a causa dell'errore di campionamento, le associazioni parziali rivelate dai dati del campione non rivelano mai una completa assenza di associazione.
- Inoltre, poche relazioni sono completamente spurie. Ci può essere un qualche nesso causale tra due variabili, anche se non così forte come suggerito dall'analisi bivariata.

Illustrazione

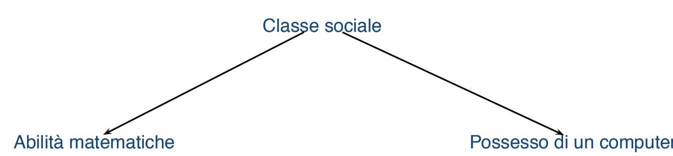
- Esempio 10.4: gli studenti che posseggono un computer (X) tendono ad avere punteggi (Y) più alti nei test che misurano le abilità matematiche (tavola 10.3). Possiamo concludere che il possesso di un computer faccia migliorare il rendimento nei test matematici?
- Dobbiamo considerare la spiegazione alternativa che ipotizza l'esistenza di una terza variabile (Z) che esercita un'influenza causale su entrambe (tavola 10.4).

Illustrazione

- La classe sociale ha un effetto sia sui punteggi di abilità matematica che sul possesso di un computer. All'aumentare della classe sociale aumenta sia la probabilità di ottenere punteggi più alti nel test matematico che la probabilità di possedere un computer.
- Questo fatto è coerente con l'esistenza di una relazione spuria tra punteggi nel test matematico e possesso di un computer.

Illustrazione

Tale relazione spuria è rappresentata di seguito.



Relazione a catena

Relazione a catena

- Le relazioni spurie non rappresentano l'unico caso in cui l'associazione tra X e Y scompare quando controlliamo una terza variabile Z.
- Si dicono *variabili intervenienti* quelle variabili che possono intervenire, influenzare o mascherare la relazione tra X e Y:

X -> Z -> Y

Relazione a catena

- Per esempio, si può osservare un'associazione tra razza e tasso di criminalità.
- Tale associazione, però, può scomparire se il reddito familiare viene controllato:

razza -> reddito familiare tasso di criminalità

Relazione a catena

- Sia nel caso delle relazioni spurie che delle relazioni a catena, l'associazione tra X e Y scompare quando la variabile Z viene controllata.
- La differenza sta nell'ordine causale tra le variabili e nell'interpretazione della relazione:
  - nel caso di una associazione spuria tra X e Y, la variabile Z costituisce un antecedente causale di entrambe.
  - nel caso di una associazione a catena, la variabile Z interviene tra X e Y.

Illustrazione

- Supponiamo che la X, livello di istruzione del padre, sia positivamente associata alla Y, il reddito del figlio all'età di quarant'anni, ma che quest'associazione scompaia se l'educazione del figlio viene controllata.
  - La correlazione tra reddito del figlio e livello di istruzione del padre è prossima a zero se viene calcolata all'interno di gruppi di figli aventi un livello omogeneo di istruzione.

Illustrazione

- In questo caso, si può pensare che il livello di istruzione del padre influenzi il livello di istruzione del figlio e, a sua volta, quest'ultimo influenzi il reddito. E' dunque naturale considerare questa situazione nei termini di un modello a catena.

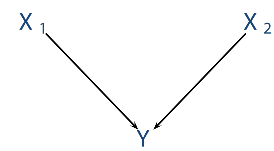
Istruzione padre -> Istruzione figlio -> Reddito

- Non ha invece senso considerare questa situazione nei termini di una relazione spuria in cui l'educazione del figlio è causalmente antecedente sia all'istruzione del padre che al reddito del figlio.

Cause multiple

Cause multiple

In psicologia, le variabili dipendenti sono solitamente influenzate da molteplici cause.



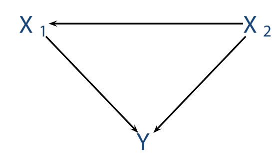
Cause multiple

Nella rappresentazione precedente, le variabili X_1 e X_2 sono incorrelate.

- Due variabili esplicative incorrelate potrebbero essere, per esempio, il genere e la razza.
- Se una variabile esplicativa X_2 influenza la variabile Y, ma X_2 è incorrelata con X_1, allora se X_2 viene esclusa dall'analisi, l'associazione misurata tra X_1 e Y non ne risulta distorta.

Cause multiple

Consideriamo ora il caso in cui le variabili X_1 e X_2 sono associate.



- In queste circostanze, se la variabile X_2 viene esclusa dall'analisi, allora l'associazione misurata tra X_1 e Y risulterà distorta, ovvero, non rispecchierà l'associazione che in realtà intercorre tra le due variabili.
- L'errore che si commette in tali circostanze si chiama errore di specificazione.

Variabili soppressorie

Variabili soppressorie

- Talvolta una relazione tra due variabili non risulta apparente fintanto che non viene controllato l'effetto di una terza variabile. Tale variabile si dice variabile soppressoria (*suppressor variable*).
- Nella tabella 10.6 sono rappresentati dei dati ipotetici in cui non vi è relazione tra reddito ed educazione.

	Reddito	
Istruzione	Alto	Basso
Alta	250	250
Bassa	250	250

Variabili soppressorie

- Nella tabella 10.5, gli stessi dati sono rappresentati in due tabelle parziali in cui viene controllata la variabile "età".

		Età bassa			Età alta		
	Reddito	Alto	Basso	% Alto	Alto	Basso	% Alto
Istruz.	Alta	125	225	35.7	125	25	83.3
	Bassa	25	125	16.7	225	125	64.3

- Se l'età viene controllata, in ciascuna tabella parziale, la percentuale di individui ad alto reddito è maggiore nel caso di un alto livello di istruzione. Il reddito, dunque, tende a crescere con il livello di istruzione.

Variabili soppressorie

- Se ignoriamo l'effetto della variabile età, però, la relazione bivariata tra reddito e livello di istruzione scompare.
- Questo esempio suggerisce che è importante studiare le associazioni parziali anche quando le analisi bivariate rivelano assenza di associazione.
- L'esempio presente rappresenta il caso di una *assenza di relazione spuria*, in quanto sia il livello di istruzione che il reddito dipendono da una terza variabile, ovvero l'età.

Interazione statistica

Interazione statistica

- Spesso la forza dell'associazione tra due variabili muta a seconda del valore assunto da un terza variabile.
- Se il grado di associazione tra Y e X_1 assume valori diversi a seconda dei diversi livelli della variabile X_2, allora si dice che le variabili X_1 e X_2 *interagiscono* nel determinare il valore della variabile Y.
- L'esistenza di interazione tra variabili si può stabilire esaminando se, per i dati di un campione, la relazione tra Y e X_1 rimane costante all'interno di ciascun livello di X_2.

Illustrazione

Razza vittima	Razza accusato	Pena di morte sì	Pena di morte no	Proporzione sì
Bianca	Bianca	227	4418	.049
	Nera	92	639	.126
Nera	Bianca	9	255	.034
	Nera	36	4392	.008
Totale	Bianca	236	4673	.048
	Nera	128	5031	.025

Illustrazione

- Per questo campione (Radelet e Pierce, 1991), la pena di morte (Y) viene inflitta più spesso agli accusati di razza bianca che a quelli di razza nera (X_1).
- Esaminiamo però quello che succede all'interno delle tabelle parziali in cui viene controllato l'effetto della variabile "razza della vittima" X_2.

Illustrazione

- Quando la vittima è bianca, la probabilità di ricevere la pena di morte è più alta per gli accusati di razza nera $\pi_n = 0.126$ che per gli accusati di razza bianca $\pi_b = 0.049$. La differenza $\pi_b - \pi_n = -0.077$ è negativa.
- Quando la vittima è nera, la probabilità di ricevere la pena di morte è più alta per gli accusati di razza bianca $\pi_b = 0.034$ che per gli accusati di razza nera $\pi_b = 0.008$. La differenza $\pi_b - \pi_n = 0.026$ è positiva.

