Statistica

Esercitazione 16

Alfonso Iodice D'Enza
iodicede@unicas.it

Università degli studi di Cassino

Studio della relazione tra due variabili

Commonly Asked Questions

- Qual'è la relazione tra la spesa sostenuta per la promozione di un prodotto e il livello di vendite nel primo mese?
- Qual'è la relazione tra la concentrazione di alcool nel sangue e il tempo di reazione di un automobilista?
- Qual'è la relazione tra il voto di laurea conseguito dagli studenti di Economia e lo stipendio da loro percepito al primo impiego?

Regressione lineare semplice

Per studiare la dipendenza lineare di una variabile di risposta (o dipendente) da una variabile indipendente (regressore, predittore) si utilizza il modello di regressione lineare semplice: tale modello, stabilisce, a meno di variazioni casuali, una relazione lineare tra risposta e predittore.

Studio della relazione tra due variabili

Galton e la regressione verso la mediocrità

Nel 1888 Francis Galton, passeggiava in campagna riflettendo sul seguente problema:

- Qual'è la relazione tra le caratteristiche fisiche e psichiche di un figlio e quelle dei genitori?

La contraddizione

Inizialmente lui credeva che l'altezza di un figlio dovesse essere, in valore atteso (in media), uguale a quella del genitore dello stesso sesso.
Dunque si attendeva che metà dei figli di genitori alti fossero ancora più alti e metà dei figli di genitori bassi fossero ancora più bassi: le generazioni successive avrebbero dovuto avere persone sempre più alte (o più basse).
Questo **tuttavia non accadeva**, perchè le altezze osservate erano stabili di generazione in generazione.

Il temporale e la soluzione

Mentre si riparava da un temporale che aveva interrotto la sua passeggiata si rese conto che l'altezza di un figlio era, in valore atteso (in media), compresa tra quella del genitore dello stesso sesso e la media della popolazione. Dunque figli di genitori particolarmente alti (bassi) erano in media meno alti (bassi) dei rispettivi genitori. Questa tendenza, confermata dai dati osservati, Galton la definì regressione verso la mediocrità.

Modello di regressione lineare semplice

In molte applicazioni il ruolo delle variabili x ed Y non è lo stesso, in particolare, assegnato un certo valore al predittore x (indicato pertanto con la lettera minuscola), il valore che Y assume dipende in qualche modo da x. La relazione più semplice tra le variabili è quella lineare, e il modello corrispondente è

$$Y = \beta_0 + \beta_1 x;$$

tale modello presuppone che, stabiliti i parametri β_0 e β_1, sia possibile determinare esattamente il valore di Y conoscendo il valore di x: salvo eccezioni, questo non si verifica mai.

Il modello

Alla determinazione del valore di Y, oltre che la componente deterministica $\beta_0 + \beta_1 x$, concorre anche una componente casuale detta errore non osservabile ϵ, una variabile casuale con media 0

$$Y = \beta_0 + \beta_1 x + \epsilon.$$

Analogamente, la relazione di regressione lineare semplice può essere espressa in termini di valore atteso

