

Statistica e Laboratorio

3. Statistica descrittiva: analisi multivariate

Paolo Vidoni

Dipartimento di Scienze Economiche e Statistiche

Università di Udine

via Tomadini 30/a - Udine

paolo.vidoni@uniud.it

<https://elearning.uniud.it/>

Sommario

- 1 **Sommario e introduzione**
- 2 Analisi di dipendenza: la connessione
- 3 Dipendenza in media
- 4 Analisi di correlazione
- 5 Analisi di regressione

Sommario

- **Introduzione**
- **Analisi di dipendenza: la connessione**
- **Dipendenza in media**
- **Analisi di correlazione**
- **Analisi di regressione**

Introduzione alle analisi multivariate

Le analisi descrittive multivariate sono relative allo studio congiunto di due o più variabili statistiche.

Si considera, in particolare, il caso di due variabili statistiche: **analisi descrittive bivariate**.

L'analisi congiunta di due variabili può fornire conclusioni interessanti sulla manifestazione di un fenomeno che si articola nell'osservazione congiunta di due suoi aspetti particolari.

In questo ambito, riveste un'importanza notevole lo studio delle potenziali **relazioni esistenti tra le due variabili** in esame.

Ad esempio, si può dare risposta a domande quali: *il voto di laurea è in relazione con il tempo impiegato per trovare lavoro? Data l'altezza del padre è possibile prevedere l'altezza del figlio?*

Distribuzioni di frequenza

Si considerano due variabili X e Y . La loro osservazione su n unità statistiche fornisce i **dati grezzi** (x_i, y_i) , $i = 1, \dots, n$.

A partire dai dati grezzi si possono determinare le **distribuzioni di frequenza assoluta e relativa** che si possono distinguere in:

- **distribuzione congiunta**, se si considerano le frequenze delle unità che presentano congiuntamente la modalità x_r , $r = 1, \dots, m$ della prima variabile e la modalità y_s , $s = 1, \dots, k$, della seconda;
- **distribuzione marginale**, se si considera la distribuzione di frequenza relativa ad una singola variabile;
- **distribuzione condizionata**, se si considera la distribuzione di frequenza relativa ad una singola variabile considerando soltanto le unità statistiche che assumono una determinata modalità dell'altra.

Si può operare allo stesso modo anche se si hanno modalità raggruppate in classi.

Rappresentazioni grafiche

Oltre all'analisi delle distribuzioni di frequenza, risultano molto utili le **rappresentazioni grafiche**.

Tenendo conto della tipologia della variabili, dei dati a disposizione e degli obiettivi dell'analisi, si possono utilizzare i diagrammi visti in precedenza o loro opportune estensioni tridimensionali.

Se si dispone dei **dati grezzi**, riferiti a due variabili quantitative, si può disegnare un **grafico di dispersione (scatterplot)**, dove le coppie (x_i, y_i) , $i = 1, \dots, n$, sono rappresentate come punti del piano, i cui assi corrispondono alle due variabili.

Per rappresentare una **distribuzione di frequenza congiunta** di due variabili quantitative si possono utilizzare **istogrammi o diagrammi a bastoncini**, disegnati in uno spazio tridimensionale.

Per rappresentare una **distribuzione di frequenza marginale o condizionata** si possono utilizzare tutte le rappresentazioni grafiche viste per il caso univariato: istogrammi, boxplot, diagrammi a bastoncini, a rettangoli, a torta, ecc.

Studio della dipendenza

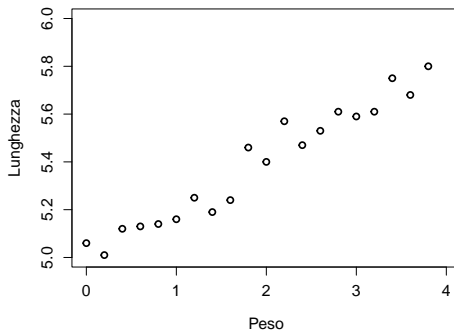
In alcuni casi le variabili X e Y vengono trattate in modo **simmetrico**.

In altri casi, come nell'esempio sottostante, è necessario individuare la **variabile dipendente (risposta)** e la **variabile indipendente (esplicativa)**; X e Y sono trattate in modo **non simmetrico**.

Esempio. *Molla.* Si considerano i dati sulla misura in cm di una molla sottoposta a $n = 20$ pesi diversi in Kg

Peso (Kg)	Lunghezza (cm)	Peso (Kg)	Lunghezza (cm)
0.0	5.06	2.0	5.40
0.2	5.01	2.2	5.57
0.4	5.12	2.4	5.47
0.6	5.13	2.6	5.53
0.8	5.14	2.8	5.61
1.0	5.16	3.0	5.59
1.2	5.25	3.2	5.61
1.4	5.19	3.4	5.75
1.6	5.24	3.6	5.68
1.8	5.46	3.8	5.80

Si considera l'associato diagramma di dispersione.



