www.sce.unimore.it

Scienze della Comunicazione
e dell'Economia

REGRESSIONE LINEARE SEMPLICE

Legacy Edition

Luca La Rocca
luca.larocca@unimore.it

UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Consideriamo il caso di un'azienda che abbia appena assunto un nuovo venditore e voglia prevederne le vendite nell'anno a venire:

supponendo per semplicità che

- *l'azienda continui a vendere gli stessi prodotti allo stesso prezzo,*
- *il venditore neoassunto sia omogeneo ai venditori già in forza all'azienda,*

la previsione potrà basarsi sulle vendite realizzate dai venditori già in forza all'azienda nel corso dell'ultimo anno.

La tabella seguente (Rees, 2001) riporta le vendite realizzate (in migliaia di sterline inglesi, nell'ultimo esercizio) da $n = 8$ venditori di una certa azienda…

Venditore	Vendite (migliaia di sterline)
Andrea	105
Barbara	120
Carlo	160
Diana	155
Elisa	70
Francesco	150
Giovanna	185
Massimo	130
Totale	1075

Vogliamo prevedere il numero aleatorio

$$Y_{n+1} \quad = \quad \text{“vendite del neoassunto”}$$

sulla base delle vendite osservate

$$y_1, \ldots, y_n$$

realizzate dagli n venditiori già in forza all'azienda...

...quale distribuzione useremo per Y_{n+1}?

In primo luogo sembra naturale assegnare

$$\begin{aligned} \mathbb{E}[Y_{n+1}] &= \bar{y} \\ sd(Y_{n+1}) &= sd_y \end{aligned}$$

dove sd_y è la deviazione standard di $y_1, \ldots, y_n$...

...per esempio, con riferimento ai dati di Rees, assegneremo

$$\begin{aligned} \mathbb{E}[Y_{n+1}] &= 134.4 \\ sd(Y_{n+1}) &= 33.8 \end{aligned}$$

Valore atteso e deviazione standard di Y_{n+1} bastano per dare:

- una previsione puntuale (con errore standard di previsione)

$$Y_{n+1} \quad = \quad 134.4 \pm 33.8$$

- una previsione per intervallo (basata sul teorema di Chebyshev)

$$66.8 = 134.4 - 2 \times 33.8 \leq Y_{n+1} \leq 134.4 + 2 \times 33.8 = 202.0$$

con probabilità almeno pari al 75%.

Si noti che un intervallo di previsione

- si riferisce a un futuro valore osservabile
- in esso contenuto con “buona” probabilità

mentre un intervallo di stima

- si riferisce a un parametro incognito
- in esso contenuto a meno di non osservare un campione “cattivo”

e dunque si tratta di due cose diverse (adottando l’approccio frequentista e non quello bayesiano).

In secondo luogo possiamo assegnare a Y_{n+1}, come approssimazione, una distribuzione normale: nella misura in cui tale approssimazione è valida, essa ci consente di migliorare la nostra previsione per intervallo:

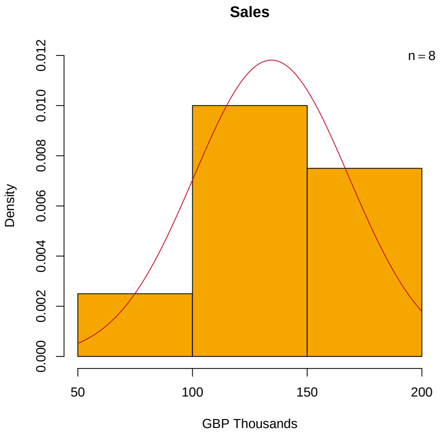
$$66.8 = 134.4 - 2 \times 33.8 \leq Y_{n+1} \leq 134.4 + 2 \times 33.8 = 202.0$$

con probabilità 95%, oppure

$$95.5 = 134.4 - 1.15 \times 33.8 \leq Y_{n+1} \leq 134.4 + 1.15 \times 33.8 = 173.3$$

con probabilità 75% (in quanto $z_{12.5\%} = 1.15$).