$$E[Y|x] = \beta_0 + \beta_1 x.$$

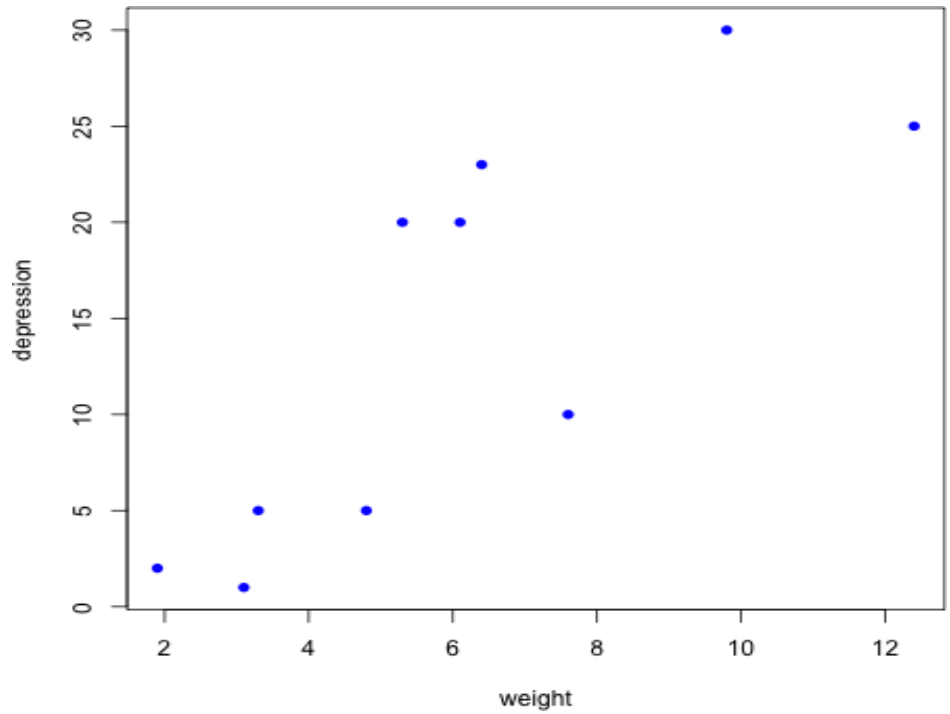
poichè $E[\epsilon] = 0$.

Modello di regressione lineare semplice

Si consideri di voler analizzare la relazione tra il peso del rullo di un taglia erba e l'entità della depressione riscontrata nel prato da tagliare. Sia Y la depressione (depression) e x il peso del rullo utilizzato (weight). Per vedere se l'utilizzo del modello di regressione lineare semplice sia ragionevole in questo caso occorre raccogliere delle coppie di osservazioni (x_i, y_i) e rappresentarle graficamente attraverso il diagramma di dispersione.

units	weight	depression
1	1.9	2.0
2	3.1	1.0
3	3.3	5.0
4	4.8	5.0
5	5.3	20.0
6	6.1	20.0
7	6.4	23.0
8	7.6	10.0
9	9.8	30.0
10	12.4	25.0

Il diagramma di dispersione (scatter plot)



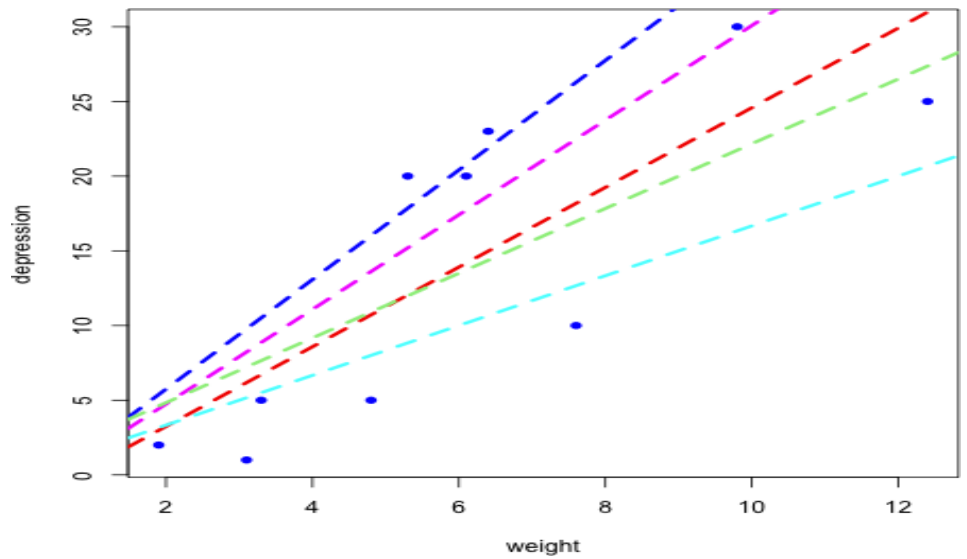
La retta di regressione

La retta di regressione

La retta di regressione fornisce una approssimazione della dipendenza dei valori di Y dai valori di X. La relazione di dipendenza non è esattamente riprodotta dalla retta; i valori $\hat{y}_i = \beta_0 + \beta_1 x_i$ sono dunque i valori teorici, ovvero i valori che la variabile Y assume, secondo il modello $Y = \beta_0 + \beta_1 x$, in corrispondenza dei valori x_i osservati.
Le differenze e_i tra i valori teorici $\hat{y}_i$ e i valori osservati y_i vengono definite residui. Questo perchè per ciascuna osservazione il modello è dato da

