# INTRODUZIONE A ANALISI DATI MICROCLIMA COVID 19

C. Fagarazzi

Una breve introduzione metodologica per capire che metodologia state sviluppando..

Nella presentazione HANDBOOK_REGRESSIONE_MULTIPLA_COVID.pdf troverete un breve remind metodologico e gli steep per lo sviluppo su excel

# Introduzione

- L'analisi della regressione multipla è una tecnica statistica che può essere impiegata per analizzare la relazione tra una **variabile dipendente** e diverse **variabili indipendenti** (**predittori**)
- La regressione lineare multipla rappresenta un'estensione del modello di regressione lineare semplice

L'**OBIETTIVO** dell'analisi è prevedere i valori assunti da una variabile dipendente a partire dalla conoscenza di quelli osservati su più variabili indipendenti

# Modello di regressione lineare multipla

La relazione tra le variabili esplicative e la variabile dipendente può essere scritta come:

$$Y = f(X_1, X_2, \ldots, X_m) + \varepsilon = f(\mathbf{X}) + \varepsilon$$

Se si esplicita una **relazione di tipo lineare** si ottiene l'equazione:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon$$

nella quale dovranno essere stimati i **parametri** $\beta_i$

**A tal scopo è necessario osservare le variabili esplicative e la variabile dipendente su un campione di** *n* **osservazioni**

Regressione lineare semplice (1 dip, 1 indip)