Si ricordi che il teorema di Chebyshev tipicamente sovrastima pesantemente la variabilità di una distribuzione (caso peggiore).


Sales
n = 8
Density
0.000
0.002
0.004
0.006
0.008
0.010
0.012
50
100
150
200
GBP Thousands

In realtà attribuendo direttamente $\bar{y}$ e sd_y alla distribuzione di Y_{n+1}, senza considerarli stime di μ e σ incognite, sottovalutiamo la nostra incertezza sul fenomeno oggetto di previsione.

Converrà allora porre $\mathbb{E}[Y_{n+1}] = \bar{y}$ e

$$\hat{se}(Y_{n+1}) = \sqrt{s_y^2 + \frac{s_y^2}{n}} = s_y\sqrt{1 + \frac{1}{n}}$$

dove $s_y = sd_y\sqrt{\frac{n}{n-1}}$ tiene conto dell'incertezza su σ e l'incertezza su μ è incorporata sommando il quadrato del suo errore standard stimato alla varianza stimata (fonti indipendenti di incertezza)...

...dopo di che assegneremo a $\frac{Y_{n+1}-\mathbb{E}[Y_{n+1}]}{\hat{se}(Y_{n+1})}$ la distribuzione t di Student con $n-1$ gradi di libertà (invece di quella normale standard).

Nell'esempio di Rees troveremo

$$\hat{se}(Y_{n+1}) = 33.8 \times \sqrt{\frac{8}{7}} \times \sqrt{1+\frac{1}{8}} = 33.8 \times \sqrt{\frac{9}{7}} = 38.3$$

e dunque (invece che $66.8 \leq Y_{n+1} \leq 202.0$) prevederemo

$$44.0 = 134.4 - 2.36 \times 38.3 \leq Y_{n+1} \leq 134.4 + 2.36 \times 38.3 = 224.8$$

con probabilità 95% (visto che $t_{0.025} = 2.36$ con 7 gradi di libertà).

Supponiamo ora che l'azienda si avvalga di un test attitudinale per la selezione dei suoi venditori:

in questo caso oltre alla variabile risposta Y (le vendite) abbiamo una variabile esplicativa X (il punteggio al test) sulla base della quale (conoscendo il punteggio ottenuto dal venditore neoassunto) è ragionevole sperare di migliorare la nostra previsione.

La tabella seguente (Rees, 2001) riporta le vendite realizzate (in migliaia di sterline inglesi, nell'ultimo esercizio) e il punteggio conseguito al test (su un'opportuna scala) da $n = 8$ venditori di una certa azienda: se il punteggio del neoassunto è $x_{n+1} = 60$, quale sarà la nostra previsione? come useremo questa informazione?

	Vendite (Y)	Punteggio (X)
Andrea	105	45
Barbara	120	75
Carlo	160	85
Diana	155	65
Elisa	70	50
Francesco	150	70
Giovanna	185	80
Massimo	130	55
Totale	1075	525

Previsione di un carattere quantitativo

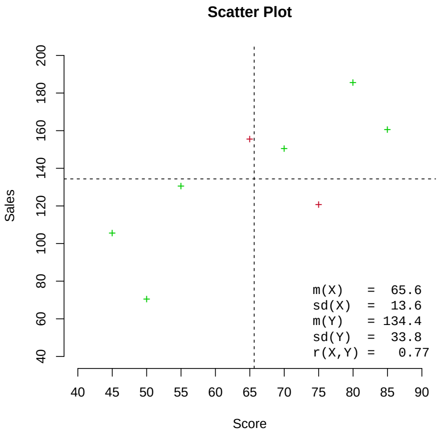
Correlazione lineare

Retta di regressione

Propensione marginale all'importazione

La distribuzione doppia (o congiunta) di due caratteri quantitativi (continui) X e Y si può rappresentare nel piano euclideo mediante un diagramma di dispersione: ogni unità statistica sarà rappresentata da un simbolo (es. una croce) centrato nel punto che ha come coordinate, rispetto a un'opportuna coppia di assi cartesiani ortogonali, i valori assunti da X e Y sull'unità rappresentata.