$$y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{comp. deterministica}} + \underbrace{\epsilon_i}_{\text{comp. casuale}}$$

rette passanti per la nube di punti



Determinazione della retta di regressione

L'identificazione della retta avviene attraverso la determinazione dei valori di b_0, e b_1, stime dell'intercetta e del coefficiente angolare o pendenza, rispettivamente. La retta 'migliore' è quella che passa più 'vicina' ai punti osservati. In altre parole, si vuole trovare la retta per la quale le differenze tra i valori teorici $\hat{y}_i$ e i valori osservati y_i siano minime.

La retta di regressione

Metodo dei minimi quadrati

La retta di regressione è tale che la somma dei residui al quadrato sia minima. Formalmente

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Il problema consiste dunque nel ricercare b_0 e b_1 che minimizzano la precedente espressione. Da un punto di vista operativo bisogna risolvere il seguente sistema di equazioni (condizioni del primo ordine o stazionarietà).

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial b_1} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

Nota: si tratta di punti di minimo perchè le derivate seconde $\partial_{b_0 b_0} f(b_0, b_1) = -2(-n)$, $\partial_{b_1 b_1} f(b_0, b_1) = -2 \sum_i^n (-x_i^2)$ sono sempre non negative.

Stimatori dei parametri della retta di regressione: (b_0)

$$-2 \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) =$$

$$\sum_{i=1}^{n} y_i - n * b_0 - b_1 \sum_{i=1}^{n} x_i = 0$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

La retta di regressione

I residui

le differenze tra i valori teorici $\hat{y}_i$ e i valori osservati y_i vengono definite residui. La retta di regressione è tale che la somma dei residui al quadrato sia minima. Formalmente

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

$$= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Il problema consiste dunque nel ricercare b_0 e b_1 che minimizzano la precedente espressione. Da un punto di vista operativo bisogna risolvere il seguente sistema di equazioni (condizioni del primo ordine o stazionarietà).

$$\frac{\partial}{\partial b_0} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

$$\frac{\partial}{\partial} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = 0$$

Stimatori dei parametri della retta di regressione: (b_1)

$$-2 \sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - b_0 \sum_{i=1}^{n} x_i - b_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \left(\frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} \right)$$

$$b_1 \left(n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 \right) = n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} = \frac{\sigma_{xy}}{\sigma_x^2}$$

Determinazione della retta di regressione

...statistiche descrittive

- $\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 6.07 \quad \bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = 14.1$
- $s_x = \sqrt{\frac{\sum_{i=1}^{10} (x_i - \bar{x})^2}{10}} = 3.04 \quad s_y = \sqrt{\frac{\sum_{i=1}^{10} (y_i - \bar{y})^2}{10}} = 10.1$
- $s_{xy} = \frac{\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y})}{10} = 24.7$
- $r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = 0.8$

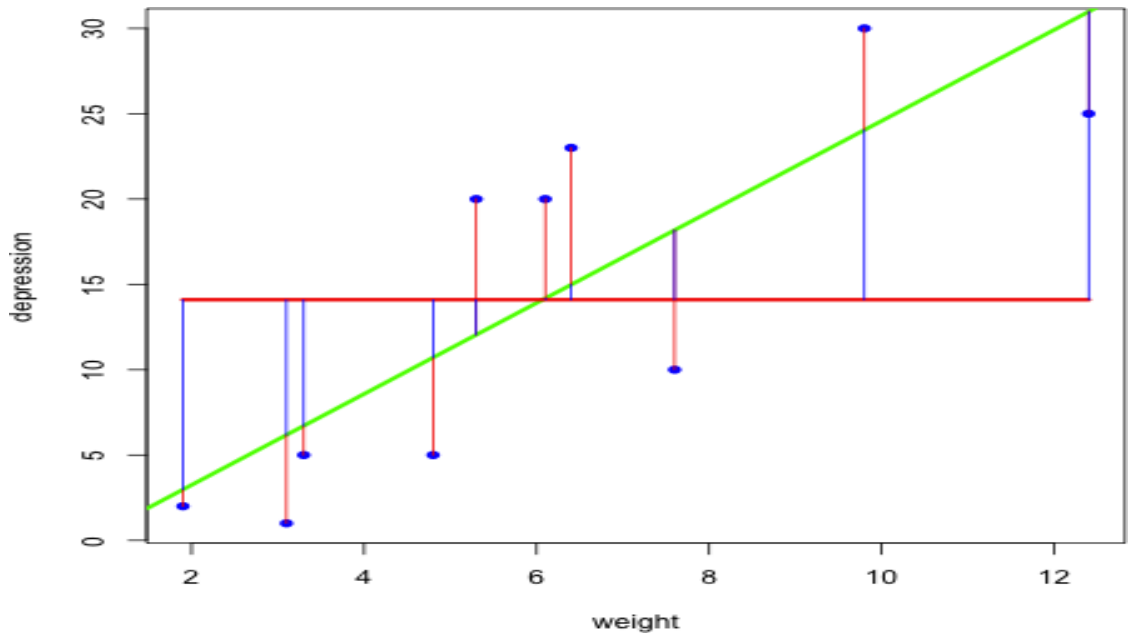
Determinazione della retta di regressione