$$Y_i = a + bX_i + \varepsilon_i$$

intercetta

pendenza

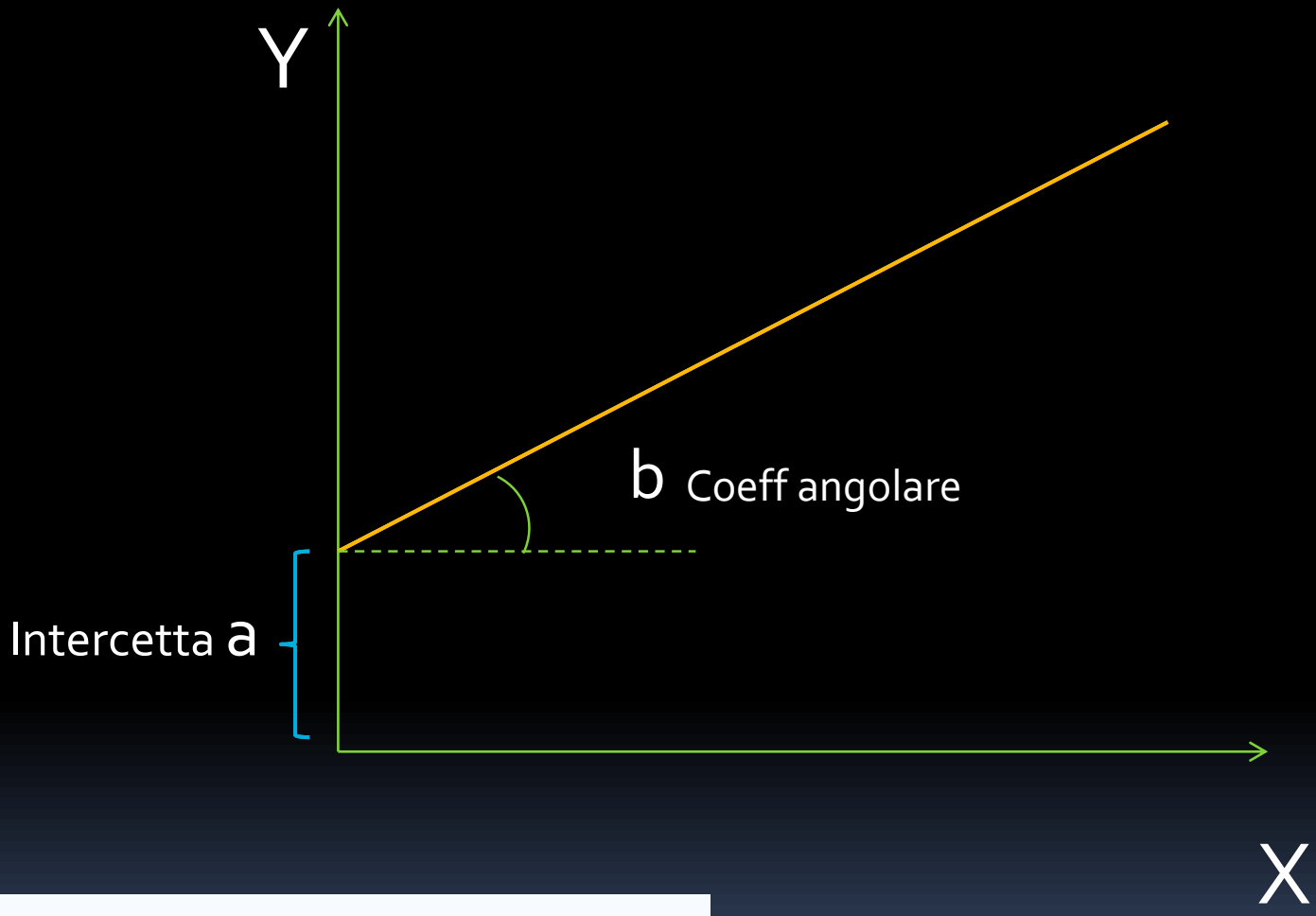
variabile indipendente

errore

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i+} \varepsilon_i$$

Regressione lineare multipla (2 indip, 1 dip)

Nel caso di una sola variabile abbiamo una relazione bidimensionale es:



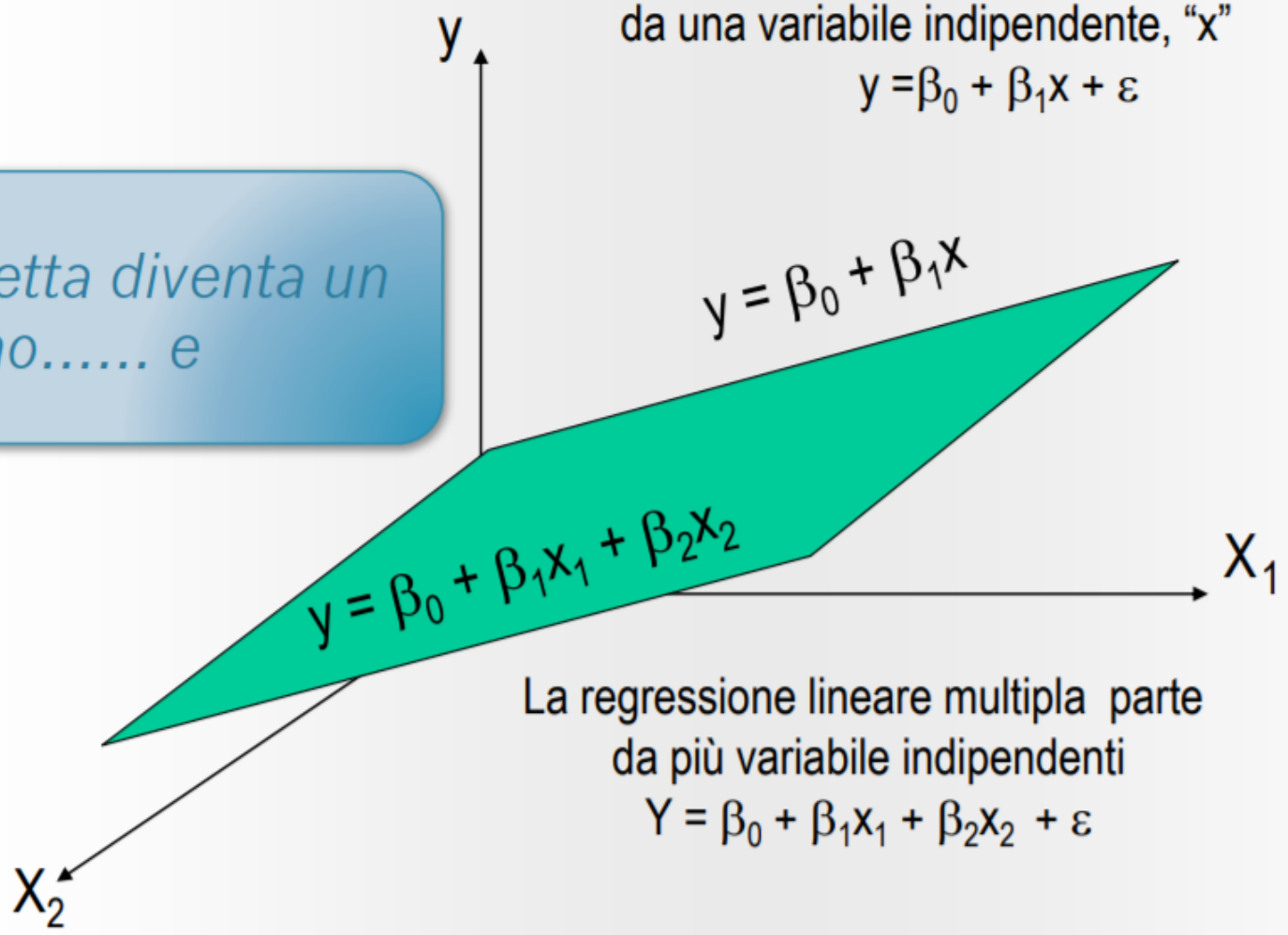

$$Y_i = a + bX_i + \varepsilon_i$$

# Se la relazione è lineare……

La regressione lineare semplice parte da una variabile indipendente, "x"

