

Analisi Dati con Excel

Giovanni Della Lunga

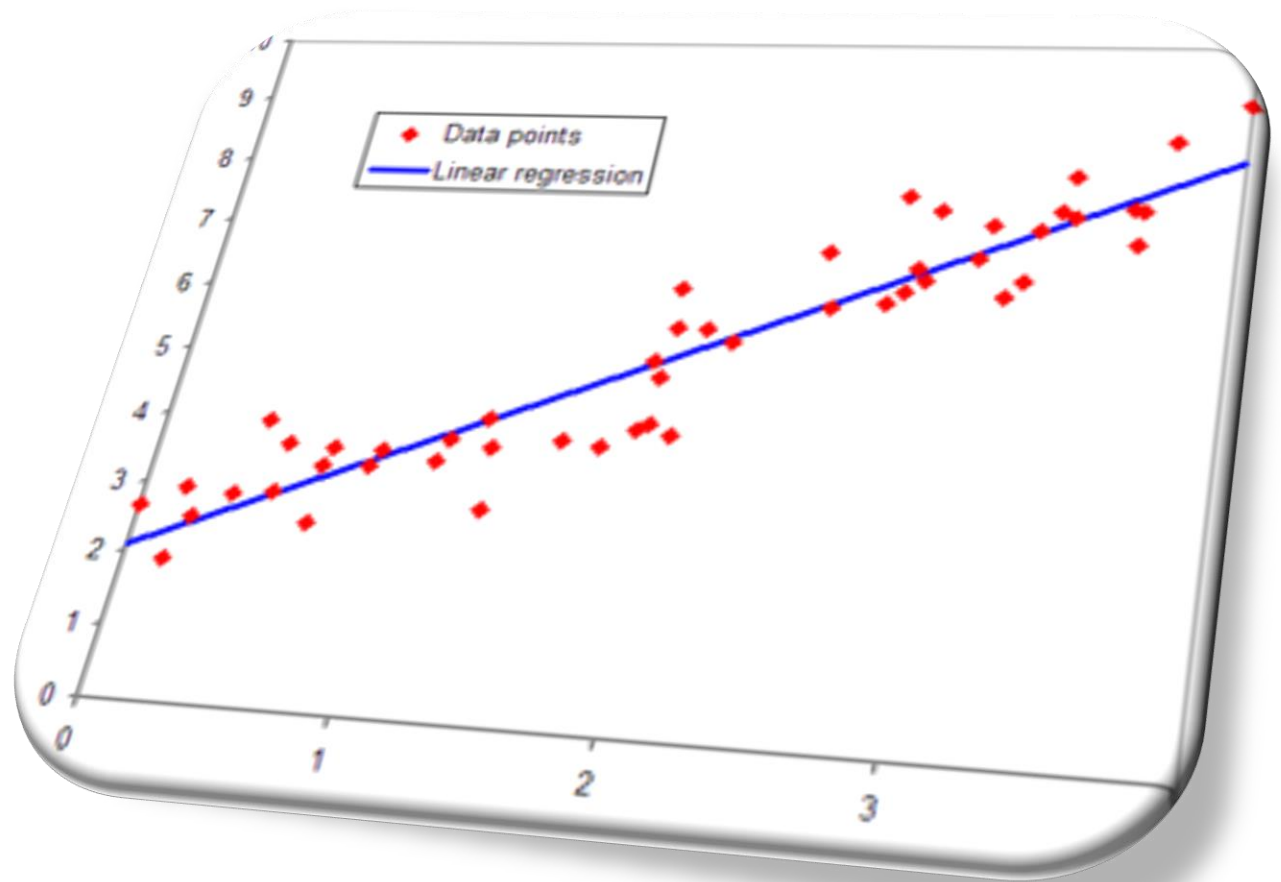
giovanni.dellalunga@gmail.com

La prima regola di ogni tecnologia è che l'automazione applicata ad un'operazione efficiente ne aumenterà l'efficienza. La seconda è che l'automazione applicata ad un'operazione inefficiente ne aumenterà l'inefficienza.

Bill Gates

Analisi di Regressione

Making Predictions



| Analisi di regressione: alcune nozioni di base

- In statistica, l'**analisi di regressione** viene utilizzata per effettuare una stima tra le relazioni tra due o più variabili.
- Possiamo fare subito una distinzione tra le variabili.
- La **variabile dipendente** (o variabile y) è la variabile risposta ovvero il fattore principale che si sta tentando di comprendere e prevedere.
- Le **variabili indipendenti** (o variabili x) sono le variabili esplicative ovvero i fattori che potrebbero influenzare la variabile dipendente.