Calcolo dei coefficienti

Richiamando le quantità calcolate in precedenza e le formule per il calcolo dei parametri si ha

- $b_1 = \frac{\sigma_{xy}}{\sigma_x^2} = 2.66 \quad b_0 = \bar{y} - b_1\bar{x} = 14.1 - (2.66 * 6.07) = -2.04$
- $Y = -2.04 + 2.66x$ rappresenta la retta di regressione stimata

La retta 'migliore'



Interpretazione dei valori dei coefficienti di regressione

- b_0 rappresenta l'intercetta della retta di regressione ed indica il valore della variabile di risposta Y quando il predittore x assume valore 0.
- b_1 rappresenta l'inclinazione della retta di regressione, ovvero la variazione della variabile di risposta Y in conseguenza di un aumento unitario del predittore x.

Assunzioni sul modello

Il modello di regressione lineare semplice è

$$Y = \beta_0 + \beta_1 x + \epsilon$$

e l'errore non osservabile ϵ è una variabile aleatoria con valore atteso pari a 0. Per poter fare inferenza sono necessarie alcune assunzioni:

- la variabile aleatoria ϵ_i si distribuisce come una Normale di parametri 0 e σ^2: dunque la varianza dell'errore non osservabile ϵ_i non dipende dal predittore x_i;
- $cov(\epsilon_i, \epsilon_j) = 0, \forall i \neq j \ (i, j = 1, \ldots, n)$, questo comporta che la risposta relativa al predittore x_i è indipendente da quella relativa al predittore x_j;
- x è nota e non stocastica (priva di errore);
- dalle precedenti assunzioni segue che $\forall i$ la variabile di risposta Y_i si distribuisce secondo una Normale di parametri

$$E[Y_i] = \beta_0 + \beta_1 x_i \quad \text{e} \quad var(Y_i) = \sigma^2.$$

Lo stimatore della varianza σ^2

La quantità σ^2 è incognita e deve essere stimata a partire dai dati. A questo scopo si consideri che la standardizzazione di Y_i si distribuisce secondo una normale

$$\frac{Y_i - E[Y_i]}{var(Y_i)} = \frac{Y_i - (\beta_0 + \beta_1 x_i)}{\sigma}.$$

La somma dei quadrati delle Y_i standardizzate è

$$\frac{\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2}$$

ed essendo la somma di n normali standardizzate indipendenti, si distribuisce come una variabile aleatoria chi-quadro con n gradi di libertà.
Sostituendo i parametri β_0 e β_1 con gli stimatori dei minimi quadrati b_0 e b_1 la precedente diventa

$$\frac{\sum_{i=1}^{n} (Y_i - b_0 - b_1 x_i)^2}{\sigma^2}$$

è un chi-quadro con n-2 gradi di libertà, in quanto si perde un grado di libertà per ogni parametro stimato.

