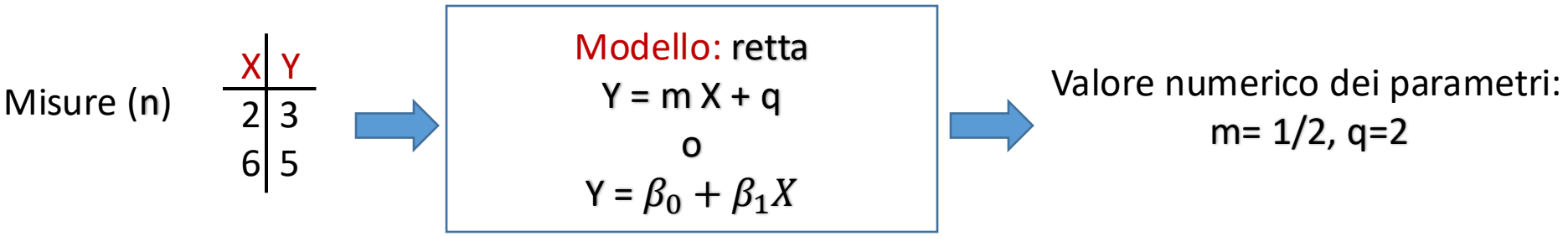
Il Modello di regressione lineare multipla: Identificazione parametrica di Sistema tramite lo stimatore dei minini quadrati

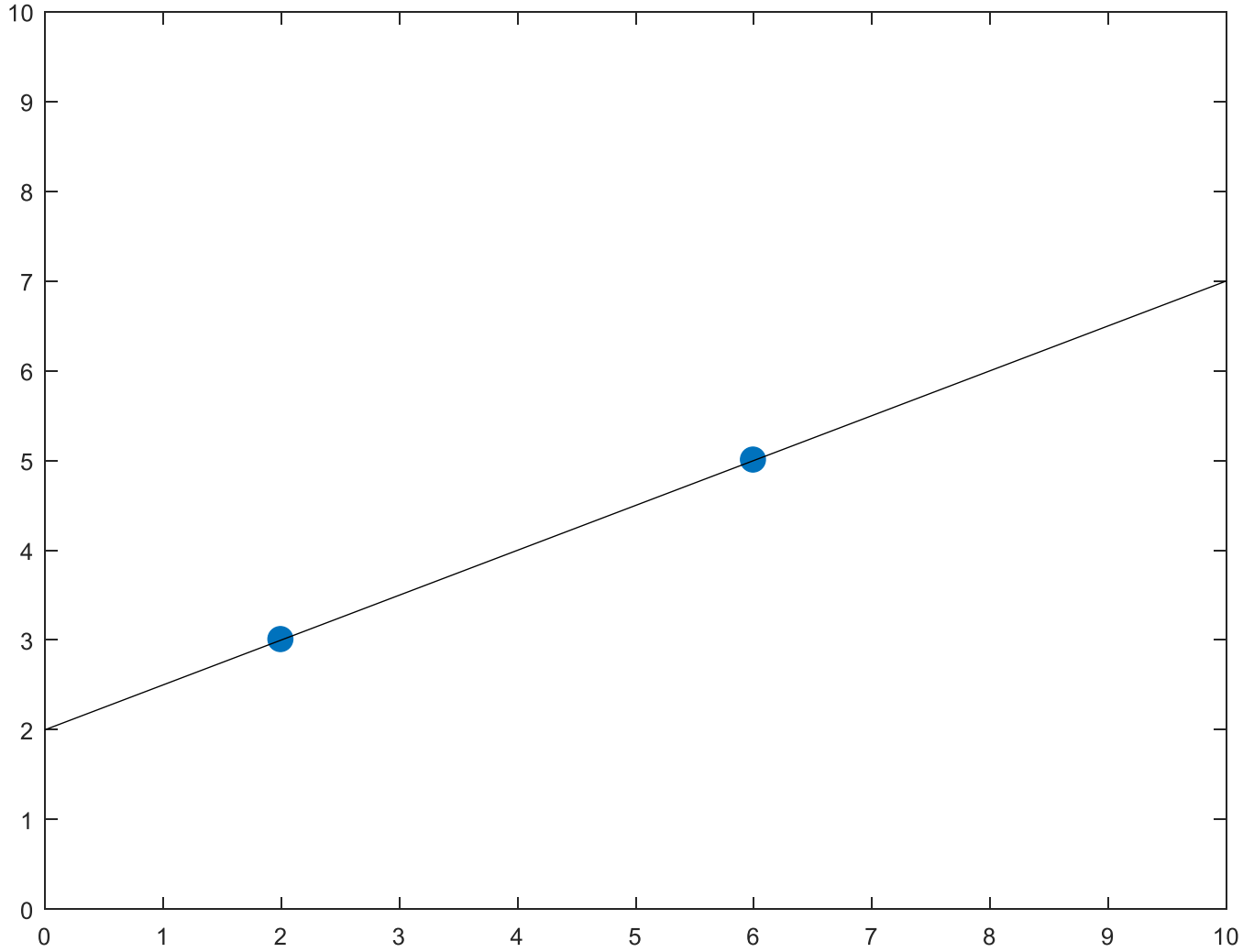
Uno dei modelli più semplici che abbiamo tutti utilizzato è quello della retta ossia il modello che esprime una relazione lineare tra una variabile X (variabile indipedente) e una variabile Y (variabile dipendente):

$$Y = \beta_0 + \beta_1 X \ \ con\, Y\, e\, X \in \mathcal{R}^{nx1}$$

$$\beta_0, \beta_1 : parametri\ del\ modello$$

Esempio: in un piano cartesiano consideriamo 2 punti (misure)





Nel formalismo matematico che useremo m e q sono detti parametri, m x + q è detto modello e **y** saranno le **osservazioni/dati/variabili**

Il modello in realtà dovrebbe tenere in considerazione il fatto che le misure non sono mai prive di errore. Quindi la sua formulazione più corretta è:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$con\ Y, X\ e\ \varepsilon\ \in \mathcal{R}^{nx1}$$

$$\beta_0, \beta_1 : parametri\ del\ modello$$
ε: errore di misura (o di modelo)

L'errore è assunto essere la realizzazione di una variabile aleatoria (**spesso l'assunzione è che la variabile aleatoria sia normale)** con momento primo e momento centrale di ordine 2 pari a:

$$E[\varepsilon] = 0$$
$$\text{Var}(\varepsilon) = \sigma^2$$

questa esprime la variabilità di Y non spiegata dal modello (in questo caso dal modello lineare)

Il modello è generalizzabile a un numero qualsiasi di variabili indipendenti nel seguente modo:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \varepsilon$$

$$con\ Y, X_i\ e\ \varepsilon\ \in \mathcal{R}^{nx1}$$

$$\beta_0, \beta_1, \ldots \beta_m : parametri\ del\ modello$$

ε: errore di misura (o di modelo)

In forma compatta:

$$Y = \beta_0 + \boldsymbol{X\beta} + \varepsilon$$

Spesso viene sottintesa la presenza del β_0 (o inglobata direttamente in X) e utilizzata la formulazione:

$$Y = \boldsymbol{X\beta} + \varepsilon$$

$con\ Y, \varepsilon\ \in \mathcal{R}^{nx1}$, $\quad \boldsymbol{\beta} \in \mathcal{R}^{mx1}$, $\boldsymbol{X} \in \mathcal{R}^{nxm}$

Nota: se le variabili inserite nelle colonne di X hanno media nulla (tipicamente sono z-scorate) allora nel modello lineare l'intercetta ossia β_0 non va inserita nel modello

Se ricordiamo che l'errore è assunto essere la realizzazione di una variabile aleatoria con momento primo e momento centrale di ordine 2 pari a:

$$E[\varepsilon] = 0$$
$$\text{Var}(\varepsilon) = \sigma^2$$

Allora Y è una variabile aleatoria con:

$$\text{E}[Y] = E[\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m] + E[\varepsilon] = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$$

$$\text{Var}(Y) = \text{Var}(\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m) + \text{Var}(\varepsilon) = \text{Var}(\varepsilon) = \sigma^2$$

La varianza di Y non dipende dal modello ma solo dall'errore di misura (o errore di modello)

Ipotizziamo che β_0=0 e scriviamo il modello nel seguente modo:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Ricordando che questo viene rappresentato:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \beta_m \end{bmatrix}$$

Per trovare il valore da dare ai parametri del modello $(\beta_1, \beta_2, \dots, \beta_m)$ dobbiamo usare uno stimatore

Esempio: ci spostiamo da un «mondo ideale» al «mondo reale» e consideriamo il fatto che i dati sono affetti da rumore di misura

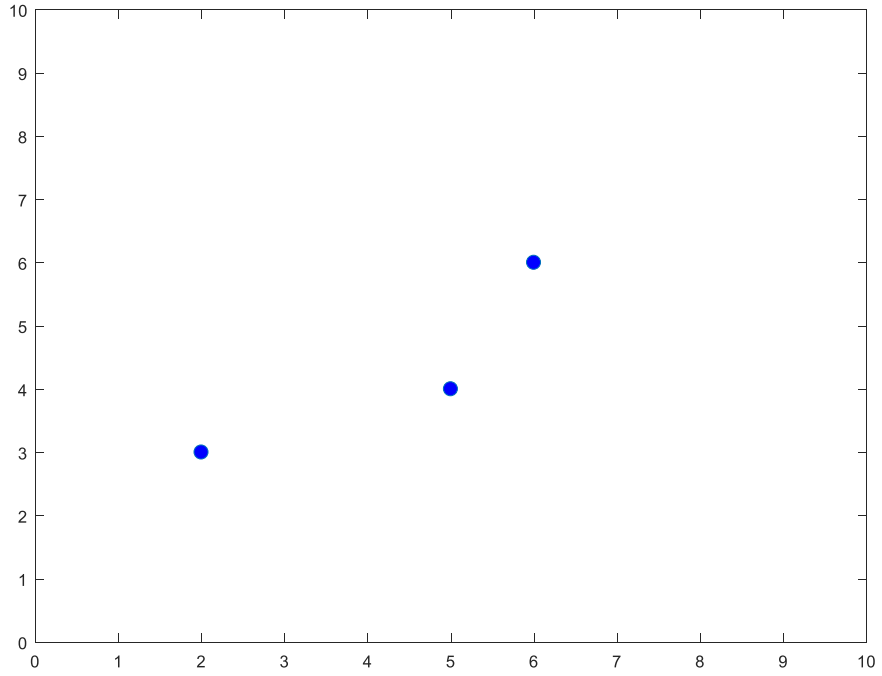
Misure (n)

