Statistica e Laboratorio

10. Inferenza statistica: modello di regressione lineare

Paolo Vidoni

Dipartimento di Scienze Economiche e Statistiche
Università di Udine
via Tomadini 30/a - Udine
paolo.vidoni@uniud.it

https://elearning.uniud.it/

Sommario

Sommario

- **Introduzione**
- **Analisi di correlazione**
- **Analisi di regressione**

Introduzione

L'analisi di regressione e l'analisi di correlazione sono state introdotte in statistica descrittiva come tecniche per studiare l'eventuale dipendenza tra variabili quantitative.

Queste procedure vengono ora estese all'ambito inferenziale, tenendo conto che i dati raccolti sono di tipo sperimentale, e quindi interpretabili come osservazioni campionarie riferite ad opportune variabili casuali.

Nell'analisi di correlazione, l'obiettivo è quello di studiare l'eventuale dipendenza lineare tra due fenomeni aleatori di tipo quantitativo, tenendo conto dell'associato campione casuale semplice costituito da osservazioni bivariate.

Nell'analisi di regressione si vuole specificare un opportuno modello statistico parametrico dove si assume che la variabile casuale di interesse dipende (con particolare riferimento al suo valore atteso) dai valori osservati di una o più variabili esplicative.

Sommario

Correlazione

Nell'analisi di correlazione si è interessati allo studio dell'eventuale **dipendenza lineare** tra due fenomeni aleatori di interesse, descritti dalla variabile casuale bivariata (X, Y).

Le due variabili X e Y vengono trattate in modo *simmetrico* e l'interesse non consiste nello studio della dipendenza causale o funzionale tra esse (una variabile spiega l'altra), ma nell'analisi della loro eventuale dipendenza (lineare).

Come si è evidenziato nella parte dedicata al calcolo delle probabilità, data una variabile casuale bivariata (X, Y), un indicatore normalizzato della dipendenza lineare tra X e Y è il **coefficiente di correlazione lineare** definito da

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \in [0, 1],$$

dove $\sigma_{XY} = Cov(X, Y)$, $\sigma_X = \sqrt{V(X)}$ e $\sigma_Y = \sqrt{V(Y)}$.

Dato un campione casuale semplice $(X_1, Y_1), \ldots, (X_n, Y_n)$, uno **stimatore** per ρ_{XY} è il **coefficiente di correlazione lineare campionario** definito da

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sum_{i=1}^{n}(Y_i - \bar{Y}_n)^2}},$$

con $\bar{X}_n$ e $\bar{Y}_n$ le medie campionarie riferite alle componenti X e Y.

L'interpretazione dell'associato valore di stima sarà analoga a quella proposta in statistica descrittiva, anche se bisogna tenere presente che l'analisi è ora di natura inferenziale e non puramente descrittiva.

Lo stimatore $\hat{\rho}_{XY}$ è molto sensibile alla presenza di valori anomali.

Con riferimento alla dipendenza lineare è possibile considerare opportune procedure di verifica di ipotesi sull'assenza o meno di correlazione.

Test di correlazione

Si assume che le variabili casuali $(X_1, Y_1), \ldots, (X_n, Y_n)$ abbiano *distribuzione normale bivariata* (X e Y sono entrambe normali con correlazione ρ_{XY}).

Si vuole verificare l'ipotesi nulla $H_0 : \rho_{XY} = 0$, a fronte di una ipotesi alternativa H_1, bilaterale o unilaterale, ad un livello di significatività α fissato. Si considera la **statistica test**

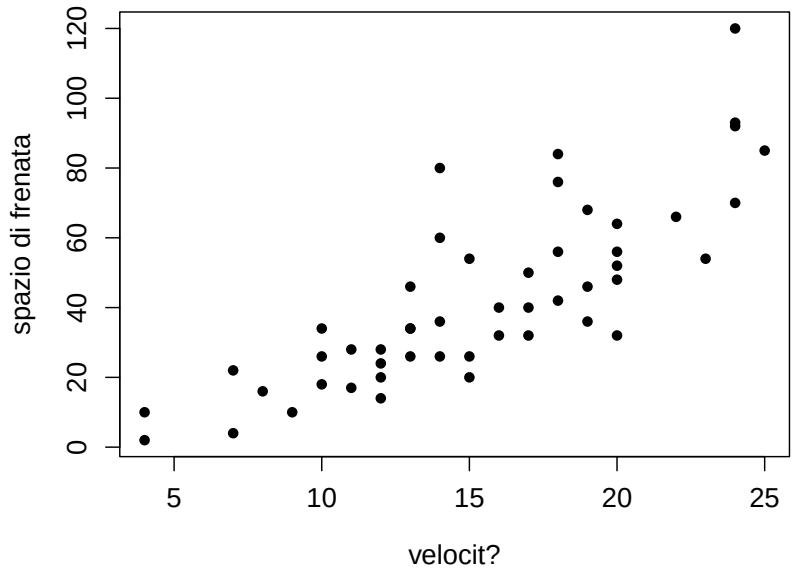
$$R = \frac{\hat{\rho}_{XY}\sqrt{n-2}}{\sqrt{1-\hat{\rho}_{XY}^2}},$$

che, sotto H_0, ha distribuzione $t(n-2)$.

La **regione di rifiuto** e il **livello di significatività osservato** si determinano come nel caso del test t sulla media, con l'unica differenza che i gradi di libertà sono $n-2$.

Nel caso di dati non gaussiani, si possono considerare versioni alternative del test (che sono anche meno sensibili alla presenza di valori anomali) basate sull'indice di correlazione di Spearman o su quello di Kendall.

