

Regressione

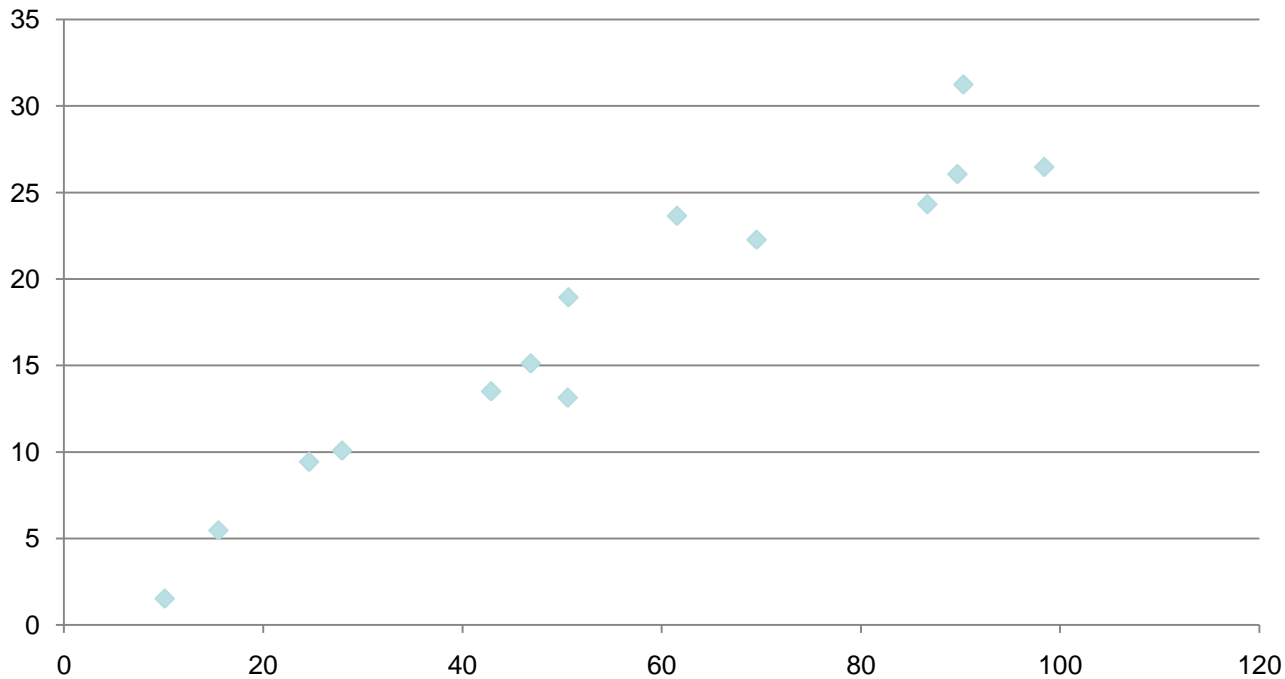
Esempio

Un'azienda manifatturiera vuole analizzare il legame che intercorre tra il **volume produttivo** X per uno dei propri stabilimenti e il corrispondente **costo mensile** Y di produzione.

Volume X (ton.)	Costo Y (K€)	Volume X (ton.)	Costo Y (K€)
10.11	1.53	42.87	13.51
50.56	13.14	61.53	23.65
90.28	31.24	24.60	9.43
15.50	5.47	46.85	15.12
69.52	22.27	50.63	18.94
98.40	26.47	89.68	26.06
86.66	24.32	27.91	10.08