X	Y
2	3
5	4
6	6

Modello: retta

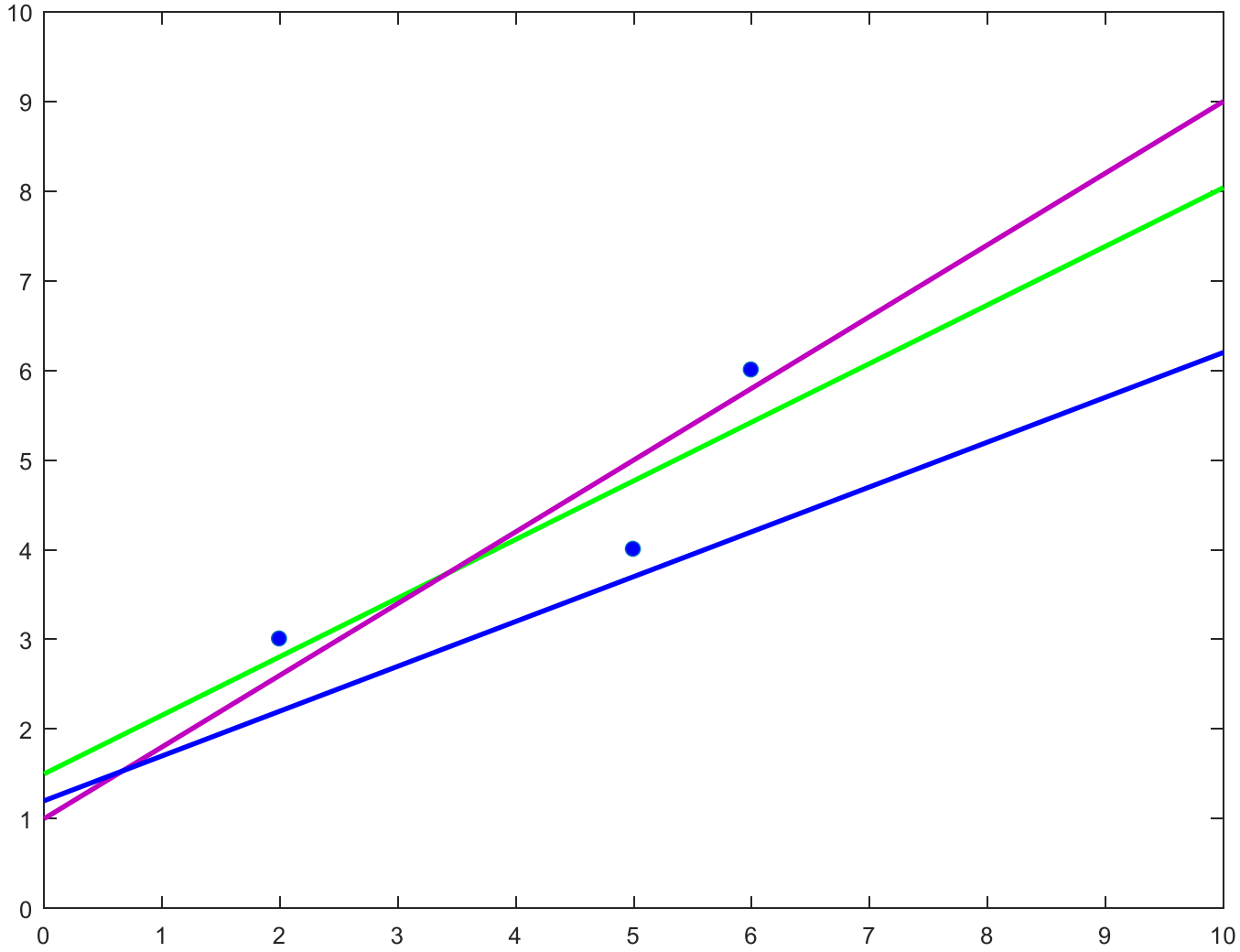
$Y = m X + q$

Valore numerico dei parametri (m, q) ?



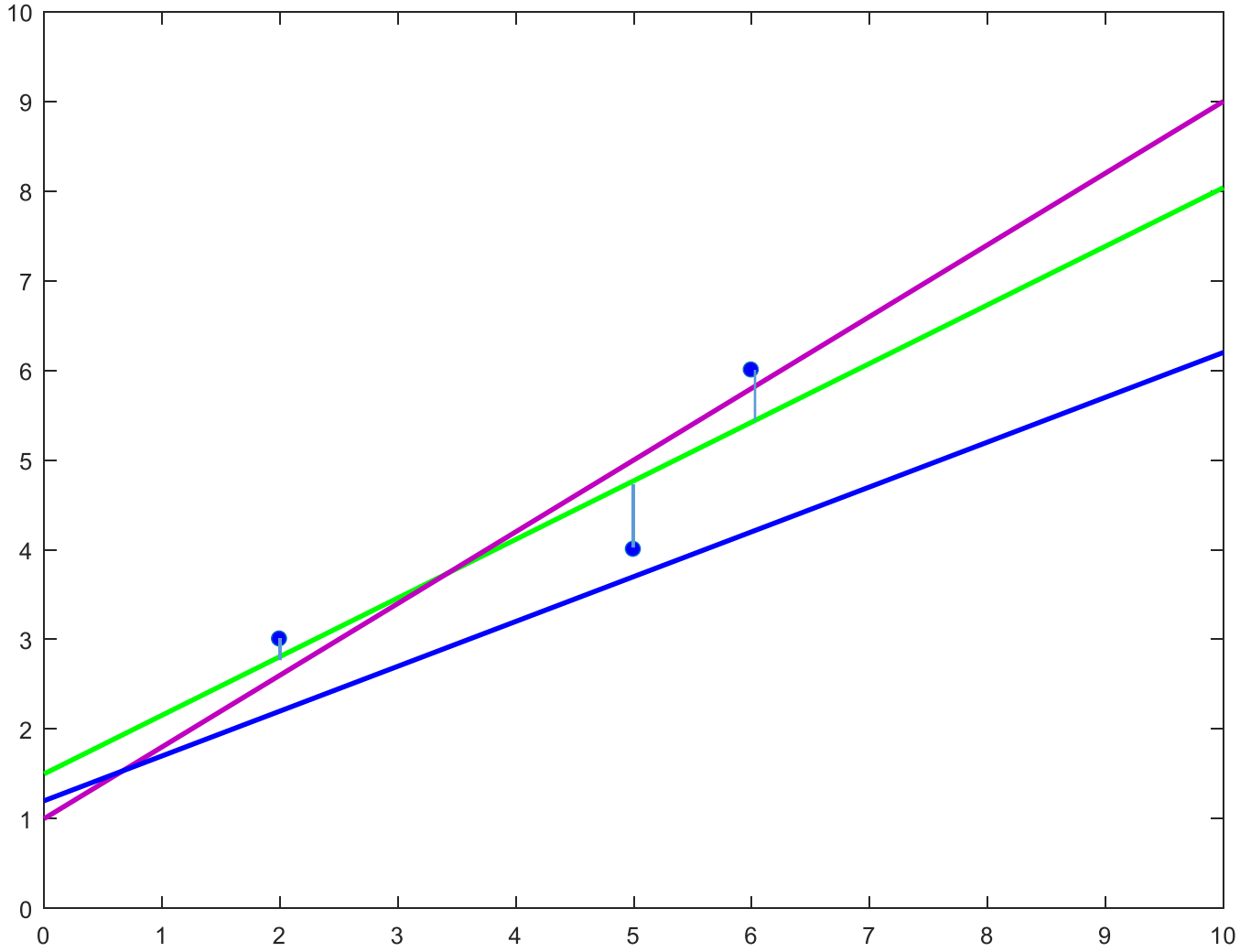
In pratica vogliamo determinare i parametri della retta che passa per questi dati. Ovviamente dobbiamo rinunciare all'idea che la retta passi esattamente per questi punti. Dobbiamo cercare un qualche compromesso

Possibili soluzioni potrebbe essere:



Molti di noi sceglierebbero la retta verde come migliore soluzione.

Se osserviamo scegliendo la retta verde abbiamo dato preferenza alla soluzione che minimizza le distanze tra dato misurato e modello:



Le distanze sono:

3 – (m 2 + q)
4 – (m 5 + q)
6 – (m 6 + q)

In generale:

Y – Modello

Traduciamo la nostra scelta in qualche metodo matematico da usare ➔ STIMATORE

Ipotizziamo che β_0=0 e scriviamo il modello nel seguente modo:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

Ricordando che questo viene rappresentato:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_m \end{bmatrix}$$

Per trovare il valore da dare ai parametri del modello $(\beta_1, \beta_2, \ldots, \beta_m)$ dobbiamo usare uno stimatore. Lo stimatore più usato è chiamato "stimatore dei minimi quadrati lineari" e la sua formula è:

$$J^{LS} = \arg\min \sum_{i=1}^{n} (y_i - X_i\beta)^2 = \arg\min[\boldsymbol{Y} - \boldsymbol{X\beta}]^T[\boldsymbol{Y} - \boldsymbol{X\beta}]$$

$$J^{LS} = \arg\min \sum_{i=1}^{n} (y_i - X_i\beta)^2 = \arg\min[\boldsymbol{Y} - \boldsymbol{X\beta}]^T[\boldsymbol{Y} - \boldsymbol{X\beta}]$$

Cioè quello che minimizza la distanza euclidea tra misure e modello (predizione del modello).

Si può dimostare che il risultato del problema di minimo è, **se la matrice $\boldsymbol{X^T X}$ non è singolare** è:

$$\hat{\beta}^{LS} = (X^T X)^{-1} X^T Y$$

Dove X e Y sono matrici e vettori (per semplicità non saranno evidenziati in grassetto).

Dimostrazione:

$$[\boldsymbol{Y} - \boldsymbol{X\beta}]^T[\boldsymbol{Y} - \boldsymbol{X\beta}] = \boldsymbol{Y^T Y} - \boldsymbol{Y^T X\beta} - \boldsymbol{\beta^T X^T Y} + \boldsymbol{\beta^T X^T X\beta} = \boldsymbol{Y^T Y} - 2\boldsymbol{\beta^T X^T Y} + \boldsymbol{\beta^T X^T X\beta}$$

in quanto $\boldsymbol{Y^T X\beta} = \boldsymbol{\beta^T X^T Y}$ poichè sono scalari

Derivando rispetto i parametri e ponendo a zero:

$$\frac{\partial J^{LS}}{\partial \beta} = -2\boldsymbol{X^T Y} + \boldsymbol{2X^T X\beta} = \boldsymbol{0}$$

Da cui si ricava la formula esplicita:

$$\hat{\beta}^{LS} = (X^T X)^{-1} X^T Y$$

Si noti l'importanza dell'esistenza della matrice $(X^T X)^{-1}$. Il rango di $(X^T X)$ è pieno, e quindi la matrice è invertibile, solo se i regressori (le colonne di X) sono linearmente indipendenti.

Ulteriori informazioni che possiamo ricavare:

1) Nel caso in cui la varianza Var(ε) = σ^2 non sia nota, possiamo stimarla tramite:

$$\hat{\sigma}^2 = \frac{\left[\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}^{\boldsymbol{LS}}\right]^T \left[\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}^{\boldsymbol{LS}}\right]}{n - m} = \frac{SSR}{n - m}$$

2) Possiamo stimare anche il limite inferiore della varianza delle stime dei parametri:

$$Var\left(\hat{\beta}^{LS}\right) = (X^T X)^{-1}$$ varianza dello stimatore

Questa è una matrice di dimensione $m \times m$: a noi interessano i valori della diagonale principale. Questi esprimono la varianza associata alla stima del rispettivo β.

