

DATA SCIENCE 4 CITIZENS

LABORATORIO DI DATA SCIENCE



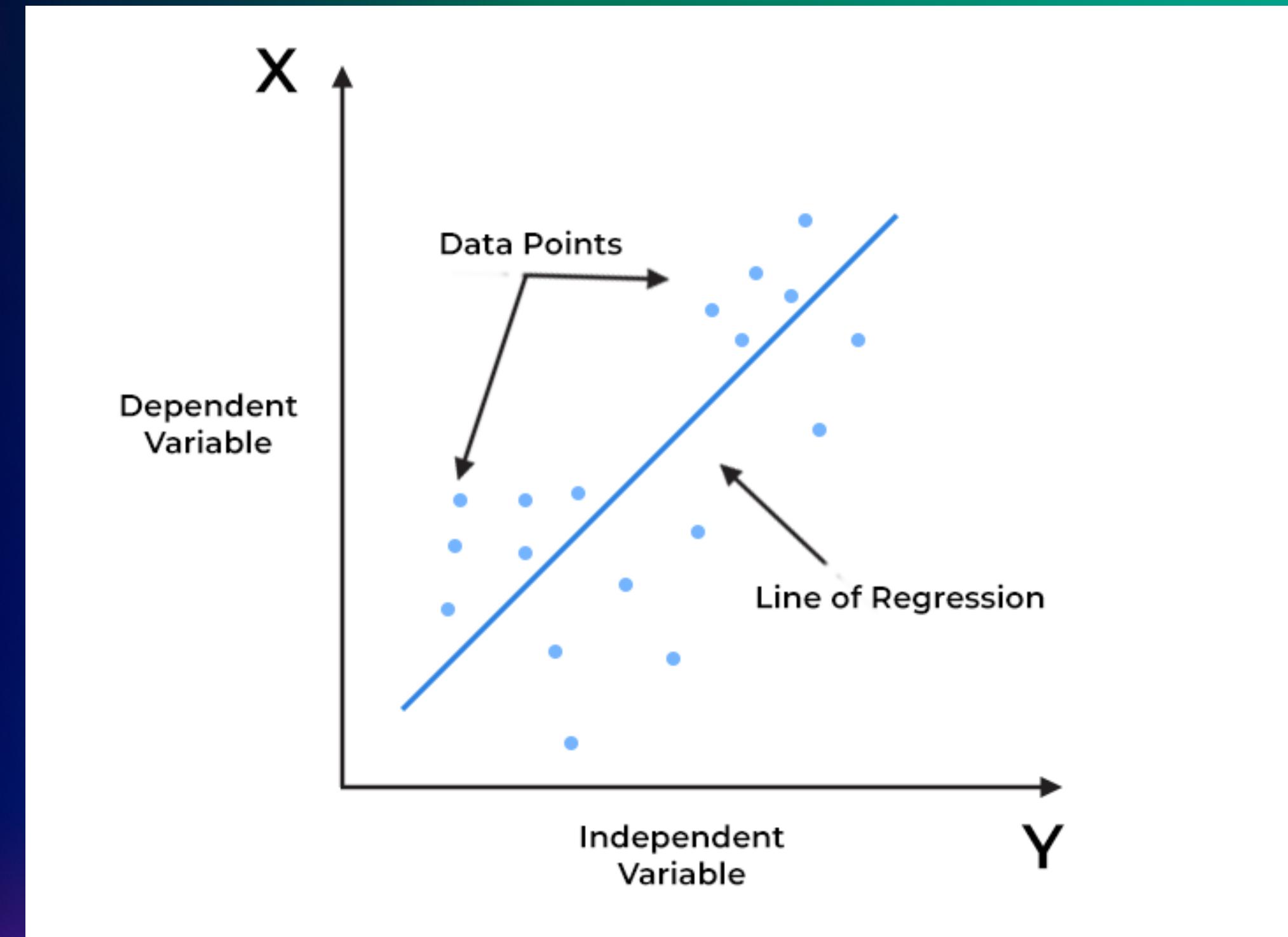
LICEO SCIENTIFICO
LORENZO RESPIGHI

REGRESSIONE LINEARE

COS'E'?

Un metodo statistico che definisce una relazione lineare tra una variabile indipendente (y) e la variabile dipendente (x).

E' l'algoritmo che usiamo per prevedere la **popolarità delle canzoni** (x), date le variabili che abbiamo osservato (y).



