

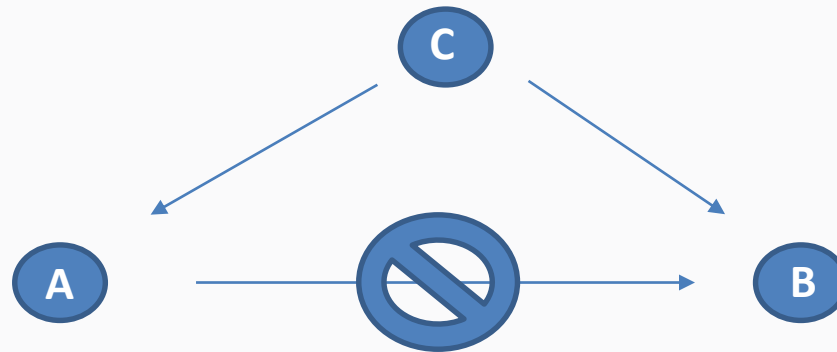
Correlazione e regressione

Il termine **associazione** è largamente usato nella letteratura scientifica ed esprime la relazione che esiste tra due variabili

Per studiare l'associazione tra due variabili bisogna pensare almeno a due livelli di analisi:

1. La relazione tra le due variabili in studio può essere spiegata da una terza variabile
(la terza variabile deve essere inclusa nello studio!!!)
2. Utilizzare i metodi statistici appropriati per studiare la relazione

1. La relazione tra le due variabili in studio può essere spiegata da una terza variabile
(la terza variabile deve essere inclusa nello studio!!!)



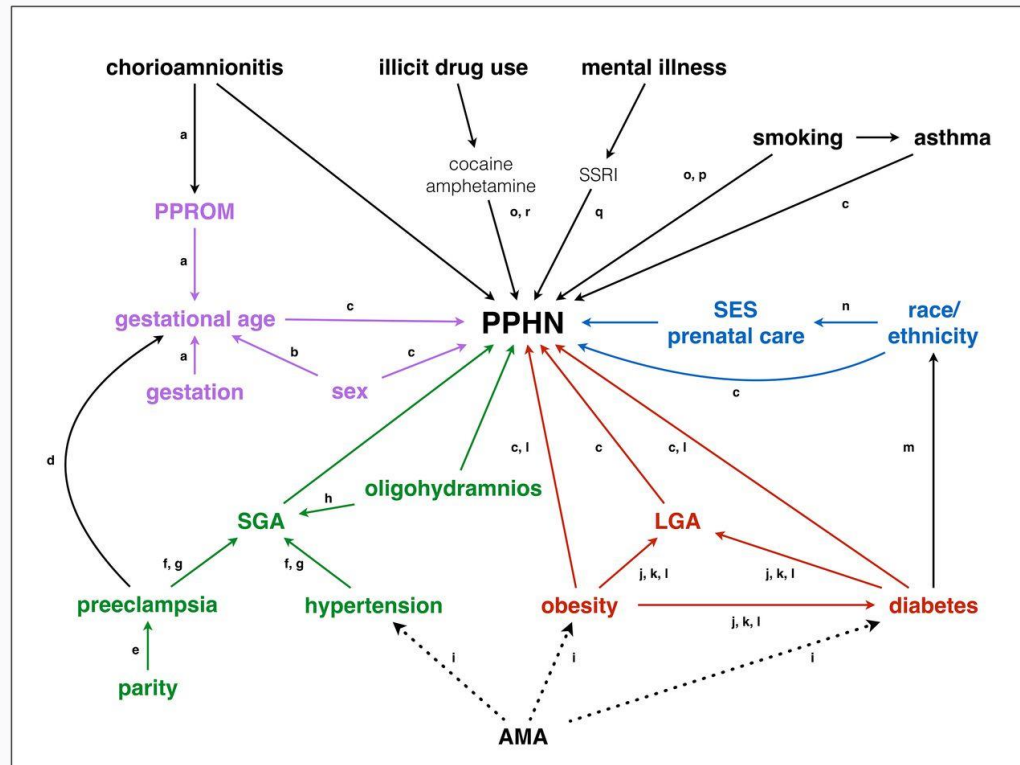
630

THE NEW ENGLAND JOURNAL OF MEDICINE

March 12, 1981

COFFEE AND CANCER OF THE PANCREAS

BRIAN MACMAHON, M.D., STELLA YEN, M.D., DIMITRIOS TRICHOPOULOS, M.D., KENNETH WARREN, M.D.,
AND GEORGE NARDI, M.D.



2. Utilizzare i metodi statistici appropriati per studiare la relazione

		Response variable	
		Categorical	Quantitative
Explanatory variable	Categorical	C-C	C-Q
	Quantitative	Q-C	Q-Q

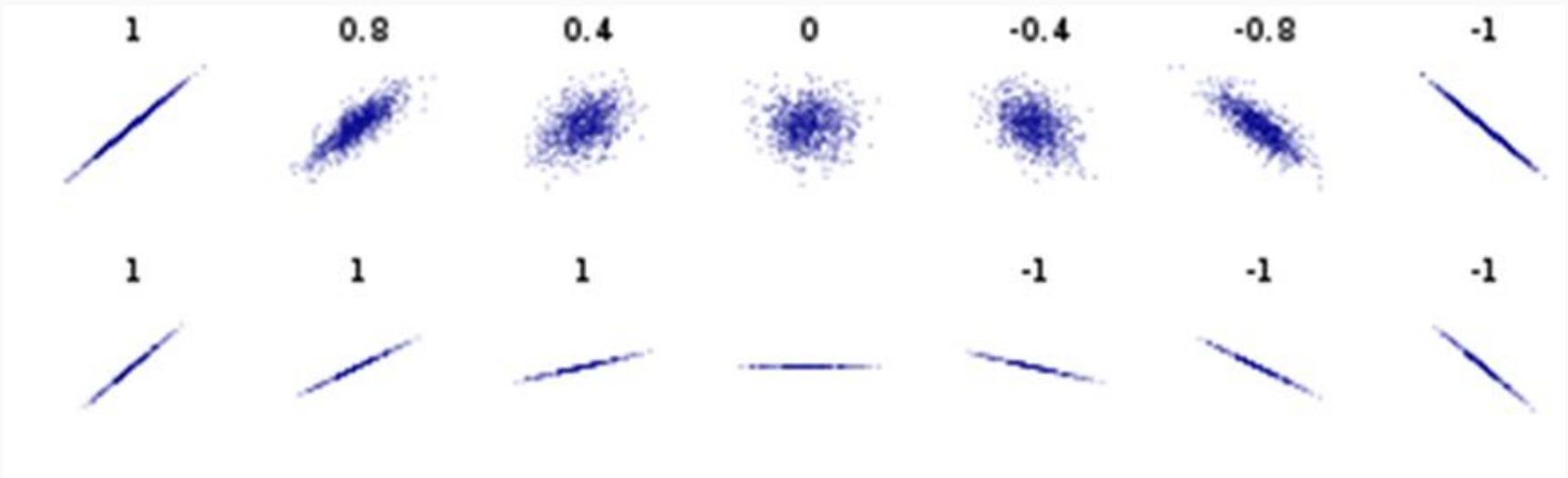
$Q \rightarrow C$
Logistic Regression

Correlazione

Il coefficiente di correlazione descrive quanto due variabili sono associate.
In altre parole la quantità di variabilità in una misura che è spiegata da un'altra misura.