Esempio. *Velocità* (continua). Si considerano i dati sulla velocità X e sullo spazio di frenata Y di $n = 50$ automobili degli anni 20.



Dai dati si ottiene il seguente valore di stima $\hat{\rho}_{XY} = 0.81$, e quindi il valore della statistica test è $r = 0.81 \cdot \sqrt{48}/\sqrt{1 - 0.81^2} = 9.46$.

Se $H_1 : \rho_{XY} > 0$, posto $\alpha = 0.01$, la regione di rifiuto è $R_{0.01} = \{r \in \mathbf{R} : r \geq 2.41\}$: si rifiuta H_0 e quindi c'è una correlazione positiva tra X e Y. Inoltre, $\alpha^{oss} = 1 - F_R(9.46) = 7.45 \cdot 10^{-13}$. $\Diamond$

Sommario

Modello di regressione lineare semplice

Il modello di regressione lineare semplice può venire considerato anche in ambito inferenziale, quando la variabile risposta viene misurata con errore o quando fa riferimento ad un campione associato ad una certa popolazione (fenomeno) di interesse.

Quindi, la **variabile risposta** Y è una variabile casuale, mentre la **variabile esplicativa** (**regressore** o **previsore**) x si suppone non casuale, ad esempio perché fissata dallo sperimentatore o misurata senza errore sulle unità campionarie.

Nonostante l'analisi di regressione presenti alcuni aspetti in comune con l'analisi di correlazione, in questo contesto l'obiettivo è studiare l'eventuale rapporto tra una variabile *dipendente* Y e un'altra variabile x assunta come *indipendente*.

Si vuole indagare se e in che misura le variazioni riferite alla variabile Y possano essere interpretate come *risposta* alle variazioni della variabile *esplicativa* x.

In molti casi questa distinzione risulta chiara dalla struttura dell'esperimento e dalla natura dei dati, come ad esempio quando si analizza la relazione tra la dose di un certo farmaco e il suo effetto sui pazienti.

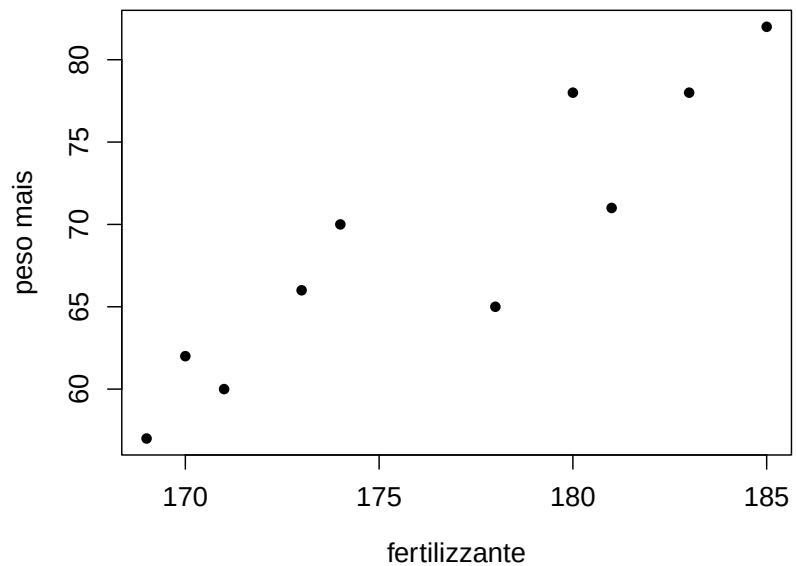
In altri contesti applicativi, questa distinzione non sussiste e non ci sono ragioni obiettive per supporre una relazione di causa ed effetto, come ad esempio quando si studia la relazione tra peso corporeo e diametro del torace.

In questo caso l'analisi di regressione può essere comunque utile per descrivere la relazione tra le variabili, senza postulare l'esistenza di una specifica relazione di causa ed effetto.

Esempio. *Mais*. Si considerano i dati sulla dose di fertilizzate utilizzata x e sulla quantità di mais prodotta Y (peso della granella in Kg), con riferimento a $n = 10$ distinte parcelle sperimentali, simili per caratteristiche e della medesima dimensione.

parcella	1	2	3	4	5	6	7	8	9	10
x	171	169	181	173	178	180	185	183	170	174
Y	60	57	71	66	65	78	82	78	62	70

In questo caso è chiara la distinzione tra la variabile risposta Y e la variabile esplicativa x.



Il grafico evidenzia una sostanziale relazione lineare tra x e Y, anche se è presente una certa quota di variabilità che non viene spiegata da tale relazione. $\Diamond$

Nel modello di regressione lineare semplice si considera un campione casuale $Y_1, \ldots, Y_n$, riferito alla variabile risposta, e si assume che

$$Y_i = a + b\, x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

dove a e b sono parametri reali non noti chiamati **coefficienti di regressione**.

I valori $x_1, \ldots, x_n$, nell'ottica tipica della regressione, sono interpretati come le corrispondenti osservazioni riferite alla variabile esplicativa, supposta non casuale.

Le quantità $\epsilon_1, \ldots, \epsilon_n$, chiamate **residui** (**errori**), sono variabili casuali *indipendenti*, con distribuzione $N(0, \sigma^2)$, dove la varianza comune $\sigma^2 > 0$ non è nota.

Quindi le osservazioni $y_1, \ldots, y_n$, riferite alla variabile risposta, sono determinazioni osservate di variabili casuali **indipendenti**, ma **non necessariamente identicamente distribuite**,

$$Y_i \sim N(a + b\, x_i, \sigma^2), \quad i = 1, \ldots, n.$$

Dal punto di vista del calcolo delle probabilità, si propone un modello per la distribuzione di probabilità della variabile risposta Y, "condizionata" al valore assunto dalla variabile esplicativa x.