| Analisi di regressione: alcune nozioni di base

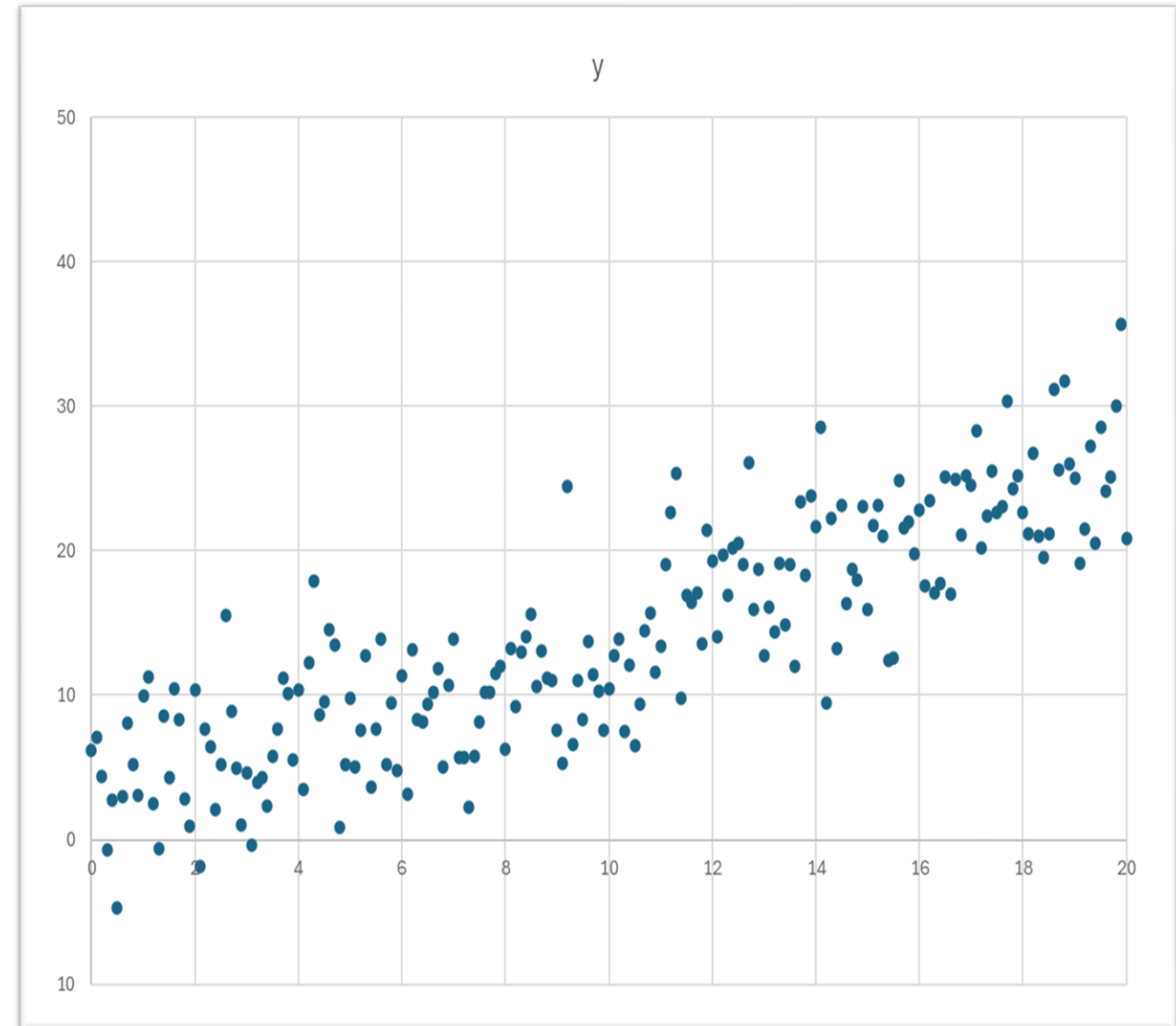
- Attraverso l'analisi di regressione possiamo capire come si comporta la variabile dipendente (y) quando varia una delle variabili indipendenti (x). Questo ci consente di determinare statisticamente quale delle variabili ha un impatto rilevante.
- Inoltre, possiamo fare una ulteriore distinzione. Possiamo distinguere tra **regressione lineare semplice** e **regressione lineare multipla**.
- La **regressione lineare semplice** consente di individuare la relazione tra una variabile dipendente e una variabile indipendente attraverso l'utilizzo di una funzione lineare.
- La **regressione lineare multipla** consente di prevedere la variabile dipendente quando si utilizzano **due o più variabili esplicative**.

| Analisi di regressione: alcune nozioni di base

In entrambi i casi, se la relazione tra i dati NON segue una linea retta, è necessario utilizzare una **regressione non lineare**!

Un Esempio di Regressione Lineare

- » Cerchiamo di capire subito il tipo di relazione tra la variabile indipendente (x) e la variabile dipendente (y).
- » Per far questo inseriamo un diagramma a dispersione.
- » A colpo d'occhio otteniamo immediatamente le informazioni relative al tipo di relazione.
- » Il grafico ci informa che il tipo di relazione è di tipo lineare. La retta è crescente. Possiamo pertanto dedurre che ad un aumento della variabile indipendente x possa corrispondere un aumento della variabile dipendente y .



Regressione lineare: l'equazione

L'equazione che definisce la regressione lineare è la seguente:

$$Y_i = \alpha_0 + \alpha_1 \cdot X_i + \epsilon$$

Dove

- i indica la generica osservazione
- Y è la variabile dipendente
- X la variabile dipendente
- α_0 è l'intercetta della retta di regressione. Rappresenta il valore di Y quando la variabile X è uguale a 0. All'interno di un grafico di regressione, è il punto in cui la retta interseca l'asse Y .
- α_1 è il coefficiente angolare della retta di regressione ovvero il tasso di variazione di Y quando X cambia. Rappresenta la pendenza della retta.
- ed infine ϵ è l'errore statistico.