Il range del coefficiente di correlazione va da -1 a +1
I valori estremi indicano una perfetta associazione lineare

Un coefficiente di correlazione positivo indica che le variabili crescono di valore insieme; negativo quando una variabile cresce l'altra decresce.
È pari a 0 quando non c'è un'associazione lineare tra due variabili



Correlazione vs regressione

La correlazione non discrimina tra tipi di variabili (y variabile dipendente, x variabile indipendente), ma misura l'associazione lineare tra le due variabili e misura quanto variano insieme x e y

La regressione lineare discrimina tra x e y e calcola la migliore retta che predice i valori di y per dati valori osservati di x

In molte analisi che indagano la relazione tra variabili continue calcoliamo entrambe le statistiche

Tipi di correlazione

Coefficiente di correlazione di Pearson

Utilizzato per misurare la relazione lineare tra due variabili continue entrambe normalmente distribuite

Se non assumiamo la normalità della distribuzione delle variabili, possiamo scegliere un coefficiente di correlazione non parametrico

Coefficiente di correlazione di Spearman

È utilizzato quando una variabile ha una distribuzione normale e l'altra è categorica o non normalmente distribuita

La variabile categorica o non normalmente distribuita è classificata in ranghi che vengono ordinati e correlati alla variabile continua

Tipi di correlazione

Il coefficiente di correlazione di popolazione ρ (rho) misura la forza dell'associazione tra variabili

Il coefficiente di correlazione campionario r è una stima di ρ ed è calcolato per misurare la forza dell'associazione lineare tra le variabili nel campione in studio

Il coefficiente di correlazione campionario r tra due variabili è ottenuto dal rapporto tra la loro covarianza divisa per il prodotto delle deviazioni standard

$$r = r_{xy} = \frac{\text{Cov}(x, y)}{S_x \times S_y}$$

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Suggerimenti per le analisi

1. Esplora le variabili

analisi descrittiva: scatter plot

le variabili hanno una distribuzione normale?

2. Calcola l'appropriato coefficiente di correlazione

la correlazione è usata per misurare la forza dell'associazione lineare tra due variabili

3. Guarda l'intervallo di confidenza

4. Guarda il pvalue

NB

La relazione lineare tra due variabili non implica una relazione causale tra due variabili

Se il pvalue è maggiore del livello di significatività scelto come soglia i dati non sono consistenti per concludere che vi sia una correlazione reale.

Ciò non significa che non ci sia una correlazione tout court!

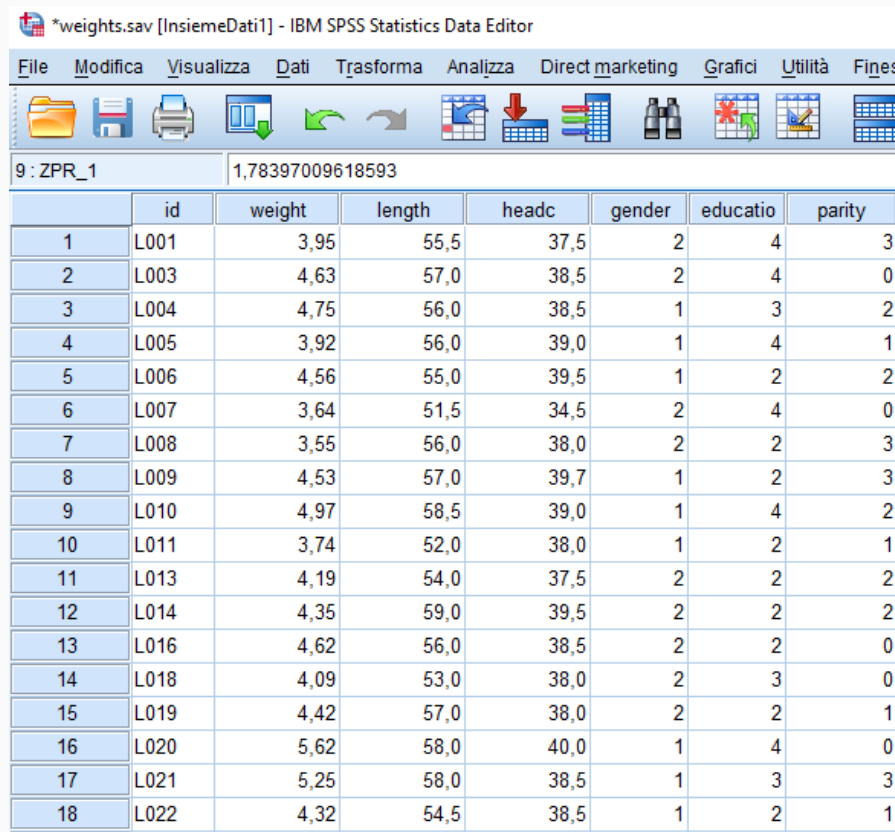
Esempio

Database weights.sav

contiene i dati di un campione di 550 bambini a cui è stato registrato il peso a un mese di vita

Vogliamo rispondere alla domanda

C'è un'associazione lineare tra peso, lunghezza e circonferenza cranica nei bambini di un mese?