In particolare, si assume che il corrispondente **valor medio** risulti specificato dalla **retta di regressione** $y = a + bx$, con intercetta a e coefficiente angolare b.

Si noti che se $b = 0$, le variabili casuali $Y_1, \ldots, Y_n$ sono anche *identicamente distribuite* e di fatto la variabile risposta Y non risulta "spiegata" dalla variabile esplicativa x.

Il modello di regressione è detto **lineare** e **semplice**, poiché si assume una *relazione lineare*, con riferimento al valore medio, e si considera *una sola variabile esplicativa*.

Il modello può venire generalizzato considerando più di una variabile esplicativa (**regressione lineare multipla**), introducendo ipotesi diverse sugli errori o definendo una relazione non lineare con riferimento alla media.

Inferenza e verifica di ipotesi nel modello regressione lineare

Sulla base delle osservazioni $y_1, \ldots, y_n$ e $x_1, \ldots, x_n$ si possono stimare i parametri non noti del modello: a, b e σ^2.

Gli stimatori per a e b, basati sul **metodo dei minimi quadrati**, si ottengono con una procedura analoga a quella vista in statistica descrittiva e corrispondono a

$$\hat{a} = \bar{Y}_n - \hat{b}\bar{x}_n, \quad \hat{b} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_n)(x_i - \bar{x}_n)}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} = \frac{\sum_{i=1}^n Y_i x_i - n\bar{Y}_n \bar{x}_n}{\sum_{i=1}^n x_i^2 - n\bar{x}_n^2},$$

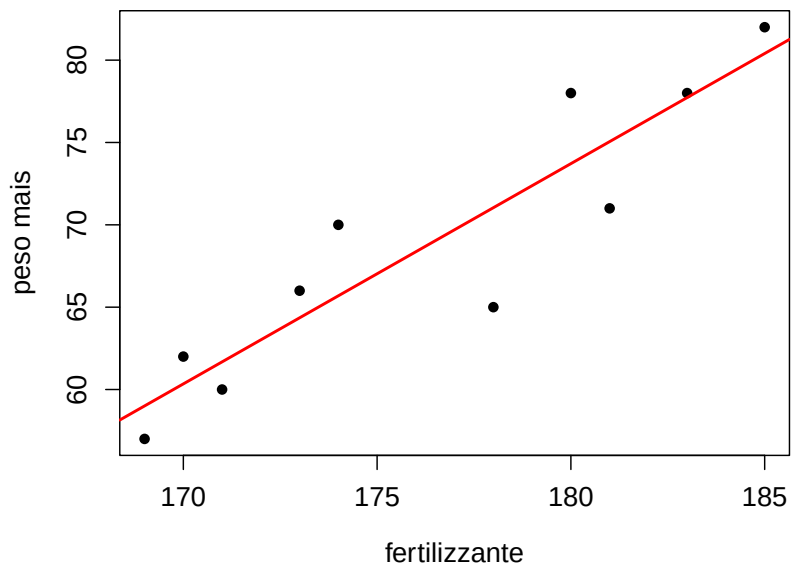
dove $\bar{Y}_n$ è la media campionaria di $Y_1, \ldots, Y_n$ e $\bar{x}_n$ è la media dei valori $x_1, \ldots, x_n$.

La retta $y = \hat{a} + \hat{b}x$ è detta **retta di regressione stimata**.

Si noti che il coefficiente angolare $\hat{b} = \hat{\sigma}_{Yx}/\hat{\sigma}_x^2$, con $\hat{\sigma}_{Yx}$ la covarianza campionaria di $Y_1, \ldots, Y_n$ e $x_1, \ldots, x_n$ e $\hat{\sigma}_x^2$ la varianza basata su $x_1, \ldots, x_n$. Il segno di $\hat{b}$ è determinato dal segno di $\hat{\sigma}_{Yx}$.

Poiché i residui sono gaussiani, gli stimatori $\hat{a}$ e $\hat{b}$ coincidono con quelli basati sul **metodo della massima verosimiglianza**.

Esempio. *Mais* (continua). Con riferimento ai dati sulla produzione di mais, si ha che $\bar{y}_n = 68.9$, $\bar{x}_n = 176.4$, $\hat{\sigma}_{Yx} = 39.64$ e $\hat{\sigma}_x^2 = 29.64$, da cui $\hat{a} = -167.01$ e $\hat{b} = 1.34$.



Nel grafico si è disegnata in rosso la retta di regressione stimata $y = -167.01 + 1.34\,x$, che si adatta bene alle osservazioni. ◊

Gli stimatori $\hat{a}$ e $\hat{b}$ hanno *distribuzione di probabilità normale*, poiché sono combinazioni lineari di variabili casuali gaussiane. Inoltre, sono *non distorti*, cioè $E(\hat{a}) = a$ e $E(\hat{b}) = b$, con varianza

$$V(\hat{a}) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2)}, \quad V(\hat{b}) = \frac{\sigma^2}{\sum_{i=1}^n x_i^2 - n\bar{x}_n^2}.$$

Sotto deboli ipotesi su $x_1, \ldots, x_n, \ldots$, si verifica che sono anche *consistenti*, con *standard error* $se(\hat{a}) = \sqrt{V(\hat{a})}$ e $se(\hat{b}) = \sqrt{V(\hat{b})}$.