Illustrazione

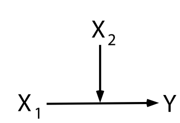
- Dato che l'associazione tra X_1 "razza dell'accusato" e Y "pena di morte" assume valori opposti a seconda del valore della variabile X_2 "razza della vittima", diciamo che la variabile X_1 interagisce con la variabile X_2 nel determinare il valore della variabile dipendente Y.
- Se ci limitiamo ad interpretare i dati forniti nella tabella in cui la variabile "razza della vittima" non viene controllata, giungiamo a delle conclusioni fuorvianti (agli accusati di razza bianca viene inflitta più spesso la pena di morte).

Illustrazione

- Tale conclusione è errata dato che una relazione di segno opposto viene rivelata quando la vittima è di razza bianca.
- In presenza di interazione tra variabili è necessario descrivere le associazioni presenti in ciascuna tabella parziale dato che, per diversi valori di una terza variabile, esse possono assumere segni diversi, come nel caso presente.

Paradosso di Simpson

- Si dice *paradosso di Simpson* il caso in cui ciascuna relazione parziale assume il segno opposto rispetto alla relazione bivariata in cui non viene controllata una terza variabile.
- La figura seguente fornisce una rappresentazione grafica dell'interazione tra le variabili X_1 e X_2.



Conclusioni

Conclusioni

- L'analisi statistica fa quasi sempre uso più di due variabili. Una relazione causale richiede
  - l'associazione statistica tra le variabili,
  - una appropriata sequenza temporale,
  - che non vi siano spiegazioni alternative.
- Spiegazioni alternative possono essere formulate nei termini di variabili di controllo che vengono mantenute costanti.

Conclusioni

- Se la relazione tra Y e X_1 viene esaminata mentre la variabile X_2 viene mantenuta costante, può essere possibile stabilire che
  - la relazione tra Y e X_1 è spuria: X_2 influenza sia X_1 che Y;
  - X_1 influenza Y indirettamente tramite X_2;
  - la relazione tra Y e X_1 risulta apparente solo dopo che la variabile soppressoria X_2 viene controllata;
  - il segno e la forza della relazione tra Y e X_1 variano al variare di X_2 (interazione).

Conclusioni

- Nel caso di relazioni spurie e di relazioni a catena, l'associazione tra X e Y scompare se la variabile Z viene controllata.
- Nel caso in cui Y sia influenzata da cause molteplici, l'associazione tra Y e X_1 può cambiare se la variabile X_2 viene controllata.
- Se vi è una variabile soppressoria Z, l'associazione tra Y e X risulta apparente solo se Z viene controllata.
- Nel caso di un'interazione tra X_1 e X_2, l'associazione tra X_1 e Y può assumere intensità e direzioni diverse a seconda dei valori della variabile X_2.

Psicometria 1 (023-PS)

Michele Grassi
mgrassi@units.it

Università di Trieste

Lezione 25 26

Sommario

Sommario

Il materiale coperto è contenuto nel capitolo 11 del testo di Agresti e Finlay (2012).

- Modello di regressione multipla
- Due variabili esplicative
- Valori osservati, adattati e residui
- Devianza totale, spiegata, residua
- Indice di determinazione e correlazione multipla
- Condizioni di validità del modello di regressione multipla
- Correlazione parziale
- Coefficienti standardizzati

Modello di regressione multipla

Modello di regressione multipla

- Limiti della regressione semplice: la variabile dipendente, in generale, dipende da molteplici variabili indipendenti.
- La regressione multipla è un'estensione della regressione semplice in cui due o più variabili esplicative sono combinate per prevedere i valori di una variabile dipendente.

Modello di regressione multipla

- La previsione dei valori Y diventa più accurata: le dimensioni dei residui diminuiscono.
- E' possibile controllare statisticamente l'effetto delle variabili confondenti.
  - Se la variabile X è associata ad altre variabili che influenzano Y, allora l'omissione di queste variabili ci porterebbe ad attribuire erroneamente il loro effetto su Y alla variabile X.

Una nota di cautela

Modello di regressione multipla

- La relazione tra Y e le variabili esplicative non è però necessariamente di tipo lineare.
- È quindi importante ispezionare visivamente i dati per verificare se tale ipotesi trova riscontro nei dati.

Due variabili esplicative

Due variabili esplicative

Due variabili esplicative

Due variabili esplicative

Vogliamo rappresentare la relazione tra il livello medio della variabile Y e due variabili esplicative X_1 e X_2 come una funzione

$$\hat{Y} = f(X_1, X_2)$$

o, in maniera equivalente, come

$$Y = f(X_1, X_2) + \text{residuo}$$

Due variabili esplicative

Due variabili esplicative

Come nel caso di un'unica variabile esplicativa, la funzione più semplice è quella lineare:

$$Y = \underbrace{\alpha + \beta_1 X_1 + \beta_2 X_2}_{\textit{funzione di regressione}} + \text{residuo}$$

Due variabili esplicative

Due variabili esplicative

- La funzione lineare

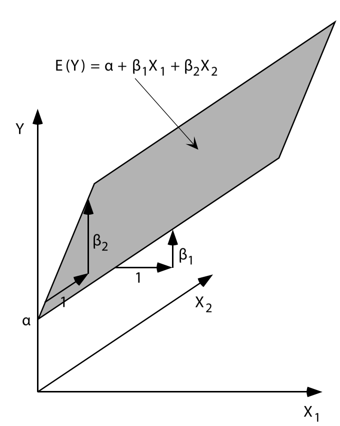
$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2$$

descrive un piano, come illustrato nella figura seguente.

- α è l'intercetta del piano con l'asse y.
- β_1 è la pendenza del piano nella direzione X_1 (tenendo X_2 costante).
- β_2 è la pendenza del piano nella direzione X_2 (tenendo X_1 costante).

Interpretazione geometrica

Due variabili esplicative



Interpretazione dei coefficienti

Due variabili esplicative

La pendenza o coefficiente angolare β_1 esprime la variazione del valore atteso della variabile dipendente per una variazione unitaria della variabile esplicativa X_1, se il valore di X_2 viene mantenuto costante.

Interpretazione dei coefficienti

Due variabili esplicative

- Supponiamo di sommare un'unità a X_1 mentre teniamo costante il valore di X_2:

$$E(Y|X_1+1, X_2) = \alpha + \beta_1(X_1+1) + \beta_2 X_2$$
$$= \alpha + \beta_1 X\ 1 + \beta_1 + \beta_2 X_2$$

- Sottraendo da $E(Y|X_1+1, X_2)$ il valore iniziale $E(Y|X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2$ otteniamo

$$E(Y|X_1+1, X_2) - E(Y|X_1, X_2) = \beta_1$$

Interpretazione dei coefficienti

Due variabili esplicative

Allo stesso modo, la pendenza β_2 può essere interpretata come il cambiamento del valore atteso della variabile dipendente per una variazione unitaria della variabile esplicativa X_2, se il valore di X_1 viene mantenuto costante.

Interpretazione dei coefficienti

Due variabili esplicative

- Con il metodo dei minimi quadrati scegliamo come stime dei parametri i valori di β_1 e β_2 per i quali la somma dei quadrati degli scarti delle osservazioni dalla funzione di regressione è minima.
- I calcoli necessari sono più complessi che nel caso della regressione semplice, ma le stime dei minimi quadrati $\hat{\beta}_1$ di β_1 e $\hat{\beta}_2$ di β_2 possono essere facilmente ottenute utilizzando un software.

Accostamento del modello ai dati

Accostamento del modello ai dati
Indice di determinazione
Coefficiente di correlazione multipla
Illustrazione

Accostamento del modello ai dati

Accostamento del modello ai dati
Indice di determinazione
Coefficiente di correlazione multipla
Illustrazione

- Dopo aver stimato il modello di regressione è opportuno verificare l'adattamento ai dati.
- L'accostamento del modello ai dati è tanto migliore quanto più elevata è la percentuale di devianza totale costituita dalla devianza spiegata.
- Di conseguenza l'adattamento può essere misurato mediante il rapporto fra la devianza spiegata e la devianza totale. Si ottiene così l'indice di determinazione R^2:

$$R^2 = \frac{\text{devianza spiegata}}{\text{devianza totale}}$$

Indice di determinazione

L'indice di determinazione è

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = 1 - \frac{\sum(Y_i - \hat{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

dove $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$.