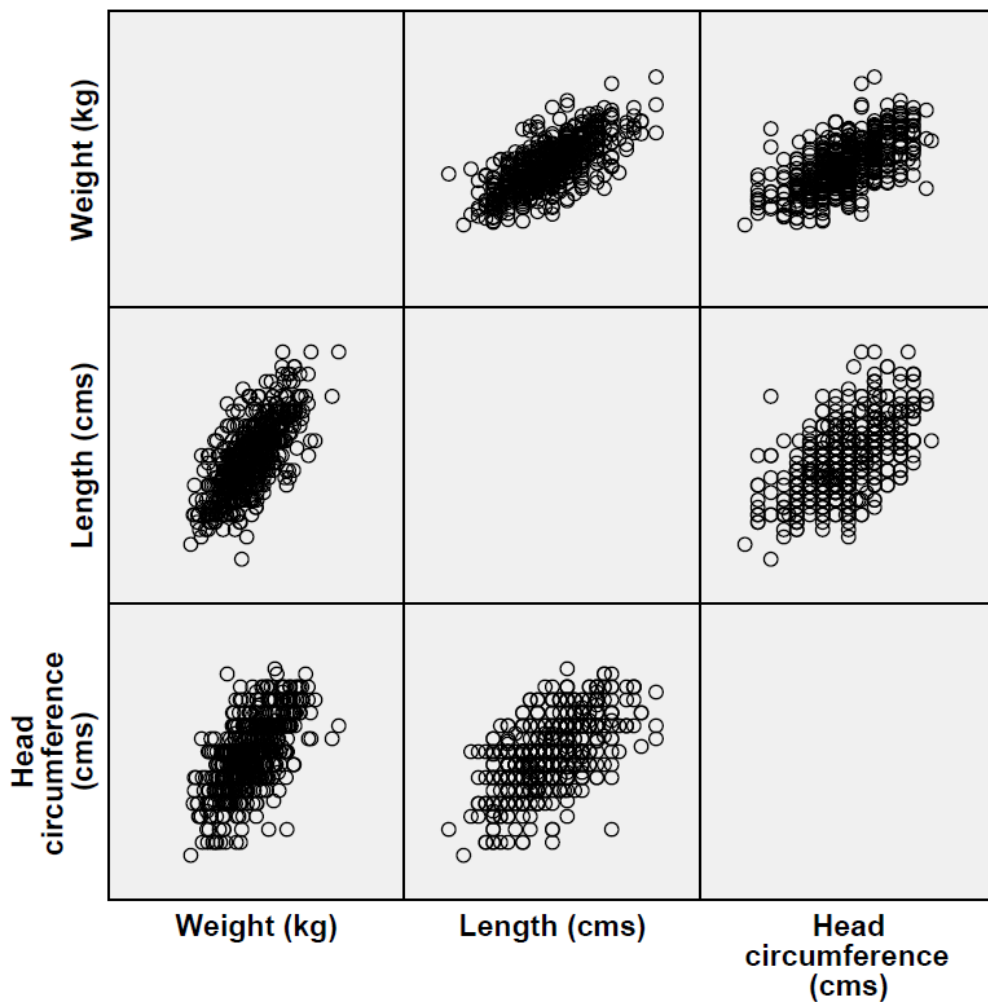


	id	weight	length	headc	gender	educatio	parity
1	L001	3,95	55,5	37,5	2	4	3
2	L003	4,63	57,0	38,5	2	4	0
3	L004	4,75	56,0	38,5	1	3	2
4	L005	3,92	56,0	39,0	1	4	1
5	L006	4,56	55,0	39,5	1	2	2
6	L007	3,64	51,5	34,5	2	4	0
7	L008	3,55	56,0	38,0	2	2	3
8	L009	4,53	57,0	39,7	1	2	3
9	L010	4,97	58,5	39,0	1	4	2
10	L011	3,74	52,0	38,0	1	2	1
11	L013	4,19	54,0	37,5	2	2	2
12	L014	4,35	59,0	39,5	2	2	2
13	L016	4,62	56,0	38,5	2	2	0
14	L018	4,09	53,0	38,0	2	3	0
15	L019	4,42	57,0	38,0	2	2	1
16	L020	5,62	58,0	40,0	1	4	0
17	L021	5,25	58,0	38,5	1	3	3
18	L022	4,32	54,5	38,5	1	2	1

Esempio

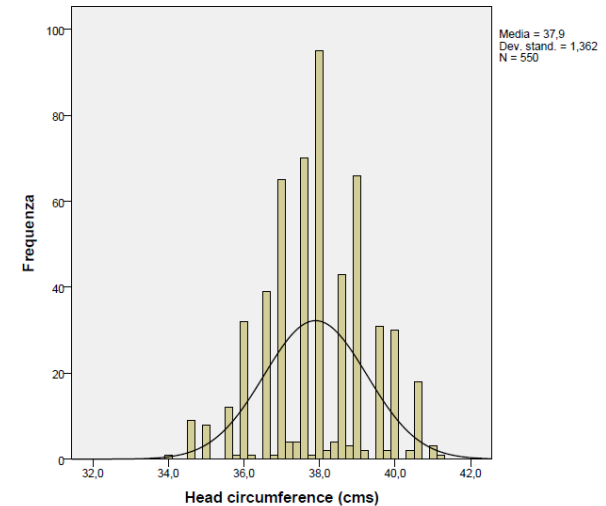
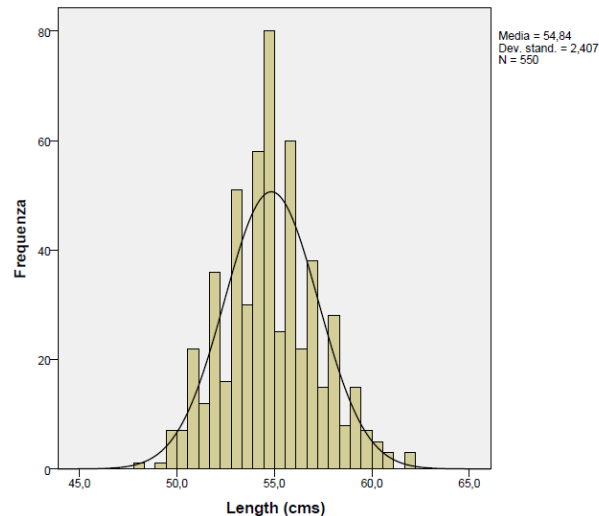
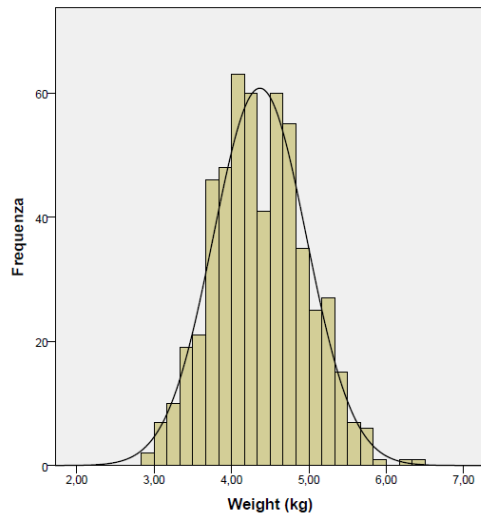
Analisi descrittiva: scatter plot

Tutti i grafici
mostrano
un'associazione
lineare positiva
per ogni
combinazione
bivariata



Esempio

Verifica della distribuzione normale delle variabili



	Asimmetria		Curtosi	
	Statistica	Errore std	Statistica	Errore std
Weight (kg)	,201	,104	-,146	,208
Length (cms)	,223	,104	-,141	,208
Head circumference (cms)	-,121	,104	-,206	,208
Validi (listwise)				

Le tre variabili
hanno una
distribuzione
approssimabile a
una normale

Esempio

Calcola l'appropriato coefficiente di correlazione

Correlazioni

		Weight (kg)	Length (cms)	Head circumference (cms)
Weight (kg)	Correlazione di Pearson	1	,713	,622
	Sig. (2-code)		,000	,000
	N	550	550	550
Length (cms)	Correlazione di Pearson	,713**	1	,598**
	Sig. (2-code)	,000		,000
	N	550	550	550
Head circumference (cms)	Correlazione di Pearson	,622**	,598**	1
	Sig. (2-code)	,000	,000	
	N	550	550	550

** La correlazione è significativa al livello 0,01 (2-code).

Guarda il pvalue

Guarda l'intervallo di confidenza

Coefficienti^a

Modello	Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Intervallo di confidenza 95,0% per B	
	B	Deviazione standard Errore	Beta			Limite inferiore	Limite superiore
1 (Costante)	-2,383E-015	,030		,000	1,000	,066	,066
Punteggio Z: Length (cms)	,713	,030	,713	23,813	,000	,654	,772

a. Variabile dipendente: Punteggio Z: Weight (kg)

Coefficienti^a

Modello	Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Intervallo di confidenza 95,0% per B	
	B	Deviazione standard Errore	Beta			Limite inferiore	Limite superiore
1 (Costante)	-1,486E-014	,033		,000	1,000	,066	,066
Punteggio Z: Head circumference (cms)	,622	,033	,622	18,618	,000	,557	,688

a. Variabile dipendente: Punteggio Z: Weight (kg)

Coefficienti^a

Modello	Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Intervallo di confidenza 95,0% per B	
	B	Deviazione standard Errore	Beta			Limite inferiore	Limite superiore
1 (Costante)	-8,620E-015	,034		,000	1,000	,067	,067
Punteggio Z: Head circumference (cms)	,598	,034	,598	17,478	,000	,531	,666

a. Variabile dipendente: Punteggio Z: Length (cms)