3) Possiamo ricavare la predizione del modello ossia:

$$predizione = \hat{Y} = X\hat{\beta}^{LS}$$

4) Possiamo ricavare il valore dei residui ossia della differenza tra valori Y misurati e valori Y predetti (predizione):

$$\hat{\varepsilon}^{LS} = Y - predizione = Y - \hat{Y} = Y - X\hat{\beta}^{LS}$$

Supponiamo di conoscere il valore della varianza dell'errore di misura. In questo caso possiamo inserirlo come conoscenza nello stimatore. Definiamo:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \vdots & 0 \\ 0 & \sigma^2 & \vdots & 0 \\ \dots & \dots & \vdots & \dots \\ 0 & 0 & \vdots & \sigma^2 \end{bmatrix}$$

Ossia l'errore è scorrelato (matrice diagonale) a media nulla e varianza nota.
In questo caso:

$$J^{LS} = \arg\min \sum_{i=1}^{n} \left(\frac{y_i - X_i \beta}{\sigma} \right)^2 = \arg\min [\boldsymbol{Y} - \boldsymbol{X\beta}]^T \boldsymbol{\Sigma}^{-1} [\boldsymbol{Y} - \boldsymbol{X\beta}]$$

Si può dimostare che il risultato del problema di minimo è in questo caso:

$$\hat{\beta}^{LS} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y = (X^T W X)^{-1} X^T W Y$$

W = matrice dei pesi

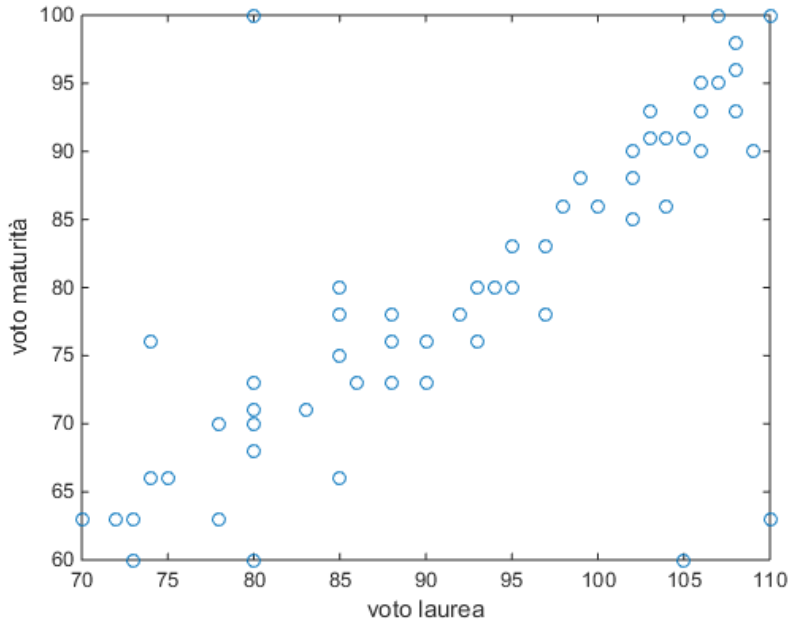
UN ESEMPIO DI APPLICAZIONE

Correlazione e regressione lineare

- Scopo dell'analisi è studiare la relazione tra due variabili X e Y.

- I due metodi più utilizzati sono l'indice di correlazione di Pearson (*parametrico*) e l'indice di correlazione per ranghi di Spearman (*non parametrico*)

- Per prima cosa, per avere una "prima idea", sul tipo di relazione esistente tra le due variabili X e Y è utile rappresentare graficamente tali variabili su un piano cartesiano attraverso il cosiddetto grafico di dispersione. Ad esempio:



IL COEFFICIENTE DI CORRELAZIONE DI PEARSON

Il coefficiente di correlazione di Pearson misura il tipo e l'intensità della relazione lineare tra due variabili X e Y. In altre parole esprime un'eventuale relazione di linearità tra esse.

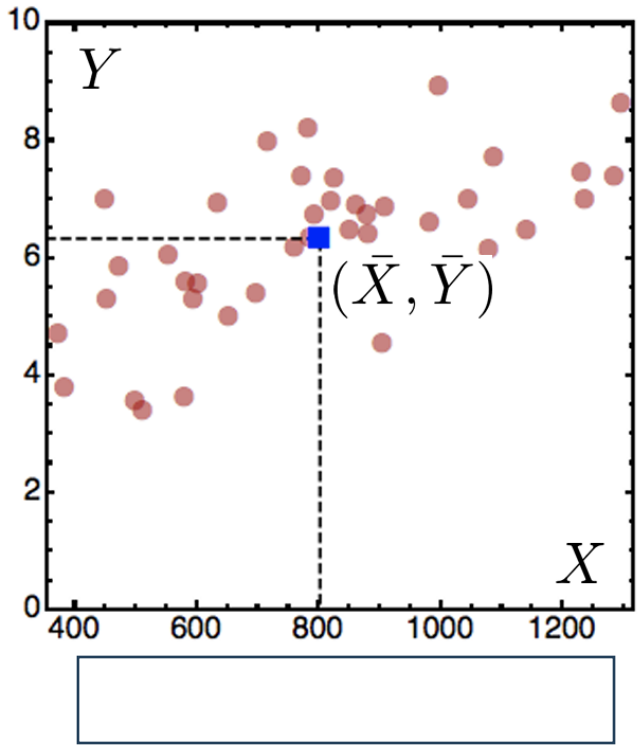
Date due variabili statistiche X e Y, l'indice di correlazione di Pearson (r) è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili:

$$r = \frac{cov(X,Y)}{dev.st(X)dev.st(Y)} \qquad -1 \leq r \leq 1$$

Misura la tendenza di due variabili numeriche a variare assieme (**co-variare**)

$$X = \{x_1, x_2, ..., x_n\}$$

$$Y = \{y_1, y_2, ..., y_n\}$$



$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}$$

IL COEFFICIENTE DI CORRELAZIONE DI PEARSON

Il coefficiente di correlazione di Pearson misura il tipo e l'intensità della relazione lineare tra due variabili X e Y. In altre parole esprime un'eventuale relazione di linearità tra esse.

Date due variabili statistiche X e Y, l'indice di correlazione di Pearson (r) è definito come la loro covarianza divisa per il prodotto delle deviazioni standard delle due variabili:

$$r = \frac{cov(X,Y)}{dev.st(X)dev.st(Y)} \qquad -1 \leq r \leq 1$$

se r > 0, le variabili X e Y si dicono direttamente correlate, oppure correlate positivamente;
se r = 0, le variabili X e Y si dicono incorrelate;
se r < 0, le variabili X e Y si dicono inversamente correlate, oppure correlate negativamente

A volte i risultati sono dati in termini di R^2:

Coefficiente di determinazione $R^2 = r^2$

Se $R^2 = 1$: tutti i valori osservati giacciono sulla retta di regressione
Se $R^2 = 0$: non c'è relazione lineare fra x e y

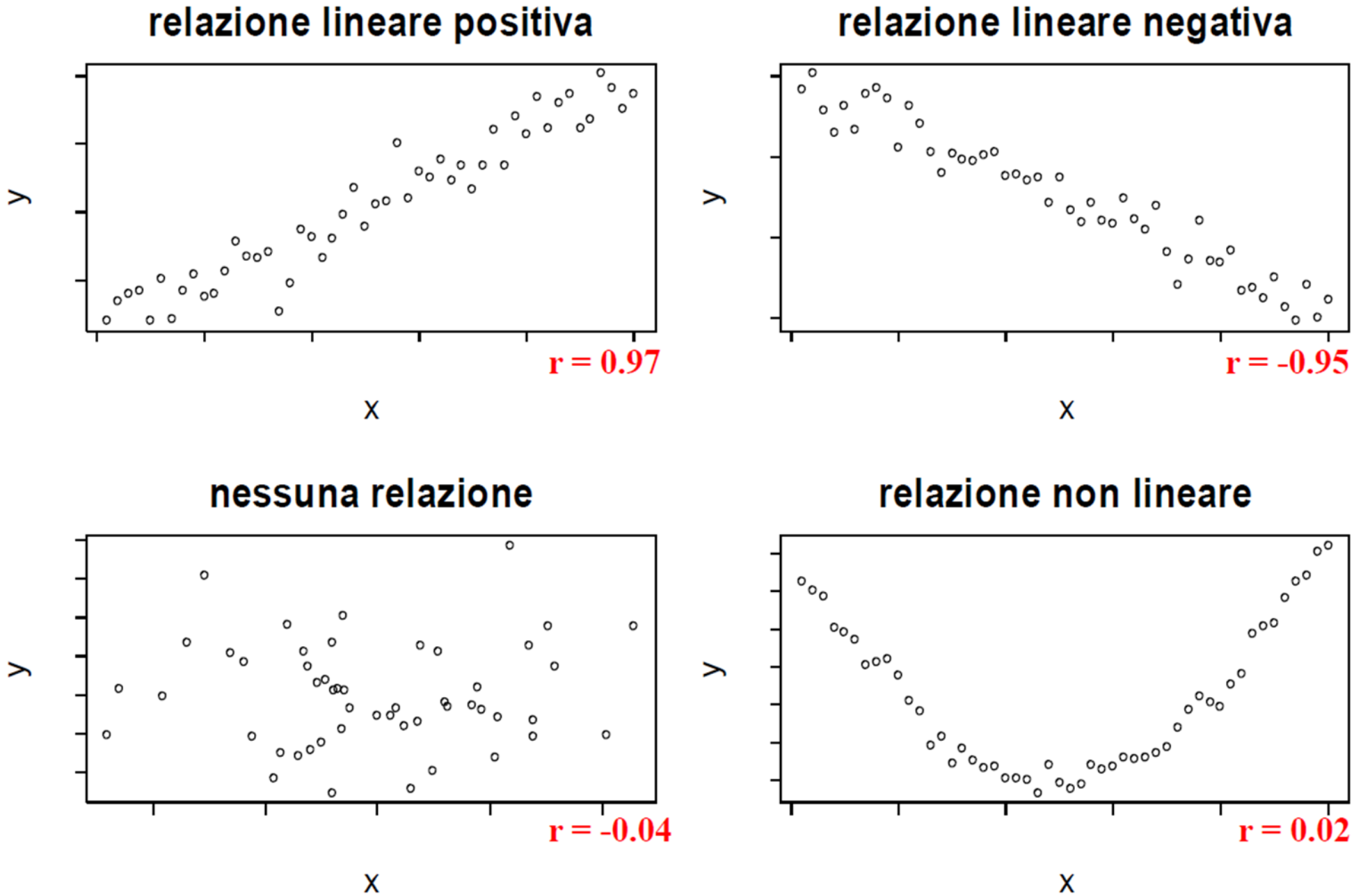
R^2 rappresenta la proporzione di variabilità tra i valori osservati di Y spiegata dalla regressione lineare di Y su X

Ad esempio: $r=0.67 \Rightarrow R^2=0.45 \Rightarrow$ il 45% della variazione di Y è spiegato dalla relazione lineare fra Y e X, il 55% della variazione di Y non è spiegato da questa relazione.

Per **r** invece abbiamo che:

Il segno di r (+ o -) da informazioni sul tipo di relazione:
- il segno positivo indica che le due variabili aumentano o diminuiscono assieme (relazione lineare positiva)
- il segno negativo indica che all'aumentare di una variabile l'altra diminuisce e viceversa (relazione lineare negativa)

Il valore assoluto di r, che varia tra 0 e 1, da informazioni sulla forza della relazione lineare:
- è massimo (assume valore 1) quando esiste una perfetta relazione lineare tra le due variabili.
- tende a ridursi al diminuire dell'intensità della relazione lineare e assume il valore 0 quando essa è nulla.


relazione lineare positiva
y
x
r = 0.97
relazione lineare negativa
y
x
r = -0.95
nessuna relazione
y
x
r = -0.04
relazione non lineare
y
x
r = 0.02

L’ipotesi sono:

H0: r = 0

La statistica del test è:

$$t = \frac{cov(X,Y)}{dev.st(X)dev.st(Y)} \sqrt{\frac{n-2}{1-\left(\frac{cov(X,Y)}{dev.st(X)dev.st(Y)}\right)^2}} \qquad \text{n} = \text{numero dati}$$

la t-student è definita da n-2 gradi di libertà

Esercizio: nel file voti.mat sono contenute le seguenti variabili rappresentative di un campione di 60 neolaureati

- Voto di Laurea (0-110)
- Voto di Maturità (0-60)
- Voto ottenuto al Test di Ingresso all'Università (0-100)

Si vuole sapere se c'è sufficiente evidenza statistica ad un livello di significatività 5% per dire che c'è un legame lineare tra le due variabili Voto di Laurea e Voto di Maturità.