Regressione lineare in Excel: Primo Metodo

- Il primo metodo per eseguire la regressione in Excel utilizza il **componente aggiuntivo** chiamato **Strumenti di analisi**.
- Questo strumento è incluso in Excel ed è **necessario attivarlo**. È disponibile in tutte le versioni di Excel (dalla versione 2003 alla versione 2019) ma, **per impostazione predefinita, non è abilitato**. Quindi, è necessario procedere alla sua attivazione. Vediamo come.
- Fate click su **File** e successivamente su **Opzioni**. Nella finestra di dialogo **Opzioni** di Excel, selezionate la voce **Componenti aggiuntivi...**

Regressione Lineare in Excel: Primo Metodo

Opzioni di Excel

Generale

Formule

Dati

Strumenti di correzione

Salvataggio

Lingua

Accessibilità

Impostazioni avanzate

Personalizzazione barra multifunzione

Barra di accesso rapido

Componenti aggiuntivi

Centro protezione

Opzioni generali per l'utilizzo di Excel.

Opzioni interfaccia utente

Quando si usano più schermi: ⓘ

- ☒ Ottimizza per l'aspetto migliore
- ☐ Ottimizza per la compatibilità (riavvio dell'applicazione necessario)

☒ Mostra barra di formattazione rapida quando si seleziona testo ⓘ

☒ Mostra opzioni di Analisi rapida per la selezione

☒ Attiva antep^rima dinamica ⓘ

☐ Comprimi automaticamente la barra multifun^zione ⓘ

Stile descrizione comando: Mostra descrizione caratteristica nelle descrizioni comandi

Dimensione: 11

Visualizzazione predefinita per i nuovi fogli: Visualizzazione Normale

Numero di fogli da includere: 1

Personalizzazione della copia di Microsoft Office in uso

Nome urente: Damiano Causale | Excel per tutti

☐ Usa sempre questi valori indipendentemente dall'accesso a Office

Sfondo Office: Cerchi e strisce

Tema di Office: A colori

Impostazioni di privacy

Impostazioni di privacy...

OK Annulla

Selezionare Componenti aggiuntivi

Regressione Lineare in Excel: Primo Metodo

Opzioni di Excel

Generale

Formule

Dati

Strumenti di correzione

Salvataggio

Lingua

Accessibilità

Impostazioni avanzate

Personalizzazione barra multifunzione

Barra di accesso rapido

Componenti aggiuntivi

Centro protezione



Consente di visualizzare e gestire i componenti aggiuntivi di Microsoft Office.

Componenti aggiuntivi

Nome ^	Posizione	Tipo
Componenti aggiuntivi attivi dell'applicazione		
Dropbox for Office	C:\...ceAddin.13.dll	Componente aggiuntivo COM
Microsoft Power Map for Excel	C:\...INSHELL.DLL	Componente aggiuntivo COM
Microsoft Power Pivot for Excel	C:\...lientAddIn.dll	Componente aggiuntivo COM
Microsoft Power View for Excel	C:\...ExcelClient.dll	Componente aggiuntivo COM
Componenti aggiuntivi inattivi dell'applicazione		
Componente ag	AM	Componente aggiuntivo di Excel
Data (XML)	.DLL	Azione
Euro Currency T	LAM	Componente aggiuntivo di Excel
Ext_Add-In_Gest	lam	Componente aggiuntivo di Excel
Inquire	C:\...inquire.dll	Componente aggiuntivo COM
Microsoft Actions Pane 3		Pacchetto di espansione XML
Strumenti di analisi	C:\...NALYS32.XLL	Componente aggiuntivo di Excel
Strumenti di analisi - VBA	C:\...VBAEN.XLAM	Componente aggiuntivo di Excel
Componenti aggiuntivi correlati a documenti		
Nessun componente aggiuntivo correlato a documenti		

Componente aggiuntivo: Strumenti di analisi

Autore: Microsoft Corporation

Compatibilità: Informazioni sulla compat

Posizione: C:\Program Files (x86)\Mi

Descrizione: Strumenti per l'analisi di d

Gestisci:

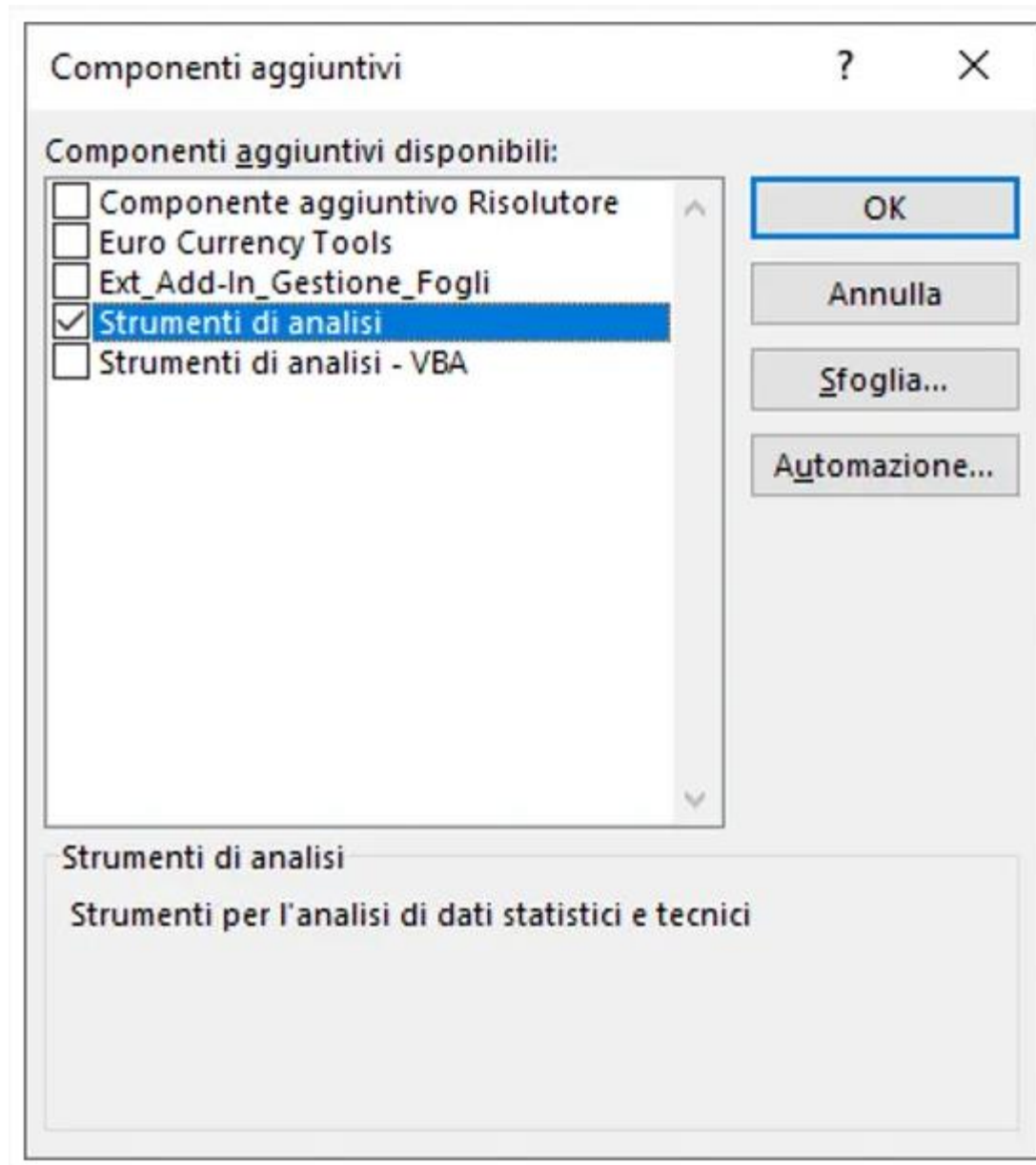
Componenti aggiuntivi di Excel

Vai...

OK

Annulla

Regressione Lineare in Excel: Primo Metodo

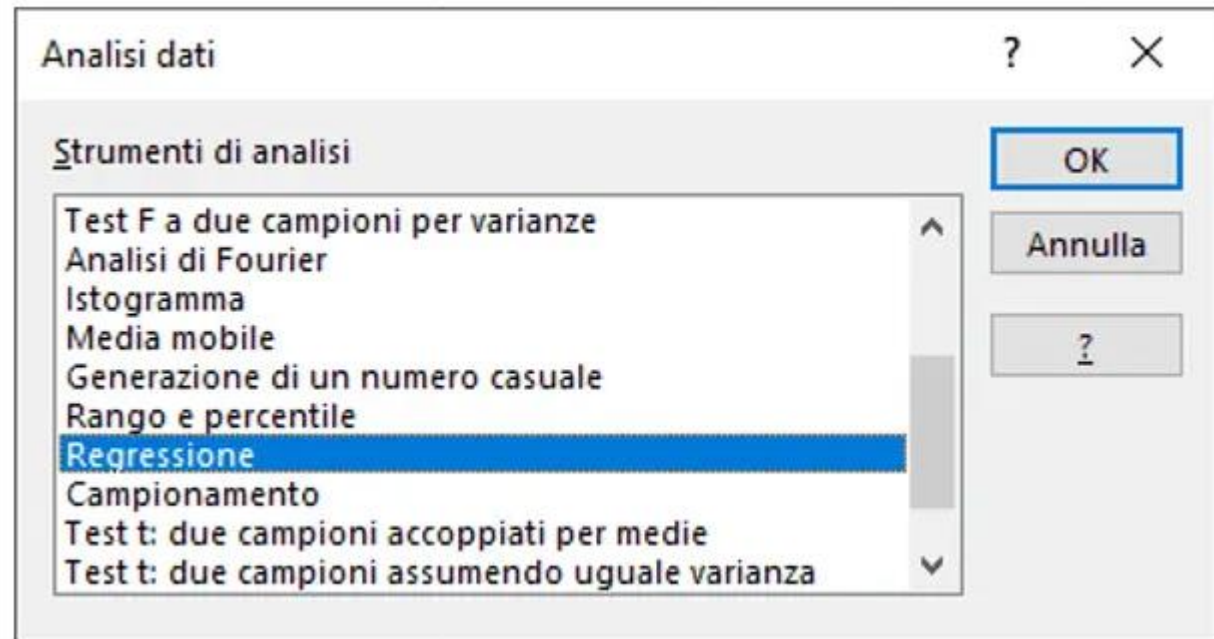


Regressione lineare in Excel: Primo Metodo

- Nella scheda Dati, fate un clic sul pulsante **Analisi dati** presente nel gruppo **Analisi**.