Coefficiente di correlazione multipla

Il coefficiente di correlazione multipla R è uguale al coefficiente di correlazione di Pearson tra i valori adattati $\hat{Y}_i$ e i valori osservati Y_i.

- E' possibile dimostrare che il coefficiente di correlazione multipla R è uguale alla radice quadrata dell'indice di determinazione R^2 ed è sempre positivo.
- L'indice di determinazione R^2 (ovvero, il quadrato del coefficiente di correlazione multipla) esprime qual è la percentuale di devianza della variabile dipendente spiegata dalla relazione lineare su X_1 e X_2.

Illustrazione

Illustrazione. Analizziamo i dati riportati nella tabella 11.1, una piccola parte del campione raccolto da Holzer (1977).

- L'indice composito "*mental impairment*" è la variabile dipendente Y e misura la presenza di sintomi quali ansietà e depressione.
- La prima variabile esplicativa, "*life events*", misura il numero e la gravità di esperienze potenzialmente traumatiche negli ultimi tre anni (morte di una persona cara, licenziamento, traslocco, ...). Valori alti della variabile X_1 indicano un numero maggiore e una maggiore severità di tali esperienze.

Illustrazione

- La seconda variabile esplicativa, "*SES*", misura lo status sociale utilizzando indici quali occupazione, reddito e istruzione. Valori alti della variabile X_2 indicano uno status sociale più alto.

Illustrazione

```
> library(smss); data(mental.impairment)
> dati<-mental.impairment
> summary(dati)
       Y               X1              X2
 Min.   :17.00   Min.   : 3.00   Min.   : 3.00
 1st Qu.:23.75   1st Qu.:32.75   1st Qu.:39.75
 Median :27.00   Median :43.00   Median :56.00
 Mean   :27.30   Mean   :44.42   Mean   :56.60
 3rd Qu.:31.00   3rd Qu.:55.50   3rd Qu.:75.75
 Max.   :41.00   Max.   :97.00   Max.   :97.00
>
> cor(dati)
            Y        X1         X2
Y   1.0000000 0.3722206 -0.3985676
X1  0.3722206 1.0000000  0.1233370
X2 -0.3985676 0.1233370  1.0000000
```

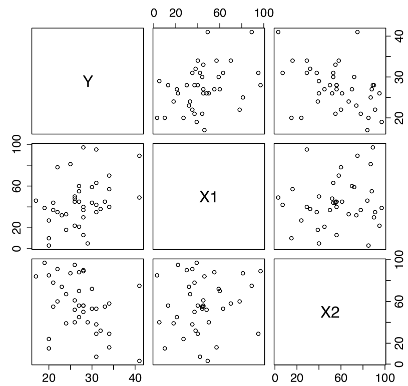

Illustrazione

- Possiamo utilizzare la funzione `pairs` di **R** per rappresentare graficamente i dati mediante una matrice di diagrammi di dispersione (*scatterplot matrix*).

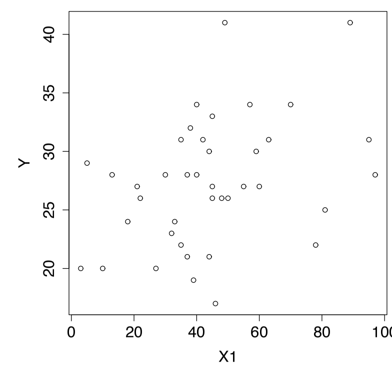
```
> pairs(dati)
```

- La prima riga della matrice riporta i grafici che pongono Y sull'ordinata e X_1 e X_2 sull'ascissa.
- L'ispezione visiva di questi diagrammi di dispersione suggerisce che un modello lineare rappresenta in maniera adeguata la relazione tra Y, X_1 e X_2.

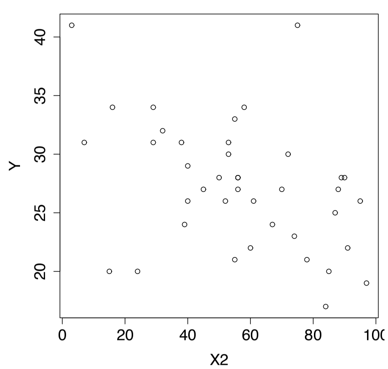
Illustrazione



Illustrazione



Illustrazione



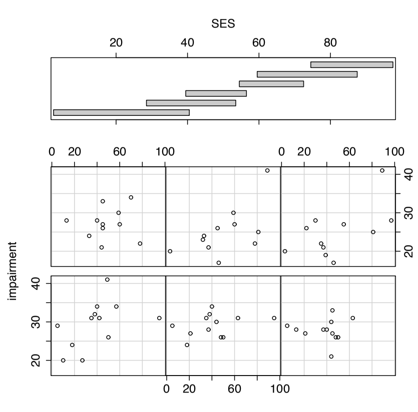
Illustrazione

- Il modello di regressione multipla $E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2$ ipotizza che l'effetto di X_1 su Y non varia a seconda del valore assunto dalla variabile X_2 – lo stesso si può dire dell'effetto di X_2 su Y.
- Per verificare questa ipotesi, ispezioniamo visivamente la relazione tra Y e X_1 in corrispondenza di valori omogenei della variabile X_2. Una tale rappresentazione grafica può essere generata con **R** mediante la funzione `coplot`:

```
> coplot(Y ~ X1 | X2,
    xlab=c("life events", "SES"), ylab=c("impairment"))
```

- Non ci sono evidenze di non linearità nella relazione tra "*mental impairment*" e "*life events*".

Illustrazione



Illustrazione

- Eseguiamo innanzitutto due regressioni bivariate che descrivono come il valore atteso di Y (*mental impairment*) dipende in media da X_1 (*life events*) o da X_2 (*SES*).
- Le due funzioni di regressione stimate sono:

$$\hat{Y} = 23.31 + 0.090X_1$$
$$\hat{Y} = 32.17 - 0.086X_2$$

Illustrazione

```
> fm1 <- lm(Y ~ X1)
> summary(fm1)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4415  -3.6899  -0.5973   3.6215  13.2890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.30949    1.80675  12.901 1.85e-15 ***
X1           0.08983    0.03633   2.472   0.0180 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.133 on 38 degrees of freedom
Multiple R-Squared: 0.1385,     Adjusted R-squared: 0.1159
F-statistic: 6.112 on 1 and 38 DF,  p-value: 0.01802
```


Illustrazione

```
> fm2 <- lm(Y ~ X2)
> summary(fm2)

Residuals:
     Min       1Q   Median       3Q      Max
-10.8808  -2.7472   0.2939   2.7382  15.2838

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 32.17201    1.98765  16.186   <2e-16 ***
X2          -0.08608    0.03213  -2.679   0.0109 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.072 on 38 degrees of freedom
Multiple R-Squared: 0.1589,     Adjusted R-squared: 0.1367
F-statistic: 7.177 on 1 and 38 DF,  p-value: 0.01085
```


Illustrazione

Nel primo modello di regressione, $\hat{Y} = 23.31 + 0.090X_1$, *mental impairment* è positivamente associato a *life events*.

- Ciò significa che, all'aumentare del numero e della gravità di eventi potenzialmente traumatici, aumenta il valore atteso della variabile che rappresenta i disturbi psicologici.
- L'indice di determinazione è $r^2_{yx_1} = 0.14$.

Illustrazione

Nel secondo modello, $\hat{Y} = 32.17 - 0.086X_2$, *mental impairment* è negativamente associato a *SES*.