Soluzione:

correlazione di Pearson r = 0.7531
t = 8.7185
Valore della t-student a t= 8.7185 ➔ 3.7019e-12 (p_value)

Risposta:

C'è un legame lineare positivo tra le due variabili con significatività statistica.

Possiamo formulare il problema anche in questo modo alternativo: riteniamo che ci sia una dipendenza causale e lineare della variabile Y dalla X.
Ossia:

$$Y = \beta_1 X + \beta_2 \qquad \boldsymbol{\varepsilon} \sim N(0, I)$$

Risolviamo con i minimi quadrati e, se esplicitiamo la formula generale di stima per i due parametri, troviamo:

$$\beta_1 = \frac{\sum_i (x_i y - x_i \bar{Y})}{\sum_i \left(x_i^2 - x_i \bar{X}\right)} = \frac{Cov(XY)}{Var(X)}$$

Si noti la relazione tra il coefficiente di correlazione di Pearson e il parametro β_1:

$$\beta_1 = \frac{\sum_i (x_i y_i - x_i \bar{Y})}{\sum_i \left(x_i^2 - x_i \bar{X}\right)} = \frac{\sum_i (x_i - \bar{X})\ (y_i - \bar{Y})}{\sum_i (x_i - \bar{X})^2} = \frac{Cov(XY)}{Var(X)}$$

$$r = \frac{\sum_i (x_i - \bar{X})\ (y_i - \bar{Y})}{\sqrt{\sum_i (x_i - \bar{X})^2}\sqrt{\sum_i (y_i - \bar{Y})^2}} = \frac{Cov(XY)}{STD(X)STD(Y)}$$

Dunque:

$$r = \beta_1 \frac{STD(X)}{STD(Y)}$$

e si può dimostrare (non riportato) che la significatività statistica della correlazione è equivalente alla significatività statistica del parametro β_1 (pendenza) della regressione.

RITORNIAMO ALLA REGRESSIONE LINEARE GENERALE

Come misura della bontà della regressione si utilizza il cosidetto **coefficiente di determinazione**:

$$R^2 = 1 - \frac{varianza\ spiegata\ dalla\ regressione}{varianza\ di\ Y}$$

$$= 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{Y})^2}$$

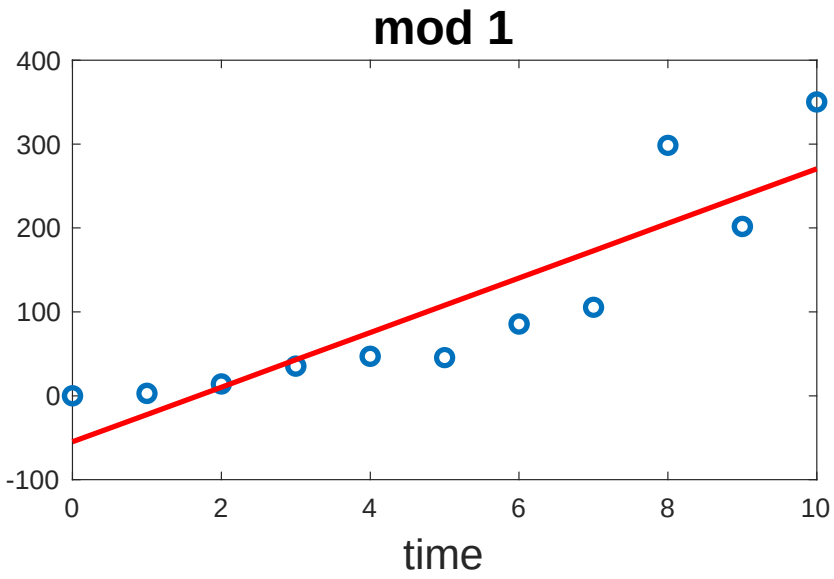
Varia tra 0 e 1 e esprime la frazione di varianza spiegata dalla regressione sul totale della varianza del fenomeno in studio

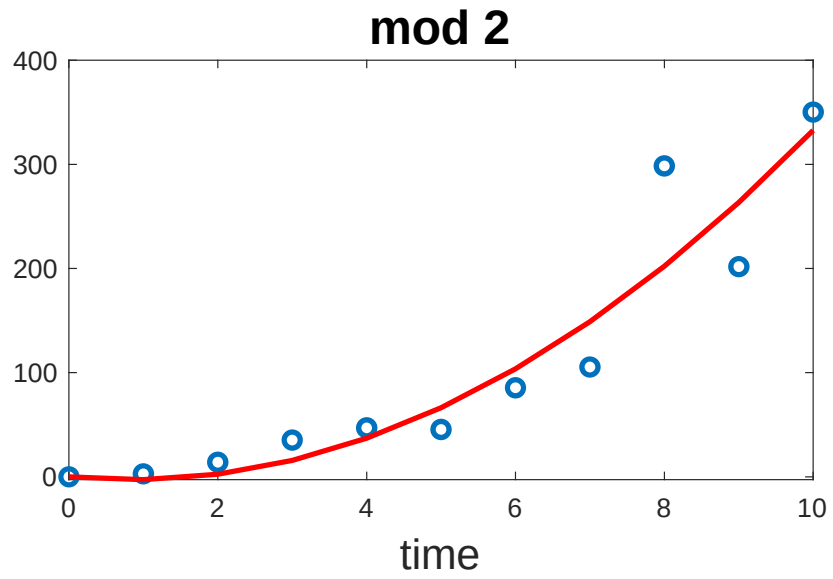
Risultati dell'identificazione numerica

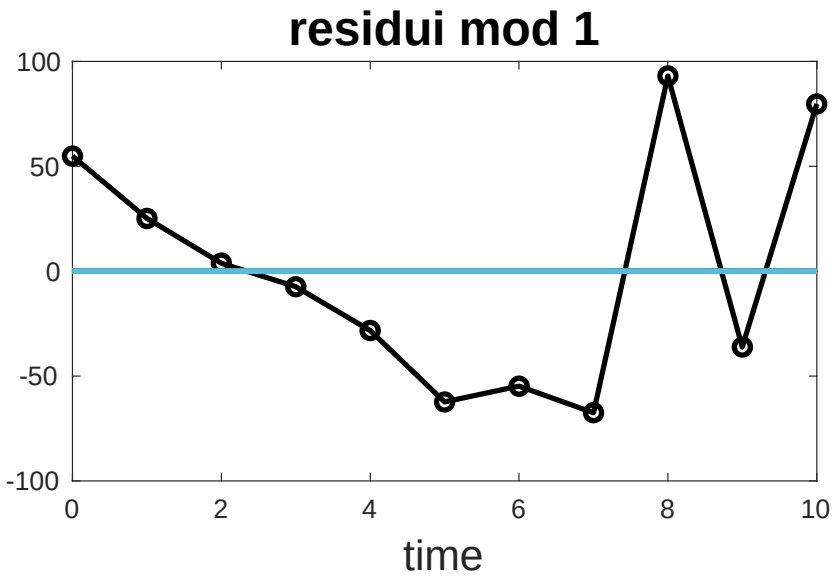
Trovato le stime dei parametri e quindi determinato il modello come si valuta la bontà dei risultati dell'identificazione ?

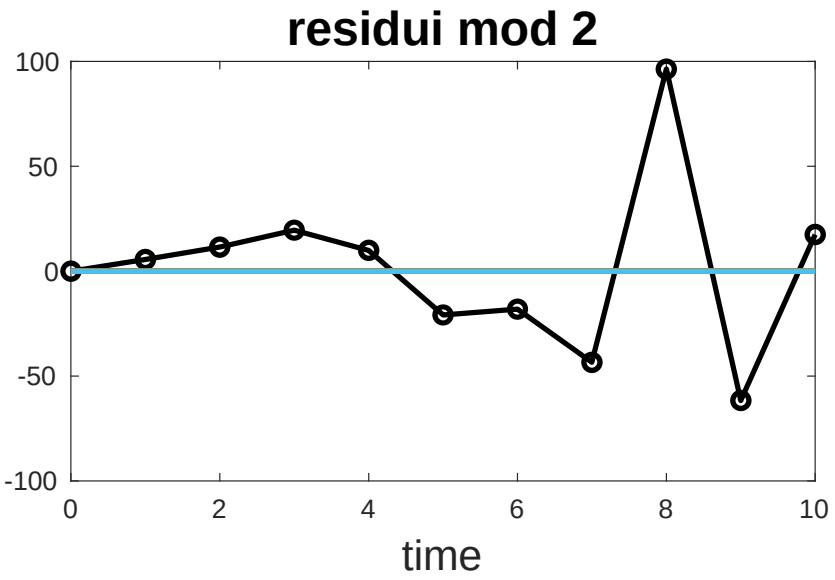
1. Analisi dei residui: se il modello (=struttura + valore numerico del parametri) è "buono" è logico attendersi che i residui siano compatibili con le proprietà statistiche dell'errore di misura

ESEMPIO: STESSI DATI MA DUE MODELLI LINEARI DIVERSI









I residui hanno media nulla?

Si può calcolare la media e verificare che l'assunzione originale sia giusta.

I residui sono scorrelati?

Possiamo usare il calcolo dell'autocorrelazione e la sua rappresentazione grafica come correlogramma. Per fare questo creiamo idealmente una tabella:

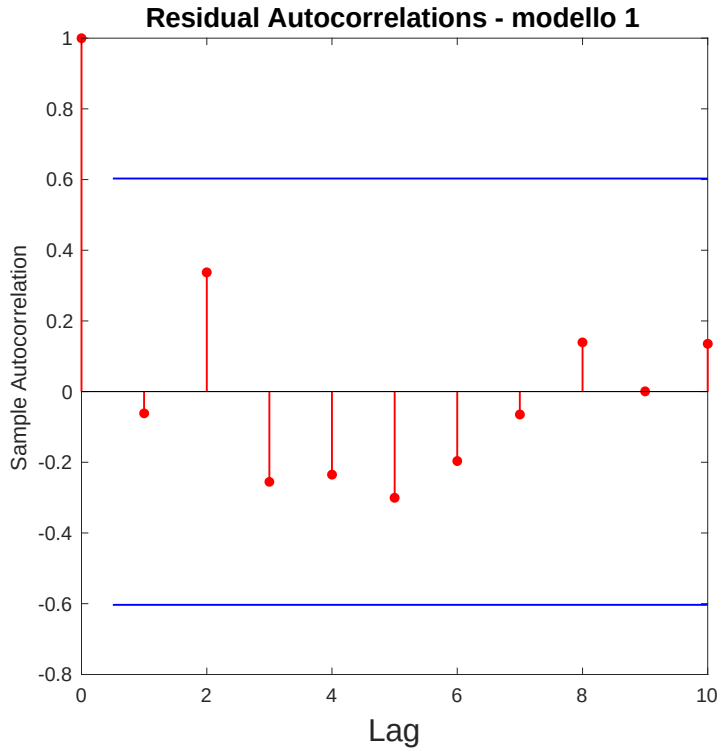
$$\begin{matrix} \text{lag0} & \text{lag1} & \text{lag2} & \text{lag3...lagK} \end{matrix}$$
$$\begin{bmatrix} \varepsilon_1 & & & & \\ \varepsilon_2 & \varepsilon_1 & & & \\ \varepsilon_3 & \varepsilon_2 & \varepsilon_1 & & \\ \vdots & \vdots & \vdots & \vdots & \\ \varepsilon_n & \varepsilon_{n-1} & \varepsilon_{n-2} & \vdots & \varepsilon_1 \end{bmatrix}$$