Esempio

Test di ipotesi per la correlazione

$H_0: \rho = 0$ (non c'è correlazione)

$H_1: \rho \neq 0$ (c'è correlazione)

Statistica test

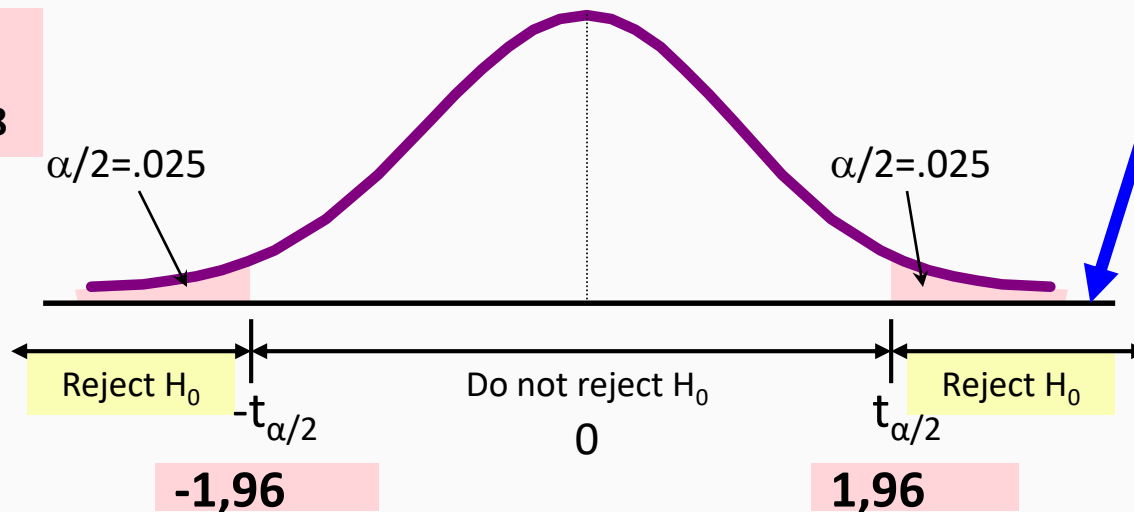
$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

con $n-2$
gradi di
libertà

$$t = \frac{0.713}{\sqrt{\frac{1 - (0.713)^2}{n - 2}}} = 23,8$$

$\alpha = .05$

d.f. = $550 - 2 = 548$



Esempio

Test di ipotesi per la correlazione

Il coefficiente di correlazione lineare di Pearson mostra che la maggiore associazione lineare è tra peso e lunghezza con un valore di r pari a 0,713 e una modesta associazione tra peso e circonferenza cranica (0,622).

Malgrado la differenza nella forza di associazione i coefficienti di correlazione sono tutti significativi $p\text{value} < 0.0001$; a sostegno di quanto il $p\text{value}$ sia poco sensibile nella selezione dei predittori di un dato outcome.

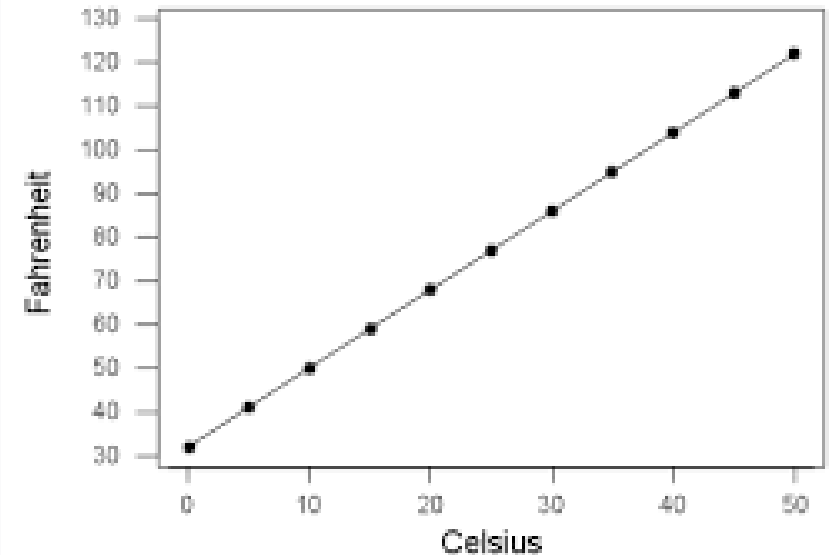
Regressione lineare

È un metodo statistico che permette di studiare e sintetizzare la relazione tra due variabili quantitative

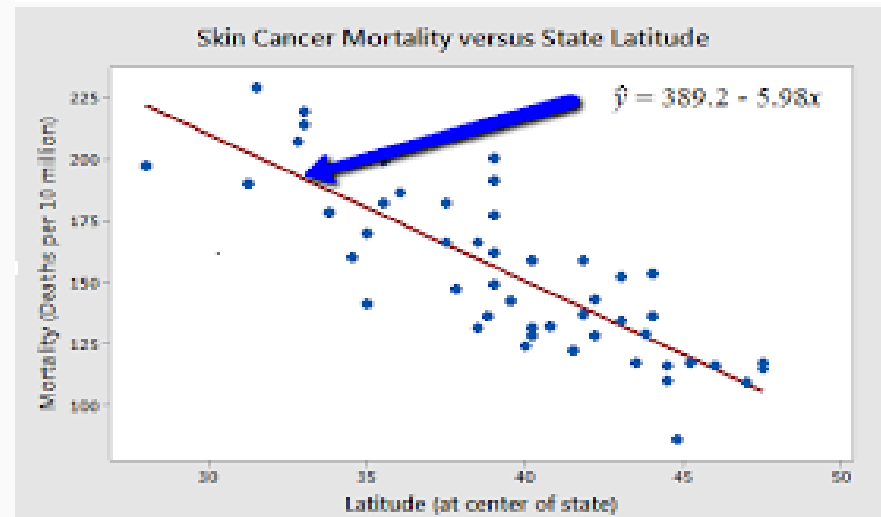
- una o più variabili sono chiamate predittori o variabili indipendenti o variabili esplicative
- Una variabile di risposta o outcome o variabile dipendente

Relazione deterministica vs relazione probabilistica

Relazione deterministica
L'equazione della retta di regressione descrive esattamente la relazione tra due variabili



Relazione probabilistica
L'associazione non è perfetta, siamo interessati a calcolare una statistica sui parametri della retta di regressione



Obiettivi della regressione lineare

1. Costruire un modello di predizione, in grado di prevedere l'outcome di interesse conoscendo altre informazioni dei soggetti inclusi nel campione (variabili esplicative)
2. Valutare l'effetto di una variabile esplicativa (es. un fattore di rischio) su un outcome di interesse

Equazione della retta di regressione

The population regression model:

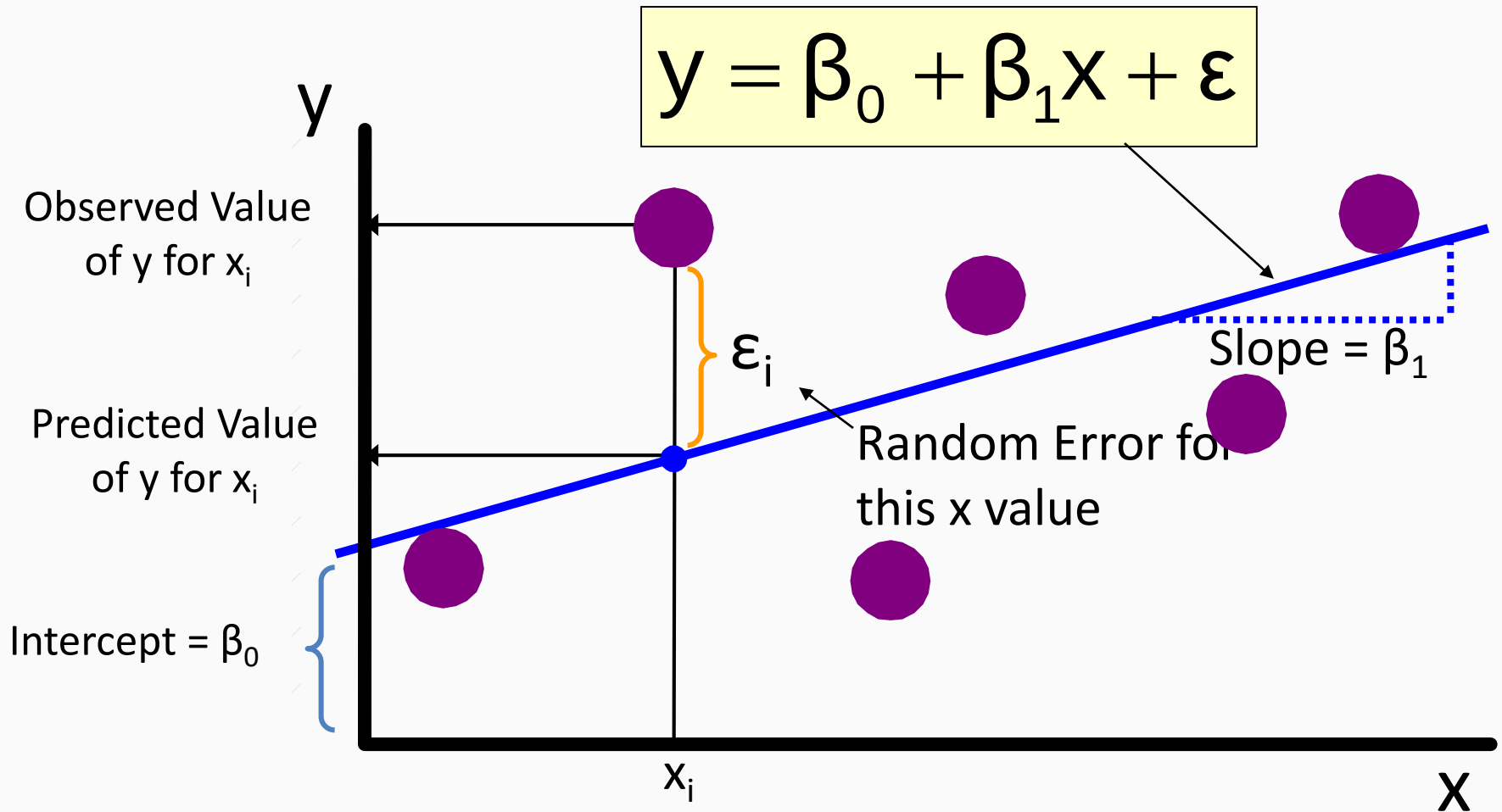
The diagram illustrates the population regression model equation $y = \beta_0 + \beta_1 X + \epsilon$ within a yellow rectangular box. Several labels with arrows point to specific parts of the equation: 'Dependent Variable' points to y ; 'Population y intercept' points to β_0 ; 'Population Slope Coefficient' points to β_1 ; 'Independent Variable' points to X ; and 'Random Error term, or residual' points to ϵ . Below the box, two purple curly braces identify the 'Linear component' (covering $\beta_0 + \beta_1 X$) and the 'Random Error component' (covering ϵ).

$$y = \beta_0 + \beta_1 X + \epsilon$$

Labels and components:

- Dependent Variable: y
- Population y intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: X
- Random Error term, or residual: ϵ
- Linear component: $\beta_0 + \beta_1 X$
- Random Error component: ϵ

Population Linear Regression



Equazione della retta di regressione

La retta di regression campionaria fornisce una stima della retta di regressione di popolazione

Estimated (or
predicted) y
value

Estimate of the
regression
intercept

Estimate of the
regression slope

Independent
variable

$$\hat{y}_i = b_0 + b_1 x$$

The individual random error terms e_i have a mean of zero

Equazione della retta di regressione

In generale quando calcoliamo la retta di regressione

$$\hat{y}_i = b_0 + b_1 x$$

Commettiamo un errore (residui) di grandezza pari a

$$e_i = y_i - \hat{y}_i$$

Criterio dei minimi quadrati

- b_0 e b_1 sono ottenuti calcolando i valori di b_0 e b_1 che minimizzano la somma residui quadrati

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1 x))^2\end{aligned}$$

Coefficienti della retta

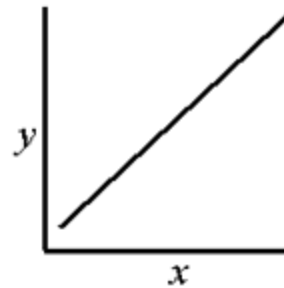
Pendenza

$$b_1 = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

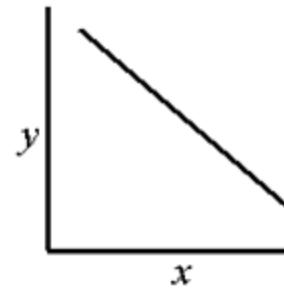
Denominatore sempre positivo. Il segno della pendenza è determinato dal numeratore.

$b_1 > 0$ quando x aumenta y tende ad aumentare

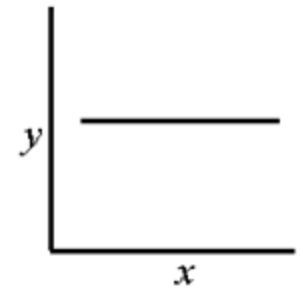
$b_1 < 0$ quando x aumenta y tende a diminuire



Positive slope



Negative slope



Zero slope

Intercetta

$$b_0 = \bar{y} - b_1 \bar{x}$$

Interpretazione della pendenza e dell'intercetta

- b_0 rappresenta l'intercetta della retta di regressione ed indica il valore della variabile di risposta Y quando il predittore x assume valore 0.
- b_1 rappresenta l'inclinazione della retta di regressione, ovvero la variazione della variabile di risposta Y in conseguenza di un aumento unitario del predittore x .

Assunzioni nella regressione lineare

Disegno dello studio

- Il campione deve essere rappresentativo della popolazione su cui viene fatta inferenza
- I dati devono essere raccolti in un periodo nel quale la relazione tra outcome e variabili esplicative rimane costante

Modello

- La relazione tra variabili esplicative e outcome è approssimabile ad una retta
- I residui sono distribuiti in modo normale
- Omoscedasticità: varianza costante su tutto il modello

Indipendenza

- Tutte le osservazioni sono indipendenti le une dalle altre
- La collinearità tra le variabili esplicative è bassa

Esempio

Utilizzando lo stesso studio (dataset weights.sav)

Domanda: La lunghezza del corpo può essere utilizzata per predire il peso a un mese di vita?