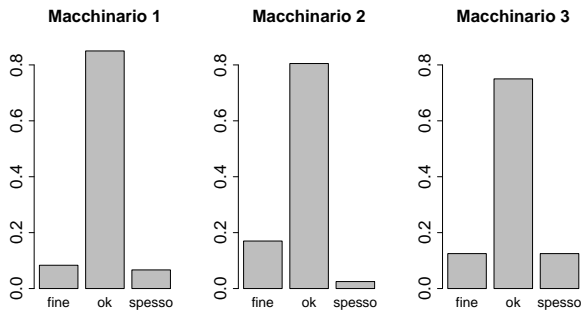
La lunghezza della molla (variabile risposta) dipende dall'entità del peso applicato (variabile esplicativa).

Dall'analisi del diagramma si vede che, in generale, un peso elevato porta ad un incremento della lunghezza della molla, sulla base di una relazione che sembra lineare.



Esempio. *Perni* (continua). Considerando i dati riferiti alla produzione dei perni, si determinano le distribuzioni di frequenza relativa riferite al diametro, condizionatamente al macchinario di produzione.

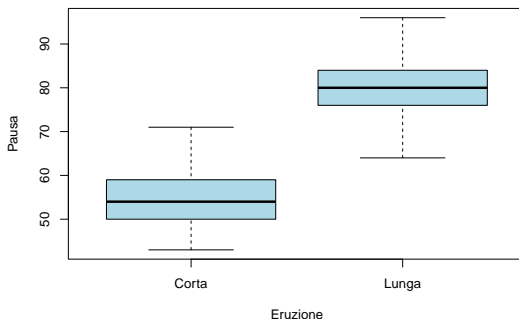
Le distribuzioni di frequenza vengono rappresentate con diagrammi a barre.



Si nota che il macchinario di produzione influenza la dimensione del diametro dei perni.



Esempio. *Geyser Old Faithful* (continua). Si considerano i dati riferiti alle eruzioni del geyser Old Faithful. Si analizzano le distribuzioni di frequenza relativa riferite alla durata delle pause, condizionatamente alla tipologia dell'eruzione precedente.



Si nota che la tipologia di eruzione che precede le pause influisce sulla loro durata.



Alla luce degli esempi presentati, può essere interessante studiare, in particolare, il caso bivariato con l'obiettivo di evidenziare le relazioni esistenti tra due variabili statistiche.

Si evidenziano tre situazioni tipiche:

- le due variabili sono qualitative: **analisi di dipendenza o (connessione)**;
- una variabile è qualitativa e l'altra quantitativa: **analisi di dipendenza in media**;
- le due variabili sono quantitative: **analisi di correlazione e analisi di regressione**.

Se si hanno più di due variabili, si può in prima battuta analizzare le relazioni tra le variabili considerate a due a due.

Per una analisi complessiva si possono utilizzare i metodi più generali, propri della statistica descrittiva multivariata.

Sommario

- 1 Sommario e introduzione
- 2 Analisi di dipendenza: la connessione**
- 3 Dipendenza in media
- 4 Analisi di correlazione
- 5 Analisi di regressione

Analisi di dipendenza

Si considerano due **variabili statistiche** X e Y **qualitative (categoriali)** e si vuole indagare l'esistenza o meno di associazione (dipendenza) tra le modalità corrispondenti.

Esempio. Attitudine. Si analizza l'attitudine musicale X e pittorica Y di $n = 15$ individui con la seguente scala di modalità: sufficiente (S), buona (B), ottima (O). I dati vengono sintetizzati nella seguente tabella di frequenza congiunta, detta tabella di contingenza

		Y			
		S	B	O	
X	S	1	3	0	4
	B	1	3	2	6
	O	2	1	2	5
		4	7	4	15

Ad esempio, il valore 3 nella prima riga indica che ci sono 3 individui con attitudine musicale sufficiente e attitudine musicale buona. ◇

Tabella di contingenza

Una **tabella di contingenza** descrive la frequenza con la quale le modalità (categorie) di due variabili qualitative X e Y vengono *congiuntamente* osservate.

Se X ha m categorie, x_1, \dots, x_m , ed Y ha k categorie, y_1, \dots, y_k , la tabella di contingenza contiene la frequenza assoluta n_{rs} delle $m \times k$ possibili coppie (x_r, y_s) , $r = 1, \dots, m$, $s = 1, \dots, k$,

	y_1	y_2	\dots	y_k	
x_1	n_{11}	n_{12}	\dots	n_{1k}	n_{1+}
x_2	n_{21}	n_{22}	\dots	n_{2k}	n_{2+}
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
x_m	n_{m1}	n_{m2}	\dots	n_{mk}	n_{m+}
	n_{+1}	n_{+2}	\dots	n_{+k}	n

dove n_{r+} e n_{+s} sono, rispettivamente, i totali di riga e di colonna e n il numero totale di osservazioni.

Distribuzioni di frequenza

Una tabella di frequenza è di fatto una distribuzione doppia di frequenza. Inoltre, risulta utile per indagare le relazioni esistenti tra le modalità delle due variabili.