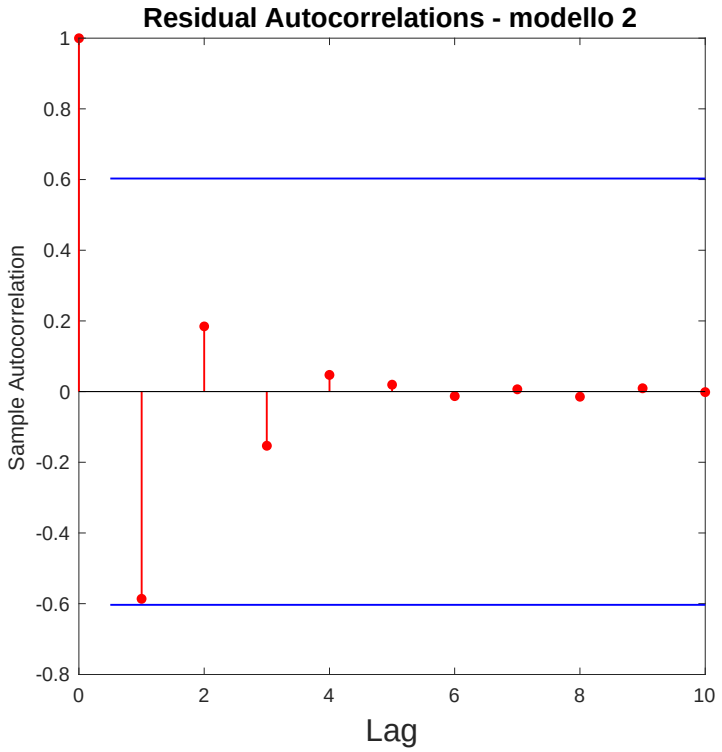
E calcoliamo tutte le correlazioni tra la prima colonna e tutte le altre K colonne:

- si parte sempre dalla (K+1)-esima riga in modo da confrontare sempre serie di uguale lunghezza
- K, il valore massimo di k, non è maggiore di $n/4$ o $n/2$ (dipende dalla numerosità), al fine di non ridurre troppo il numero di confronti

I valori poi sono rappresentati su un grafico.

A esempio per i nostri dati:





Se valori variano ma oscillano sempre entro una banda ristretta ciò vuol dire che gli errori non sono significativamente correlati con le serie ritardate, ovvero che il passato non "spiega" il presente e che le variazioni da un istante o periodo ad un altro sono sostanzialmente casuali.

La banda è determinata dallaBartlett two-standard-error bands for white noise, given by the blue lines.

La banda è determinata dall'applicazione della formula di Bartlett's: nello stesso grafico vengono rappresentati i limiti superiori e inferiori per un valore di significatività statistica α dei valori di autocorrelazione

$$Banda = \pm(1-\alpha) \cdot SE(r_{lag\_k})$$

Dove r_{lag_k} è l'autocorrelazione calcolata la lag k esimo.
SE è ottenuto tramite:

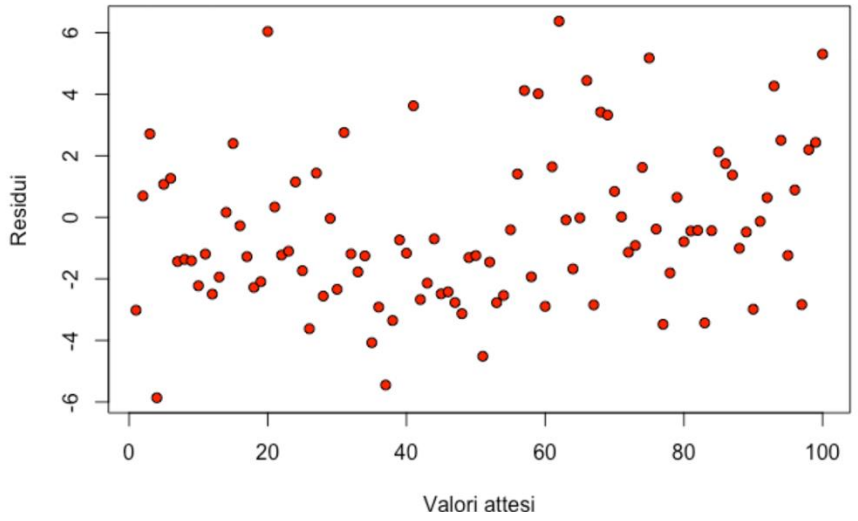
$$SE(r_{lag_0}) = \frac{1}{\sqrt{n}}$$

$$SE(r_{lag_k}) = \sqrt{\frac{1 + 2\sum_{i=1}^{K} r_{lag\_i}^2}{N}}$$

L'assunzione è che i residui siano variabili aleatorie gaussiane.
Se i valori di autocorrelazione sono contenuti nelle bande allora accettiamo l'ipotesi nulla (H0: non c'è autocorrelazione) altrimenti la rifiutiamo.

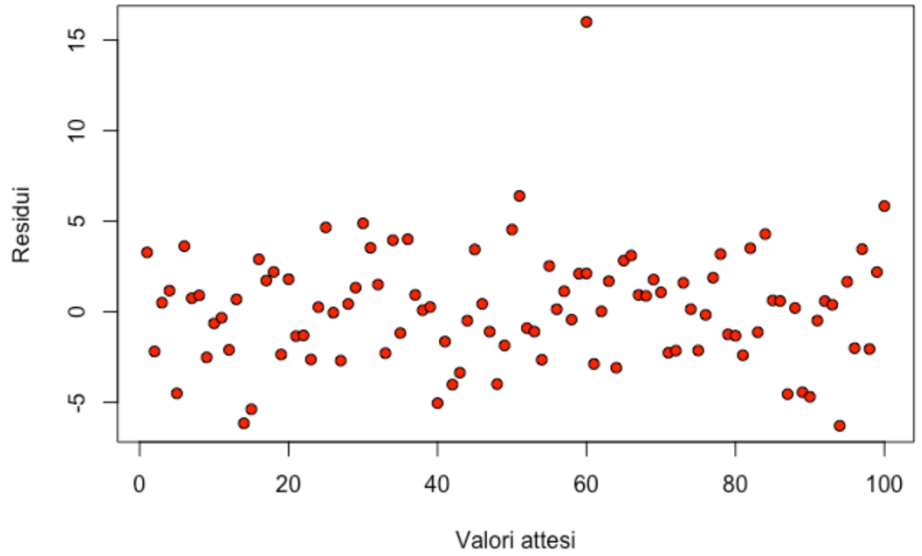
Ci sono degli outliers che influenzano la stima?

Il grafico a dispersione tra valori predetti e residui permette di individuare anche i possibili outliers, ovvero i punti isolati nel grafico. Se non vi sono problemi, i punti nel grafico dovrebbero essere distribuiti in modo assolutamente casuale:

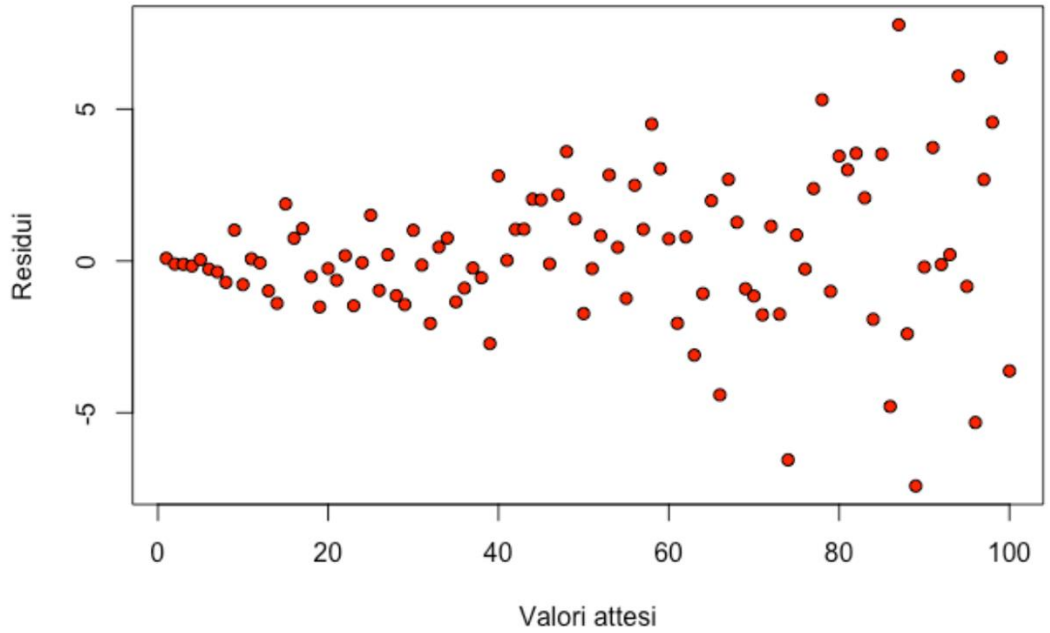


➔ Nessun outlier

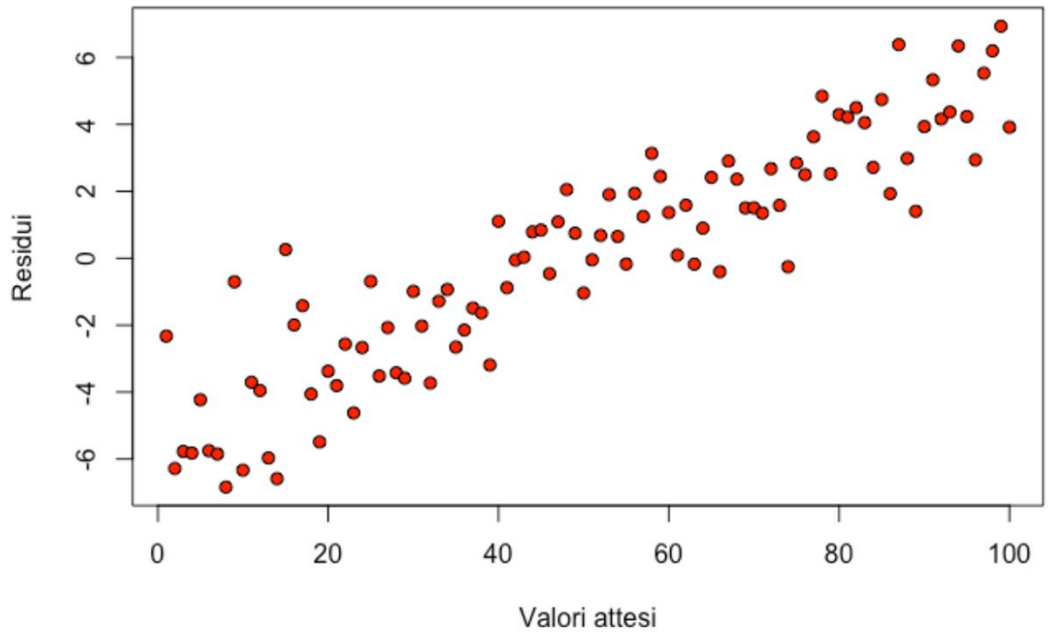
Invece in questo caso



➔ un outlier evidente



➔ I residui hanno una varianza eterogenea ossia non costante



➔ l'andamento non è più casuale, ma mostra un qualche evidente pattern. Ad esempio, nella figura i residui sono tendenzialmente negativi per bassi valori attesi e positivi per alti valori, chiaro segnale che il modello sovrastima le osservazioni quando sono basse e le sottostima quando sono alte.