- Ciò significa che, all'aumentare del livello socio-economico, diminuisce il valore atteso della variabile che rappresenta i disturbi psicologici.
- L'indice di determinazione è $r^2_{yx_2} = 0.16$.

Illustrazione

Esaminamo ora la devianza spiegata e la devianza residua dei due modelli di regressione. Per il modello $\hat{Y} = 23.31 + 0.090X_1$ abbiamo

```
> anova(fm1)
Analysis of Variance Table

Response: Y
          Df  Sum Sq Mean Sq F value  Pr(>F)
X1         1  161.05  161.05  6.1116 0.01802 *
Residuals 38 1001.35   26.35
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Illustrazione

Per il modello $\hat{Y} = 32.17 - 0.086X_2$ abbiamo

```
> anova(fm2)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X2         1 184.65  184.65  7.1766 0.01085 *
Residuals 38 977.75   25.73
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Illustrazione

Consideriamo ora il modello di regressione multipla stimato

$$\hat{Y} = 28.230 + 0.103X_1 - 0.097X_2$$

```
> fm3 <- lm(Y ~ X1 + X2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.22981    2.17422  12.984 2.38e-15 ***
X1           0.10326    0.03250   3.177  0.00300 **
X2          -0.09748    0.02908  -3.351  0.00186 **

Residual standard error: 4.556 on 37 degrees of freedom
Multiple R-Squared: 0.3392,     Adjusted R-squared: 0.3034
F-statistic: 9.495 on 2 and 37 DF,  p-value: 0.0004697
```


Illustrazione

- Il coefficiente di correlazione multipla R è il coefficiente di correlazione di Pearson tra i valori adattati $\hat{Y}$ e i valori osservati Y.

```
> cor(fm3$fitted.values, Y)
[1] 0.5823736
```

- Il quadrato del coefficiente di correlazione multipla R^2 (*indice di determinazione*) rappresenta la porzione di devianza della variabile dipendente spiegata dalle variabili esplicative X_1 e X_2.

```
>  0.5823736^2
[1] 0.339159
```


Illustrazione

- L'associazione tra *mental impairment* e *life events* è positiva dato che $\hat{\beta}_1$ è positivo.
  - Se la variabile *SES* viene controllata, l'aumento di un'unità della variabile *life events* produce un'aumento medio di 0.103 punti sulla scala *mental impairment*.
- L'associazione tra *mental impairment* e *SES* è negativa dato che $\hat{\beta}_2$ è negativo.
  - Se la variabile *life events* viene controllata, l'aumento di un'unità della variabile *SES* produce una diminuzione media di 0.097 punti sulla scala *mental impairment*.

Illustrazione

Per stabilire l'entità di questi effetti, esaminiamo la gamma di valori entro cui variano i punteggi Y, X_1 e X_2:

```
> summary(dati)
       Y               X1               X2
 Min.   :17.00   Min.   : 3.00   Min.   : 3.00
 1st Qu.:23.75   1st Qu.:32.75   1st Qu.:39.75
 Median :27.00   Median :43.00   Median :56.00
 Mean   :27.30   Mean   :44.42   Mean   :56.60
 3rd Qu.:31.00   3rd Qu.:55.50   3rd Qu.:75.75
 Max.   :41.00   Max.   :97.00   Max.   :97.00
```


Illustrazione

- $\hat{\beta}_1 = 0.103$ e *life events* varia nella gamma (0 – 100, in questo campione 3 – 97). La variabile *life events* ha dunque un effetto sui punteggi di *mental impairment* pari a $100 \times 0.103 = 10.3$.
- Dato che *mental impairment* varia nell'intervallo 17 – 41, questo effetto è considerevole.
- Le stesse considerazioni possono essere fatte a proposito di X_2.

Valori osservati e residui

Valori osservati e residui

- Il valore osservato del primo soggetto è $Y = 17$ per $X_1 = 46$ e $X_2 = 84$.
- Il valore adattato del primo soggetto è

$$\begin{aligned}\hat{Y}_1 &= a + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2\\ &= 28.22981 + 0.10326 \times 46 - 0.09748 \times 84\\ &= 24.79180\end{aligned}$$

- Il residuo del primo soggetto è

$$\begin{aligned}E_1 &= Y_1 - \hat{Y}_1\\ &= 17 - 24.79180\\ &= -7.7918042\end{aligned}$$

Devianza residua

La devianza residua è $\sum(Y_i - \hat{Y}_i)^2 = \sum(E_i)^2$.

```
> sum( residuals(fm3)^2 )
[1] 768.1616
> anova(fm3)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X1         1 161.05  161.05  7.7572 0.008385 **
X2         1 233.19  233.19 11.2320 0.001863 **
Residuals 37 768.16   20.76
```


Devianza totale

La devianza totale è $\sum(Y_i - \bar{Y})^2$.

```
> sum( (Y - mean(Y))^2 )
[1] 1162.4
> anova(fm3)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X1         1 161.05  161.05  7.7572 0.008385 **
X2         1 233.19  233.19 11.2320 0.001863 **
Residuals 37 768.16   20.76


> 161.05 + 233.19 + 768.16
[1] 1162.4
```


Devianza spiegata

La devianza spiegata è $\sum(\hat{Y}_i - \bar{Y})^2$.

```
> sum( (fitted.values(fm3) - mean(Y))^2 )
[1] 394.2384
> anova(fm3)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X1         1 161.05  161.05  7.7572 0.008385 **
X2         1 233.19  233.19 11.2320 0.001863 **
Residuals 37 768.16   20.76


> 161.05 + 233.19
[1] 394.24
```


Indice di determinazione

Il quadrato del coefficiente di correlazione multipla (*indice di determinazione*) è uguale al rapporto tra la devianza spiegata e la devianza totale:

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

Per i dati dell'esempio avremo:

$$R^2 = \frac{394.24}{1162.4} = 0.3392$$

```
Residual standard error: 4.556 on 37 degrees of freedom
Multiple R-Squared: 0.3392,     Adjusted R-squared: 0.3034
F-statistic: 9.495 on 2 and 37 DF,  p-value: 0.0004697
```


Indice di determinazione

- Dato che rappresenta la proporzione della devianza spiegata dal modello di regressione, l'indice di determinazione varia fra 0 e 1,

$$0 \le R^2 \le 1$$

- E' pari ad 1 quando vi è un perfetto accostamento del modello ai dati. In tal caso tutte le osservazioni si trovano sul piano di regressione stimato e la devianza residua è nulla.

Indice di determinazione

- L'indice di determinazione invece è nullo quando il piano stimato è parallelo al piano definito dal X_1 e X_2, cioè $\hat{\beta}_1 = \hat{\beta}_2 = 0$, e la devianza spiegata è nulla. In tal caso le variabili X_1 e X_2 non sono in grado di spiegare le variazioni della Y.
- R^2 non può diminuire se aggiungiamo una o più variabili esplicative al modello di regressione multipla. Di conseguenza, R^2 è $\geq r^2_{yx1}$ e r^2_{yx2}.

Coefficiente di correlazione multipla

- Il coefficiente di correlazione multipla R è uguale alla radice quadrata positiva del coefficiente di determinazione multipla:

$$R = +\sqrt{R^2}$$

```
> sqrt( 0.3392 )
[1] 0.5824088
```

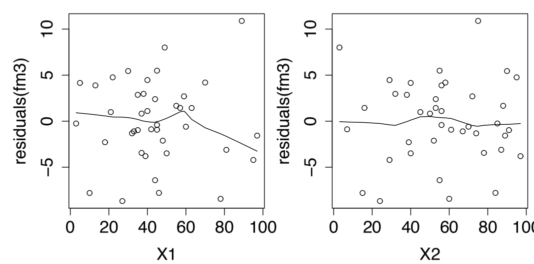
lo stesso risultato si trova calcolando la correlazione tra Y e $\hat{Y}$:

```
> cor(fm3$fitted.values, Y)
[1] 0.5823736
```


Condizioni di validità

- L'analisi grafica dei residui permette di stabilire se il modello di regressione è adeguato ai dati sperimentali.
- Se non esistono violazioni delle ipotesi di validità del modello ci aspettiamo che
  - la media dei residui sia uguale a zero per tutti gli intervalli di valori delle variabili esplicative,
  - la variabilità dei residui attorno allo zero sia costante per tutti i valori delle variabili esplicative.