Una volta calcolate le stime $\hat{a}$ e $\hat{b}$, si possono determinare i **valori stimati dal modello**

$$\hat{y}_i = \hat{a} + \hat{b}x_i, \quad i = 1, \ldots, n,$$

cioè i valori stimati della variabile risposta Y per ogni valore osservato x_i e i **residui stimati**

$$\hat{\epsilon}_i = y_i - \hat{a} - \hat{b}x_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n,$$

cioé la stima degli errori (residui) basata sulle osservazioni.

Una stima per la varianza dei residui σ^2 coinvolge i residui stimati elevati al quadrato e corrisponde a

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n \hat{\epsilon}_i^2}{n-2}.$$

Lo stimatore corrispondente S^2 è *non distorto* e *consistente*. Inoltre, si verifica facilmente che la media dei residui stimati è $\sum_{i=1}^n \hat{\epsilon}_i/n = 0$

Gli **standard error stimati** di $\hat{a}$ e $\hat{b}$ si ottengono rimpiazzando σ^2 con s^2:

$$\hat{se}(\hat{a}) = \sqrt{\frac{s^2 \sum_{i=1}^n x_i^2}{n(\sum_{i=1}^n x_i^2 - n\bar{x}_n^2)}}, \quad \hat{se}(\hat{b}) = \sqrt{\frac{s^2}{\sum_{i=1}^n x_i^2 - n\bar{x}_n^2}}.$$

Esempio. *Mais* (continua). I valori stimati dal modello $\hat{y}_i$ e i residui stimati $\hat{\epsilon}_i$ corrispondono a:

$\hat{y}_i$	61.68	59.00	75.06	64.35	71.04	73.71	80.40	77.73	60.34	65.69
$\hat{\epsilon}_i$	-1.68	-2.00	-4.05	1.65	-6.04	4.29	1.60	0.27	1.66	4.31

Quindi, la stima per la varianza è $s^2 = 13.09$, mentre gli standard error stimati di $\hat{a}$ e $\hat{b}$ sono $\hat{se}(\hat{a}) = 37.10$ e $\hat{se}(\hat{b}) = 0.21$. $\Diamond$

Valgono i seguenti risultati, che risultano utili per definire *intervalli di confidenza* e procedure di *verifica di ipotesi* per a e b:

$$\frac{\hat{a} - a}{\hat{se}(\hat{a})} \sim t(n-2), \quad \frac{\hat{b} - b}{\hat{se}(\hat{b})} \sim t(n-2).$$

In particolare, si può determinare facilmente un **intervallo di confidenza per** b con livello $1 - \alpha$, che corrisponde a

$$[\hat{b} - t_{\alpha/2}\, \hat{se}(\hat{b}), \hat{b} + t_{\alpha/2}\, \hat{se}(\hat{b})],$$

con $t_{\alpha/2}$ **valore critico** tale che $P(T \geq t_{\alpha/2}) = \alpha/2$, dove $T \sim t(n-2)$.

In modo analogo si definiscono gli **intervalli di confidenza per** a.

È possibile considerare opportune procedure di verifica di ipotesi su a e b. Con riferimento a b si considera l'ipotesi nulla

$$H_0 : b = 0,$$

a fronte di una ipotesi alternativa H_1, bilaterale o unilaterale, ad un livello di significatività α fissato.

L'ipotesi H_0 indica assenza di dipendenza lineare tra Y e x ed è molto importante da analizzare perché, se accettata, evidenzia la non adeguatezza del modello di regressione in esame.

Si considera la **statistica test**

$$T = \frac{\hat{b}}{\hat{se}(\hat{b})},$$

che, sotto H_0, ha distribuzione $t(n-2)$. Quindi la **regione di rifiuto** e il **livello di significatività osservato** si determinano come nel caso del test t sulla media, con l'unica differenza che i gradi di libertà sono $n-2$.

In modo analogo si può definire una procedura di verifica di ipotesi su a, considerando l'ipotesi nulla $H_0 : a = 0$ e la statistica test $T = \hat{a}/\hat{se}(\hat{a})$.

Esempio. *Mais* (continua). Si considerano le seguenti ipotesi nulle $H_0 : a = 0$ e $H_0 : b = 0$, riferite all'intercetta e al coefficiente angolare della retta di regressione.

Dai dati campionari si ottengono i seguenti valori osservati per le corrispondenti statistiche test: $t = -4.50$ e $t = 6.36$.

In tutti e due i casi, se si considera un'alternativa bilaterale e si pone $\alpha = 0.01$, la regione di rifiuto è $R_{0.01} = \{t \in \mathbf{R} :| t | \geq 3.36\}$, dove $t_{0.005} = 3.36$ è l'opportuno valore critico riferito alla $t(8)$.

Si rifiutano entrambe le ipotesi nulle e i livelli di significatività osservati sono $\alpha^{oss} = 0.002$ e $\alpha^{oss} = 0.0002$, rispettivamente.

Infine, gli intervalli di confidenza per a e b con livello $1 - \alpha = 0.95$ sono $[-252.56, -81.47]$ e $[0.85, 1.82]$, rispettivamente. $\Diamond$

Utilizzando la retta di regressione stimata $y = \hat{a} + \hat{b}\,x$ si possono determinare i **valori previsti dal modello**, che fanno riferimento a nuovi valori per x, diversi da $x_1, \ldots, x_n$.

Le quantità ottenute sono utili per fare previsioni o ricostruzioni di valori mancanti per Y, a fronte di nuovi valori per x.

Occorre fare molta attenzione quando si estrapola la retta di regressione stimata, cioè quando si fanno previsioni al di fuori dell'intervallo dei valori osservati per la variabile esplicativa x.