H0: Non c'è relazione tra lunghezza del corpo e peso a un mese

Variabili:

Variabile di outcome: peso corporeo (Kg)

Variabile esplicativa= lunghezza del corpo (cm)

Output del modello di regressione lineare in SPSS: riepilogo del modello

Riepilogo del modello

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima
1	,713 ^a	,509	,508	,42229

a. Predittori: (Costante), Length (cms)

Il valore di r^2 (**coefficiente di determinazione**) di 0,509 indica che il 50,9% della variazione nel peso è spiegata dalla lunghezza

Output del modello di regressione lineare in SPSS: tabella ANOVA

La varianza del modello può essere divisa in due parti

$$SST = SSE + SSR$$

Total sum of
Squares

Sum of
Squares Error

Sum of Squares
Regression

$$SST = \sum (y - \bar{y})^2$$

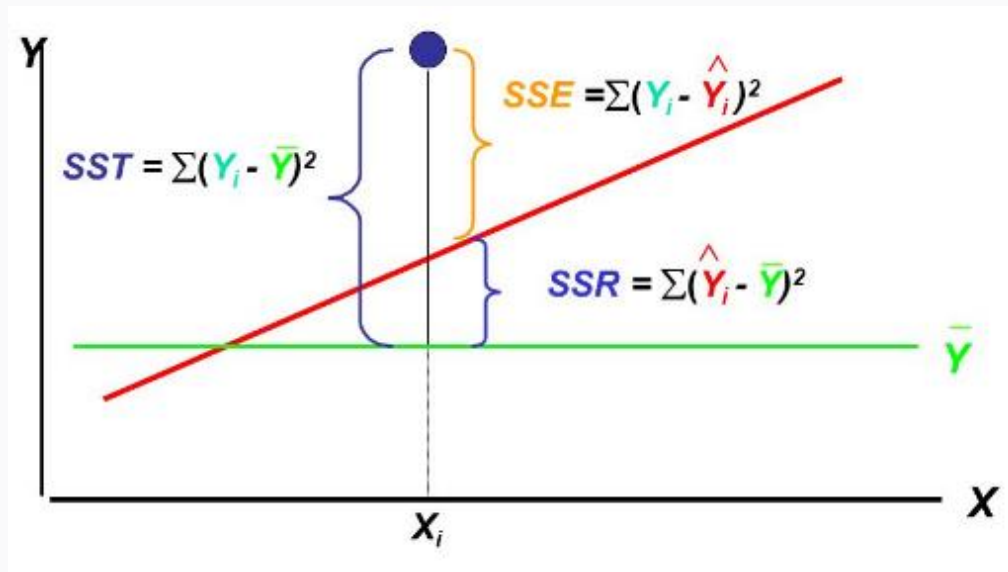
$$SSE = \sum (y - \hat{y})^2$$

$$SSR = \sum (\hat{y} - \bar{y})^2$$

Misura la variazione dei
valori y_i intorno alla loro
media

Misura la variazione
attribuibile a fattori
esterni alla relazione
di x e y

Misura la variazione
attribuibile alla
relazione di x e y



Output del modello di regressione lineare in SPSS: tabella ANOVA

Anova^a

Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	101,119	1	101,119	567,043	,000 ^b
	Residuo	97,723	548	,178		
	Totale	198,842	549			

a. Variabile dipendente: Weight (kg)

b. Predittori: (Costante), Length (cms)

La statistica F, che si ottiene dividendo la varianza spiegata dalla regressione per la varianza non spiegata dalla relazione x/y , è largamente significativa $p < .0001$.

C'è una significativa relazione lineare tra lunghezza e peso. Questa statistica indica anche che il modello di regressione lineare complessivamente predice l'outcome

Output del modello di regressione lineare in SPSS: i coefficienti

Coefficienti ^a							
Modello		Coefficienti non standardizzati		Coefficienti standardizzati		Intervallo di confidenza 95,0% per B	
		B	Deviazione standard Errore	Beta		Limite inferiore	Limite superiore
1	(Costante)	-5,412	,411		-13,167	,000	-6,220
	Length (cms)	,178	,007	,713	23,813	,000	-4,605

a. Variabile dipendente: Weight (kg)

$$y = b_0 + b_1x$$

$$\text{Peso} = -5.412 + (0,178 \times \text{lunghezza})$$

b_0 = serve solo per calcolare i valori sulla retta di regressione; ha uno scopo strumentale e nessun significato biologico

b_1 = coefficiente di regressione indica che per un incremento unitario della lunghezza il peso aumenta di 0,178 Kg

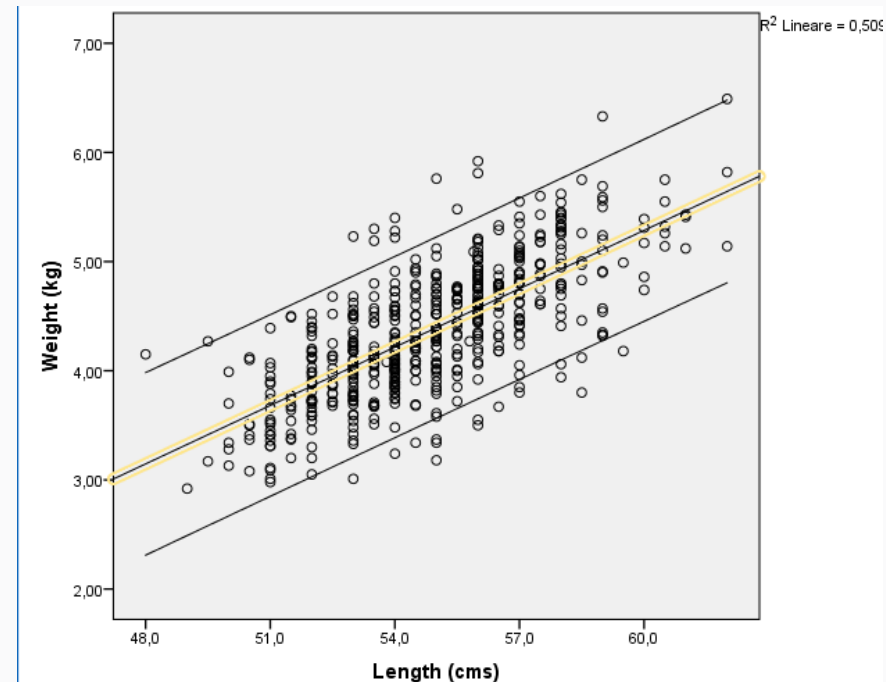
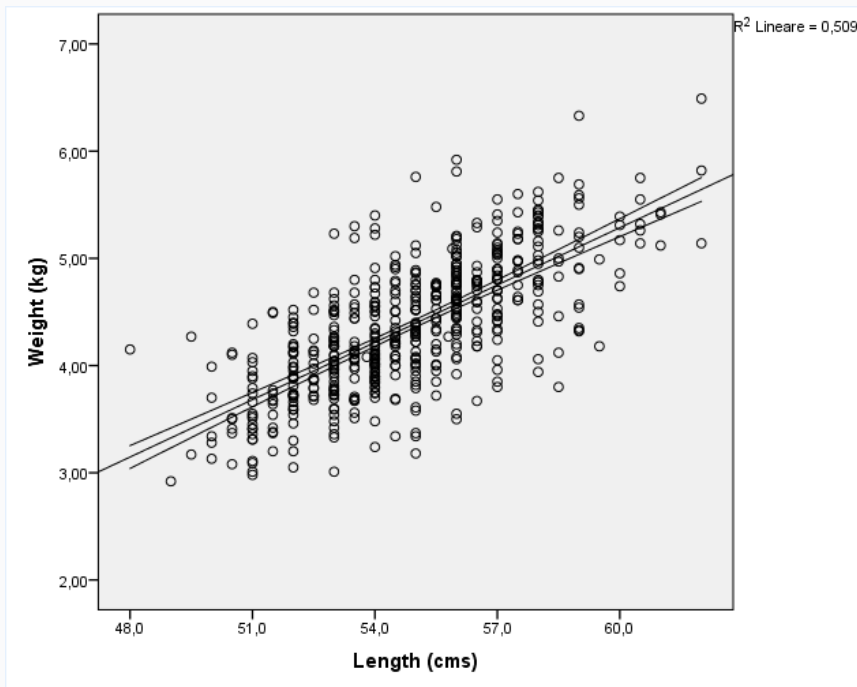
Test di ipotesi sui coefficienti della retta ($H_0 : b_0 \text{ e } b_1 = 0$) : i valori di t sono calcolati dividendo i valori dei b per i loro errori standard