Illustrazione



Più di due variabili esplicative

Più di due variabili esplicative

- Il modello di regressione multipla con k variabili esplicative $X_1, X_2, \ldots, X_k$ è

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \text{residui}$$

- Il metodo dei minimi quadrati consente di trovare i coefficienti $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ che minimizzano $\sum \text{residui}_i^2$.

Più di due variabili esplicative

I coefficienti del modello di regressione multipla hanno la seguente interpretazione:

- α è il valore atteso della variabile Y quando $X_1 = X_2 = \cdots = X_k = 0$;
- β_1 è il cambiamento medio della variabile Y che accompagna un cambiamento unitario in X_1, tenendo costante il valore delle altre variabili esplicative $(X_2, \ldots, X_k)$.
- $\beta_2, \ldots, \beta_k$ hanno un'interpretazione simile.

Coefficienti standardizzati

Spesso lo sperimentatore vuole confrontare i coefficienti del modello di regressione multipla.

- Quando le variabili esplicative sono commensurabili, il confronto è immediato.
- I coefficienti di regressione standardizzati consentono di confrontare gli effetti relativi di variabili esplicative incommensurabili.
- Per trovare i coefficienti standardizzati è necessario stimare il modello di regressione dopo avere standardizzato la variabile dipendente e le variabili esplicative.

Più di due variabili esplicative

$$Y_i = \hat{\alpha} + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_k X_{ik} + E_i$$

$$\bar{Y} = \alpha + \hat{\beta}_1 \bar{X}_1 + \cdots + \hat{\beta}_k \bar{X}_k + E_i$$

$$Y_i - \bar{Y} = \hat{\beta}_1 (X_{i1} - \bar{X}_1) + \cdots + \hat{\beta}_k (X_{ik} - \bar{X}_k) + E_i$$

$$\frac{Y_i - \bar{Y}}{s_y} = \left(\hat{\beta}_1 \frac{s_1}{s_y}\right) \frac{X_{i1} - \bar{X}_1}{s_1} + \cdots + \left(\hat{\beta}_k \frac{s_k}{s_y}\right) \frac{X_{ik} - \bar{X}_k}{s_k} + \frac{E_i}{s_y}$$

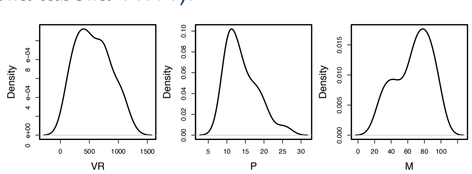
$$z_{iy} = \hat{\beta}_1^* z_{i1} + \cdots + \hat{\beta}_k^* z_{ik} + E_i^*$$

Più di due variabili esplicative

- $z_y = (Y - \bar{Y})/s_y$ è la variabile dipendente standardizzata (trasformata linearmente in modo tale da avere media zero e varianza unitaria).
- $z_1, \ldots, z_k$ sono le variabili esplicative standardizzate.
- $E^* = E/s_y$ è il residuo trasformato – non ha varianza unitaria.
- $\hat{\beta}_j^* = \hat{\beta}_j (s_j/s_y)$ è il coefficiente parziale di regressione standardizzato per la j-esima variabile esplicativa.
- $\hat{\beta}_j^*$ può essere interpretato come il cambiamento medio della variabile Y, espresso nei termini della deviazione standard s_y, che risulta associato al cambiamento di una deviazione standard della variabile X_j, mantenendo costanti le altre variabili esplicative.

Illustrazione

- Analizziamo i dati della tabella 9.1 con Y = *violent crime rate*, X_1 = *poverty rate*, X_2 = *percent living in metropolitan areas* escludendo il dato relativo a Washington D.C.
- Nella prima analisi, il modello viene stimato utilizzando i dati grezzi. Nella seconda analisi, il modello viene stimato dopo avere standardizzato le variabili (si confrontino i risultati con quelli della tabella 11.11).



Illustrazione

```
> fm3 <- lm(VR ~ P + M, data=murder2)
> summary(fm3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -498.683    140.988  -3.537 0.000922 ***
P             32.622      6.677   4.885 1.24e-05 ***
M              9.112      1.321   6.900 1.16e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 197.9 on 47 degrees of freedom
Multiple R-Squared: 0.5708,     Adjusted R-squared: 0.5525
F-statistic: 31.25 on 2 and 47 DF,  p-value: 2.335e-09
```

Illustrazione

```
> fm4 <- lm( z.VR ~ z.P + z.M)
> summary(fm4)

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) -9.396e-17  9.460e-02 -9.93e-16        1
z.P          4.726e-01  9.674e-02     4.885 1.24e-05 ***
z.M          6.675e-01  9.674e-02     6.900 1.16e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.669 on 47 degrees of freedom
Multiple R-Squared: 0.5708,     Adjusted R-squared: 0.5525
F-statistic: 31.25 on 2 and 47 DF,  p-value: 2.335e-09
```


Una nota di cautela

- E' inappropriato usare i coefficienti standardizzati per confrontare gli effetti della stessa variabile esplicativa in campioni estratti da popolazioni con varianze diverse.
- L'uso delle deviazioni standard per standardizzare le variabili esplicative, inoltre, è inappropriato se le variabili esplicative non sono distribuite in modo normale.

Interazione tra variabili

Interazione statistica

- Nel modello di regressione

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + E_i$$

la relazione parziale tra Y e ciascun X_j è lineare e la pendenza β_j che esprime tale relazione è costante per qualsiasi combinazione delle altre variabili esplicative.
- Tale modello, però, è inadeguato in presenza di interazione statistica.

Nel caso di variabili esplicative quantitative, si dice che esiste interazione statistica quando l'associazione tra la variabile dipendente e una variabile esplicativa cambia al mutare dei valori di altre variabili esplicative.

Termini di interazione

- Il metodo usato per rappresentare le interazioni tra variabili è quello di introdurre degli ulteriori termini di interazione nel modello di regressione multipla.
- Nel caso di due variabili indipendenti, si avrà

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + E_i$$

- In questo modello, il termine $X_1 X_2$ è una variabile artificiale creata facendo il prodotto delle variabili X_1 e X_2.

Termini di interazione

- Se una delle due variabili esplicative, diciamo X_2, fosse qualitativa, con due livelli, allora il coefficiente β_3 rappresenterebbe la differenza tra la pendenza della retta di regressione tra Y e X_1 nei due gruppi.
- Se X_2 è quantitativa, tale interpretazione può essere estesa, dicendo che β_3 rappresenta la differenza tra la pendenza della retta di regressione tra Y e X_1 tra ciascuno dei diversi valori che X_2 può assumere.

Termini di interazione

Scriviamo il modello di regressione nel modo seguente:

$$E(Y) = (\alpha + \beta_2 X_2) + (\beta_1 + \beta_3 X_2) X_1 = \alpha' + \beta_1' X_1$$

dove $\alpha' = (\alpha + \beta_2 X_2)$ e $\beta_1' = (\beta_1 + \beta_3 X_2)$.

- Quando la X_2 viene mantenuta costante, la media di Y cambia come funziona lineare di X_1.
- Sia l'intercetta α' che la pendenza β' dipendono dal valore assunto da X_2.
- Dunque, al mutare del valore di X_2 l'effetto di X_1 su Y cambia.

Termini di interazione

- Anche se nella popolazione non vi è interazione tra le variabili, per gli effetti degli errori di campionamento, nei dati del campione non succederà mai che l'effetto di X_1 su Y è esattamente uguale per tutti i valori di X_2 (e viceversa).
- Dobbiamo dunque sottoporre a verifica l'ipotesi nulla di assenza di interazione tra X_1 e X_2 nella popolazione.
- Tale ipotesi nulla viene verificata mediante un test sul coefficiente parziale di regressione β_3 che è associato al termine di interazione $X_1 X_2$.