Ad esempio, nel caso della produzione di mais, non è detto che, per valori dalla dose di fertilizzante che siano inferiori a 169 e superiori a 185, il modello di regressione considerato sia ancora valido.

Se x_{n+1} è un nuovo valore per la variabile esplicativa, allora $\hat{y}_{n+1} = \hat{a} + \hat{b}\,x_{n+1}$ è un valore di previsione per la variabile casuale non osservata Y_{n+1}.

Oltre ad una previsione puntuale, è anche possibile definire un opportuno **intervallo di previsione** per Y_{n+1}.

Valutazioni sul modello di regressione

Dopo aver definito, con riferimento ad un determinato insieme di dati, un opportuno modello di regressione lineare, è utile *valutare la bontà del modello* che si è ottenuto.

Il modello lineare è giustificato solo nel caso di relazioni sostanzialmente lineari tra Y, e in particolare $E(Y)$, e la variabile esplicativa x. Inoltre, sarà un buon modello se la variabile indipendente x spiega in modo efficace le variazioni della variabile risposta Y.

La varianza associata alla variabile Y (**varianza totale**) può essere vista come somma della quota descritta dal modello (**varianza spiegata**) e della quota rimanente (**varianza residua**).

Come già considerato nella parte dedicata alla statistica descrittiva, tanto maggiore è la varianza spiegata dal modello, tanto migliore sarà l'adattamento dei dati al modello teorico.

Per valutare l'adattamento globale del modello ai dati, e quindi anche la sua capacità esplicativa per il fenomeno Y, si considera l'**indice di determinazione** R^2, che corrisponde alla proporzione di varianza di Y spiegata dal modello di regressione

$$R^2 = \frac{\sum_i (\hat{y}_i - \bar{y}_n)^2}{\sum_i (y_i - \bar{y}_i)^2} = \frac{\text{varianza spiegata}}{\text{varianza totale}} = 1 - \frac{\text{varianza residua}}{\text{varianza totale}}.$$

Si ha che $0 \leq R^2 \leq 1$ e un valore per R^2 vicino a 1 (vicino a 0) indica un buon (pessimo) adattamento del modello ai dati.

La *varianza residua* corrisponde alla varianza calcolata sui residui stimati $\hat{\epsilon}_i$, $i = 1, \ldots, n$. Poiché la media dei residui è nulla per costruzione, corrisponde semplicemente a $\sum_{i=1}^{n} \hat{\epsilon}_i^2 / n$.

Si dimostra che vale la seguente relazione tra R^2 e l'indice di correlazione campionario basato sui dati $y_1, \ldots, y_n$ e $x_1, \ldots, x_n$: $R^2 = \hat{\rho}_{Yx}^2$.

Esempio. *Mais* (continua). Se si considerano i residui stimati $\hat{\epsilon}_i$, $i = 1, \ldots, n$, calcolati in precedenza, si ottiene la varianza residua $\sum_{i=1}^{n} \hat{\epsilon}_i^2 / n = 10.48$.

Quindi, poiché la varianza riferita alle osservazioni di Y è 63.49, $R^2 = 1 - (10.48/63.49) = 0.83$. In alternativa, si può ottenere R^2 considerando il quadrato di $\hat{\rho}_{Yx} = 0.91$. $\Diamond$

Per migliorare l'adattamento del modello ai dati, a volte può essere utile trasformare i valori della variabile esplicativa x e/o della variabile risposta Y.

Le trasformazioni tipiche sono il logaritmo e l'elevamento a potenza.

Dopo aver trasformato i dati, occorre stimare e controllare nuovamente il modello.

Oltre a R^2 esistono altri indici per valutare la bontà di un modello e per confrontare modelli alternativi.

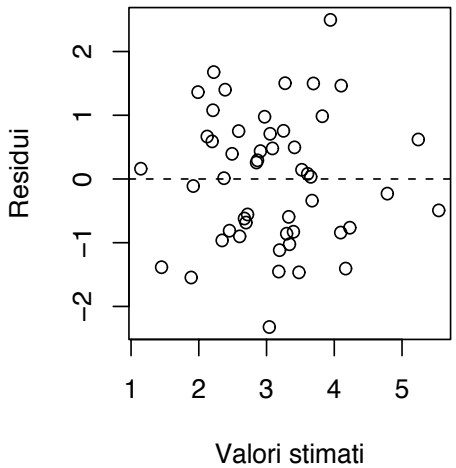
Diagnostiche

Esistono delle semplici procedure, sostanzialmente di tipo grafico, che permettono di verificare l'ipotesi di linearità e le ipotesi fatte sui residui, ipotesi che risultano necessarie per la validità del modello.

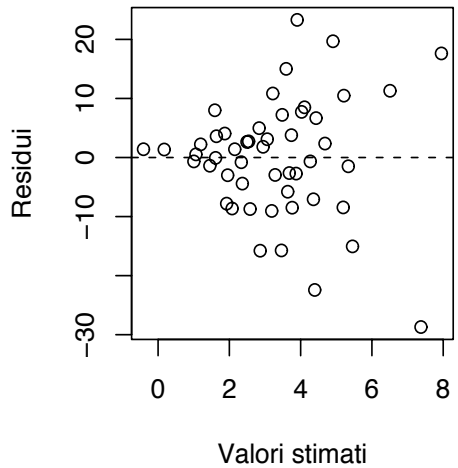
Tali procedure coinvolgono i residui stimati $\hat{\epsilon}_i$, $i = 1, \ldots, n$, e in particolare si considera:

- il grafico di $(\hat{y}_i, \hat{\epsilon}_i)$, $i = 1, \ldots, n$: per evidenziare eventuali andamenti sistematici nei residui, dovuti alla non linearità, ad una varianza non costante (eteroschedasticità) o ad una non indipendenza;
- il grafico di $(x_i, \hat{\epsilon}_i)$, $i = 1, \ldots, n$: per evidenziare eventuali andamenti sistematici nei residui, dovuti alla non linearità;
- il q-q plot sui residui stimati $\hat{\epsilon}_i$, $i = 1, \ldots, n$: per verificare l'ipotesi che i residui siano gaussiani;
- il grafico dei residui stimati $\hat{\epsilon}_i$, $i = 1, \ldots, n$, rispetto al tempo (nel caso di osservazioni con ordine temporale): per evidenziare l'eventuale dipendenza temporale tra le osservazioni.