Indicano il contributo relativo di ciascuna variabile esplicativa al modello (in questo caso è uguale a r perché c'è una sola variabile indipendente)

Output del modello di regressione lineare in SPSS: grafico della retta di regressione

Possiamo disegnare la retta di regressione con gli intervalli di confidenza intorno al valore predetto che rappresenta l'area nella quale c'è il 95% della «vera» retta di regressione

NB da non confondere con l'intervallo di confidenza in cui si trovano il 95% delle osservazioni



Regressione multipla

Un modello lineare nel quale sono presenti due o più variabili esplicative è chiamato regressione lineare multipla

Le variabili esplicative possono essere continue o categoriche

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 \dots$$

b_0 è l'intercetta

b_i sono i coefficienti di regressione di ciascuna variabile esplicativa, o, in altre parole, i pesi che assegniamo a ciascuna variabile esplicativa inclusa nel modello

Tre metodi per costruire il modello

- **Standard:** tutte le variabili entrano insieme nel modello
- **Stepwise:** l'ordine di ingresso delle variabili nel modello è determinato dalla forza della loro correlazione
 - Forward: le variabili sono aggiunte una alla volta fino a che l'ingresso di una variabile spiega solo una quantità di varianza nel modello non significativa
 - Backward: tutte le variabili entrano insieme nel modello e viene tolta una variabile per volta se non contribuisce significativamente alla predizione dell'outcome
- **Sequential:** l'ordine di ingresso delle variabili è determinato dal ricercatore secondo criteri teorici o statistici

Un esempio di modello di regressione multipla

Utilizzando lo stesso studio (dataset weights.sav)

Domanda: La lunghezza del corpo e il genere possono essere utilizzate per predire il peso a un mese di vita?

H0: Non c'è relazione tra lunghezza del corpo, genere e peso a un mese

Variabili:

Variabile di outcome: peso corporeo (Kg)

Variabile esplicatoria: lunghezza del corpo (cm)

Variabile esplicatoria: genere

Esempio

Riepilogo del modello									
Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima	Variazione dell'adattamento				
					Variazione di R-quadrato	Variazione di F	df1	df2	Sig. Variazione di F
1	,713 ^a	,509	,508	,42229	,509	567,043	1	548	,000
2	,741 ^b	,549	,548	,40474	,041	49,543	1	547	,000

a. Predittori: (Costante), Length (cms)

b. Predittori: (Costante), Length (cms), Gender

Poiché ci sono due variabili è appropriato valutare il coefficiente di determinazione r^2 aggiustato

Nel secondo modello r^2 aumenta di 0,041

L'aumento è significativo e il modello è migliore nella predizione dell'outcome

Anova ^a						
Modello		Somma dei quadrati	df	Media dei quadrati	F	Sig.
1	Regressione	101,119	1	101,119	567,043	,000 ^b
	Residuo	97,723	548	,178		
	Totale	198,842	549			
2	Regressione	109,235	2	54,617	333,407	,000 ^c
	Residuo	89,607	547	,164		
	Totale	198,842	549			

a. Variabile dipendente: Weight (kg)

La statistica F mostra una significativa relazione lineare tra lunghezza e peso e tra lunghezza, genere e peso. Questa statistica indica che i modelli di regressione lineare complessivamente predicono l'outcome

Esempio

Coefficienti ^a							
Modello	Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Intervallo di confidenza 95,0% per B	
	B	Deviazione standard Errore	Beta			Limite inferiore	Limite superiore
1 (Costante)	-5,412	,411		-13,167	,000	-6,220	-4,605
Length (cms)	,178	,007	,713	23,813	,000	,164	,193
2 (Costante)	-4,312	,424		-10,173	,000	-5,144	-3,479
Length (cms)	,165	,007	,660	22,259	,000	,151	,180
Gender	-,251	,036	-,209	-7,039	,000	-,321	-,181

a. Variabile dipendente: Weight (kg)

L'errore standard della lunghezza resta costante anche aggiungendo al modello la variabile genere, ciò depone per una buona stabilità del modello

Possiamo riscrivere l'equazione della retta
 $y = -4,56 + 0,165 \times \text{lunghezza} - (0,251 \times \text{genere})$

Poiché nel db i maschi sono codificati con zero, l'ultimo termine può essere rimosso per i maschi. Una femmina in media, dopo aggiustamento per la lunghezza, ha un peso più basso di 0,251 Kg rispetto ad un maschio.

I coefficienti standardizzati ci dicono che la lunghezza è il maggiore predittore del peso (0,660 lunghezza, 0,209 genere)

Esempio

Variabili escluse ^a					
Modello	Beta In	t	Sig.	Correlazioni parziali	Statistiche di collinearità
					Tolleranza
1 Gender	-,209 ^b	-7,039	,000	-,288	,936

a. Variabile dipendente: Weight (kg)

b. Predittori nel modello : (Costante), Length (cms)

Statistiche di collinearità

Una tolleranza vicino 1 indica assenza di multicollinearità tra le variabili

NB

Tolleranza <0,2 multicollinearità

Interazione

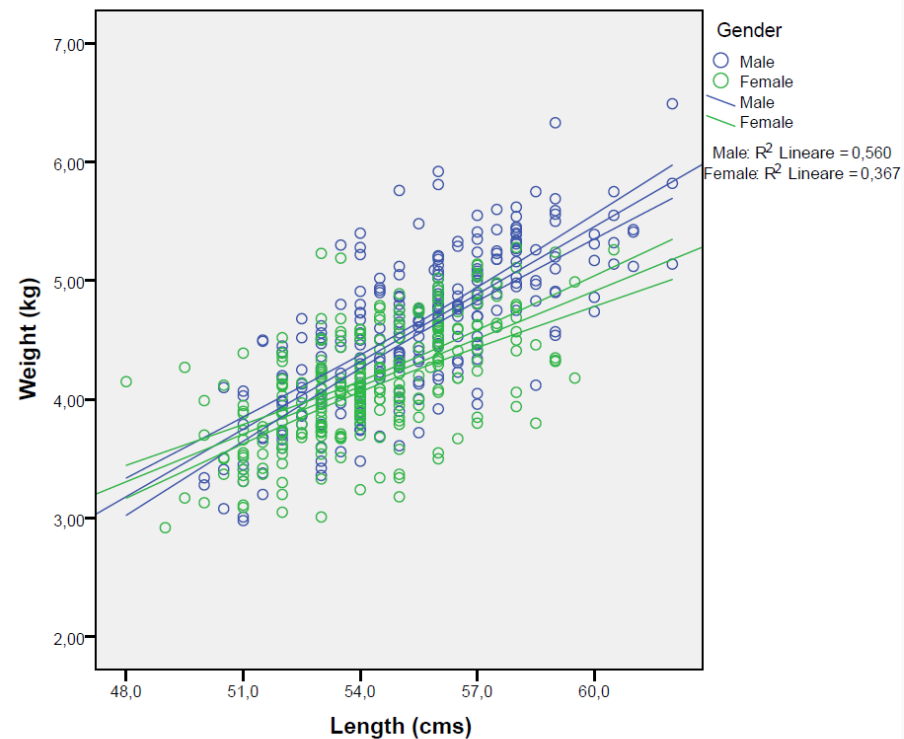
Una interazione si verifica quando c'è una relazione moltiplicativa piuttosto che additiva tra le variabili esplicative