Esempio

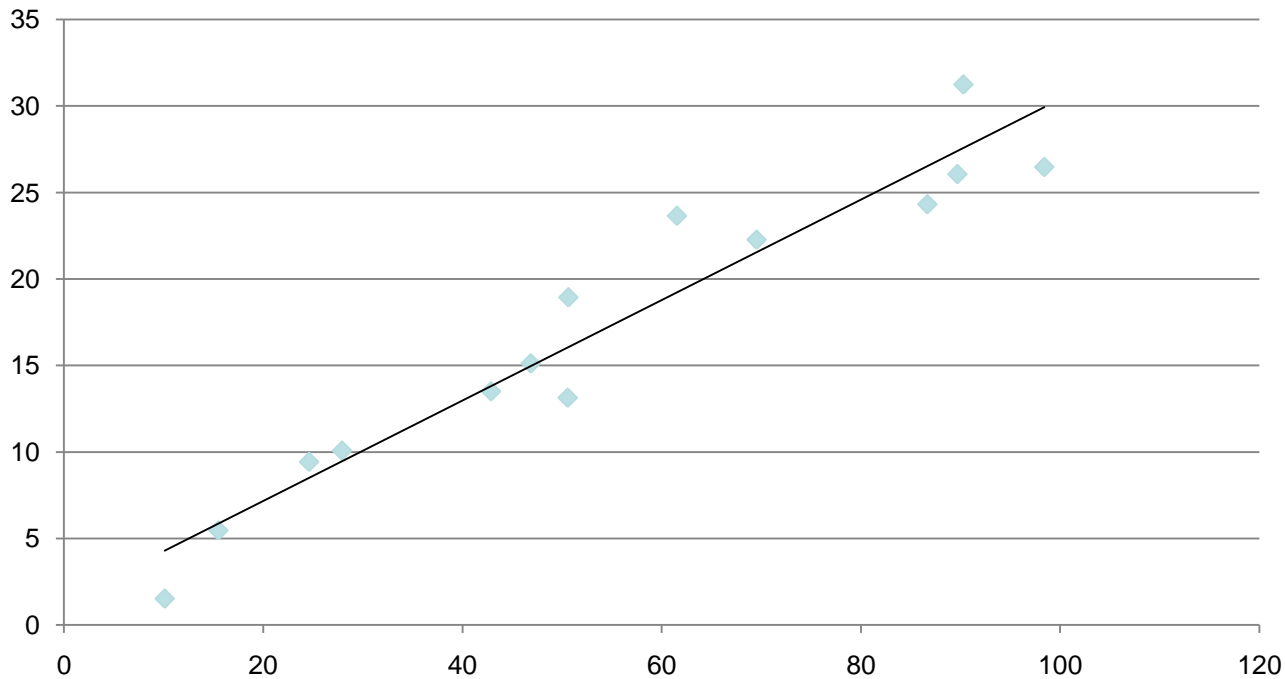
Volume - Costo



- È possibile trovare un legame semplice e tendenziale tra la variabile dipendente Y (costo) e la variabile indipendente X (volume)?

Esempio

Volume - Costo



- Tale funzione permette di:
 - Evidenziare il legame della variabile dipendente da quella indipendente
 - Predire il valore futuro dell'attributo target.

Modelli di stima

- Si ipotizza l'esistenza di una funzione $f: R^n: \rightarrow R$ che esprime il legame tra la variabile dipendente Y e le n variabili esplicative X_j

$$Y = f(X_1, X_2, \dots, X_n).$$

- Tale funzione va cercata in una classe di funzioni che assicuri:
 - l'identificazione accurata tra target e variabili indipendenti, per garantire errori di modesta entità per le osservazioni disponibili
 - Una buona capacità di predire i valori futuri (generalizzazione)

Modelli di stima

- Si ipotizza l'esistenza di una funzione $f: R^n \rightarrow R$ che esprime il legame tra la variabile dipendente Y e le n variabili esplicative X_j

$$Y = f(X_1, X_2, \dots, X_n).$$

- La funzione f può essere:

- Lineare: $Y = b + \omega X$

- Quadratica: $Y = b + \omega X + d X^2$

- Posto $Z = X^2$, il modello è $Y = b + \omega X + d Z$

- Esponenziale: $Y = e^{b + \omega X}$

- Posto $Z = \log Y$, il modello è $Z = b + \omega X$

Modello probabilistico

- E' improbabile che le coppie (X, Y) si dispongano esattamente lungo una retta del piano.
- E' più realistico supporre un legame di natura approssimata tra X e Y , espresso dal modello

$$Y = \omega X + b + \varepsilon$$

con ε variabile casuale detta *scarto* o *errore*, che deve soddisfare alcune ipotesi di natura stocastica.