Regressione Lineare in Excel: Primo Metodo




Regressione Lineare in Excel: Primo Metodo

Regressione

Input


Intervallo di input Y: 

Intervallo di input X: 

☐ Etichette ☐ Passa per l'origine

☐ Livello di confidenza %

Opzioni di output

☐ Intervallo di output: 

☒ Nuovo foglio di lavoro:

☐ Nuova cartella di lavoro

Residui

☐ Residui

☐ Residui standardizzati

☐ Tracciati dei residui

☐ Tracciati delle approssimazioni

Probabilità normale

☐ Tracciati delle probabilità normali

OK

Annulla

?

Regressione Lineare in Excel: Primo Metodo

Regressione

Input

Intervallo di input y:

Intervallo di input x:

☐ Etichette ☐ Passa per l'origine

☐ Livello di confidenza %

Opzioni di output

☒ Intervallo di output:

☐ Nuovo foglio di lavoro:

☐ Nuova cartella di lavoro

Residui

☐ Residui

☐ Residui standardizzati

☐ Tracciati dei residui

☐ Tracciati delle approssimazioni

Probabilità normale

☐ Tracciati delle probabilità normali

OK

Annulla

?

Regressione lineare in Excel: Primo Metodo

Analisi dei Risultati



- » **R multiplo.** Questo valore rappresenta il coefficiente di **correlazione** che misura la forza di una relazione lineare tra due variabili.
- » Il coefficiente di correlazione può essere qualsiasi valore compreso tra 1 e -1. Il suo valore assoluto indica la forza della relazione.
- » Maggiore è il valore, più forte è la relazione.

Statistica della regressione

R multiplo	0.94299071
R al quadrato	0.889231479
R al quadrato corretto	0.888674853
Errore standard	3.930705341
Osservazioni	201

Regressione lineare in Excel: Primo Metodo

Analisi dei Risultati

- » **R al quadrato:** È il valore che misura la proporzione della variazione della variabile dipendente che viene spiegata dalla retta di regressione.
- » Questa proporzione deve essere un valore compreso tra zero e uno ed è spesso espresso come percentuale.
- » Rappresenta il coefficiente di determinazione che viene utilizzato come indicatore della bontà dell'adattamento.

<i>Statistica della regressione</i>	
R multiplo	0.94299071
R al quadrato	0.889231479
R al quadrato corretto	0.888674853
Errore standard	3.930705341
Osservazioni	201

Regressione lineare in Excel: Primo Metodo

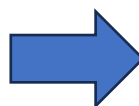
Analisi dei Risultati

- » **R al quadrato corretto.** È una versione modificata dell' R^2 , progettata per tenere conto del numero di variabili indipendenti nel modello.
- » Mentre l' R^2 standard indica la proporzione di variazione nella variabile dipendente spiegata dal modello, l' R^2 corretto penalizza i modelli con un numero eccessivo di variabili, riducendo la probabilità di **overfitting**.



Formula dell' R^2 corretto:

$$R^2_{\text{corretto}} = 1 - \left(\frac{1 - R^2}{n - k - 1} \right) \times (n - 1)$$



Statistica della regressione	
R multiplo	0.94299071
R al quadrato	0.889231479
R al quadrato corretto	0.888674853
Errore standard	3.930705341
Osservazioni	201

- R^2 è l' R^2 standard,
- n è il numero di osservazioni,
- k è il numero di variabili indipendenti.

L' R^2 corretto tende a essere inferiore all' R^2 standard, specialmente quando il modello include molte variabili rispetto al numero di osservazioni.

Regressione lineare in Excel: Primo Metodo

Analisi dei Risultati

- » **Errore Standard.** Questo valore mostra la precisione dell'analisi di regressione: più piccolo è questo valore, più è precisa l'equazione di regressione.
- » È da considerarsi un'altra misura di bontà di adattamento.
- » Mentre R al quadrato rappresenta la percentuale della varianza delle variabili dipendenti spiegata dal modello, l'Errore standard è una misura assoluta che mostra la distanza media attorno alla retta di regressione.