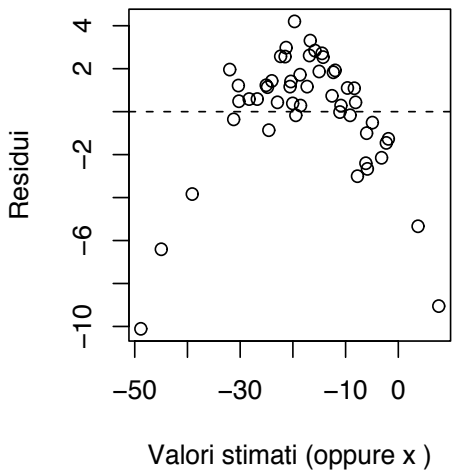
OK



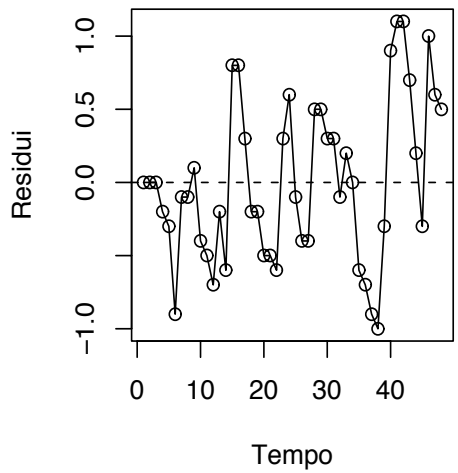
Varianza non costante



Andamento non casuale (non linearità)



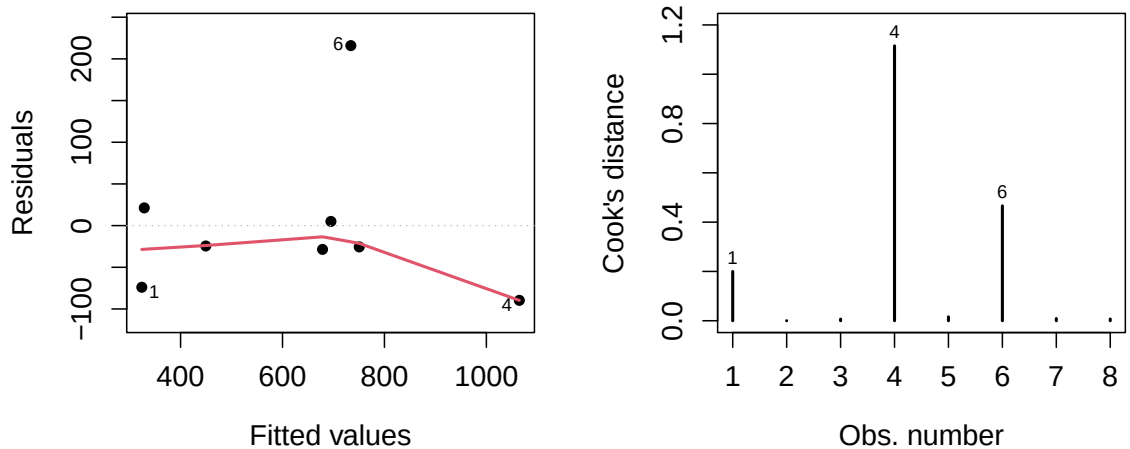
Dipendenza temporale



Un ulteriore aspetto da considerare riguarda la presenza di **valori anomali** (**outliers**), che corrispondono ad osservazioni campionarie che appaiono lontane dalla maggior parte dei dati osservati.

Corrispondono a osservazioni che presentano un residuo stimato molto alto oppure a osservazioni che producono un effetto rilevante sulle stime dei parametri (e quindi sulla retta di regressione stimata).

Si possono individuare considerando i valori dei residui stimati e calcolando il valore della distanza di Cook (un valore prossimo o superiore a 1 indica la presenza di un possibile outlier).