Illustrazione

Illustrazione. Consideriamo nuovamente l'esempio relativo alle variabili mental impairment, life events e SES e analizziamo tali dati mediante un modello che contiene il termine di interazione $X_1 X_2$. (tabella 11.7)

```
> fm <- lm(Y  ~ X1*X2)
> summary(fm)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.036648   3.948826   6.594 1.13e-07 ***
X1           0.155865   0.085338   1.826   0.0761 .
X2          -0.060493   0.062674  -0.965   0.3409
X1:X2       -0.000866   0.001297  -0.668   0.5087
```


Illustrazione

Se $SES = 0$, la relazione stimata tra Y e X_1 è

$$E(Y) = (\alpha + \beta_2 X_2) + (\beta_1 + \beta_3 X_2) X_1$$

ovvero

$$E(Y) = (26.037 + (-0.060)0) + (0.156 + (-0.00087)0) X_1$$

$$E(Y) = 26.037 + 0.156 X_1$$

Illustrazione

In maniera corrispondente, se $SES = 50$, la relazione stimata tra Y e X_1 è

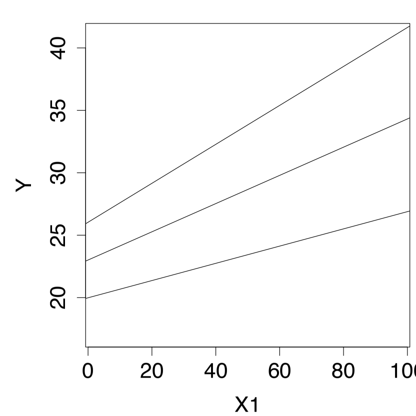
$$E(Y) = 23.012 + 0.113 X_1$$

e quando $SES = 100$

$$E(Y) = 19.987 + 0.069 X_1$$

Più alto è il valore di SES, minore è la pendenza della retta di regressione tra life events e mental impairment. Questo indica che gli individui che possiedono maggiori risorse (valori alti di SES) sono meglio in grado di assorbire gli eventi potenzialmente traumatici.

Illustrazione



Illustrazione

- La discussione precedente è stata presentata solo per scopi didattici, dato che l'analisi dei dati non consente di rifiutare l'ipotesi nulla di assenza di interazione [$t(36) = -0.668$, n.s.].
- Se l'ipotesi di assenza di interazione non può essere rifiutata, possiamo interpretare gli effetti principali delle variabili X_1 e X_2.
- In presenza di un interazione statisticamente significativa, invece, non ha senso interpretare gli effetti delle altre variabili. Dobbiamo invece interpretare l'effetto dell'interazione (come abbiamo fatto nell'illustrazione precedente, per esempio).

Correlazione parziale

Definizione

- Il modello di regressione multipla descrive la relazione tra due variabili, controllando l'effetto di altre variabili di interesse.
- Considereremo ora un indice che misura l'intensità dell'associazione tra due variabili, controllando l'effetto di altre variabili.
  - **Illustrazione**. Per l'esempio precedente, potremmo chiederci: se la variabile *SES* viene controllata, che proporzione della varianza della variabile *mental imparement* viene spiegata dalla variabile *life events*?
- L'indice usato per rispondere a questa domanda viene chiamato correlazione parziale.

Definizione

Per calcolare la correlazione parziale tra Y e X_1 al netto dell'effetto lineare di X 2 è necessario trovare le componenti di Y e di X_1 linearmente indipendenti da X_2.

La componente di Y linearmente indipendente da X_2 è data dai residui del modello

$$Y = A_1 + B_1 X_2 + E_1$$

e lo stesso può dirsi di X_1:

$$X_1 = A_2 + B_2 X_2 + E_2$$

Definizione

La correlazione parziale $r_{Y,X_1|X_2}$ non è altro che la correlazione di Pearson tra E_1 e E_2, ovvero la correlazione tra le componenti di Y e X_1 linearmente indipendenti da X_2.

Definizione

Si ricordi infatti che, nel modello di regressione

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

i coefficienti dei minimi quadrati $\beta_0, \ldots, \beta_p$ sono stati calcolati in modo tale da rendere linearmente indipendenti le due componenti in cui viene scomposta la Y_i:

- $\hat{Y}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$
- ε_i

Illustrazione

Consideriamo nuovamente i dati della tabella 11.5 e poniamoci il problema di calcolare la correlazione parziale tra Y e X_1 al netto dell'effetto di X_2.

Eseguiamo la regressione semplice di Y e di X_1 su X_2:

```
> fm1 <- lm(Y  ~ X2)
> fm2 <- lm(X1 ~ X2)
```


Illustrazione

In questo modo ciascun valore Y_i viene scomposto in due componenti indipendenti, il valore adattato $\hat{Y}_i$ e il residuo, E_i:

```
> d <- cbind(Y, fm1$fit, fm1$res)
> d[1:4, ]
   Y
1 17 24.94146  -7.941464
2 19 23.82245  -4.822451
3 20 30.10614 -10.106141
4 20 24.85539  -4.855386
```

Lo stesso può dirsi di X_1.

Illustrazione

I residui E_{1i} sono incorrelati con la componente di Y predicibile da X_2:

```
> cor(fm1$fit, fm1$res)
[1] -1.982922e-17
```

I residui E_{2i}, inoltre, sono incorrelati con la componente di X_1 predicibile da X_2:

```
> cor(fm2$fit, fm2$res)
[1] -1.746154e-16
```


Illustrazione

La correlazione parziale tra Y e X_1 al netto dell'effetto lineare di X_2 su ciascuna di esse è uguale alla correlazione di Pearson tra i residui E_1 e E_2:

```
> cor(fm1$res, fm2$res)
[1] 0.4629842
```


Illustrazione

Si noti come la correlazione parziale $r_{Y,X_1|X_2}$ sia diversa della correlazione tra Y e X_1:

```
> cor(Y, X1)
[1] 0.3722206
```

La correlazione parziale può avere intensità diversa, e anche cambiare segno, rispetto alla correlazione semplice.

Quadrato della correlazione parziale

Il quadrato del coefficiente di correlazione parziale può essere interpretato nei termini della riduzione proporzionale dell'errore di previsione di Y.

$r^2_{Y,X_1|X_2}$ è la proporzione della varianza di Y che viene spiegata da X_1, dopo avere controllato l'effetto di X_2.

Dopo avere rimosso dalla devianza totale l'effetto di X_2, $r^2_{Y,X_1|X_2}$ indica la porzione della devianza rimanente che può essere attribuita alla componente di X_1 linearmente indipendente da X_2.

Calcolo di $r_{X_j X_k | F}$

La correlazione parziale tra Y e X_1 al netto dell'effetto di X_2, $r_{Y X_1 | X_2}$, può anche essere calcolata direttamente dalle correlazioni semplici tra le tre variabili X_1, X_2 e Y:

$$r_{Y X_1 | X_2} = \frac{r_{Y X_1} - r_{Y X_2} r_{X_1 X_2}}{\sqrt{(1 - r^2_{Y X_2})(1 - r^2_{X_1 X_2})}}$$

$$= \frac{0.372 - (-0.399) \times 0.123}{\sqrt{(1 - 0.399^2)(1 - 0.123^2)}} =$$

$$= 0.463$$

Interpretazione di $r_{X_j X_k | F}$

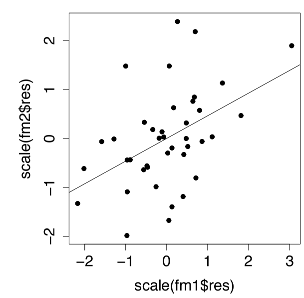
- Se potessimo isolare nella popolazione tutte le osservazioni che hanno un valore costante di X_2, allora la correlazione parziale parziale $r_{Y X_1 | X_2}$ corrisponderebbe alla correlazione di Pearson all'interno di ciascuno di questi sottogruppi.
- Come la correlazione di Pearson, $r_{Y X_1 | X_2}$ assume valori compresi tra -1 e +1.
- La correlazione parziale $r_{Y X_1 | X_2}$ ha lo stesso segno del coefficiente parziale di regressione B_1 nell'equazione di regressione $Y = B_0 + B_1 X_1 + B_2 X_2 + E$.
- Il valore del coefficiente di correlazione parziale non dipende dall'unità di misura delle variabili.