Dalla tabella di contingenza si ricavano le seguenti **distribuzioni di frequenza assoluta**

- **congiunta:** (x_r, y_s) , n_{rs} , $r = 1, \dots, m$, $s = 1, \dots, k$;
- **marginale di X :** x_r , n_{r+} , $r = 1, \dots, m$;
- **marginale di Y :** y_s , n_{+s} , $s = 1, \dots, k$;
- **condizionata di X dato $Y = y_s$:** x_r , n_{rs} , $r = 1, \dots, k$;
- **condizionata di Y dato $X = x_r$:** y_s , n_{rs} , $s = 1, \dots, m$.

Le frequenze marginali di X e di Y corrispondono, rispettivamente, ai totali di riga e di colonna.

Le frequenze condizionate di X e di Y corrispondono, rispettivamente, ai valori di colonna e di riga individuati dalle condizioni $Y = y_s$ e $X = x_r$.

Le **distribuzioni di frequenza relativa** si ottengono dividendo per i corrispondenti totali. In particolare, le distribuzioni congiunta e marginali si dividono per n , le distribuzioni condizionate per i totali di riga o di colonna corrispondenti alla condizione.

Esempio. *Attitudine* (continua). Con riferimento alla analisi delle attitudini musicale e pittorica, si ottengono le seguenti distribuzioni di frequenza marginale assoluta e relativa

X	S	B	O	Totale
f	4	6	5	15
p	0.267	0.400	0.333	1

Y	S	B	O	Totale
f	4	7	4	15
p	0.267	0.466	0.267	1

e le seguenti distribuzioni di frequenza condizionata, assoluta e relativa, di Y dato $X = x_r$

$Y X = S$	S	B	O	Totale
f	1	3	0	4
p	0.250	0.750	0	1


$Y X = B$	S	B	O	Totale
f	1	3	2	6
p	0.167	0.500	0.333	1

$Y X = O$	S	B	O	Totale
f	2	1	2	5
p	0.400	0.200	0.400	1

In modo analogo si ottengono le distribuzioni di frequenza condizionata, assoluta e relativa, di X dato $Y = y_s$.

Indipendenza statistica

Si nota che la distribuzione condizionata di Y varia al variare di X , e viceversa. Ad esempio, se l'attitudine musicale è sufficiente, l'attitudine pittorica non può essere ottima.

Quindi si può concludere che esiste una qualche forma di dipendenza fra le due variabili. 

Si parla di **indipendenza statistica** quando tutte le distribuzioni condizionate di Y dato $X = x_r$, $r = 1, \dots, m$, sono uguali, e quindi uguali alla distribuzione marginale di Y .

Analoghe considerazioni si possono fare per le distribuzioni condizionate di X dato $Y = y_s$, $s = 1, \dots, k$, e per la distribuzione marginale di X .

In tal caso, il valore assunto da una variabile non influenza il valore assunto dall'altra.

Dall'uguaglianza delle distribuzioni di frequenza relativa condizionata di X dato $Y = y_s$, $s = 1, \dots, k$ alla distribuzione marginale di X (o viceversa), si ha che

$$\frac{n_{rs}}{n_{+s}} = \frac{n_{r+}}{n}, \quad r = 1, \dots, m, \quad s = 1, \dots, k,$$

ovvero

$$n_{rs} = \frac{n_{r+}n_{+s}}{n}, \quad r = 1, \dots, m, \quad s = 1, \dots, k,$$

che corrisponde alla definizione di variabili X e Y **statisticamente indipendenti**.

Dividendo per n si ottiene la seguente specificazione alternativa basata sulle frequenze relative

$$\frac{n_{rs}}{n} = \frac{n_{r+}}{n} \frac{n_{+s}}{n}, \quad r = 1, \dots, m, \quad s = 1, \dots, k.$$

Indice di connessione

La *distanza* fra le frequenze osservate in una tabella di contingenza e le frequenze attese nel caso di indipendenza è misurata dall'**indice di connessione** χ^2

$$\chi^2 = \sum_{r=1}^m \sum_{s=1}^k \frac{(n_{rs} - n_{rs}^*)^2}{n_{rs}^*},$$

dove $n_{rs}^* = (n_{r+}n_{+s})/n$ è la frequenza attesa nel caso di indipendenza.

L'indice χ^2 vale 0 quando tutte le frequenze osservate coincidono con quelle attese, e quindi vi è **indipendenza fra le due variabili**.

Viceversa, tanto maggiori sono i valori osservati di χ^2 , tanto più le due **variabili** saranno **connesse (statisticamente dipendenti)**. Il valore massimo dell'indice è $n \min(m-1, k-1)$. I valori m e k indicano, rispettivamente, il numero di righe e di colonne della tabella di contingenza.

Per valutare la forza dell'eventuale dipendenza tra X e Y si può determinare l'**indice χ^2 normalizzato**, che si ottiene dividendo l'indice assoluto per il suo massimo.

Quindi si ottiene una quantità che assume valori nell'intervallo $[0, 1]$. Se vale 0 le variabili sono statisticamente indipendenti, se vale 1 si ha dipendenza statistica piena tra X e Y .