Risultati dell'identificazione numerica

Trovato le stime dei parametri e quindi determinato il modello come si valuta la bontà dei risultati dell'identificazione ?

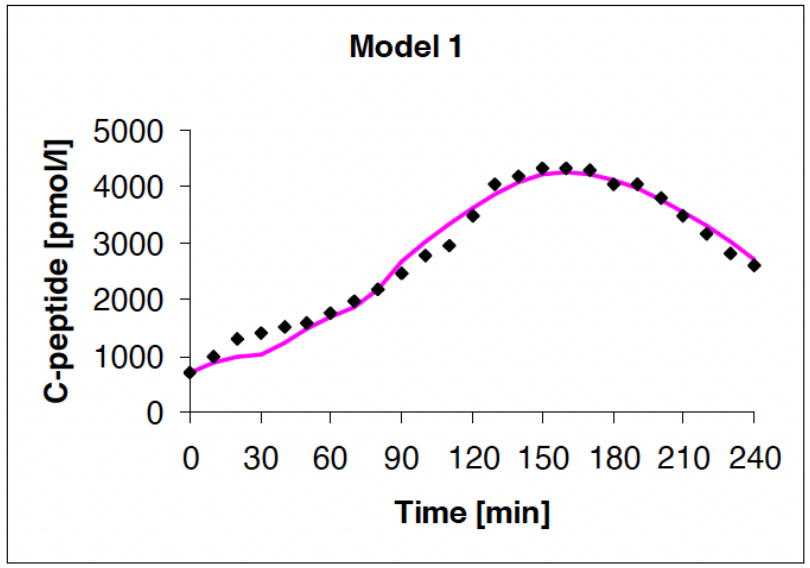
1. Analisi dei residui: se il modello (=struttura + valore numerico del parametri) è "buono" è logico attendersi che i residui siano compatibili con le proprietà statistiche dell'errore di misura

2. Analisi della varianza dell'errore di stima dei parametri

$$Var\left(\hat{\beta}^{LS}\right) = (X^T X)^{-1} \quad \rightarrow \quad CV\% = \frac{\sqrt{diag\left((X^T X)^{-1}\right)}}{\hat{\beta}^{LS}} \times 100$$

Se la deviazione standard dell'erroe di predizione delle stime è troppo elevate allora i CV% avranno valori elevate, ad es: >100%

3. Analisi della bontà della predizione (strettamente legata all'analisi dei residui):



= è un buon fit?

Per rispondere in modo quantitativo dobbiamo prima introdurre queste osservazioni basate sul calcolo della devianza della Y ossia della variabile da predire:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left(y_i - y_i^{pred} + y_i^{pred} - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_i - y_i^{pred}\right)^2 + \sum_{i=1}^{n}\left(y_i^{pred} - \bar{y}\right)^2$$

Questo perché l'errore di misura NON è correlato con $Y = X\beta + \varepsilon$

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\textbf{SST}} = \underbrace{\sum_{i=1}^{n}\left(y_i - y_i^{pred}\right)^2}_{\textbf{SSE}} + \underbrace{\sum_{i=1}^{n}\left(y_i^{pred} - \bar{y}\right)^2}_{\textbf{SSR}}$$

$\bar{y}$ = media dei dati in Y

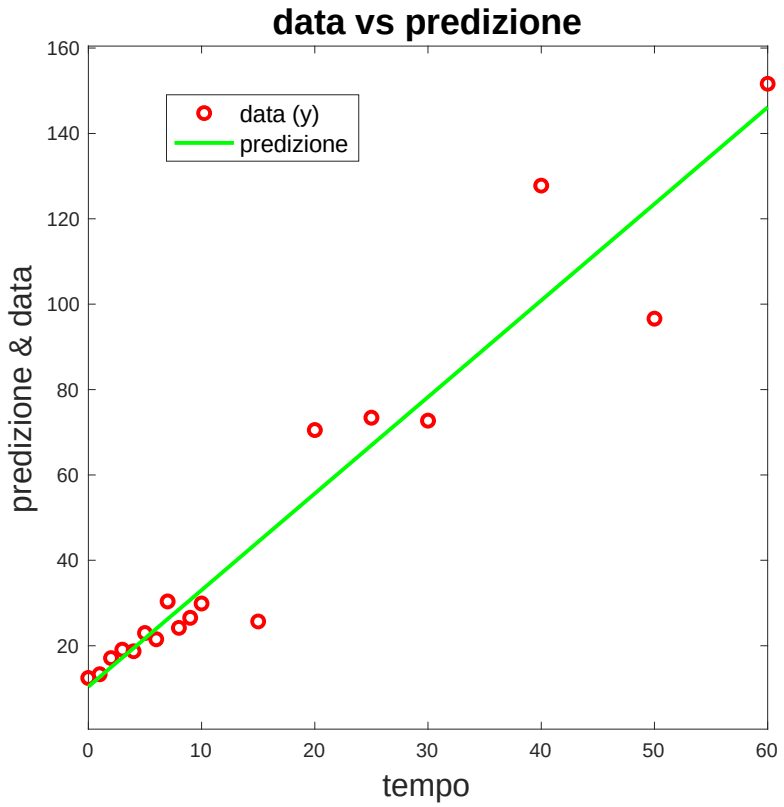
Dividendo le devianze per i gradi di libertà, otteniamo, nel caso in cui i predittori comprendono il β_0 $(ossia\ sono\ in\ totale\ m+1)$, le varianze:

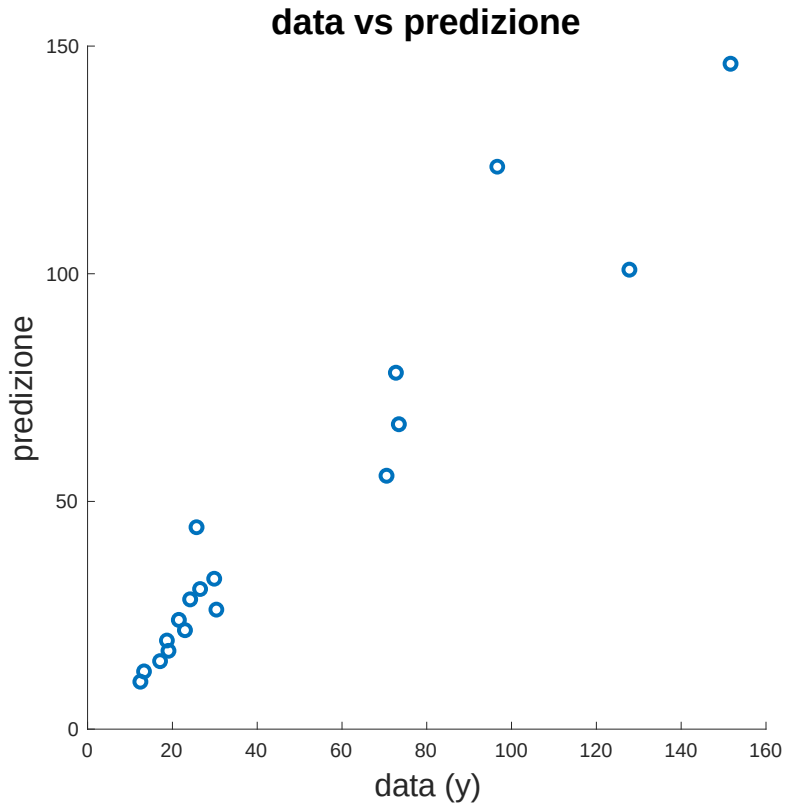
$$var(Y) = \frac{\textbf{SST}}{n-1} = s_Y^2$$

$$var(\epsilon) = \frac{\textbf{SSE}}{n-m-1} = s_\varepsilon^2$$

$$var\left(Y^{pred}\right) = var(modello) = \frac{\textbf{SSR}}{m} = s_{pred}^2$$

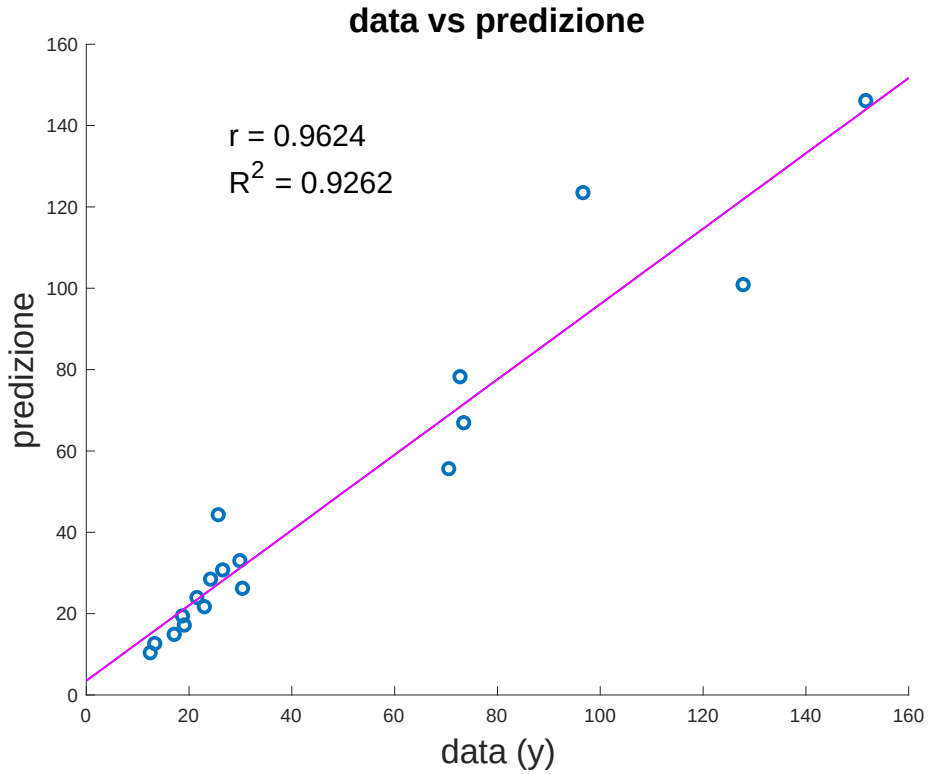
Possiamo rappresentare la predizione del modello (Xβ) verso il dato originale (Y) in due modi:





Consideriamo la rappresentazione tramite scatter plot: più la predizione del modello si avvicina ai dati acquisiti, più lo scatter plot dei dati si avvicina ad una bisettrice.

Possiamo quindi calcolare la correlazione tra Y e predizione (r) e l'indice di determinazione (R^2):



Ricordiamo che R^2 rapprsenta la proporzione di variabilità tra i valori osservati di Y (SST) spiegata dalla regressione lineare di Y su X (SSR):

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SST - SSE}{SST} = \frac{SSR}{SST}$$

Per la sua facilità di calcolo, R^2 viene spesso usato come grandezza per valutare la bontà del modello: attenzione valuta solo l'aderenza dei dati ottenuta con il modello, non dice se il modello è il migliore

Quando R quadro è significativo?

Per capire se il coefficiente di determinazione è statisticamente significativo si usa un test statistico, in particolare il test sulla varianza.

Se Y è quindi rappresentabile come variabile aleatoria gaussiana e le osservazioni (yi) sono indipendenti:

$$F = \frac{s^2_{pred}}{s^2_{\varepsilon}}$$

Segue la distribuzione F di Fisher con $F(m, n-m-1)$.