<i>Statistica della regressione</i>	
R multiplo	0.94299071
R al quadrato	0.889231479
R al quadrato corretto	0.888674853
Errore standard	3.930705341
Osservazioni	201

Regressione lineare in Excel: Primo Metodo

ANALISI VARIANZA					
	<i>gdl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>Significatività F</i>
Regressione	1	24682.692	24682.692	1597.539283	4.97301E-97
Residuo	199	3074.638451	15.45044448		
Totale	200	27757.33045			

- » Questa parte è dedicata all'ANALISI VARIANZA. Viene usata raramente per l'analisi della regressione lineare semplice.
- » Tuttavia, è importante osservare l'ultimo valore: la **Significatività F**.
- » Questo valore dà un'idea di quanto sono **statisticamente significativi** (ovvero affidabili) i risultati.
- » **Se il valore della Significatività F è inferiore a 0,05 (5%), il modello utilizzato è buono. Se è maggiore di 0,05, probabilmente è meglio scegliere un'altra variabile indipendente.**

Regressione lineare in Excel: Primo Metodo

	<i>Coefficienti</i>	<i>Errore standard</i>	<i>Stat t</i>	<i>Valore di significatività</i>	<i>Inferiore 95%</i>	<i>Superiore 95%</i>	<i>Inferiore 95.0%</i>	<i>Superiore 95.0%</i>
Intercetta	4.586433381	0.552438528	8.302160596	1.54588E-14	3.49704861	5.675818152	3.49704861	5.675818152
Variabile X 1	1.909845227	0.047782889	39.96922921	4.97301E-97	1.815619448	2.004071007	1.815619448	2.004071007

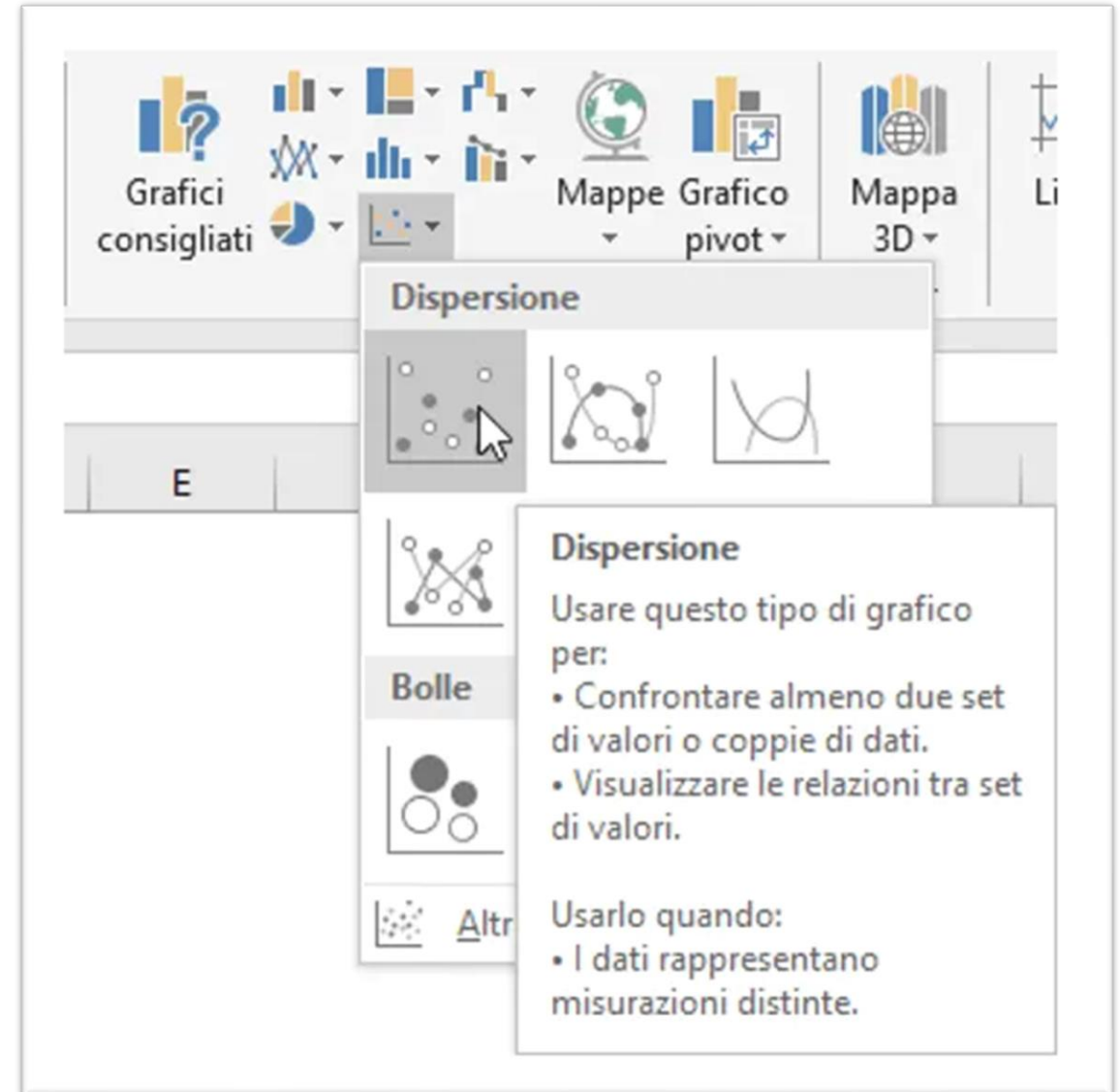
- L'aspetto più interessante di questa sezione sono i **coefficienti**.
- Attraverso i **coefficienti** è possibile creare l'equazione di regressione lineare in Excel.
- Questa equazione può essere quindi utilizzata per la **previsione** di nuovi risultati a partire da nuovi dati della variabile indipendente.