Esempio. *Attitudine* (continua). Con riferimento alla analisi delle attitudini musicale e pittorica, se ci fosse indipendenza tra le due, ferme restando le distribuzioni marginali, si avrebbe la seguente tabella di contingenza

	S	B	O	
S	1.067	1.867	1.066	4
B	1.600	2.800	1.600	6
O	1.333	2.333	1.334	5
	4	7	4	15

Si può calcolare facilmente l'indice χ^2 che è pari a 3.527. Quindi, si può dunque escludere l'indipendenza tra attitudine musicale e pittorica.

Dal momento che tale valore è lontano dal valore massimo $15 \cdot (3 - 1) = 30$, ed inoltre l'indice normalizzato vale 0.117, si conclude che i dati indicano una moderata connessione (dipendenza) tra le due variabili. \diamond

Esempio. *Casco.* La seguente tabella di contingenza illustra i risultati di uno studio sull'efficacia dei caschi protettivi per ciclisti nella prevenzione dei traumi cranici. Si considerano $n = 793$ soggetti coinvolti in incidenti.

Trauma cranico	Casco		Totale
	si	no	
si	17	218	235
no	130	428	558
Totale	147	646	793

Se ci fosse indipendenza tra uso del casco e trauma cranico, ferme restando le distribuzioni marginali, si avrebbe la seguente tabella di contingenza

Trauma cranico	Casco		Totale
	si	no	
si	43.56	191.44	235
no	103.44	454.56	558
Totale	147	646	793

Confrontando le due tabella si calcola facilmente l'indice di connessione χ^2 che vale 28.26.

Dal momento che il suo valore massimo è $793 \cdot (2 - 1) = 793$ e l'indice normalizzato è $28.26/793 = 0.036$, si conclude che esiste una lieve relazione tra le due variabili in esame.



Sommario

- 1 Sommario e introduzione
- 2 Analisi di dipendenza: la connessione
- 3 Dipendenza in media**
- 4 Analisi di correlazione
- 5 Analisi di regressione

Dipendenza in media

Le variabili vengono analizzate in modo **asimmetrico** perché si studia la dipendenza in media della **variabile quantitativa** Y dai livelli della **variabile qualitativa** X .

L'indipendenza in media è una forma debole di indipendenza nella quale non si considera la distribuzione di frequenza di Y ma solo la sua media.

Due variabili Y ed X si diranno **indipendenti in media** se la media condizionata di Y dato X è la stessa per ogni valore assunto da X , ovvero se

$$E(Y|X = x_r) = E(Y),$$

per ogni possibile x_r , $r = 1, \dots, m$.

Viceversa, se le varie medie condizionate sono diverse, allora le due variabili si diranno **dipendenti in media**.

Se due variabili sono indipendenti allora sono anche indipendenti in media, mentre non è vero il viceversa.

È possibile estendere tale nozione di indipendenza considerando l'**indipendenza in distribuzione**. In questo caso, oltre al calcolo delle medie condizionate, si confrontano anche altri indici sintetici oppure i grafici (ad esempio, istogrammi o boxplot) ottenuti per i sottogruppi definiti le varie modalità di X .

Esempio. *Geyser Old Faithful* (continua). Si considerano i dati riferiti alle durate delle eruzioni del geyser Old Faithful e si indica con X il tipo di eruzione e con Y la durata della pausa.

Modalità x di X	Media condizionata di Y dato $X = x$
Corta	54.49
Lunga	79.99

Le medie condizionate sono molto diverse, quindi non si può parlare di indipendenza in media tra X e Y .

Se si confrontano gli istogrammi per la variabile statistica Y riferita ai due gruppi, si conclude che c'è, più in generale, dipendenza in distribuzione.



Sommario

- 1 Sommario e introduzione
- 2 Analisi di dipendenza: la connessione
- 3 Dipendenza in media
- 4 Analisi di correlazione**
- 5 Analisi di regressione

Covarianza

Si vuole misurare l'intensità del **legame lineare** tra due **variabili quantitative** e la direzione della relazione.

Una misura della dipendenza lineare fra due variabili quantitative X e Y , con media $E(X)$ e $E(Y)$, è data dalla **covarianza**

$$\begin{aligned} Cov(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - E(X))(y_i - E(Y)). \end{aligned}$$

In alternativa, si può calcolare utilizzando la **formula per il calcolo**

$$Cov(X, Y) = E(XY) - E(X)E(Y) = \frac{1}{n} \sum_{i=1}^n x_i y_i - E(X)E(Y).$$

Spesso si indica con σ_{XY} , che ne richiama il legame con la varianza che corrisponde a $V(X) = \sigma_X^2 = \sigma_{XX}$

Coefficiente di correlazione lineare

Vale la **diseguaglianza di Cauchy-Schwarz**:

$$-\sigma_X\sigma_Y \leq \sigma_{XY} \leq \sigma_X\sigma_Y.$$

Una misura normalizzata della dipendenza lineare è il **coefficiente di correlazione lineare** definito da

$$\rho_{XY} = \text{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

Dalla diseguaglianza di Cauchy-Schwarz si ha che $-1 \leq \rho_{XY} \leq 1$.