$$y = \beta_0 + \beta_1 x + \varepsilon$$

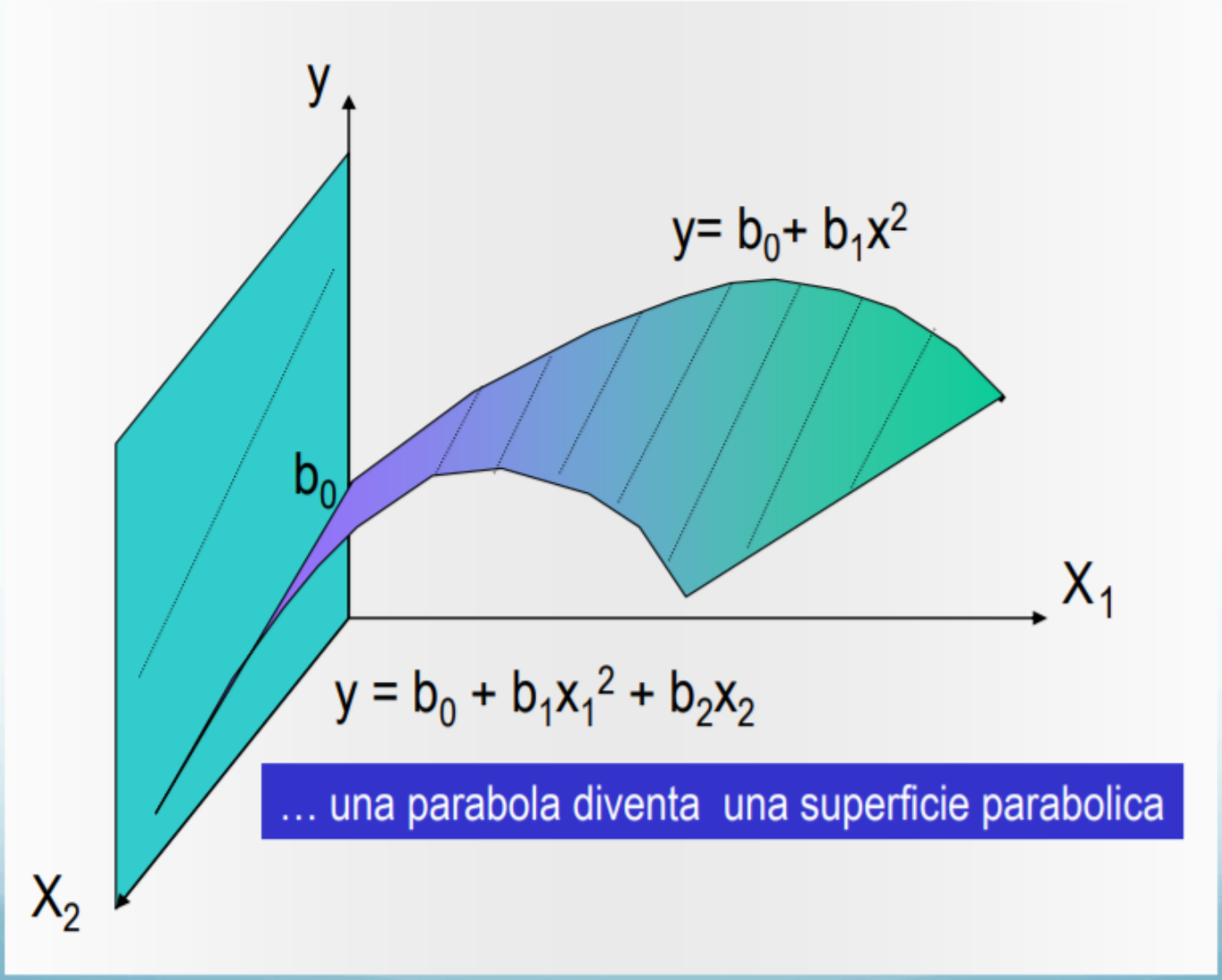
*La retta diventa un piano…… e*




La regressione lineare multipla parte da più variabile indipendenti

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

# Se la relazione *non* è lineare.....



# Esempio....

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \varepsilon_i$$

| y | $X_1$ | X2 | |
|---|---|---|---|
| 3 | 2 | 1 | $3=1\beta_0+2\beta_1+1\beta_2+e_1$ |
| 2 | 3 | 5 | $2=1\beta_0+3\beta_1+5\beta_2+e_2$ |
| 4 | 5 | 3 | $4=1\beta_0+5\beta_1+3\beta_2+e_3$ |
| 5 | 7 | 6 | $5=1\beta_0+7\beta_1+6\beta_2+e_4$ |
| 8 | 8 | 7 | $8=1\beta_0+8\beta_1+7\beta_2+e_5$ |

# Indice di determinazione lineare

L'indice di determinazione lineare $R^2$ rappresenta la frazione di varianza di Y che è spiegabile dai regressori X inclusi nel modello

$$R^2 = \frac{Dev(\hat{Y})}{Dev(Y)} = 1 - \frac{Dev(E)}{Dev(Y)}$$

L'indice $R^2$ può presentare alcuni problemi calcolatori e di interpretazione:

- in **assenza** di relazione lineare non è pari a zero
- $R^2$ **tende ad aumentare** all'aumentare del numero di variabili esplicative X
- Un aumento di $R^2$ non significa necessariamente che il nuovo regressore concorre in modo significativo a spiegare Y. L'$R^2$ fornisce quindi un indice "per eccesso"

# Indice $R^2$ corretto

L'indice **$R^2$ corretto**, indicato con $\bar{R}^2$, corregge tale eccesso deflazionando l'$R^2$ per un termine che aumenta con il numero di regressori inclusi nel modello.
La logica è molto semplice: se l'aumento dell'indice $R^2$ eccede la penalità indotta dall'avere un regressore in più nel modello, **$R^2$ corretto** cresce.
In caso contrario, **$R^2$ corretto** decresce.

$$\bar{R}^2 = 1 - \frac{n-1}{n-m-1}\left(1-R^2\right)$$

Tale indice può essere **negativo** nel caso in cui gli m regressori presi nel complesso aumentino $R^2$ di un ammontare piccolo rispetto al fattore n-1/n-m-1

Tuttavia nella maggior parte dei casi:

$$0 \geq \bar{R}^2 \geq 1$$

Se poi risulta $R^2 \leq \frac{m-1}{n-1}$ allora $\bar{R}^2 \leq 0$

# Osservazioni

- Un aumento di **$R^2$ corretto**, o di $R^2$, non significa necessariamente che la variabile aggiunta sia statisticamente significativa (la risposta a tale domanda passa per un test t)

- Dal momento che **$R^2$ corretto** è interpretabile come compromesso tra bontà di adattamento e penalità dovuta al soprannumero di regressori "**utili**", una procedura ragionevole nella specificazione del modello consiste nel continuare ad includere regressori fino al momento in cui l' **$R^2$ corretto** inizia a decrescere.

# Analisi dei residui

**Una volta eseguita la regressione multipla**
**Potete controllare la distribuzione d residui di ciascuna variabile**

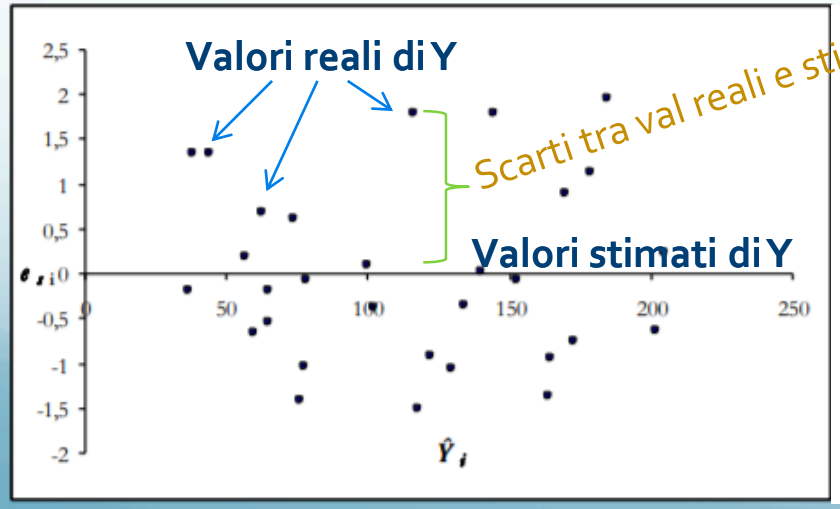
**A cosa vi serve?**

L'analisi grafica dei residui consente di valutare, a posteriori, se il modello ipotizzato è corretto. In tal caso, infatti, gli errori dovrebbero distribuirsi in modo normale.