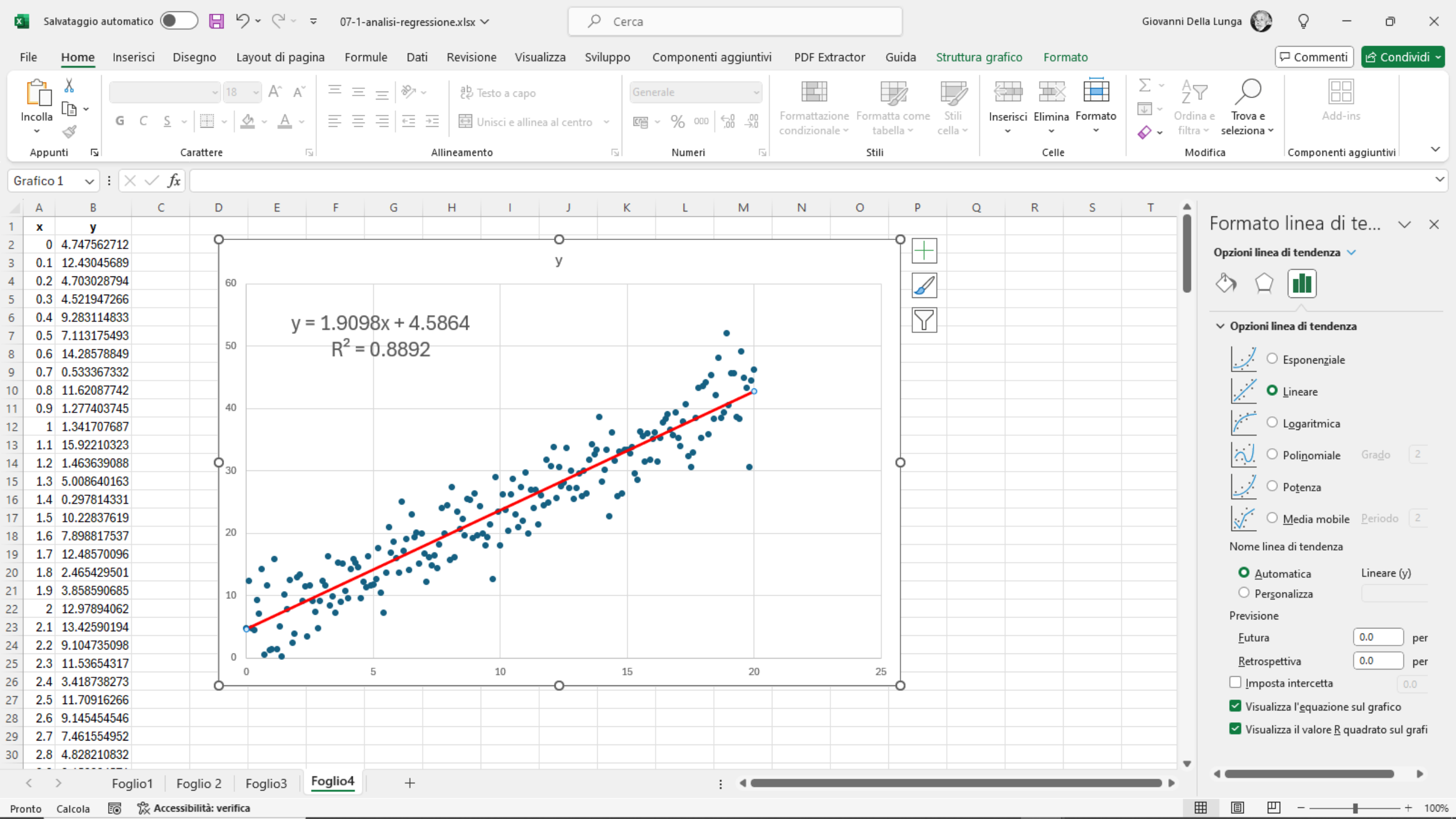
$$Y_i = \alpha_0 + \alpha_1 \cdot X_i$$

Regressione lineare in Excel: Secondo Metodo

Grafico a dispersione con una linea di tendenza

- » Il secondo metodo per eseguire la regressione in Excel è inserire un grafico di regressione lineare.
- » Il grafico consente di visualizzare rapidamente la relazione tra le due variabili.
- » Vediamo come
- » Prima di tutto inseriamo un grafico a dispersione ...





Relazione fra Correlazione e Coefficiente Angolare Retta Regressione

- Il coefficiente di correlazione lineare r e il coefficiente angolare della retta di regressione b sono strettamente collegati.
- La relazione tra questi due è data dalla formula:

$$b = r \left(\frac{s_Y}{s_X} \right)$$

Dove:

- r è il coefficiente di correlazione lineare.
- s_Y è la deviazione standard della variabile Y .
- s_X è la deviazione standard della variabile X .