Se $\rho_{XY} > 0$ c'è **relazione lineare crescente** fra X e Y ; nel caso in cui $\rho_{XY} = 1$ i punti (x_i, y_i) sono allineati su una retta di pendenza positiva.

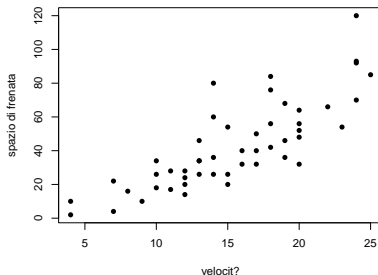
Se $\rho_{XY} < 0$ c'è **relazione lineare decrescente** fra X e Y ; nel caso in cui $\rho_{XY} = -1$ i punti (x_i, y_i) sono allineati su una retta di pendenza negativa.

Il valore assoluto $|\rho_{XY}|$ indica la *forza* del legame lineare.

Se $\rho_{XY} = 0$, c'è assenza di legame lineare tra X e Y , che sono dette **incorrelate** (ma non necessariamente indipendenti).

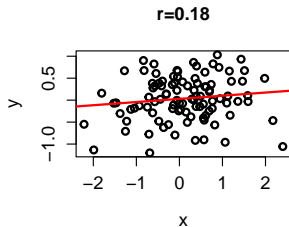
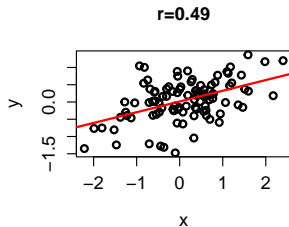
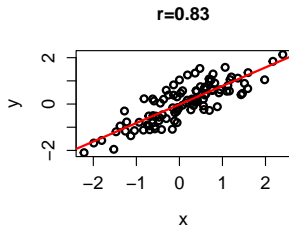
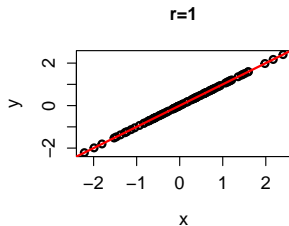
L'incorrelazione è una forma di indipendenza più debole dell'indipendenza statistica: la seconda implica la prima, ma non vale necessariamente il viceversa.

Esempio. *Velocità* (continua). Si considerano i dati sulla velocità X e sullo spazio di frenata Y di automobili degli anni 20.

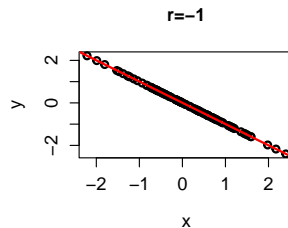
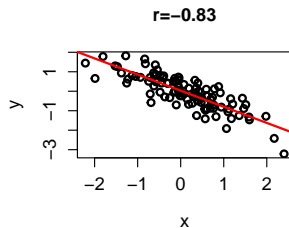
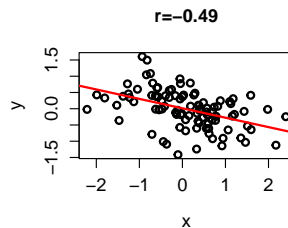
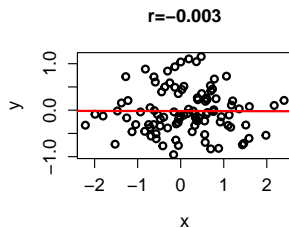


Si ha $\rho_{XY} = 0.81$, che indica un significativo legame lineare positivo. ◇

Si considerano alcuni esempi di correlazione positiva

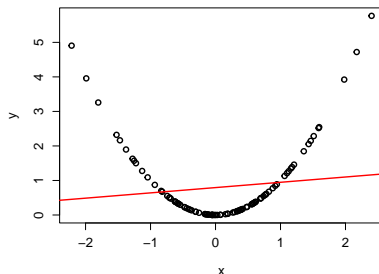


e di correlazione negativa



Nel grafico seguente si rappresentano punti (x_i, y_i) , $i = 1, \dots, n$, tali che $y_i = x_i^2$.

Tra X e Y c'è un *legame quadratico* perfetto, che il coefficiente di correlazione lineare, che vale 0.12, non misura.



Il coefficiente di correlazione lineare è influenzato dalla presenza di valori anomali.

Esempio. *Molla* (continua). Si considerano i dati sulla lunghezza della molla Y e sugli $n = 20$ diversi pesi X a cui viene sottoposta.

Dai dati riportati nella tabella presentata in precedenza si ha che

$$E(X) = 1.9, E(Y) = 5.388, V(X) = 1.33,$$

$$V(Y) = 0.059, Cov(X, Y) = 0.272.$$

Da cui si ottiene che

$$\rho_{XY} = 0.272 / \sqrt{1.33 \cdot 0.059} = 0.97,$$

valore che indica una correlazione positiva molto forte tra X e Y .