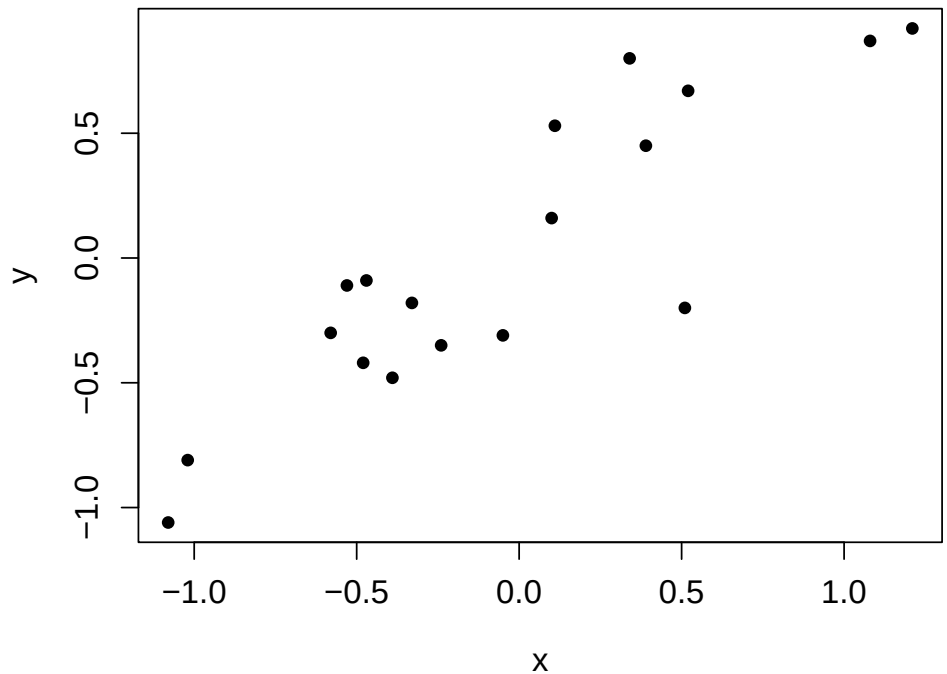
Le osservazioni 4 e 6 sono potenziali outlier: discutere la loro presenza e effettuare l'analisi con e senza questi valori.

Esempio. *Misurazioni*. Per valutare la qualità di un prodotto si può utilizzare una procedura precisa ma costosa, descritta dalla variabile casuale Y, oppure una procedura meno precisa ma anche meno costosa, descritta dalla variabile x.

Si considerano le misurazioni, effettuate con entrambe le procedure, riferite a $n = 18$ prodotti

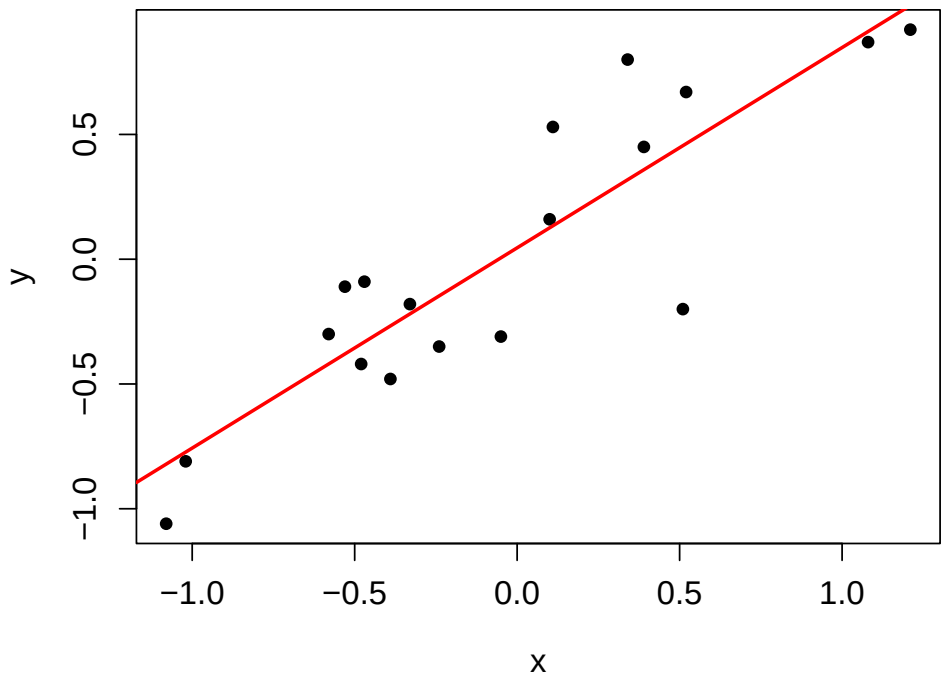
i	y	x	i	y	x
1	-1.06	-1.08	10	-0.11	-0.53
2	-0.81	-1.02	11	-0.09	-0.47
3	-0.48	-0.39	12	0.16	0.10
4	-0.42	-0.48	13	0.45	0.39
5	-0.30	-0.58	14	0.53	0.11
6	-0.35	-0.24	15	0.67	0.52
7	-0.31	-0.05	16	0.80	0.34
8	-0.18	-0.33	17	0.87	1.08
9	-0.20	0.51	18	0.92	1.21

Si vuole studiare la relazione tra le due procedure e, in particolare, quanto la variabile esplicativa x riesce a spiegare le variazioni nella qualità di un prodotto, come descritta dalla variabile risposta Y.



Il grafico evidenzia una sostanziale relazione lineare tra x e Y, quindi si può pensare di analizzare i dati considerando un opportuno modello di regressione lineare semplice.

Dai dati si ottiene che $\bar{y}_n = 0.005$, $\bar{x}_n = -0.051$, $\hat{\sigma}_{Yx} = 0.312$ e $\hat{\sigma}_x^2 = 0.389$, da cui $\hat{a} = 0.046$ e $\hat{b} = 0.803$.



Nel grafico si è disegnata in rosso la retta di regressione stimata $y = 0.046 + 0.803\,x$, che sembra adattarsi abbastanza bene alle osservazioni.