Interpretazione di $r_{X_j X_k | F}$

- Il quadrato di $r_{Y X_1 | X_2}$ ha un'interpretazione nei termini della riduzione proporzionale dell'errore di previsione: è la proporzione della varianza di Y che viene spiegata da X_1, dopo avere controllato l'effetto di X_2 su entrambe le variabili.

Partial plots

Il coefficiente di correlazione parziale $r_{Y X_1 | X_2}$ è uguale alla pendenza della retta di regressione tra i residui standardizzati E_1 e E_2.



Partial plots

```
> plot(scale(fm1$res), scale(fm2$res))
> fit <- lm(scale(fm2$res) ~ scale(fm1$res))
> abline(fit)
> summary(fit)

Call: lm(formula = scale(fm2$res) ~ scale(fm1$res))

Residuals:
      Min        1Q    Median        3Q       Max
-1.699575 -0.418341 -0.009051  0.486768  2.266031

Coefficients:
                 Estimate Std. Error   t value Pr(>|t|)
(Intercept)    -2.736e-17  1.420e-01 -1.93e-16  1.00000
scale(fm1$res)  4.630e-01  1.438e-01      3.22  0.00263 **
```


Conclusioni

Conclusioni

- Il modello di regressione multipla che mette in relazione una variabile dipendente con un insieme di k variabili esplicative è

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

- Il coefficiente parziale di regressione β_j rappresenta il cambiamento medio della variabile Y per un cambiamento unitario della variabile X_j, se le altre variabili esplicative sono mantenute costanti.

Conclusioni

- Il coefficiente di correlazione multipla R è uguale al coefficiente di correlazione di Pearson tra i valori osservati Y e i valori adattati $\hat{Y}$. R rappresenta dunque l'intensità della relazione lineare tra Y e le variabili esplicative.
- Il quadrato del coefficiente di correlazione multipla $0 \leq R^2 \leq 1$ rappresenta la riduzione proporzionale dell'errore di previsione che si commette utilizzando $\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k$ anzichè anziché $\bar{Y}$.

Conclusioni

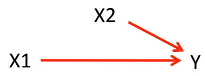
- I coefficienti standardizzati $\hat{\beta}_j^*$ non dipendono dall'unità di misura delle variabili esplicative e descrivono il cambiamento atteso della variabile Y, espresso in termini di deviazioni standard di Y, per il cambiamento di una deviazione standard della variabile X_j, se le altre variabili esplicative sono mantenute costanti.

Errore di specificazione, metodo dei residui e correlazione parziale:

approfondimento

Errore di specificazione

Popolazione



$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon;$

- Supponiamo che *nella popolazione* la relazione tra y e le due variabili esplicative x_1 e x_2 sia lineare;
- dove ε è incorrelato con x_1 e x_2 .
- Se i dati di un campione estratto dalla popolazione vengono analizzati utilizzando entrambe le variabili x_1 e x_2 allora i coefficienti dei minimi quadrati forniranno delle stime non distorte (prive di errore sistematico) dei parametri β_1 e β_2.

Campione

$y = a + bx_1 + e'$

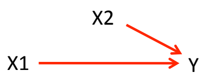
- Supponiamo però di analizzare i dati del campione con il modello di regressione bivariata, utilizzando solo x_1;
- implicitamente, l'effetto di x_2 su y viene assorbito nell'errore

$$e' = e + \beta_2 x_2$$

e se x_1 e x_2 sono correlate, allora parte dell'effetto di x_2 su y sarà erroneamente attribuito a x_1.

Errore di specificazione

Popolazione



$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon;$

- Supponiamo che *nella popolazione* la relazione tra y e le due variabili esplicative x_1 e x_2 sia lineare;

- Vediamo di quantificare il *bias*…

Campione

$y = a + bx_1 + e'$

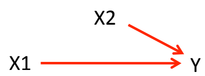
- Supponiamo però di analizzare i dati del campione con il modello di regressione bivariata, utilizzando solo x_1;

- La conseguenze è che la stima dei minimi quadrati *b* sarà distorta, ossia avremo che:

$$b = \beta_1 + bias$$

Errore di specificazione

Popolazione



$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e;$

consideriamo il valore atteso di y :

$E(y) = \alpha + \beta_1 E(x_1) + \beta_2 E(x_2) + E(e)$

visto che $E(e) = 0$, otteniamo

$\mu_y = \alpha + \beta_1 \mu_1 + \beta_2 \mu_2$

Campione

$y = a + bx_1 + e'$

Errore di specificazione

Popolazione

X2 → Y

X1 → Y

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e;$

consideriamo ora la differenza $y - \mu_y$:

$y - \mu_y = \cancel{\alpha} + \beta_1 x_1 + \beta_2 x_2 + e - \cancel{\alpha} + \beta_1 \mu_1 + \beta_2 \mu_2$

dalla quale otteniamo:

$(y - \mu_y) = \beta_1 (x_1 - \mu_1) + \beta_2 (x_2 - \mu_2) + e$

Campione

X1 → Y

$y = a + bx_1 + e'$

Errore di specificazione

Popolazione

X2 → Y

X1 → Y

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e;$

moltiplichiamo entrambi i lati per $(x_1 - \mu_1)$:

$(x_1 - \mu_1)(y - \mu_y) = \beta_1 (x_1 - \mu_1)(x_1 - \mu_1) + \beta_2 (x_1 - \mu_1)(x_2 - \mu_2) + e(x_1 - \mu_1)$

e consideriamo il valore atteso $E\left[(x_1 - \mu_1)(y - \mu_y)\right]$,

$E\left[(x_1 - \mu_1)(y - \mu_y)\right] = \beta_1 E\left[(x_1 - \mu_1)(x_1 - \mu_1)\right] + \beta_2 E\left[(x_1 - \mu_1)(x_2 - \mu_2)\right] + \cancel{E\left[(x_1 - \mu_1)e\right]}$

che possiamo riscrivere come:

$Cov(x_1 y) = \beta_1 Var(x_1) + \beta_2 Cov(x_1 x_2).$

Campione

X1 → Y

$y = a + bx_1 + e'$

Errore di specificazione

Popolazione

X2 → Y

X1 → Y

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e;$

risolvendo $Cov(x_1 y) = \beta_1 Var(x_1) + \beta_2 Cov(x_1 x_2)$ per β_1 otteniamo:

$\beta_1 = \frac{Cov(x_1 y)}{Var(x_1)} - \beta_2 \frac{Cov(x_1 x_2)}{Var(x_1)};$

dato che (come risulta dai minimi quadrati) $\frac{Cov(x_1 y)}{Var(x_1)} = b,$

$\beta_1 = b - \beta_2 \frac{Cov(x_1 x_2)}{Var(x_1)}$ ⟶

Campione

X1 → Y

$y = a + bx_1 + e'$

$b = \beta_1 + \boxed{\beta_2 \frac{Cov(x_1 x_2)}{Var(x_1)}}$

$b = \beta_1 + bias$

Errore di specificazione

Popolazione

$Cov(x_1 x_2) \neq 0$

(b) X2 $\beta_2 \neq 0$ (a)

X1 → Y

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + e;$

- L'interpretazione del bias nella stima del coefficiente di regressione bivariato dipende dalla natura della relazione tra x_1 e x_2 :

Campione

X1 → Y

$y = a + bx_1 + e'$

$b = \beta_1 + bias$

$bias = \beta_2 \frac{Cov(x_1 x_2)}{Var(x_1)}$

$bias \neq 0$ se x_2

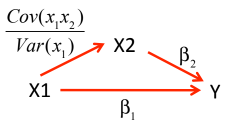
(a) è una variabile rilevante $(\beta_2 \neq 0)$