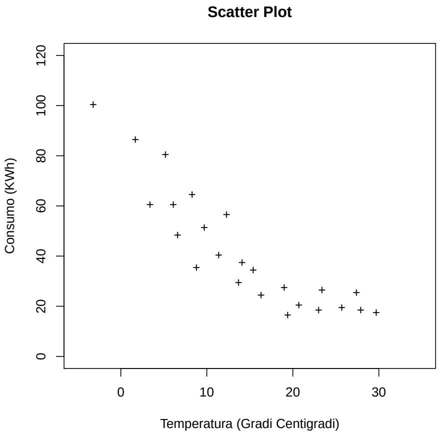
La figura seguente rappresenta i dati di Rees (in verde i venditori per i quali i caratteri sono entrambi "sopra media" o entrambi "sotto media", in rosso i venditori per i quali un carattere è "sopra media" e l'altro "sotto media")...


Scatter Plot
Sales
200
180
160
140
120
100
80
60
40
Score
40 45 50 55 60 65 70 75 80 85 90
m(X) = 65.6
sd(X) = 13.6
m(Y) = 134.4
sd(Y) = 33.8
r(X,Y) = 0.77

. . . diremo che c'è

- concordanza tra X e Y se modalità "grandi" di X si associano più frequentemente a modalità "grandi" di Y, mentre modalità "piccole" di X si associano più frequentemente a modalità "piccole" di Y (come nel precedente diagramma);
- discordanza tra X e Y se modalità "grandi" di X si associano più frequentemente a modalità "piccole" di Y, mentre modalità "piccole" di X si associano più frequentemente a modalità "grandi" di Y (es. temperatura e consumo di elettricità in una località fredda, come mostra il diagramma successivo. . . ovvero la Figura 16.2.2 di Borra & Di Ciaccio, 2008).

Questo tipo di associazione tra caratteri (ordinali) si dice anche cograduazione/contrograduazione.


Scatter Plot
Consumo (KWh)
0
20
40
60
80
100
120
0
10
20
30
Temperatura (Gradi Centigradi)

Diremo che c'è interdipendenza lineare tra X e Y se le loro modalità osservate tendono a disporsi lungo una retta, ovvero se la loro "nuvola di punti" ha forma (approssimativamente) "ellissoidale":

- questa tendenza è esibita in una certa misura dal diagramma di dispersione che rappresenta i dati di Rees;
- la Figura 16.2.2 di Borra & Di Ciaccio (2008) esibisce questa tendenza limitatamente ai valori bassi della temperatura (oppure ai valori elevati, ma con diversa pendenza... infatti in questo secondo caso il consumo di elettricità si mantiene praticamente costante... complessivamente si tratta di interdipendenza lineare "a tratti").

La concordanza/discordanza tra caratteri (quantitativi) linearmente interdipendenti si dice correlazione lineare...

...se è stato riscontrato per via grafica che qualitativamente X e Y sono linearmente correlate, una quantificazione della forza della loro associazione è fornita dal coefficiente di correlazione lineare tra X e Y:

$$-1 \quad \leq \quad r_{xy} \quad \leq \quad +1$$

dove il segno tiene conto della direzione in cui l'associazione si sviluppa (concordanza/discordanza).

In altre parole: il coefficiente di correlazione lineare è una misura algebrica di quanto una nuvola di punti ellissoidale (ellissoide) sia allungata (abbia un asse più lungo dell'altro) in diagonale.

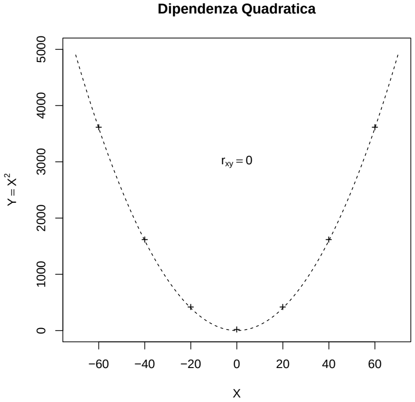
Il coefficiente di correlazione lineare r_{xy} gode delle seguenti proprietà:

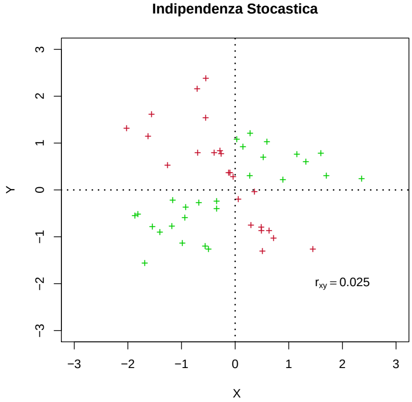
- $r_{xy} > 0$ se e soltanto se l'ellissoide è allungato in direzione Sud Ovest - Nord Est (concordanza);
- $r_{xy} < 0$ se e soltanto se l'ellissoide è allungato in direzione Nord Ovest - Sud Est (discordanza);
- $r_{xy} = 0$ se e soltanto se l'ellissoide ha gli assi paralleli agli assi coordinati (assenza di correlazione lineare);
- $|r_{xy}| = 1$ se e soltanto se le modalità osservate si dispongono esattamente lungo una retta inclinata (ellissoide degenere);
- tanto più grande è $|r_{xy}|$, tanto più l'ellissoide sarà allungato e i suoi punti si disporranno approssimativamente lungo una retta (vedremo che r_{xy}^2 potrà interpretarsi come una proporzione).

Osserviamo che:

- il caso dell'ellissoide degenere individuato da $r_{xy} = 1$ o $r_{xy} = -1$ è il caso di una "relazione funzionale lineare" tra X e Y (caso che sarà illustrato più avanti);
- il caso dell'ellissoide con assi paralleli agli assi coordinati ($r_{xy} = 0$) è il caso dell'indipendenza stocastica tra X e Y (nell'ambito dell'assunto qualitativo di interdipendenza lineare tra X e Y);
- il coefficiente di correlazione lineare non misura altri tipi di associazione tra X e Y (es. non misura la dipendenza quadratica di Y da X).

I due diagrammi di dispersione seguenti illustrano queste due ultime situazioni. . .


Dipendenza Quadratica
$r_{xy} = 0$
$Y = X^2$
X
0
1000
2000
3000
4000
5000
−60
−40
−20
0
20
40
60


Indipendenza Stocastica
$r_{xy} = 0.025$
Y
X
−3
−2
−1
0
1
2
3

Il coefficiente di correlazione lineare tra X e Y è definito come

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{sd_x}\right)\left(\frac{y_i - \bar{y}}{sd_y}\right)$$

ovvero come “media del prodotto delle variabili standardizzate”; si avrà

- un contributo positivo per ogni coppia di scarti concordi

$$x_i > \bar{x} \quad \text{e} \quad y_i > \bar{y} \quad \text{oppure} \quad x_i < \bar{x} \quad \text{e} \quad y_i < \bar{y},$$

- un contributo negativo per ogni coppia di scarti discordi

$$x_i > \bar{x} \quad \text{e} \quad y_i < \bar{y} \quad \text{oppure} \quad x_i < \bar{x} \quad \text{e} \quad y_i > \bar{y}.$$

Per il calcolo del coefficiente di correlazione lineare tra X e Y conviene avvalersi della formula

$$r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{sd_x sd_y}$$

dove $\overline{xy}$ è la media del prodotto delle due variabili:

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i;$$

la tabella seguente illustra il calcolo di $\overline{xy} = 73350/8 = 9168.8$ per l'esempio del venditore neoassunto e permette di trovare

$$r_{xy} = \frac{9168.8 - 134.4 \times 65.6}{33.8 \times 13.6} = 0.77.$$

	Vendite (Y)	Punteggio (X)	Prodotto (XY)
Andrea	105	45	4725
Barbara	120	75	9000
Carlo	160	85	13600
Diana	155	65	10075
Elisa	70	50	3500
Francesco	150	70	10500
Giovanna	185	80	14800
Massimo	130	55	7150
Totale	1075	525	73350