- Deve essere assimilabile ad un errore casuale, oppure assimilabile all'effetto su Y di variabili non considerate
- Supponiamo ε variabile aleatoria con distribuzione normale di media 0 e deviazione standard σ

Calcolo della retta di regressione

- L'identificazione della retta di regressione si riduce all'identificazione del coefficiente angolare ω e dell'intercetta b . della retta $Y = \omega X + b + \varepsilon$
- Minimizzazione della funzione SSE (*sum of squared errors*):

$$SSE = \sum_{i=1}^m e_i^2 = \sum_{i=1}^m [y_i - f(x_i)]^2 = \sum_{i=1}^m [y_i - \omega x_i - b]^2$$

➤ In Excel: REGR.{LIN,LOG}

Esempio regressione

Cartel1 - Microsoft Excel uso non commerciale

Home Inserisci Layout di pagina Formule Dati Revisione Visualizza Componenti aggiuntivi

Tabella pivot Tabella Tabelle Immagine ClipArt Forme SmartArt Illustrazioni Istogramma Grafico a linee Grafico a torta Grafico a barre Grafico ad area Grafici Grafico a dispersione Altri grafici Collegamento ipertestuale Casella di testo Intestazione e piè di pagina WordArt Riga della firma Oggetto Simbolo Testo

A1 fx Tempo

	A	B	C	D	E	F	G	H	L	M	N	O	P	Q
1	Tempo	Quantità												
2	1	2,698703												
3	2	3,927361												
4	3	3,509222												
5	4	4,018662												
6	5	5,263152												
7	6	5,474225												
8	7	6,261665												
9	8	6,965404												
10	9	6,527422												
11	10	7,986237												
12	11	7,697039												
13	12	8,812096												
14	13	9,245222												
15	14	9,101311												
16	15	10,0392												
17	16	10,88013												
18	17	11,15419												
19	18	11,14093												
20	19	11,60103												
21	20	12,70734												
22														
23														
24														
25														

Dispersione

Tutti i tipi di grafico...

Esempio regressione

Cartel1 - Microsoft Excel uso non commerciale

Strumenti grafico

Layout

Linea di tendenza

Nome grafico: Grafico 2

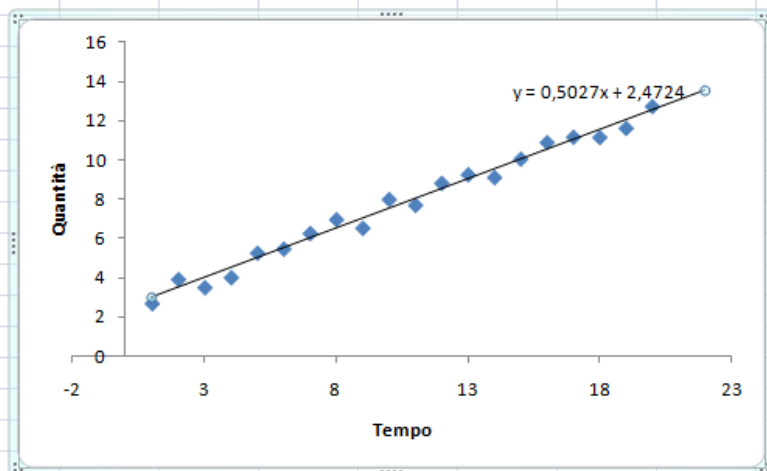
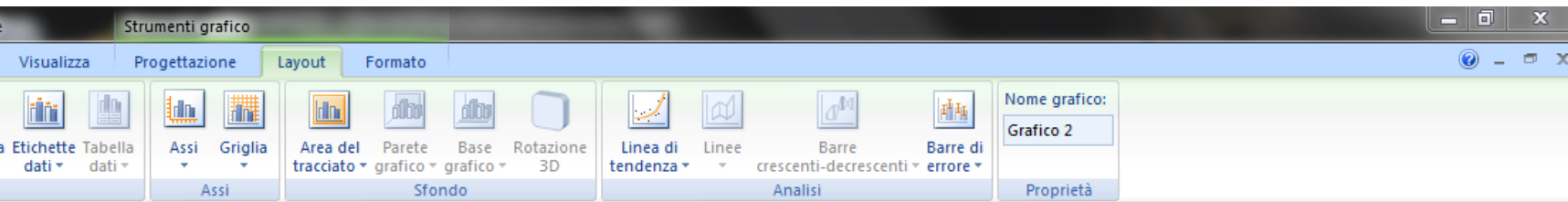
- Nessuna**
Rimuove la linea di tendenza selezionata o tutte le linee di tendenza se nessuna è selezionata
- Linea di tendenza lineare**
Aggiunge/imposta una linea di tendenza lineare per la serie del grafico selezionata
- Linea di tendenza esponenziale**
Aggiunge/imposta una linea di tendenza esponenziale per la serie del grafico selezionata
- Linea di tendenza lineare di previsione**
Aggiunge/imposta una linea di tendenza lineare con previsione su 2 periodi per la serie del grafico selezionata
- Media mobile su due periodi**
Aggiunge/imposta una linea di tendenza a media mobile su 2 periodi per la serie del grafico selezionata