**Diagramma di dispersione dei residui**:

**in ordinata**: $e_{is}$

**in ascissa**: $\hat{Y}_i$ (i valori stimati della variabile dipendente) o $X_{ji}$

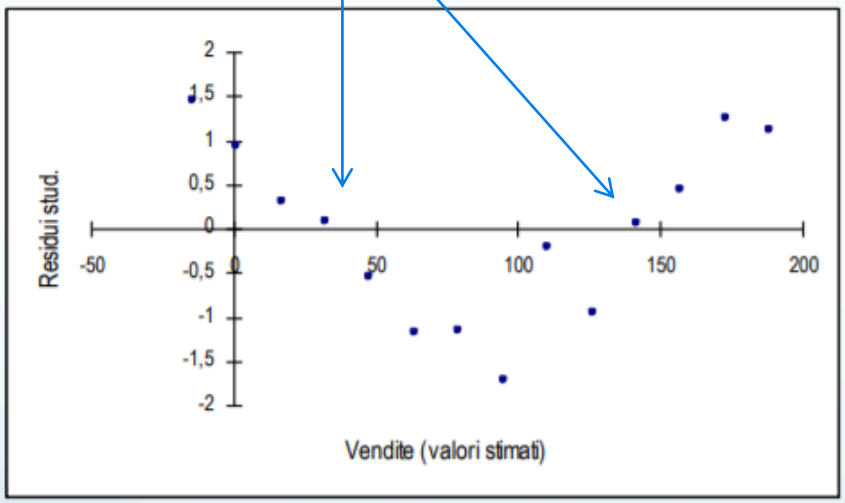


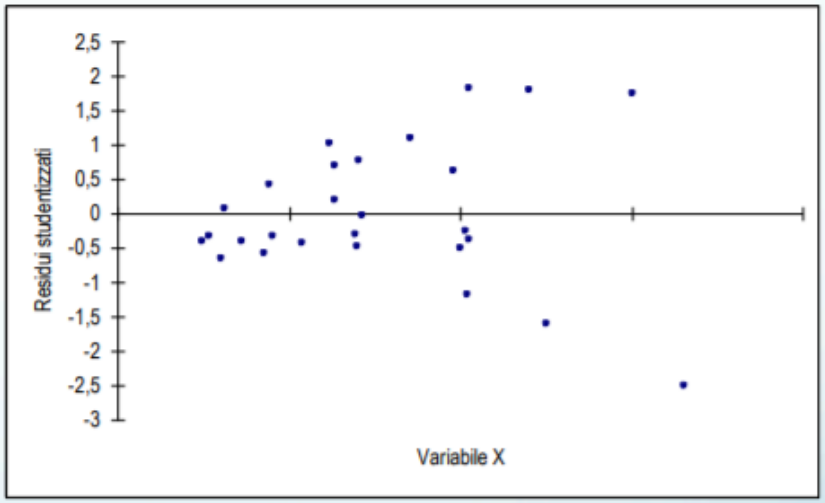

**I valori degli scarti si distribuiscono in modo casuale in questo intervallo**

**SIGNIFICA CHE L'HP DI RELAZIONE LINEARE E' CORRETTA**

# Analisi dei residui

**Lo scarto tra i valori reali di Y e quelli stimati non è casuale.**
**L'Hp di relazione lineare tra X e Y NON E' CORRETTA**