Se il p-value relativo al test F è molto basso, allora puoi affermare che l'R^2 è statisticamente significativo.
Se invece il valore del p-value del test F è oltre la soglia prefissata allora si dice che l'R^2 non è statisticamente significativo.

Le variabili indipendenti sono incorrelate con l'errore?

Se una variabile X_i è correlata con il termine d'errore, si potrebbe utilizzare X_i per predire quale sarà l'errore del modello di misura/regressione. Questo non è un buon indice perché la componente di errore di un modello di previsione deve essere imprevedibile.

Per verificare questo, si generano per ogni variabile X_i uno scatter plot ➔ Sull'asse orizzontale si mettono i valori di X_i, mentre sull'asse verticale i valori dei residui.

Si calcola per ogni scatter plot il valore dell'indice di determinazione R^2.

Le variabili indipendenti sono incorrelate con l'errore?

Possiamo a questo punto capire se le variabili $X_1, X_2, \dots, X_m$ abbiano tutte un sufficiente potere predittivo. Possiamo anche calcolare:

$$R^2_{Y-X_1} = tra\ Y\ e\ la\ predizione\ ottenuta\ con\ un\ modello\ che\ include\ la\ sola\ variabile\ X_1$$

$$R^2_{Y-X_1,X_2} = tra\ Y\ e\ il\ modello\ che\ include\ X_1 \text{ e } X_2$$

ecc… sappiamo che:

$$R^2_{Y-X_1} \leq R^2_{Y-X_1,X_2} \leq \cdots \leq R^2_{Y-X_1,X_2,\dots,X_m}$$

Possiamo pensare di investigare se eliminare dal modello un gruppo di variabili, ad esempio chiederci se un gruppo di variabili, ad esempio: $X_{i+1}, X_{i+2}, \dots, X_m$ apportano un aumento significativo all'indice di devianza R^2.

Per far questo costruiamo una statistica F:

$$F = \frac{\left(R^2_{Y-X_1,X_2,\dots,X_m} - R^2_{Y-X_1,X_2,\dots,X_i}\right)/(m-i)}{\left(1 - R^2_{Y-X_1,X_2,\dots,X_m}\right)/(n-m-1)}$$

con l'ipotesi nulla.

$$H0:\ \beta_{i+1} = \beta_{i+2} = \cdots = \beta_m = 0$$

In pratica si testa se la differenza in R2 tra il modello con tutti gli m regressori e quella del modello con solo i regressori è simile.

Se l'ipotesi nulla NON viene rifiutata, allora si può optare per il modello con solo i regressori.

La pdf di riferimento è una v.a. $F(m-i, n-m-1)$

I residui hanno una distribuzione normale?

La distribuzione normale degli errori può essere verificata attraverso un grafico dei quantili, detto anche q-q plot (quantile-quantile plot).

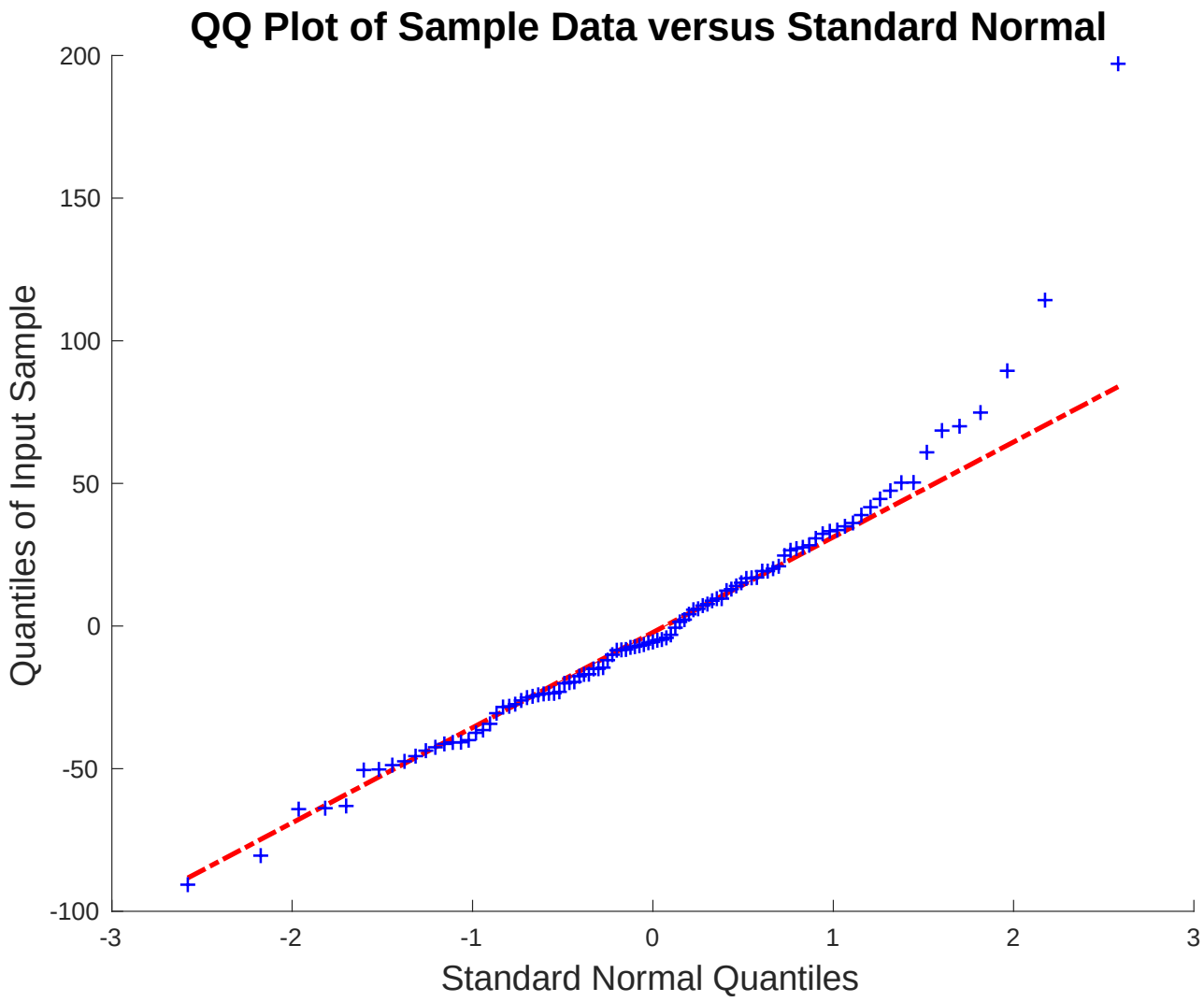
Il q-q plot seleziona I quantili sulla base del numero di valori che abbiamo ossia n.

Ricordiamo che i quantili sono espressi in frazione. I quantili di ordine 1/n, dividono la popolazione in n parti ugualmente popolate:

la mediana è il quantile di ordine 1/2 (50%).
I quartili sono i quantili di ordine 1/4, 2/4 e 3/4 (25%, 50% e 75%)

In questa tipologia di grafico, **i quantili teorici** di una distribuzione Normale sono riportati **sull'asse orizzontale.** I **quantili dei residui** sono invece riportati **sull'asse verticale.**

L'idea è che se i residui hanno una distribuzione normale, i loro quantili dovrebbero seguire una linea retta con quelli della distribuzione Normale. A livello visivo, questo significa che i punti dovrebbero disporsi lungo la retta.


QQ Plot of Sample Data versus Standard Normal
Quantiles of Input Sample
Standard Normal Quantiles
200
150
100
50
0
-50
-100
-3
-2
-1
0
1
2
3

Multicollinearità tra variabili

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

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

Esempio: In un modello di regressione

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

se X_2 è in relazione lineare con X_1 e quindi sussiste una relazione del tipo:

$$X_2 = a + bX_1$$

le due variabili sono perfettamente correlate.

In un caso del genere lo stimatore:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

con $X = [X_1\ X_2]$ non dà risultati attendibili. Infatti se c'è una correlazione perfetta tra le due variabili, la matrice $(X^T X)^{-1}$ diventa singolare ha determinante uguale a 0 e perciò non esiste la matrice inversa.

Se le variabili sono fortemente correlate vuol dire che danno la stessa informazione e il modello di regressione non riesce più ad attribuire una stima ai β ossia ai parametri del modello:

1. gli errori di stima dei coefficienti di regressione (β) delle variabili correlate diventano più ampi; le stime, cioè, diventano meno precise;
2. i coefficienti (β) sono instabili e possono cambiare (anche di segno) in seguito a lievi modifiche della struttura del modello.

Soluzione:
eliminare dal modello le variabile esplicative che risultano combinazione lineare delle alter. In alter parole, eliminare la collinearità tra le variabili.

Diagnosticare la quasi multicollinerità (QM). Sono stati sviluppati molti metodi:

TABLE A. Listing of collinearity diagnostics

Diagnostic	Description, formula and cutoff	Criteria	References
Correlation Matrix	High zero order or Pairwise correlation between regressors	$r_{ij} > 0.8$	Adnan et al. 2006; Gujarati & Porter 2008; Maddala 1988
Determinant	Determinant of normalized correlation matrix without intercept, while $0 \le \lvert X'X \rvert \le 1$	$\lvert X'X \rvert \sim 0$	Asteriou & Hall 2007
R^2, Var(β)'s & t-ratios	High R^2 value, conversely high variance of β's and low t-ratios		Gujarati & Porter 2008; Maddala 1988
Farrar χ^2	$\chi^2 = -\left[n-1-\frac{1}{6(2p+5)}\right] \times \log_e\left[X'X\right] \sim \psi^2_{v=\frac{1}{2}p(p-1)}$	$\chi^2 > \psi^2_{\frac{1}{2}p(p-1)}$	Farrar & Glauber 1967
Farrar w_i	$w_i = \frac{R_j^2}{1-R_j^2}\left(\frac{n-p}{p-1}\right) \sim F(n-p, p-1)$	$w_i > F_{(n-p,\,p-1)}$	Farrar & Glauber 1967
Klein's Rule	If $R^2_{xj.x_1,x_2,\ldots,x_p} > R^2_{y.x_1,x_2,\ldots,x_p}$, multicollinearity may be troublesome.		Klein 1962
VIF and TOL	$\left(X'X\right)^{-1}_{jj} = VIF_j = \frac{1}{1-R_j^2}$ $TOL_j = \frac{1}{VIF_j} = 1 - R_j^2$	VIF > 3, 5, 10 TOL ~ 0	Kutner et al. 2004; Marquardt 1970
Eigenvalues	Smaller eigenvalues of $X'X$ or its related correlation matrix indicate collinearity.	Relatively smaller than other eigenvalues	Kendall 1957; Silvey 1969
CI	$CI_j = \sqrt{\frac{\max(\lambda_j)}{\lambda_j}}; j = 1,2,\ldots,p;\ \lambda_1 \ge \lambda_2 \ldots \lambda_p$	$CI_j > 10, 15, 30$	Belsley 1980; Chatterjee & Hadi 2006; Maddala 1988
Sum of λ_j^{-1}	$\sum_{j=1}^{p} \frac{1}{\lambda_j}; j = 1,2,\ldots,p$	five times the number of predictors	Chatterjee & Hadi 2006; Dillon & Goldstein 1984
CVIF	$CVIF_j = VIF_j \times \frac{1-R^2}{1-R_0^2};$ $R_0^2 = R^2_{yx_1} + R^2_{yx_2} + \ldots + R^2_{yx_p}$	$CVIF_j \ge 10$	Curto & Pinto 2011
Leamer	$C_j = \left\{\frac{\left(\sum_i^n \left(X_{ij} - \bar{X}_j\right)^2\right)^{-1}}{\left(X'X\right)^{-1}_{jj}}\right\}^{\left(\frac{1}{2}\right)}$	$C_j \sim 0$	Greene 2002

Tra i più usati:

➔ Correlazione tra coppie di regressori

➔ VIF (Fattore di Inflazione della Varianza – Variance Inflation Factor)

➔ Tolerance Limit (TOL)

➔ Klein's rule of thumb: la multicolllinearità può essere un problema se Rj2 > R2 (R2 della regressione principale di Y su tutte le X)

➔ Valutazione degli autovalori

➔ CN (Numero di condizionamento – condition number)

➔ CI (Indice di condizionamento – condition index)

Diagnosticare la quasi multicollinerità (QM):

➔ **Correlazione tra coppie di regressori**: se una coppia ha una correlazione più alta di 0.8 allora è molto collineare. Caveat: alta multicollinearità può essere presente anche se la correlazione tra coppie di regressori è più bassa (meno di 0.5).