Se al posto dell'osservazione $(x_{19}, y_{19}) = (3.6, 5.68)$, si avesse il valore anomalo (*outlier*) $(x_{19}, y_{19}) = (3.6, 5.01)$, il coefficiente di correlazione lineare, che risente della presenza di valori anomali, risulterebbe pari a 0.76.



Anche per **variabili qualitative ordinali** X e Y è possibile definire un indice che misura l'intensità (come l'indice χ^2) e il verso dell'associazione.

Dati i valori osservati (x_i, y_i) , $i = 1, \dots, n$, si considerano i **ranghi** (posizione dell'unità statistica dopo aver ordinato i valori in senso crescente), calcolati separatamente per ciascuna delle due variabili.

Si definisce **indice di correlazione tra i ranghi di Spearman** ρ_{XY}^S , il coefficiente di correlazione lineare calcolato sui ranghi invece che sulle osservazioni (e ci sono osservazioni uguali, si considera come rango il valore medio delle loro posizioni).

Poiché si ha un indice di correlazione, $-1 \leq \rho_{XY}^S \leq 1$ e

- $\rho_{XY}^S = 1$ ($\rho_{XY}^S = -1$) indica una perfetta concordanza (discordanza) tra i ranghi di X e di Y ;
- $\rho_{XY}^S = 0$ indica che i ranghi di X e di Y non mostrano alcuna associazione.

Il coefficiente ρ_{XY}^S si può utilizzare anche per **variabili quantitative**, come alternativa *robusta* a ρ_{XY} . Un'ulteriore alternativa è rappresentata dall'**indice di correlazione di Kendall**.

Sommario

- 1 Sommario e introduzione
- 2 Analisi di dipendenza: la connessione
- 3 Dipendenza in media
- 4 Analisi di correlazione
- 5 Analisi di regressione**

Regressione lineare semplice

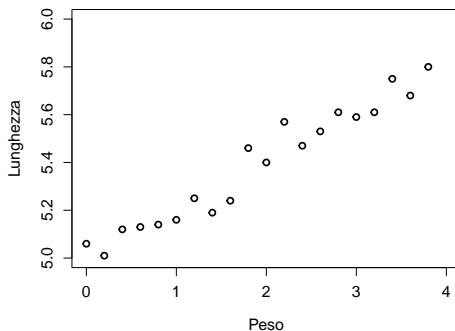
Si analizzano congiuntamente di due o più **variabili quantitative**. È una generalizzazione dell'analisi di dipendenza in media.

In generale, con l'analisi di regressione si studia la media condizionata di una **variabile risposta** Y in funzione di una (*regressione semplice*) o più (*regressione multipla*) **variabili esplicative** X_1, \dots, X_p , $p \geq 1$.

Si considera la **regressione lineare semplice**, dove tra la variabile risposta Y e l'unica variabile esplicativa X si ipotizza una relazione lineare.

Esempio. *Molla* (continua). Si considerano i dati sulla lunghezza della molla Y e sugli $n = 20$ diversi pesi X a cui viene sottoposta.

Si vuole studiare la relazione tra X e Y e, più precisamente, verificare se X spiega Y .



In particolare, per descrivere il comportamento in media di Y in funzione di X si può considerare l'equazione della retta

$$y_i = a + b x_i + \text{errore}, \quad i = 1, \dots, 20,$$

dove a indica la lunghezza attesa della molla nel caso in cui il peso sia nullo e b determina il verso e l'intensità della relazione lineare tra X e Y .

Il termine di errore evidenzia il fatto che la relazione lineare non si adatta perfettamente ai dati (x_i, y_i) , $i = 1, \dots, 20$.

L'errore racchiude ciò che la retta, e quindi X , non spiega del fenomeno Y e l'eventuale errore di misura associato a Y . \diamond

Il **modello di regressione lineare semplice (modello lineare)** è definito dall'equazione

$$y_i = a + b x_i + \epsilon_i, \quad i = 1, \dots, n,$$

dove (x_i, y_i) , $i = 1, \dots, n$, sono i valori osservati per la **variabile dipendente** Y e per la **variabile esplicativa** X .

I valori ϵ_i , $i = 1, \dots, n$, specificano gli **errori**, mentre a e b sono i **coefficienti di regressione**, con a l'intercetta e b il coefficiente angolare della **retta di regressione** $y = a + bx$.

L'interesse è rivolto al comportamento complessivo e non a ciò che avviene per le singole coppie di osservazioni.

Il modello si intende **lineare nei parametri**, non nella variabile esplicativa. Quindi,

$$y_i = a + b \log(x_i) + \epsilon_i,$$

$$y_i^2 = a + b \exp(x_i) + \epsilon_i,$$

$$y_i = a + b^2 x_i + \epsilon_i,$$

sono modelli lineari per Y (o per Y^2) nelle variabili esplicative $\log(X)$, $\exp(X)$ e X , rispettivamente (una volta effettuata la riparametrizzazione $c = b^2$).

Invece,

$$y_i = a + a^2 x_i + \epsilon_i$$

non è un modello lineare in quanto il parametro a compare anche elevato al quadrato.