Lo stimatore della varianza σ^2

Il numeratore della precedente rappresenta la somma dei quadrati dei residui

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{n} e^2 = SS_e;$$

da quanto trovato in precedenza, la quantità $\frac{SS_e}{\sigma^2}$ è un chi-quadro con n-2 gradi di libertà.
Poichè il valore atteso di un chi-quadro è uguale ai gradi di libertà possiamo scrivere

$$\frac{E[SS_e]}{\sigma^2} = n-2 \quad \text{da cui} \quad E\left[\frac{SS_e}{n-2}\right] = \sigma^2,$$

lo stimatore della varianza σ^2 è dunque $\frac{SS_e}{n-2}$.

Verifica dell'ipotesi che $\beta_1 = 0$

Un'ipotesi molto importante da verificare nel modello di regressione lineare semplice è che il coefficiente angolare della retta di regressione sia pari a 0: se infatti $\beta_1 = 0$ allora la variabile di risposta non dipende dal predittore, in altre parole non c'è regressione sul predittore.
Per ottenere il test $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ è necessario studiare la distribuzione dello stimatore b_1 di β_1: se b_1 si discosta da 0 allora si rifiuta H_0, altrimenti non si rifiuta. **Ma di quanto b_1 deve discostarsi da 0?**
A questo scopo si consideri che b_1 si distribuisce come una Normale di parametri

$$E[b1] = \beta_1 \quad \text{e} \quad var(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

la versione standardizzata di b_1 è dunque

$$\frac{b_1 - \beta_1}{\sqrt{\sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2}} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{\sigma^2}} (b_1 - \beta_1)$$

ed ha una distribuzione Normale standard.

Verifica dell'ipotesi che $\beta_1 = 0$

la Normale standard

$$\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}}(b_1 - \beta_1)$$

non consente ancora di costruire una statistica test perchè è ancora presente il parametro incognito σ^2: tuttavia si può stimare tale parametro attraverso $\frac{SSe}{n-2}$ che, come visto in precedenza, si distribuisce secondo un chi-quadrato con n-2 gradi di libertà; sostituendo a σ^2 il suo stimatore si ha

$$\sqrt{\frac{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}{SSe}}(b_1 - \beta_1).$$

Poichè questa quantità ha al numeratore una Normale standard ed al denominatore un chi-quadro rapportato ai propri gradi di libertà, si distribuisce come una distribuzione t di student con n-2 gradi di libertà.

Verifica dell'ipotesi che $\beta_1 = 0$

A questo punto la statistica test da utilizzare sotto H_0 $(\beta_1 = 0)$ è

$$ST = \sqrt{\frac{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}{SSe}} b_1 \sim t_{n-2}$$

Il test di livello α di H_0 è ha la seguente regola di decisione:

se $\mid ST \mid \geq t_{n-2,\alpha/2}$ allora si rifiuta H_0

se $\mid ST \mid < t_{n-2,\alpha/2}$ allora non si rifiuta H_0

Bontà di adattamento e diagnostica

Una volta stimato il modello di regressione, è necessario misurare la bontà dell'adattamento del modello ai dati e analizzare i residui per controllare che le assunzioni di normalità con media nulla e varianza costante dei residui siano rispettate.