Previsione di un carattere quantitativo

Correlazione lineare

Retta di regressione

Propensione marginale all'importazione

Tornando al problema di prevedere Y_{n+1}, sapendo che $x_{n+1} = 60$, osserviamo innnazitutto che

se le osservazioni si disponessero esattamente lungo una retta, troveremmo una relazione funzionale lineare tra le vendite Y e il punteggio al test X, ovvero avremmo

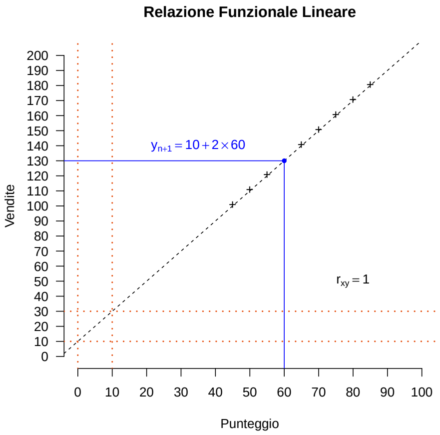
$$Y = \beta_0 + \beta_1 X$$

dove

- β_0 *è l'intercetta e*
- β_1 *è il coefficiente angolare (o "pendenza")*

della retta in questione (univocamente determinata dall'ellissoide degenere). . .

Retta di regressione



La figura precedente illustra la situazione limite di una relazione funzionale lineare tra Y e X prendendo $\beta_0 = 10$ e $\beta_1 = 2$, conservando $x_1, \ldots, x_8$ e sostituendo y_i con $\beta_0 + \beta_1 x_i$, per ogni $i = 1, \ldots, 8$.

Si noti come l'esistenza di una relazione funzionale lineare tra Y e X ci permetterebbe di ottenere, conoscendo il punteggio x_{n+1} ottenuto dal neoassunto al test attitudinale, una previsione esatta delle sue vendite:

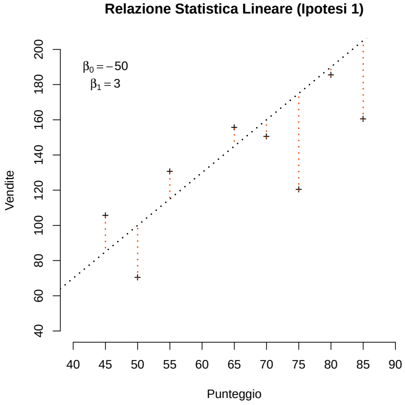
$$y_{n+1} = \beta_0 + \beta_1 x_{n+1}$$

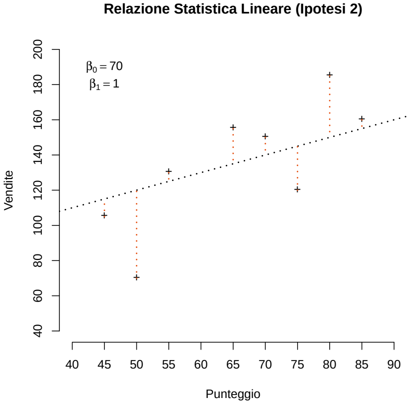
In realtà le osservazioni non si dispongono esattamente lungo una retta, ma si limitano a esibire una certa correlazione lineare (formando un ellissoide con $r_{xy} = 0.77$) di modo che non esiste una retta che passi per tutte le osservazioni, mentre ne esistono diverse che passano attraverso la nube (più o meno “vicine” alle varie osservazioni).

In questo caso possiamo ipotizzare una relazione statistica lineare tra le vendite Y e il punteggio al test X:

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

dove ϵ è un termine di errore (casuale) specifico del venditore; le due figure seguenti mostrano due possibili relazioni statistiche lineari da cui i dati dell’esempio potrebbero avere avuto origine...