(b) è correlata con x_1 $(Cov(x_1 x_2) \neq 0)$

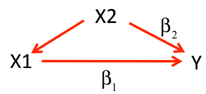
Errore di specificazione

Popolazione

① x_2 **variabile interveniente**

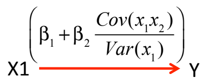


② **relazione spuria (non causale) tra x_1 e y**



- L'interpretazione del bias nella stima del coefficiente di regressione bivariato dipende dalla natura causale della relazione tra x_1 e x_2

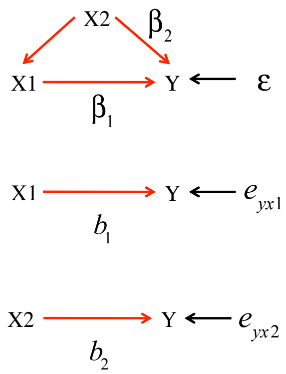
Campione



① il bias è l'effetto indiretto di x_1 su y trasmesso attraverso x_2, dal momento che $Cov(x_1x_2)/Var(x_1)$ è l'inclinazione della retta di regressione di x_2 su x_1 nella popolazione.

② x_2 è la causa comune di x_1 e y; il bias rappresenta la componente *spuria* dell'associazione empirica tra x_1 e y.

Correlazione parziale



Consideriamo nuovamente il modello di *regressione lineare multipla:*

- $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

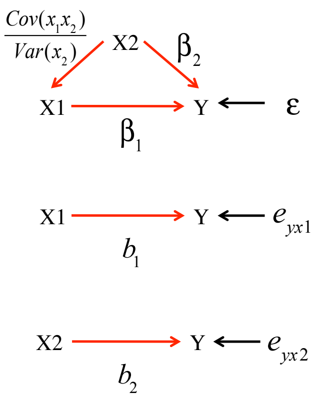
e i due modelli di *regressione lineare bivariata* per X1

- $y = a + b_1 x_1 + e_{yx_1}$

e per X2

- $y = a + b_2 x_2 + e_{yx_2}$

Correlazione parziale



La regressione multipla non corrisponde a tante regressioni bivariate.
Infatti, come abbiamo visto per l'errore di specificazione, le stime degli effetti della regressione multipla sono diverse da quelle della regressione bivariata:

- per b_1 avremo che (*dalle regole di scomposizione della varianza - covarianza*)*

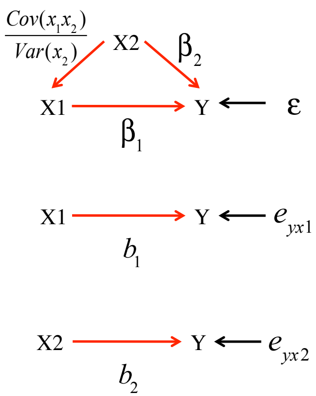
$$Cov(x_1y) = Var(x_1)\beta_1 + \frac{Cov(x_1x_2)}{\cancel{Var(x_2)}}\cancel{Var(x_2)}\beta_2$$

da cui

$$b_1 = \frac{Cov(x_1y)}{Var(x_1)} = \beta_1 + \frac{Cov(x_1x_2)}{Var(x_1)}\beta_2$$

* Sewall Wright (1921, 1934)

Correlazione parziale



La regressione multipla non corrisponde a tante regressioni bivariate.
Infatti, come abbiamo visto per l'errore di specificazione, le stime degli effetti della regressione multipla sono diverse da quelle della regressione bivariata:

- e per b_2

$$Cov(x_2y) = Var(x_2)\beta_2 + \cancel{Var(x_2)}\frac{Cov(x_1x_2)}{\cancel{Var(x_2)}}\beta_1$$

da cui

$$b_2 = \frac{Cov(x_2y)}{Var(x_2)} = \beta_2 + \frac{Cov(x_1x_2)}{Var(x_2)}\beta_1$$

* Sewall Wright (1921, 1934)

Correlazione parziale

• nella regressione bivariata viene stimato l'effetto MARGINALE di X1 (o X2) su Y;

• nella regressione multipla vengono misurati gli effetti PARZIALI di ciascuna variabile indipendente (ad es. X1) AL NETTO degli effetti della altre variabili esplicative (ad es. X2).

• I coefficienti di regressione multipla sono detti *coefficienti parziali di regressione* perché quantificano l'effetto di ciascuna variabile indipendente depurato dagli effetti attribuibili alle altre variabili inserite nel modello.

• Per chiarire il concetto di *parzializzazione* - "*depurazione dagli effetti lineari delle altre variabili*"- cerchiamo di eliminare il *bias* associato al coefficiente di regressione bivariato b_1 che si determina quando omettiamo dal modello la *variabile rilevante & correlata X2*.

$$b_1 = \beta_1 + \underbrace{\frac{Cov(x_1x_2)}{Var(x_1)}\beta_2}_{bias}$$

Correlazione parziale

• METODO dei RESIDUI: utilizzeremo gli errori associati a due regressioni bivariate:

• la covarianza tra X1 e X2, *Cov* (x_1x_2), si può eliminare se utilizziamo la parte del punteggio di X1 linearmente indipendente da X2, ossia l'errore associato al seguente modello di regressione semplice:

$$x_1 = a + \frac{Cov(x_1x_2)}{Var(x_2)}x_2 + e_{x1x2}$$

• allo stesso modo, l'effetto di X2 su Y, β_2, si può eliminare utilizzando la parte di punteggio di Y linearmente indipendente da X2, e quindi l'errore del modello di regressione:

$$y = a + \frac{Cov(x_2y)}{Var(x_2)}x_2 + e_{yx2}$$

Correlazione parziale

• se stimiamo l'effetto (marginale) di X1 su Y nella regressione bivariata, utilizzando però le quantità indipendenti da X2, ossia e_{x1x2} ed e_{yx2}:

$$e_{yx2} = a + b_1 e_{x1x2} + e$$

• **(1)** il coefficiente b_1 coinciderà con il coefficiente parziale β_1, dal momento che il bias

$$\frac{Cov(x_1x_2)}{Var(x_1)}\beta_2,$$

diventa nullo, essendo $Cov(e_{x1}x_2) = 0$, e $\beta_2 = 0$. (In virtù dell'utilizzo degli errori dei due modelli precedenti)

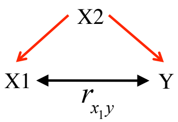
• In altre parole, ogni parametro di regressione multipla (ad es. β_1) misura l'effetto parziale di una variabile AL NETTO delle altre variabili inserite nel modello. Infatti, per ottenere questo parametro nella regressione bivariata ($b_1 = \beta_1$) è necessario utilizzare i punteggi delle variabili Y e X1 *depurati dall'effeto di X2*.

•Naturalmente, si tiene conto solo degli effetti delle variabili inserite nel modello, quindi gli effetti sono sempre AL LORDO di tutto ciò che non è stato considerato.

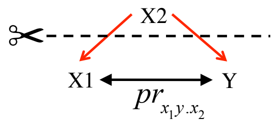
Correlazione parziale

• se stimiamo l'effetto (marginale) di X1 su Y nella regressione bivariata, utilizzando però le quantità indipendenti da X2, ossia e_{x1x2} ed e_{yx2}:

$$e_{yx2} = a + b_1 e_{x1x2} + e$$



• **(2)** la correlazione r di Pearson tra e_{yx2} ed e_{x1x2} è detta correlazione PARZIALE e rappresenta la forza (e direzione) dell'associazione lineare tra Y e X1 quando l'effetto di X2 su entrambe viene eliminato.



$$pr_{x_1y.x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{1-r^2_{yx_2}}\sqrt{1-r^2_{x_1x_2}}}$$

• il concetto di correlazione parziale sarà di fondamentale importanza *nell'analisi fattoriale*.