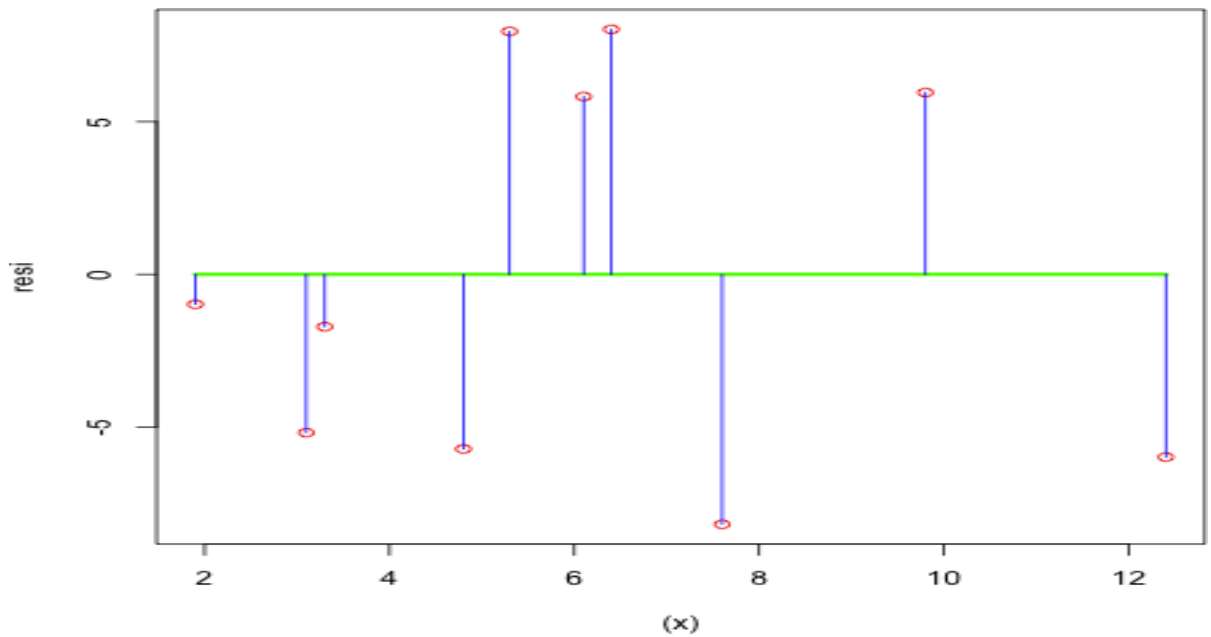
- Strumenti analitici: coefficiente di determinazione lineare R^2
- Strumenti grafici: plot dei residui
  - plot variabili esplicative vs. residui: in caso di relazione non lineare nella configurazione dei punti allora la relazione con la variabile esplicativa potrebbe non essere di primo grado (lineare), ma di grado superiore;
  - plot valori stimati dal modello vs. residui: se i residui aumentano all'aumentare dei valori stimati dal modello, allora potrebbe essere necessario effettuare una trasformazione della variabile di risposta;
  - Normal probability plot: confronto tra i quantili della distribuzione dei residui osservati e quella di una normale standardizzata;

Plot dei residui

Perchè la retta possa essere considerata una buona approssimazione della relazione che intercorre tra Y ed X è necessario che i residui abbiano un andamento casuale rispetto ai valori della x. Se, ad esempio, all'aumentare dei valori della x aumentassero sistematicamente anche i residui, allora la relazione potrebbe non essere non lineare: la retta di regressione ne sarebbe dunque una cattiva approssimazione.

variabili esplicative vs residui

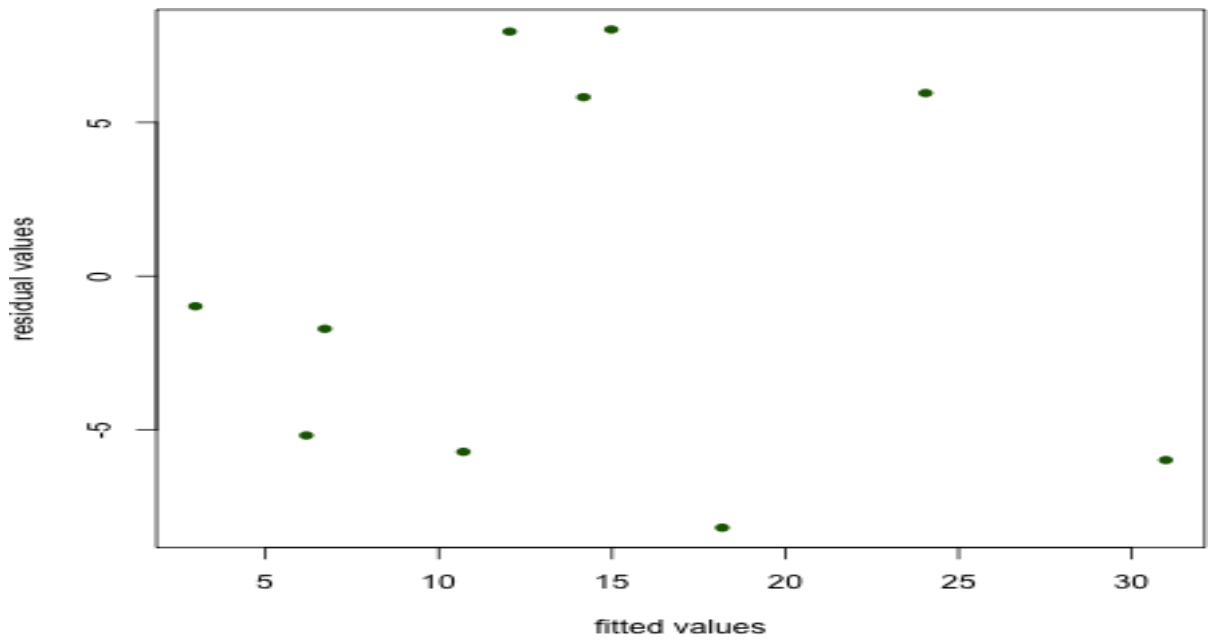
Per verificare che l'andamento dei residui sia effettivamente casuale rispetto ad x, è possibile utilizzare un diagramma di dispesione tra i valori x_i ed i corrispondenti residui $e_i (i = 1, \ldots, n)$