Altre opzioni linea di tendenza...

The scatter plot displays a series of data points showing a clear upward trend. The x-axis is labeled 'Tempo' and ranges from -2 to 23. The y-axis is labeled 'Quantità' and ranges from 0 to 14. A blue regression line is fitted to the data points, indicating a positive linear relationship.

Tempo	Quantità
1	2.5
2	4.0
3	3.5
4	4.0
5	5.5
6	6.0
7	6.5
8	7.0
9	6.5
10	8.0
11	7.5
12	9.0
13	9.5
14	9.0
15	10.0
16	11.0
17	11.5
18	11.0
19	12.0
20	13.0

Esempio regressione



Formato linea di tendenza

Opzioni linea di tendenza

Colore linea

Stile linea

Ombreggiatura

Opzioni linea di tendenza

Tipo di tendenza/regressione

- Esponenziale
- Lineare
- Logaritmica
- Polinomiale Ordine: 2
- Potenza
- Media mobile Periodo: 2

Nome linea di tendenza

- Automatico: Lineare (Quantità)
- Personalizzato:

Previsione

Eutera: 2 periodi

Verifica: 0,1 periodi

Imposta intercetta = 0,0

Visualizza l'equazione sul grafico

Visualizza il valore R^2 al quadrato sul grafico

Chiudi

Assunzioni relative ai residui

- Minimizzando SSE, la variabile aleatoria ε deve seguire una distribuzione normale di media 0 e deviazione standard σ .
- Si richiede inoltre che i residui ε_i e ε_k , corrispondenti a due distinte osservazioni x_i e x_k siano indipendenti per ogni scelta di i e k .
- Un modello è tanto più accurato quanto più la deviazione σ risulta prossima a zero.

Esercizi

- Determinare un modello di regressione lineare per il dataset <http://statmaster.sdu.dk/courses/st111/data/data/tvads.txt>
- Cosa accade se si usa una scala logaritmica?
- Cosa si può dire per il dataset <http://statmaster.sdu.dk/courses/st111/data/data/velocity.txt>

Valutazione dei modelli di regressione

- Normalità e indipendenza dei residui
- Significatività dei coefficienti
- Coefficiente di correlazione lineare
- Multi-collinearità delle variabili indipendenti
- Limiti di confidenza e predizione
 - In Excel, `REGR.LIN()`

Normalità e indipendenza dei residui

- Diagramma di dispersione dei residui rispetto ai valori predetti.
 - Un andamento regolare dei residui indica l'esistenza di fattori esplicativi non considerati nel modello.
- Diagramma di dispersione della radice dei residui
 - I valori sono tutti positivi ed attenuati rispetto ai precedenti

Significatività dei coefficienti

- REGR.LIN ci fornisce:
 - $s_1; s_2; \dots; s_n$ I valori di errore standard per i coefficienti $m_1; m_2; \dots; m_n$ s_b
 - Il valore di errore standard per la costante b .
 - r^2 Il coefficiente di determinazione.
 - Confronta i valori y stimati con quelli effettivi e può avere un valore compreso tra 0 e 1.
 - $r^2=1$, correlazione perfetta nel campione
 - $r^2=0$ previsione di un valore y non corretta.
 - ...