Analisi Regressione Multipla

- L'analisi di regressione multipla in Excel è estremamente semplice
- Partiamo dal foglio con il nostro dataset, costituito da tre serie di valori.
- Cliccate su (Barra Multifunzione) **Dati** > **Analisi dati** > **Regressione**.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
X1	X2	X3	Y												
10.99342831	35.84629258	107.1557472	45.25862694												
9.723471398	45.79354677	111.2156905	34.45801428												
11.29537708	46.57285483	121.6610249	41.47796888												
13.04605971	41.97722731	121.076041	48.71094944												
9.531693251	48.38714288	72.44661264	13.09728178												
9.531726086	54.04050857	81.2434992	13.12053518												
13.15842563	68.86185901	110.3007053	17.70206538												
11.53486946	51.74577813	110.275719	31.14787737												
9.061051228	52.57550391	110.3009537	26.01982652												
11.08512009	49.25554084	177.0546298	66.69738434												
9.073164614	30.81228785	111.4178102	47.26557444												
9.068540493	49.73486125	122.7113128	35.00688413												
10.48392454	50.6023021	119.0800353	35.39655515												
6.173439511	74.63242112	113.027825	-0.294725496												
6.550164335	48.07639035	93.69461511	15.55700849												
8.875424942	53.01547342	115.1793844	29.12753083												
7.974337759	49.6528823	84.54349571	14.24924441												
10.62849467	38.31321962	95.26362787	35.31890522												
8.183951849	61.42822815	90.29272904	4.62285573												
7.175392597	57.51933033	101.6374828	11.68441916												
12.93129754	57.91031947	146.2931713	47.07887227												
9.548447399	40.90612545	62.65469615	14.56300186												
10.13505641	64.02794311	113.7252038	18.18578213												

Ho creato un dataset di esempio per una regressione lineare multipla con 100 osservazioni e tre variabili indipendenti (X1, X2, X3), insieme a una variabile dipendente (Y). I dati sono generati secondo le seguenti specifiche:

- X1: Variabile indipendente con media 10 e deviazione standard 2.
- X2: Variabile indipendente con media 50 e deviazione standard 10.
- X3: Variabile indipendente con media 100 e deviazione standard 20.
- Y: Variabile dipendente calcolata come $Y = 5 + 2 \times X1 - 1 \times X2 + 0.5 \times X3$ + rumore , con un po' di rumore aggiunto con deviazione standard 5.

Analisi Regressione Multipla

- L'analisi di regressione multipla in Excel è estremamente semplice
- Partiamo dal foglio con il nostro dataset, costituito da tre serie di valori.
- Cliccate su (Barra Multifunzione) **Dati** > **Analisi dati** > **Regressione**.

The screenshot shows the 'Regressione' (Regression) dialog box in Excel. The dialog is titled 'Regressione' and has a question mark icon and a close button (X) in the top right corner. It is divided into several sections:

- Input:**
 - Intervallo di input Y:** A text box containing '\$D\$2:\$D\$101' with an upward arrow icon to its right.
 - Intervallo di input X:** A text box containing '\$A\$2:\$C\$101' with an upward arrow icon to its right.
 - ☐ **Etichette**
 - ☐ **Passa per l'origine**
 - ☐ **Livello di confidenza** 95 %
- Opzioni di output:**
 - ☒ **Intervallo di output:** A text box containing '\$F\$18' with an upward arrow icon to its right.
 - ☐ **Nuovo foglio di lavoro:** An empty text box.
 - ☐ **Nuova cartella di lavoro**
- Residui:**
 - ☐ **Residui**
 - ☐ **Residui standardizzati**
 - ☐ **Tracciati dei residui**
 - ☐ **Tracciati delle approssimazioni**
- Probabilità normale:**
 - ☐ **Tracciati delle probabilità normali**

On the right side of the dialog, there are three buttons: 'OK' (highlighted with a blue border), 'Annulla' (Cancel), and a help button with a question mark icon.

Analisi Regressione Multipla

- Cliccando sul tasto OK, viene prodotto il report di riepilogo con tutti i dati di output

OUTPUT RIEPILOGO									
<i>Statistica della regressione</i>									
R multiplo	0.998067								
R al quadrato	0.996137								
R al quadrato corretto	0.996016								
Errore standard	1.009954								
Osservazioni	100								
ANALISI VARIANZA									
	<i>gdl</i>	<i>SQ</i>	<i>MQ</i>	<i>F</i>	<i>significatività F</i>				
Regressione	3	25249.89	8416.63	8251.537	1.2E-115				
Residuo	96	97.92073	1.020008						
Totale	99	25347.81							
	<i>Coefficiente</i>	<i>errore standard</i>	<i>Stat t</i>	<i>p di signific.</i>	<i>inferiore 95%</i>	<i>superiore 95%</i>	<i>inferiore 95.0%</i>	<i>superiore 95.0%</i>	
Intercetta	4.773539	0.913673	5.224559	1.01E-06	2.959912	6.587166	2.959912	6.587166	
Variabile X 1	1.993221	0.057432	34.70547	3.96E-56	1.879218	2.107223	1.879218	2.107223	
Variabile X 2	-0.99976	0.010745	-93.0478	6.85E-96	-1.02109	-0.97844	-1.02109	-0.97844	
Variabile X 3	0.502893	0.004769	105.4586	4.6E-101	0.493427	0.512359	0.493427	0.512359	