Plot dei residui

Perchè la retta possa essere considerata una buona approssimazione della relazione che intercorre tra Y ed x è necessario che i residui abbiano un andamento casuale rispetto ai valori della x. Se, ad esempio, all'aumentare dei valori della x aumentassero sistematicamente anche i residui, allora la relazione potrebbe non essere non lineare: la retta di regressione ne sarebbe dunque una cattiva approssimazione.

valori stimati $\hat{y}$ vs residui

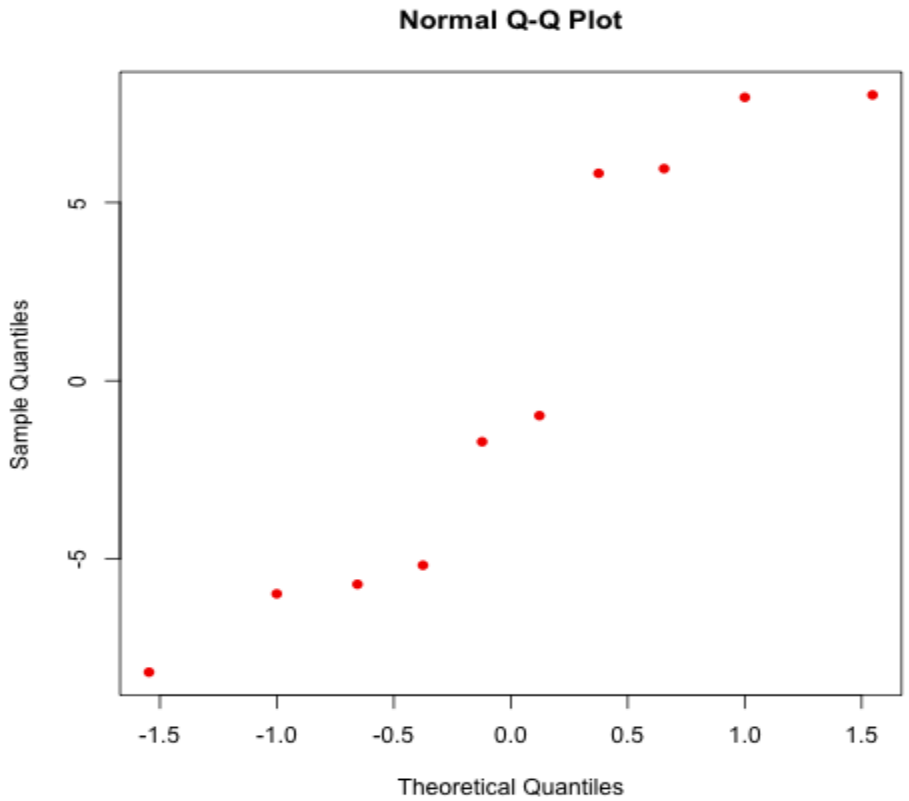


Quantile-quantile plot

Per controllare che l'assunzione della normalità dei residui sia rispettata si ricorre al confronto tra i quantili della distribuzione Normale standard ed i quantili della distribuzione dei residui osservati.

Q-Q plot

Quanto più i punti del grafico risultano allineati lungo la bisettrice del primo quadrante, tanto migliore sarà l'adattamento dei residui osservati alla distribuzione normale.



coefficiente di determinazione lineare R^2

Ricordando che la devianza il numeratore della varianza...