Covarianza

- La *covarianza* quantifica la forza della relazione tra due insiemi di valori, ovvero misura quanto lineare è la dipendenza tra i due attributi;
- La covarianza è la media del prodotto delle deviazioni dei valori dalla media degli insiemi dei dati

$$v_{jk} = cov(a_j, a_k) = \frac{1}{m-2} \sum_{i=1}^m (x_{ij} - \bar{\mu}_j)(x_{ik} - \bar{\mu}_k)$$

➤ In Excel COVARIANZA()

- un valore positivo indica una variazione di X e Y nella stessa direzione, un valore negativo l'opposto

Correlazione

- Un limite della covarianza è la sua dipendenza dall'unità di misura.
- Per esempio possiamo aumentare il fattore covarianza di 1000, semplicemente usando come unità di misura € in luogo di K€
 - Nel caso le unità sono appropriate
- La misura di *correlazione* risolve il problema producendo un risultato indipendente dalle unità di misura e compreso tra -1 e 1

$$r_{jk} = \text{corr}(a_j, a_k) = \frac{v_{jk}}{\bar{\sigma}_j \bar{\sigma}_k}$$

Correlazione

- Un valore della correlazione vicino a -1 indica che i due insiemi di valori tendono a variare in senso opposto
- Un valore della correlazione vicino a $+1$ indica che i due insiemi di valori tendono a variare nello stesso senso
- Una indipendenza nelle variazioni dei due valori produce un indice di correlazione uguale a 0
- Ma, attenzione: l'indice di correlazione è rilevante solo per relazioni *lineari*
- L'indice può risultare vicino a 0 anche se esiste una relazione non lineare tra i due insiemi di valori.

Multi-collinearità

- Si parla di multi-collinearità quando sono presenti relazioni lineari tra le variabili indipendenti.
- Si parla di multi-collinearità esatta quando almeno una delle variabili esplicative è correlata con altre variabili indipendenti.
 - Esempio: la produzione settimanale è la somma delle produzioni giornaliere, e tutte le variabili sono incluse nel modello.
- In presenza di multi-collinearità esatta la matrice $(X^T X)$ è singolare e non ammetta inversa.
- La multi-collinearità esatta è piuttosto rara e tipicamente causata da un errore nella definizione del modello.

Limiti di confidenza e di predizione

- Conseguenze della multi-collinearità nelle variabili indipendenti:
 - difficoltà di determinare i contributi individuali delle variabili, perché i loro effetti vengono mescolati o confusi;
 - alta variabilità delle stime con conseguente bassa significatività dei coefficienti di regressione;
 - forte instabilità delle stime dei coefficienti di regressione (piccole variazioni nei dati o l'aggiunta/eliminazione di una variabile dal modello possono portare a grandi variazioni nella stima).
- La multicollinearità non invalida il modello, ma l'interpretazione dei singoli coefficienti di regressione.

Selezione delle variabili predittive

- Per identificare la multi-collinearità si possono calcolare i coefficienti di correlazione tra tutte le coppie di variabili esplicative.
 - Valori elevati ($> 0,90$) indicano la forte collinearità
 - Valori bassi, non assicurano l'assenza di multi-collinearità
 - Effetto congiunto di due o più variabili esplicative.
- Rimedio:
 - Eliminare una o più variabili indipendenti altamente correlate, senza eliminare variabili significative;

Passi per la costruzione del modello

1. Individuazione dei valori anomali
2. Scelta del modello
3. Individuazione dei parametri
4. Significatività dei coefficienti
5. Previsione per diversi valori della variabile indipendente

Esercizi

- Determinare i modelli di regressione per gli esercizi della lezione scorsa

Sommario

- La regressione lineare semplice e multipla permette di determinare semplici modelli.
- È possibile valutare la bontà di tali modelli valutando la normalità, l'indipendenza dei residui e la significatività dei coefficienti
- Tramite il coefficiente di correlazione è possibile stabilire se ci sono dipendenze lineari tra le variabili indipendenti.
- Le conseguenze della multi-linearità vanno affrontate alla luce delle soluzioni esistenti.