Metodo dei minimi quadrati

I *coefficienti di regressione* non sono noti; sono parametri *da stimare sulla base dei dati osservati*, di modo che la retta di regressione si adatti bene alle osservazioni.

Avendo osservato n coppie di valori (y_i, x_i) , $i = 1, \dots, n$, si hanno n valori osservati anche per l'errore di regressione

$$\epsilon_i = y_i - (a + bx_i), \quad i = 1, \dots, n.$$

I valori ϵ_i , $i = 1, \dots, n$, detti **residui di regressione**, rappresentano gli scostamenti fra le osservazioni e il modello teorico.

Per stimare i coefficienti di regressione può essere ragionevole cercare i valori per a e b che minimizzano (non è l'unica possibilità) la **somma dei quadrati dei residui**

$$Q(a, b) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - a - bx_i)^2.$$

Il metodo presentato è detto **metodo dei minimi quadrati** e le stime ottenute, indicate con \hat{a} e \hat{b} , sono le **stime dei minimi quadrati** che corrispondono a

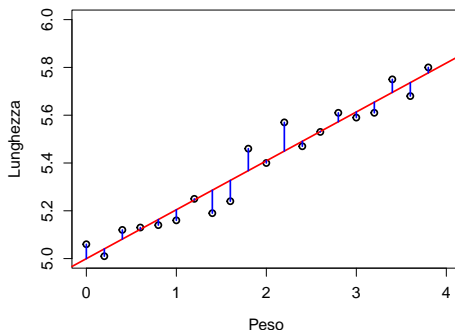
$$\hat{a} = E(Y) - \hat{b} E(X), \quad \hat{b} = \text{Cov}(X, Y)/V(X).$$

La retta $y = \hat{a} + \hat{b}x$ è detta **retta di regressione stimata (retta dei minimi quadrati)**.

Esempio. *Molla* (continua). Con riferimento ai dati sulla lunghezza della molla e sui diversi pesi a cui viene sottoposta, si ottengono le seguenti stime per i coefficienti di regressione

$$\hat{b} = 0.27215/1.33 = 0.2046, \quad \hat{a} = 5.3885 - 0.2046 \cdot 1.9 = 4.9997.$$

Nel grafico seguente si riporta, oltre ai dati osservati, la **retta di regressione stimata** $y = 4.9997 + 0.2046x$ e i **residui stimati** ϵ_i , $i = 1, \dots, 20$.

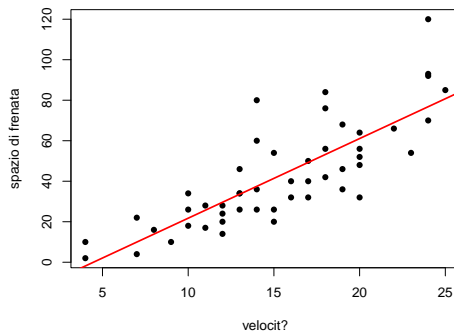


Dall'analisi della retta di regressione stimata si conclude che:

- con un peso di 2.5 Kg, si potrebbe prevedere un allungamento della molla pari a $4.9997 - 0.2046 \cdot 2.5 = 5.51$;
- con un peso di 7.5 Kg, si potrebbe prevedere (facendo un'estrapolazione) un allungamento della molla pari a $4.9997 - 0.2046 \cdot 7.5 = 6.53$ (che forse non è un valore realistico).



Esempio. *Velocità* (continua). Con riferimento ai dati su velocità e spazio di frenata, si ottengono le seguenti stime per i coefficienti di regressione $\hat{a} = -17.579$, $\hat{b} = 3.932$.



Nel grafico si riportano i dati osservati e la **retta di regressione stimata**
 $y = -17.579 + 3.932 x$.



Tra l'**analisi di regressione** e l'**analisi di correlazione** ci sono differenze e punti di contatto.

- Nella correlazione, c'è **simmetria** tra le due variabili, mentre nella regressione c'è **asimmetria**: si suppone di fissare i valori x_i per vedere come variano i valori y_i , $i = 1, \dots, n$.
- Nella regressione si opera come se i valori x_i , $i = 1, \dots, n$, fossero stati fissati a priori e ottenuti senza errore (anche se nella pratica a volte i valori di X e Y sono generati simultaneamente).
- Il coefficiente angolare stimato della retta di regressione \hat{b} è direttamente proporzionale al coefficiente di correlazione lineare ρ_{XY} . Infatti,

$$\hat{b} = \frac{\text{Cov}(X, Y)}{V(X)} = \rho_{XY} \frac{\sigma_Y}{\sigma_X}.$$

Valori stimati dal modello e residui stimati

Una volta calcolate le stime \hat{a} e \hat{b} per i coefficienti di regressione, si possono determinare i **valori stimati dal modello**

$$\hat{y}_i = \hat{a} + \hat{b}x_i, \quad i = 1, \dots, n,$$

cioè i valori della variabile risposta Y per ogni valore osservato x_i .

Nel caso in cui si considerino valori per X che non corrispondono ai valori osservati, si ottengono i **valori previsti dal modello**, che sono utili per fare previsioni o ricostruzione di valori mancanti per Y .