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

Il metodo dei minimi quadrati assicura che $\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$, quindi

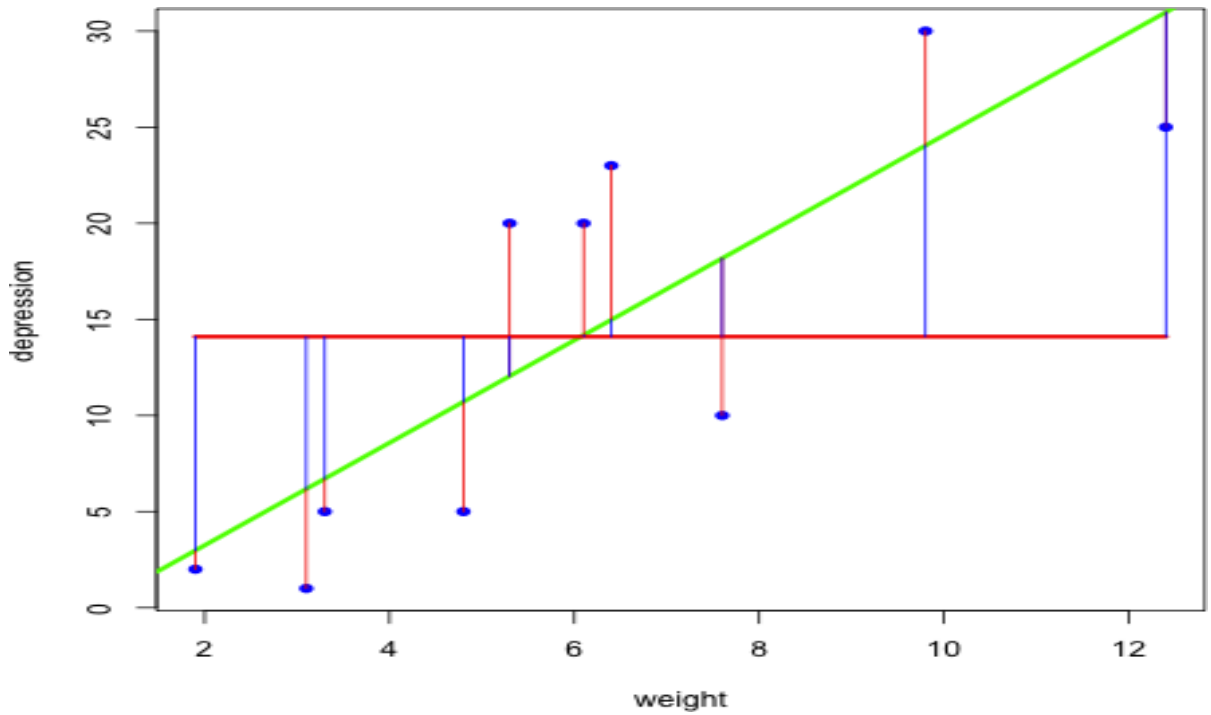
$$SS_y = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + 2 * 0 * (\sum_{i=1}^{n} \hat{y}_i - n\bar{y})$$
$$= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SS_r + SS_e$$

Decomposizione della devianza

La devianza può essere decomposta dunque nelle seguenti quantità $SS_y = SS_r + SS_e$

- $SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2$ devianza totale
- $SS_r = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ devianza di regressione
- $SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ devianza dei residui

Interpretazione grafica



Bontà dell'adattamento

Intituitivamente, l'adattamento della retta è migliore quanto maggiore sarà proporzione di variabilità totale che la retta di regressione riesce a spiegare; ovvero, l'adattamento della retta è migliore quanto minore sarà la variabilità residua. Una misura di come il modello approssima i dati osservati è data dal coefficiente di determinazione lineare R^2, dato da

$$R^2 = \frac{SS_r}{SS_y} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \mu_y)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

ovvero

$$R^2 = 1 - \frac{SS_e}{SS_y} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \mu_y)^2}$$

esempio di calcolo R^2

- $SS_y = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 1020.9$
- $SS_r = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 657.97$
- $SS_e = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 362.93$

$$R^2 = \frac{SS_r}{SS_y} = \frac{657.97}{1020.9} = 0.64$$

ovvero

$$R^2 = 1 - \frac{SS_e}{SS_y} = 1 - \frac{282.1862}{5058.4} = 1 - 0.36 = 0.64$$