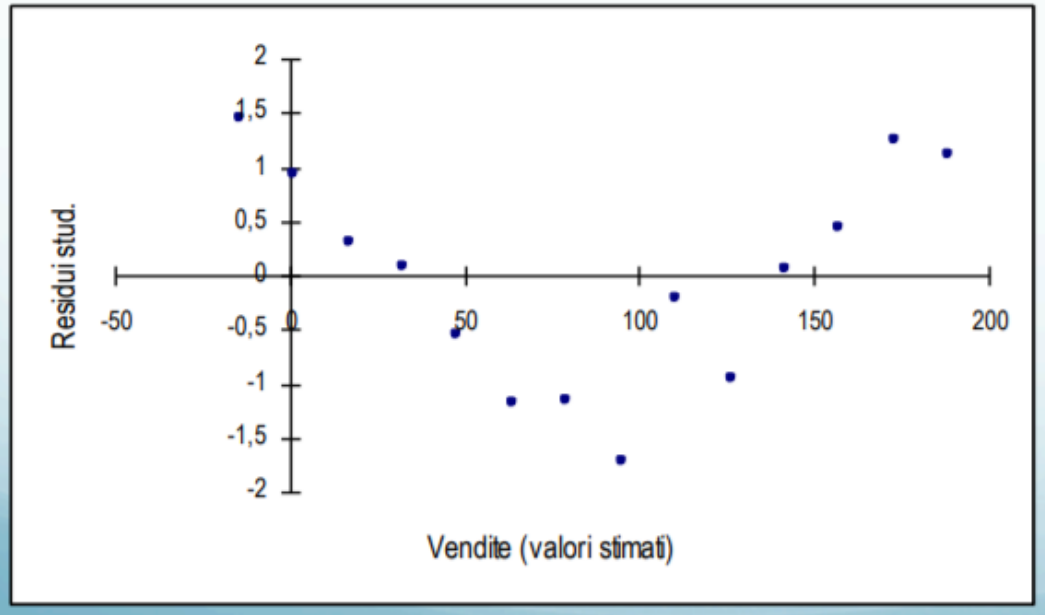

**Violazione dell'ipotesi di linearità**

**Violazione dell'ipotesi di omoschedasticità**

# Violazione dell'ipotesi di **linearità**

2. dalla struttura nel diagramma di dispersione dei residui (es. crescente o decrescente)

*Diagramma di dispersione dei residui*



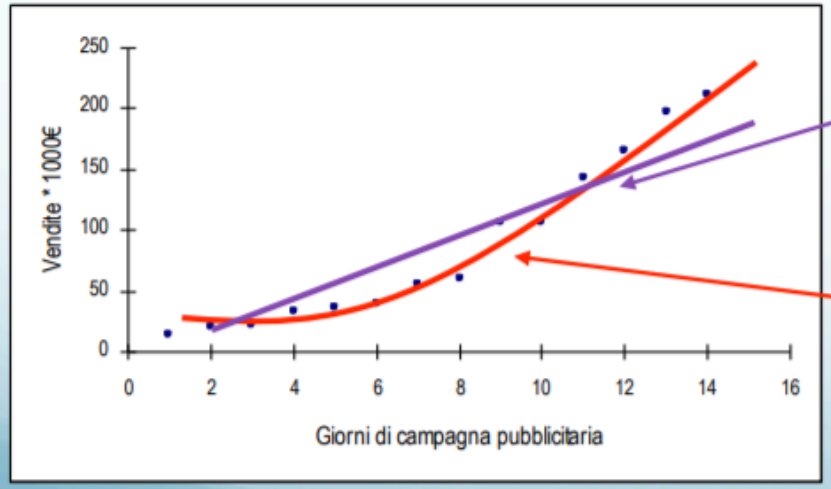

mostra non una disposizione casuale intorno allo zero ma una struttura curvilinea che indica una relazione non lineare

# Violazione dell'ipotesi di **linearità**

Si diagnostica principalmente in due modi:

1. dalla struttura dei punti campionari (nel caso bivariato)

*Diagramma di dispersione dei punti campionari*




Si può stimare un modello lineare

Ma il diagramma fa supporre una relazione non lineare, presumibilmente **esponenziale**

# Violazione dell'ipotesi di **linearità**

Si può risolvere con opportune trasformazioni di variabili

In particolare:

- trasformazione logaritmica della variabile esplicativa

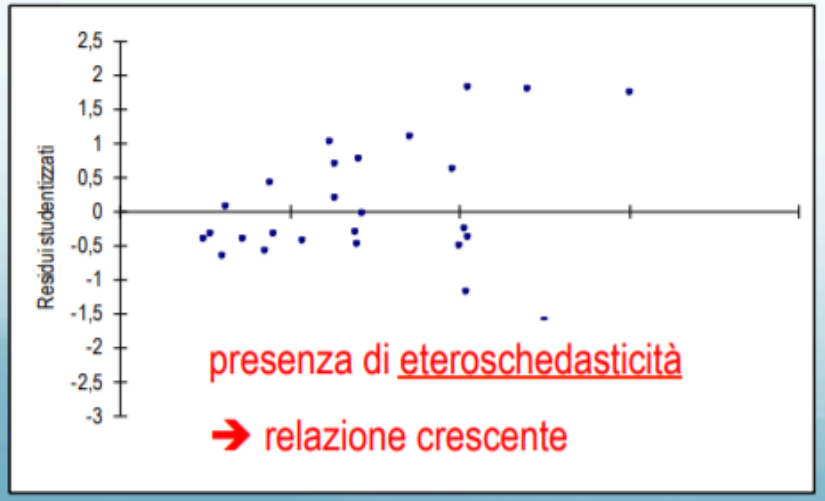
  (o di una o più delle variabili esplicative)

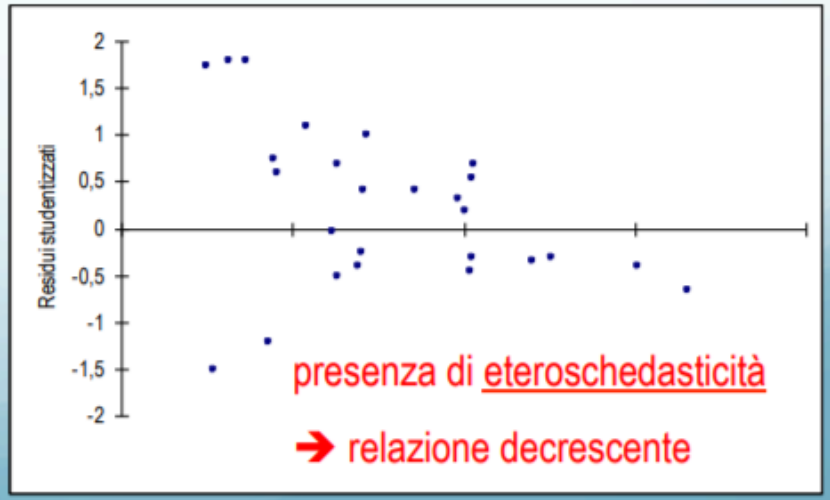
- trasformazione logaritmica della variabile dipendente

# Violazione dell'ipotesi di **omoschedasticità**

Diagnosticata attraverso l' analisi del diagramma di dispersione dei residui:

- se la banda in cui giacciono i punti tende ad allargarsi o a restringersi la varianza degli errori tende a crescere o a decrescere al crescere della variabile esplicativa
- se invece i punti giacciono tra due parallele non si riscontra alcuna evidenza di violazione dell'assunzione



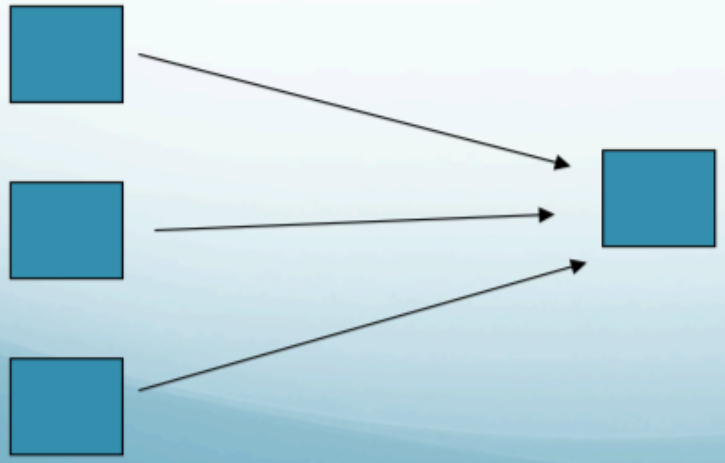



# Multicollinearità

La situazione ideale per una regressione multipla dovrebbe essere: ogni X è altamente correlata con Y, ma le X non sono correlate fra loro

**Test di Goldfeld e Quandt [1965]**

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| Y | .60 | .50 | .70 |
| $X_1$ | | .20 | .30 |
| $X_2$ | | | .20 |