I valori stimati dal modello $\hat{y}_i$ e i residui stimati $\hat{\epsilon}_i$ corrispondono a

i	$\hat{y}_i$	$\hat{\epsilon}_i$	i	$\hat{y}_i$	$\hat{\epsilon}_i$
1	-0.821	-0.239	10	-0.380	0.270
2	-0.773	-0.037	11	-0.332	0.242
3	-0.267	-0.213	12	0.126	0.035
4	-0.340	-0.080	13	0.359	0.091
5	-0.420	0.120	14	0.134	0.396
6	-0.147	-0.203	15	0.463	0.207
7	0.005	-0.315	16	0.318	0.482
8	-0.219	0.039	17	0.912	-0.042
9	0.455	-0.655	18	1.017	-0.097

Quindi, la stima per la varianza è $s^2 = 0.080$, mentre gli standard error stimati di $\hat{a}$ e $\hat{b}$ sono $\hat{se}(\hat{a}) = 0.067$ e $\hat{se}(\hat{b}) = 0.107$.

Si considerano le seguenti ipotesi nulle $H_0 : a = 0$ e $H_0 : b = 0$, riferite all'intercetta e al coefficiente angolare della retta di regressione.

Dai dati campionari si ottengono i seguenti valori osservati per le corrispondenti statistiche test: $t = 0.682$ e $t = 7.511$.

In tutti e due i casi, se si considera un'alternativa bilaterale e si pone $\alpha = 0.01$, la regione di rifiuto è $R_{0.01} = \{t \in \mathbf{R} :| t |\geq 2.92\}$, dove $t_{0.005} = 2.92$ è l'opportuno valore critico riferito alla $t(16)$.

Si accetta $H_0 : a = 0$ e si rifiuta $H_0 : b = 0$; i livelli di significatività osservati sono $\alpha^{oss} = 0.505$ e $\alpha^{oss} = 1.25 \cdot 10^{-6}$, rispettivamente.

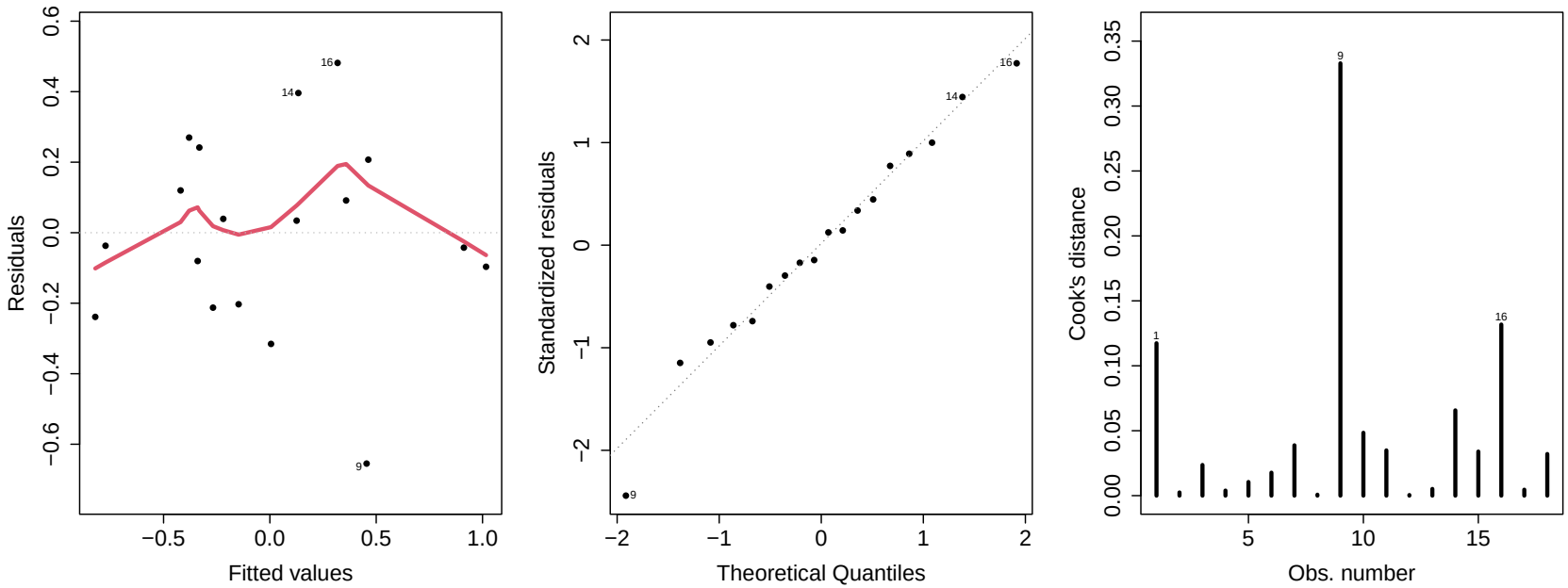
Nonostante si sia concluso che $a = 0$, l'intercetta viene lasciata nel modello di regressione lineare anche se non sembra significativamente diversa da zero.

Infine, gli intervalli di confidenza per a e b con livello 0.95 sono $[-0.096, 0.187]$ e $[0.575, 1.029]$, rispettivamente.

Considerando i residui stimati $\hat{\epsilon}_i$, $i = 1, \ldots, n$, calcolati in precedenza, si ottiene la varianza residua $\sum_{i=1}^{n} \hat{\epsilon}_i^2/n = 0.071$.

Quindi, poiché la varianza riferita alle osservazioni y è 0.322, si può calcolare il coefficiente di determinazione lineare, che risulta pari a $R^2 = 1 - (0.071/0.322) = 0.779$.

Il modello si adatta abbastanza bene ai dati, anche se la qualità dei prodotti non risulta determinata in modo preciso dai valori di misurazione x.



Dal grafico di $(\hat{y}_i, \hat{\epsilon}_i)$, $i = 1, \ldots, n$, e, in particolare, dal q-q plot sui residui stimati, sembra che le ipotesi sui residui siano sostanzialmente soddisfatte. Inoltre, non sembra che ci siano valori anomali. $\Diamond$