Relazione Statistica Lineare (Ipotesi 1)
$\beta_0 = -50$
$\beta_1 = 3$
Vendite
40 60 80 100 120 140 160 180 200
40 45 50 55 60 65 70 75 80 85 90
Punteggio


Relazione Statistica Lineare (Ipotesi 2)
$\beta_0 = 70$
$\beta_1 = 1$
Vendite
40 60 80 100 120 140 160 180 200
40 45 50 55 60 65 70 75 80 85 90
Punteggio

Nell'ipotesi di relazione statistica lineare tra Y e X, dati i coefficienti di regressione β_0 e β_1 e l'errore standard di regressione $\sigma = sd(\epsilon)$, assumendo $\mathbb{E}[\epsilon] = 0$, la previsione di Y_{n+1} conoscendo x_{n+1} sarà

$$Y_{n+1} \quad = \quad \beta_0 + \beta_1 x_{n+1} \, \pm \, \sigma$$

e si tratterà di una previsione migliorata perché

$$\sigma < sd(Y_{n+1})$$

in quanto, supponendo ϵ_{n+1} e X_{n+1} indipendenti, ovvero supponendo la variabile X esogena, si trova

$$Var(Y_{n+1}) \quad = \quad \beta_1^2 \, Var(X_{n+1}) + \sigma^2.$$

Quali valori prenderemo per β_0, β_1 e σ?

Non conoscendo i valori veri (di popolazione) ricorreremo a delle stime $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\sigma}$ il cui calcolo (sulla base delle modalità osservate di X e Y) ci accingiamo a discutere...

...supponiamo, in prima battuta, di avere già trovato $\hat{\beta}_0$ e $\hat{\beta}_1$: si dice residuo dell'i-esima osservazione y_i rispetto alla retta con intercetta $\hat{\beta}_0$ e pendenza $\hat{\beta}_1$ la quantità empirica

$$e_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \hat{y}_i$$

ovvero l'errore di previsione stimato per l'i-esima osservazione, usando $\hat{\beta}_0$ e $\hat{\beta}_1$ come stime di β_0 e β_1; le quantità $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $i = 1, \ldots, n$, si dicono valori predetti.

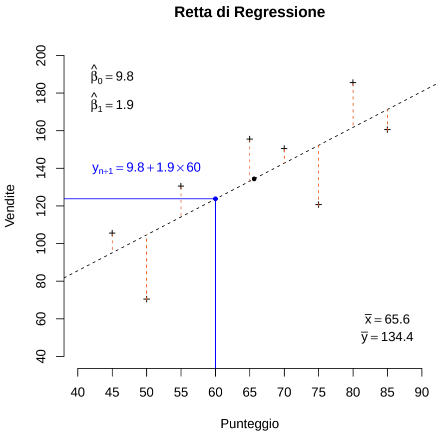
Poiché il residuo e_i stima l'errore di previsione ϵ_i, per $i = 1, \ldots n$, appare naturale (ed è sempre possibile) richiedere che $\hat{\beta}_0$ e $\hat{\beta}_1$ determinino residui a media nulla:

$$\bar{e} = \frac{e_1 + \cdots + e_n}{n} = 0$$

Stimeremo allora l'errore standard di regressione σ mediante la deviazione standard dei residui:

$$\hat{\sigma} = \sqrt{\frac{e_1^2 + \cdots + e_n^2}{n}}$$

Perché dunque non fare tutto il possibile per rendere piccola $\hat{\sigma}$?
In effetti è il criterio che si segue per trovare $\hat{\beta}_0$ e $\hat{\beta}_1$...



Retta di Regressione
$\hat{\beta}_0 = 9.8$
$\hat{\beta}_1 = 1.9$
$y_{n+1} = 9.8 + 1.9 \times 60$
$\bar{x} = 65.6$
$\bar{y} = 134.4$
Vendite
Punteggio
40 60 80 100 120 140 160 180 200
40 45 50 55 60 65 70 75 80 85 90

...più in dettaglio, definito lo Scarto Quadratico Medio della retta con intercetta β_0 e pendenza β_1 rispetto alle osservazioni $(x_1, y_1) \ldots (x_n, y_n)$ come