ASSUNTI DELLA REGRESSIONE LINEARE

- 1 Esiste una relazione *lineare* tra variabile dipendente (*Song Popularity*) e le varibili indipendenti (tutte le altre nel nostro dataset).
- 2 I dati osservati sono indipendenti tra di loro.
- 3 Assenza di auto-correlazione tra gli errori.
- 4 Omoschedasticità degli errori.
- 5 Gli errori hanno una distribuzione normale.
- 6 Non c'è *multicollinearità* tra le variabili dipendenti.

LINEARITA' DELLA DIPENDENZA

La relazione di dipendenza tra la variabile di interesse e i *regressori* (i.e., le variabili indipendenti) è lineare? O ha una forma diversa?

Questa condizione si può verificare semplicemente dando un'occhiata agli scatterplot!

ASSENZA DI AUTOCORRELAZIONE DEGLI ERRORI

Gli errori devono essere anche indipendenti da sé stessi!

L'errore deve essere casuale, dovuto ad errori di calcolo, o di altro tipo, ma non sistematici!

Se invece c'è *autocorrelazione*, significa che c'è qualche variabile che ci siamo dimenticati di includere, quindi il nostro modello è *biased*, ossia i nostri coefficienti non rappresentano correttamente la realtà.

Per verificare assenza di autocorrelazione usiamo il test di Durbin-Watson.

OMOSCHEDASTICITA' DEGLI ERRORI

Omoschedasticità degli errori significa che l'errore del modello (che noi calcoliamo come la differenza tra i valori predetti e i valori reali), è costante.

Al contrario, eteroschedasticità degli errori implica che gli errori del modello cambiano in base ai dati del campione statistico.

Se gli errori sono eteroschedastici, il nostro modello lineare sarà *inefficiente* (ovvero ci saranno dei modelli più efficienti che si potrebbero utilizzare), ma le stime non sono biased.

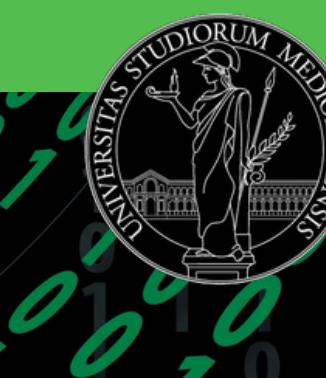
Per verificare ciò utilizziamo il test di Breusch-Pagan.

ASSENZA DI MULTICOLLINEARITÀ

Tra i principali assunti del modello lineare abbiamo l'assenza di **multicollinearità**.

Si ha multicollinearità quando le variabili Y sono tra di loro correlate (ovvero sussiste una relazione di tipo lineare anche tra di loro).

Quando un modello presenta multicollinearità i suoi risultati possono essere "biased", ovvero sistematicamente diversi dai valori reali.



C'E' MULTICOLLINEARITA'?

STEP 1

Guardiamo la matrice di correlazione: *possiamo vedere se ci sono variabili correlate tra loro?*

STEP 2

Utilizzare il VIF test: con questo test possiamo verificare se la varianza associata ad una variabile indipendente viene "gonfiata" dalla presenza di multicollinearità.

Se il VIF associato alle variabili è maggiore di 5, possiamo considerare il nostro modello "*biased*".



STEPWISE REGRESSION

Se il modello non presenta *multicollinearità* preoccupante (sotto 5), ma le variabili indipendenti presentano leggera correlazione possiamo utilizzare un ulteriore metodo: la *stepwise regression*.

Questo metodo ci permette di selezionare le variabili più significative e ridurre il bias complessivo.

Esistono tre tipi di stepwise regression:

1. **Forward selection**: parte dal modello vuoto e aggiunge una variabile per volta.
2. **Backward selection**: parte dal modello completo e rimuove una variabile per volta.
3. **Stepwise selection**: una combinazione dei due. Il modello di partenza è vuoto e vengono aggiunte variabili come nella forward selection. Man mano che vengono aggiunte variabili altre vengono eliminate se non migliorano il modello.



NORMALITA' DEGLI ERRORI

Si presume che gli errori abbiano una distribuzione *Normale*. Questa assunzione è alla base di determinati processi, quindi prendetela un po' per buona.

Per verificare la normalità degli errori possiamo utilizzare il test di **Shapiro-Wilk**.