➔ VIF (Fattore di Inflazione della Varianza – Variance Inflation Factor)

Si considerino gli elementi sulla diagonale della la matrice $m \times m$ $(X^T X)^{-1}_{jj}$
e si calcoli:

$$VIF_j = (X^T X)^{-1}_{jj} = \frac{1}{1-R_j^2}$$

Si può dimostrare che, preso un elemento della diagonale della matrice $(X^T X)^{-1}$, allora questo corrisponde al valore di $1/\left(1 - R_j^2\right)$ con R_j^2 pari all'indice di determiniazione per il regressore X_j con tutti gli altri predittori (regressori) inclusa l'intercetta se presente nel modello.

Se VIF > 5 (o 10 a seconda di quanto restrittivi volgiamo essere) significa presenza di collinearità.

Nota: non esiste un valore univocamente accettato.

➔ Tolerance Limit (TOL):

$$TOL_j = \frac{1}{VIF_j} \qquad \text{con TOL}\sim 0 = multicollinearità$$

➔ Klein's rule of thumb: la multicolllinearità può essere un problema se la correlazione tra due variabili (rij) è > a quella che si ottiene dalla regressione di Y con la predizione del modello con tutte le variabili.
Ad esempio:

- Metric: collinearity harmful if any $r_{ij} \geq$ [or ≈] R_Y,
  where
  r_{ij} = simple correlation between X_i and X_j
  R_Y = multiple correlation coefficient between Y and all the X's

- Calculations:
  $R^2_{Y|X_1, X_2, X_3, X_4} = 0.95$
  $R_{Y|} = \sqrt{R^2} = 0.97$

- Result:
  - In all cases, $r_{ij} <= R_Y$
  - r_{X_1,X_4} closest to R_Y

Simple Correlation Matrix of X's

	X_1	X_2	X_3	X_4
X_1	1.00	0.77	-0.52	0.94
X_2		1.00	-0.26	0.73
X_3			1.00	-0.41
X_4				1.00

Source: Lawrence Klein, Nobel Laureate in Economics, 1980. "Intercorrelation ... is not necessarily a problem unless it is high relative to the over-all degree of multiple correlation ..."; *An Introduction to Econometrics*, Prentice-Hall, 1962, p.101.

Considerando che possiamo calcolare gli autovalori e autovettori della matrice $X^T X$. Siano $\lambda_1, \lambda_2, \ldots, \lambda_m$ gli autovalori.

➔ Valutazione degli autovalori (Kendall): L'idea è che se alcuni di questi sono bassi allora si può supporre l'esistenza di correlazioni forti tra colonne di X.

➔ CN (Numero di condizionamento – condition number)

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}} \quad \text{se } \kappa < 100 \; non\ ci\ sono\ problemi\ di\ multicollinearità$$

si considera presente collinearità se il valore è >1000.

➔ CI (Indice di condizionamento – condition index)
è semplicemente la radice del valore:

$$CI_i = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$

valori >10 sono considerati come presenza di collinearità

Soluzione:

eliminare dal modello le variabile esplicative che risultano combinazione lineare delle altre. In alter parole, eliminare la collinearità tra le variabili.

I Test sui parametri

Un ultimo controllo statistic può essere fatto sui parametri del modello. Per estensione di quanto abbiamo visto nell'analisi dei regressori, possiamo valutare l'ipotesi nulla:

$$\text{H0: } \beta_i = 0$$

La statistica test è:

$$t = \frac{|\hat{\beta}_i|}{s_e^2 \sqrt{c_{ii}}} \quad con\ c_{ii} = i - esimo\ elemento\ della\ diagonale\ di\ (X^T X)^{-1}$$

Per testare l'ipotesi nulla si confronta con una t di Student con $n - m - 1$ gradi di libertà.

PROPRIETA' DELLO STIMATORE DEI MINIMI QUADRATI LINEARI GENERALIZZATI

$$E[\varepsilon] = 0$$

$$\text{Var}(\varepsilon) = \sigma^2 V \text{ con V matrice diagonale nota, } \sigma^2 \text{ non noto}$$

1) Lo stimatore è unbias:

$$E\left[\hat{\beta}^{LS}\right] = E[(X^T V^{-1} X)^{-1} X^T V^{-1} Y] = \beta$$

2) Lo stimatore è preciso e ha varianza:

$$\text{Var}\left[\hat{\beta}^{LS}\right] = \sigma^2 (X^T V^{-1} X)^{-1}$$

con $\sigma^2 = \frac{\varepsilon^T \varepsilon}{n-m}$

Riassumendo gli step che sono normalmente implementati :

STIMA CON I MINIMI QUADRATI LINEARI CON IPOTESI SU ERRORE:

$$E[\varepsilon] = 0$$
$$\text{Var}(\varepsilon) = \sigma^2 \text{ con } \sigma^2 \text{ non noto}$$

1° step: si ottengono le stime $E\left[\hat{\beta}^{LS}\right] = E[(X^T X)^{-1} X^T Y]$

2° step: si calcolano i residui (corrispondono ai "Residuals raw" se si usa fitlm): $\text{residui} = Y - X\hat{\beta}^{LS}$

3° step: si calcola il valore di σ^2: $\sigma^2 = \frac{[Y - X\hat{\beta}^{LS}]^T [Y - X\hat{\beta}^{LS}]}{n-m}$ $n = numero\ dati, m = numero\ parametri$

a questo punto possiamo calcolare i residui "pesati" ossia $wresidui = \frac{residui}{\sqrt{\sigma^2}} = \frac{residui}{\sqrt{MSE}}$

questi corrispondono ai "Residuals Pearson" se si usa fitlm

4° step: si calcola la matrice di covarianza delle stime: $\text{Var}(\hat{\beta}^{LS}) = \sigma^2 (X^T X)^{-1}$ e lo SE: $\text{SE} = \sqrt{\sigma^2 (X^T X)^{-1}}$

5° step: tra i vari test di calcola il VIF come:

se le variabili in X sono normalizzate con zscore: $VIF = (n-1) * diag\left(inv(X' * X)\right)$

se le variabili in X **non sono normalizzate** con zscore: $VIF = (n-1) * diag(inv(X' * X)) .* var(X)^T$

STIMA CON I MINIMI QUADRATI LINEARI CON IPOTESI SU ERRORE:

$$E[\varepsilon] = 0$$

$$\text{Var}(\varepsilon) = \sigma^2 V \text{ con } \sigma^2 \text{ non noto ma V matrice diagonale nota}$$

1° step: si ottengono le stime $E\left[\hat{\beta}^{LS}\right] = E[(X^T V^{-1} X)^{-1} X^T V^{-1} Y]$

2° step: si calcolano i residui (corrispondono ai "Residuals raw" se si usa fitlm): $\text{residui} = Y - X\hat{\beta}^{LS}$

3° step: si calcola il valore di σ^2: $\sigma^2 = \frac{[Y - X\hat{\beta}^{LS}]^T [Y - X\hat{\beta}^{LS}]}{n-m}$ $n = numero\ dati, m = numero\ parametri$

a questo punto possiamo calcolare i residui "pesati" ossia $wresidui = \frac{residui}{\sqrt{\sigma^2}} = \frac{residui}{\sqrt{MSE}}$

questi corrispondono ai "Residuals Pearson" se si usa fitlm

4° step: si calcola la matrice di covarianza delle stime: $\text{Var}(\hat{\beta}^{LS}) = \sigma^2 (X^T V^{-1} X)^{-1}$ e lo SE: $\text{SE} = \sqrt{\sigma^2 (X^T V^{-1} X)^{-1}}$

5° step: tra i vari test di calcola il VIF come:

se le variabili in X sono normalizzate con zscore: $VIF = (n-1) * diag\left(inv(X' * V^{-1} * X)\right)$

se le variabili in X **non sono normalizzate** con zscore: $VIF = (n-1) * diag(inv(X' * V^{-1} * X)) .* var(X)^T$

SELEZIONE DI MODELLI IN COMPETIZIONE
(alternative al F-test)

Ipotesi: abbiamo 2 o più modelli (modello A, modello B, …, modello M), tutti superano i controlli sui residui, la predizione e la credibilità (precisione) delle stime. Vogliamo uno strumento che ci permetta di scegliere il migliore.

Potremmo pensare di usare il valore dell'indice di determinazione ottenuto confrontando variabile da predire e predizione. Però sappiamo che più repressori si considerano, più questo valore sarà alto (ossia migliore adesione del dato alla predizione).

Inoltre in presenza di pochi dati e un numero di regressori elevato (in relazione al numero di dati), R^2 tende ad essere ad avere un bias positivo ossia tende ad una valutazione ottimistica dell'abilità del mdodello ad adattarsi ai dati.

Per ovviare a tali inconvenienti è stato introdotto nel calcolo dell'indice di determinazione un coefficiente correttivo (R^2 aggiustato) che tiene conto del numero di variabili nel modello:

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

Dove $m = numero\ di\ regressori\ (escludendo\ l'intercetta)$ e $n = numero\ di\ dati$

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

Dove $m = numero\ di\ regressori\ (escludendo\ l'intercetta)$ e $n = numero\ di\ dati$

La correzione fa sì che se l'aumento di R^2 eccede la penalità indotta allora allora $R^2_{Adjusted}$ cresce. In caso contrario $R^2_{Adjusted}$ decresce.

Sebbene questo valore ci aiuta nella valutazione della bontà della predizione del modello in considerazione della numerosità dei dati e dei rgressori, i criteri in assoluto più usati sono due:

- AIC (Akaike Information Criterion)
- BIC (Bayesian Information Criterion)

due criteri che forniscono una misura della distanza tra il modello e la distribuzione teorica dei dati in relazione al numero di parametri stimati.

AIC e BIC bilanciano la complessità del modello con la capacità di descrivere i dati.

Nel caso di regressione multipla ossia di uso dei minimi quadrati lineari:

$$AIC = n \log \frac{\varepsilon^T \varepsilon}{n} + 2m = n \log \frac{SSE}{n} + 2m$$

Aderenza ai dati

Complessità del modello

Nel confronto tra modelli si preferisce quello che ha AIC minimo.
Lo stesso vale per BIC:

$$BIC = n \log \frac{\varepsilon^T \varepsilon}{n} + m \log n = n \log \frac{SSE}{n} + m \log n$$

Anche per BIC si sceglie il modello che da il valore più basso.

Spesso si usano entrambi. AIC tende a selezionare modelli più complessi di quelli reali, BIC ha un comportamento imprevedibile nel caso in cui il vero modello non sia presente tra quelli testati