Occorre fare molta attenzione quando si estrapola la retta di regressione stimata, cioè quando si fanno previsioni al di fuori dell'intervallo dei valori osservati per la variabile esplicativa X .

Infine si possono calcolare i **residui stimati**

$$\hat{\epsilon}_i = y_i - \hat{a} - \hat{b}x_i = y_i - \hat{y}_i, \quad i = 1, \dots, n,$$

cioè la stima degli errori (residui) basata sulle osservazioni.

Coefficiente di determinazione

Il modello lineare è utile solo nel caso di relazioni sostanzialmente lineari tra Y e X .

Con l'obiettivo di **valutare la bontà del modello** di regressione, si vuole individuare un indice in grado di valutare l'adattamento globale del modello ai dati, oltre che la sua capacità esplicativa per il fenomeno Y .

La varianza $V(Y)$ associata alla variabile statistica Y (**varianza totale**) può essere vista come somma della quota $V(\hat{Y})$ descritta dal modello (**varianza spiegata**) e della quota $V(\hat{\epsilon})$ rimanente (**varianza residua**)

$$V(Y) = V(\hat{Y}) + V(\hat{\epsilon}),$$

dove \hat{Y} e $\hat{\epsilon}$ sono, rispettivamente, i valori stimati dal modello e i residui stimati.

Tanto maggiore è la varianza spiegata dal modello, tanto migliore sarà l'adattamento dei dati al modello teorico.

Un indice di bontà di adattamento del modello lineare è dato dal **coefficiente di determinazione** R^2 , che corrisponde alla proporzione di varianza di Y spiegata dal modello di regressione

$$\begin{aligned} R^2 &= \frac{\sum_i (\hat{y}_i - E(\hat{Y}))^2 / n}{\sum_i (y_i - E(Y))^2 / n} = \frac{\text{varianza spiegata}}{\text{varianza totale}} \\ &= 1 - \frac{\sum_i (\hat{\epsilon}_i - E(\hat{\epsilon}))^2 / n}{\sum_i (y_i - E(Y))^2 / n} = 1 - \frac{\text{varianza residua}}{\text{varianza totale}}, \end{aligned}$$

dove $E(\hat{Y}) = E(Y)$ è la media dei valori stimati dal modello e $E(\hat{\epsilon}) = 0$ è la media dei residui stimati.

Vale che $0 \leq R^2 \leq 1$ e un valore per R^2 vicino a 1 (vicino a 0) indica un buon (pessimo) adattamento del modello ai dati.

Si dimostra inoltre che vale la seguente relazione con l'indice di correlazione lineare

$$R^2 = \rho_{XY}^2.$$

Nel caso di modelli con $a = 0$, l'indice R^2 va modificato opportunamente.

Esempio. *Molla* (continua). Con riferimento ai dati sulla lunghezza della molla e sui pesi, dal momento che $\rho_{XY} = 0.97$, si conclude che $R^2 = 0.97^2 = 0.9409$. Quindi il modello di regressione presenta una elevata capacità esplicativa per il fenomeno in esame. \diamond

Un adattamento poco soddisfacente della retta di regressione ai dati può essere migliorato ricorrendo ad un **cambiamento di scala** della variabile risposta e/o della variabile esplicativa.

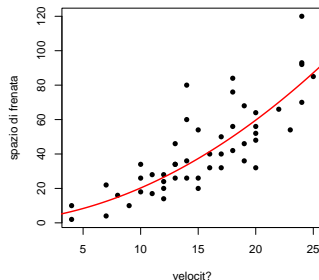
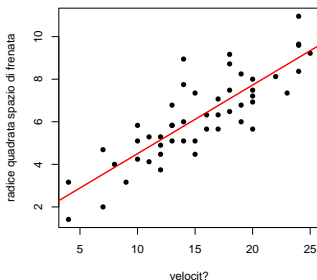
Le trasformazioni più comuni sono $\log(Y)$, \sqrt{Y} e $1/Y$.

In alcuni casi, un cattivo adattamento può essere dovuto alla presenza di **valori anomali**, riconducibili a errori di misurazione o ad unità con caratteristiche particolari.

Dopo aver individuato (tipicamente per via grafica) le osservazioni potenzialmente anomale si prova a ripetere l'analisi di regressione senza tali valori.

Infine, un adattamento non ottimale può essere spiegato dal fatto che la componente d'errore del modello risulta essere elevata.

Esempio. *Velocità* (continua). Con riferimento ai dati su velocità e spazio di frenata, si ottiene $R^2 = 0.651$. È possibile migliorare l'adattamento della retta di regressione considerando \sqrt{Y} come risposta.



Nel grafico di sinistra si riportano i dati osservati $(\sqrt{y_i}, x_i)$, $i = 1, \dots, n$, e la **retta di regressione stimata** $\sqrt{y} = 1.277 + 0.322x$. L'adattamento sembra migliorato, infatti $R^2 = 0.709$.

Nel grafico di destra si ha la rappresentazione sulla scala originaria (la retta di regressione diventa una parabola).