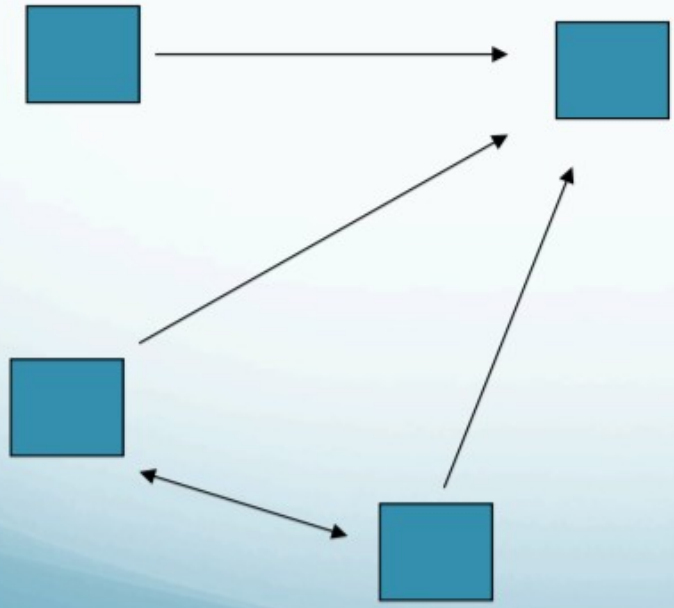


Idealmente, le correlazioni tra le X, dovrebbero essere 0

# Multicollinearità

**Test di Goldfeld e Quandt [1965]**

Spesso però, due o più X sono correlate fra loro....



| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| Y | .60 | .50 | .70 |
| $X_1$ | | **.70** | .30 |
| $X_2$ | | | .20 |

Quando due variabili X o più, sono tra loro correlate (moderatamente o più), parliamo di "**multicollinearità**".

# Definizione

*Con il termine multicollinearità ci si riferisce alla correlazione fra le variabili indipendenti di un modello di regressione*

Il suo effetto consiste nel ridurre la capacità previsiva di ogni singola variabile indipendente in modo proporzionale alla forza della sua associazione con le altre variabili indipendenti.

L' effetto della multicollinearità può interessare:

- la capacità di *spiegazione* del modello (capacità della procedura di regressione e del ricercatore di rappresentare e capire l'influenza di ciascuna variabile indipendente) sia
- la *stima* dei parametri (la sua presenza rende problematica la determinazione dei contributi individuali delle variabili indipendenti, perché i loro effetti vengono "mescolati" o confusi).

# Diminuire la multicollinearità

- combinare fra loro i predittori altamente correlati (ad esempio sommandoli)
- se ci sono molti predittori altamente correlati, usare un' analisi delle componenti principali per ridurre il numero delle X e ottenere delle componenti incorrelate tra loro
- adottare come tecnica di analisi una regressione PLS

**NEL NOSTRO CASO, E PLAUSIBILE CHE TEMPERATURE MIN, MED E MAX Siamo multicollineari…**

# Selezione delle variabili esplicative

*Per la scelta di quali e quante variabili inserire nel modello bisogna giungere ad un compromesso tra il VANTAGGIO di inserire quante più variabili possibili in modo da ridurre la componente erratica e lo SVANTAGGIO dovuto all'aumento dei costi e della varianza delle stime.*

Dato un insieme *q* di predittori esistono varie tecniche per selezionare il **numero ottimale di predittori** da inserire in un modello di regressione multipla:

- Usare la teoria (ricerca bibliografica)
- Metodi semi-automatici sequenziali
  - Regressione stepwise progressiva (avanti – forward)
  - Regressione stepwise a ritroso (indietro – backward)
  - Regressione stepwise convenzionale

# Regressione standard

*Tutte le variabili X vengono considerate assieme e tutti i coefficienti di regressione stimati contemporaneamente*

- Tutte le variabili indipendenti vengono inserite nel modello
- Non si procede quindi ad alcuna selezione
- Per valutare l'importanza di ogni singolo predittore si fa riferimento al *test t*

# Regressione Stepwise convenzionale

- È una combinazione delle due tecniche precedenti
- Si procede come in una regressione stepwise forward, ossia un predittore viene incluso nel modello se dà il contributo più significativo alla spiegazione della variabilità di Y.
- Aggiungendo successivamente una nuova variabile, i coefficienti di regressione delle variabili già incluse potrebbero risultare singolarmente non significativi a causa della forte correlazione con la nuova variabile.
- Pertanto, ad ogni interazione si rimettono in discussione i predittori già inseriti verificando la loro significatività attraverso il test F parziale
- Un predittore può essere rimosso nelle fasi successive se la sua capacità esplicativa viene surrogata da altri predittori.
- La regressione stepwise convenzionale (nota semplicemente come "regressione stepwise") è la più utilizzata nelle applicazioni pratiche.