Rapportandoci al nostro esempio se non ci fosse interazione tra le variabili, il coefficiente di regressione delle rette nei gruppo dei maschi e delle femmine dovrebbe essere uguale.

A livello grafico, quindi, nel caso di assenza di interazione tra le variabili lunghezza e genere, le rette dei maschi e delle femmine dovrebbero essere approssimativamente parallele.

Se ci fosse interazione ci aspettiamo due rette con diversa pendenza e che, quindi, si intersecano

Interazione



Coefficienti^{a,b}

Modello	Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Intervallo di confidenza 95,0% per B	
	B	Deviazione standard Errore	Beta			Limite inferiore	Limite superiore
1 (Costante)	-5,913	,564		-10,491	,000	-7,022	-4,803
Length (cms)	,189	,010	,749	18,657	,000	,169	,209

a. Variabile dipendente: Weight (kg)

b. Selezione in corso solo dei casi per cui Gender = Male

Coefficienti^{a,b}

Modello	Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Intervallo di confidenza 95,0% per B	
	B	Deviazione standard Errore	Beta			Limite inferiore	Limite superiore
1 (Costante)	-2,113	,577		-5,394	,000	-4,249	-1,977
Length (cms)	,134	,011	,606	12,580	,000	,113	,155

a. Variabile dipendente: Weight (kg)

b. Selezione in corso solo dei casi per cui Gender = Female

Interazione

Per testare la presenza di interazione, le due variabili si moltiplicano calcolando una terza variabile di interazione. Includiamo la nuova variabile nel modello

Riepilogo del modello^c

Modello	R	R-quadrato	R-quadrato corretto	Deviazione standard Errore della stima	Variazione dell'adattamento				
					Variazione di R-quadrato	Variazione di F	df1	df2	Sig. Variazione di F
1	,741 ^a	,549	,548	,40474	,549	333,407	2	547	,000
2	,749 ^b	,561	,558	,39994	,011	14,215	1	546	,000

a. Predittori: (Costante), Gender, Length (cms)

b. Predittori: (Costante), Gender, Length (cms), lengthxgender

c. Variabile dipendente: Weight (kg)

Coefficienti^a

Modello		Coefficienti non standardizzati		Coefficienti standardizzati	t	Sig.	Intervallo di confidenza 95,0% per B	
		B	Deviazione standard Errore	Beta			Limite inferiore	Limite superiore
1	(Costante)	-4,312	,424		-10,173	,000	-5,144	-3,479
	Length (cms)	,165	,007	,660	22,259	,000	,151	,180
	Gender	-,251	,036	-,209	-7,039	,000	-,321	-,181
2	(Costante)	-8,713	1,240		-7,026	,000	-11,149	-6,277
	Length (cms)	,245	,022	,981	10,914	,000	,201	,289
	Gender	2,800	,810	2,328	3,457	,001	1,209	4,391
	lengthxgender	-,056	,015	-2,478	-3,770	,000	-,085	-,027

a. Variabile dipendente: Weight (kg)

Variabili escluse^a

Modello	Beta In	t	Sig.	Correlazioni parziali	Statistiche di collinearità
					Tolleranza
1 lengthxgender	-2,478 ^b	-3,770	,000	-,159	,002

a. Variabile dipendente: Weight (kg)

b. Predittori nel modello : (Costante), Gender, Length (cms)

Residui

I residui rappresentano le distanze tra ciascun valore osservato e valore predetto dal modello di regressione

Tra le assunzioni che avevamo enunciato per il modello di regressione lineare:

- I residui sono distribuiti in modo normale
- Omoscedasticità: varianza costante su tutto il modello

Analisi dei residui sono importanti per valutare se le assunzioni sono violate

E' utile convertire i residui in residui standardizzati: distanze espresse in unità di deviazione standard

Residui

- I residui sono distribuiti in modo normale?

Calcolare e salvare i residui standardizzati del modello di regressione

Statistiche descrittive > esplora

Descrittive			
		Statistica	Errore std.
Standardized Residual	Media	0E-7	,04252348
	Intervallo di confidenza per la media al 95%	Limite inferiore	-,0835286
		Limite superiore	,0835286
	Media 5% trim	-,0010561	
	Mediana	-,0194572	
	Varianza	,995	
	Deviazione std.	,99726402	
	Minimo	-2,65810	
	Massimo	3,13650	
	Intervallo	5,79460	
	Distanza interquartilica	1,28216	
	Asimmetria	,084	,104
	Curtosi	,223	,208

Test di normalità						
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Statistica	df	Sig.	Statistica	df	Sig.
Standardized Residual	,030	550	,200*	,994	550	,033

*. Limite inferiore della significatività effettiva.

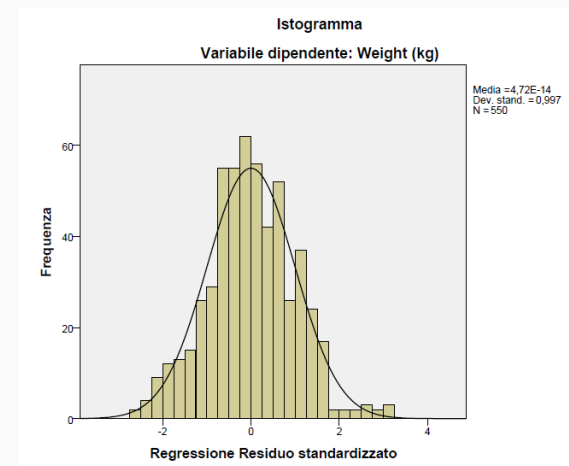
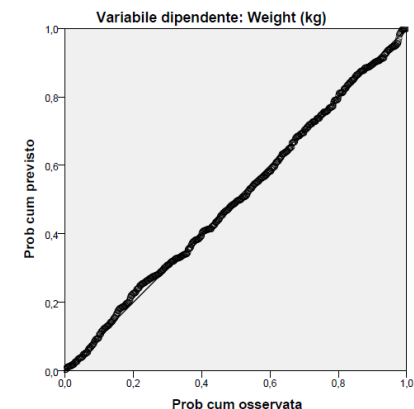


Grafico di normalità P-P di regressione Residuo standardizzato



Residui

- Omoscedasticità: varianza costante su tutto il modello

Possiamo visualizzare lo spread della varianza nel modello attraverso uno scatter plot dei residui sui valori attesi

La varianza è approssimativamente costante, il modello è omoscedastico