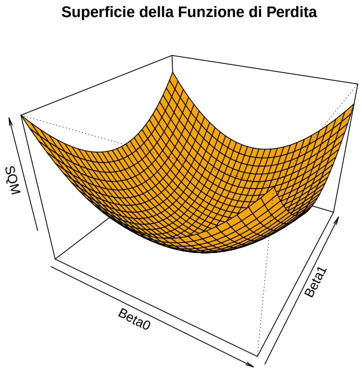
$$SQM(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

di modo che $\hat{\sigma} = \sqrt{SQM(\hat{\beta}_0, \hat{\beta}_1)}$, determineremo $\hat{\beta}_0$ e $\hat{\beta}_1$ minimizzando $SQM(\beta_0, \beta_1)$ sotto il vincolo (in realtà superfluo)

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = \bar{y} - \beta_0 - \beta_1 \bar{x} = 0.$$

La figura seguente mostra la superficie individuata da $SQM(\beta_0, \beta_1)$ al variare di β_0 e β_1 (nell'esempio dei venditori)...

Retta di regressione



Esiste un unico minimo (tale che $\bar{y} = \beta_0 + \beta_1 \bar{x}$)

...sono dunque univocamente determinate le stime ai minimi quadrati

$$\hat{\beta}_1 = r_{xy}\frac{sd_y}{sd_x}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

che nel caso del venditore neoassunto valgono

$$\hat{\beta}_1 = 0.77 \times \frac{33.8}{13.6} = 1.9$$
$$\hat{\beta}_0 = 134.4 - 1.9 \times 65.6 = 9.8$$

come già riportato nel grafico della retta di regressione.

Con riferimento all'esempio dei venditori, i coefficienti di regressione (stimati) godono della seguente interpretazione:

$$\hat{\beta}_1 = 1.9 \quad \text{migliaia di sterline / punto della scala del test}$$

è l'incremento nelle vendite corrispondente all'incremento di un punto nel risultato del test, mentre

$$\hat{\beta}_0 = 9.8 \quad \text{migliaia di sterline}$$

sono le vendite corrispondenti a un venditore che consegue punteggio nullo nel test...

...poiché i punteggi osservati vanno da un minimo di 45 a un massimo di 85, l'interpretazione di $\hat{\beta}_0$ non è affidabile (posto che abbia senso considerare un venditore...assunto...con 0 punti al test attitudinale).

A questo punto siamo <u>quasi</u> pronti a dare, per il venditore neoassunto con punteggio $x_{n+1} = 60$, la previsione

$$Y_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \pm \hat{\sigma}$$

dove $\hat{\beta}_0 = 9.8$, $\hat{\beta}_1 = 1.9$ e

$$\hat{\sigma} = \sqrt{SQM(\hat{\beta}_0, \hat{\beta}_1)} = \sqrt{\frac{e_1^2 + \cdots + e_n^2}{n}}$$

dove il “quasi” è relativo al fatto che ci conviene prima introdurre una formula alternativa per la stima dell’errore standard di regressione. . .

...basata sulla seguente decomposizione della varianza totale (varianza marginale di Y):

$$\begin{aligned} sd_y^2 &= \frac{1}{n}\sum_{i=1}^{n}(\hat\beta_0 + \hat\beta_1 x_i - \bar y)^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat\beta_0 - \hat\beta_1 x_i)^2 \\ &= R^2 \times sd_y^2 + \hat\sigma^2 \end{aligned}$$

dove R^2 si dice coefficiente di determinazione della retta di regressione (è la proporzione di varianza totale spiegata da X).

Si dimostra che $R^2 = r_{xy}^2$ e se ne ricava la seguente formula alternativa per la stima dell'errore standard di regressione:

$$\hat\sigma = sd_y \times \sqrt{1 - r_{xy}^2}$$

di modo che $\hat\sigma \leq sd_y$ (uguaglianza solo nel caso in cui $r_{xy} = 0$).

Nell'esempio del venditore neoassunto si trova

$$R^2 = (0.77)^2 = 0.5929$$

di modo che la retta di regressione di Y (vendite) su X (punteggio al test attitudinale) spiega il 59% della varianza totale; ne segue

$$\hat{\sigma} = \sqrt{1 - 0.5929} \times 33.8 = 0.638 \times 33.8 = 21.6$$

migliaia di sterline e quindi (se $x_{n+1} = 60$) prevederemo

$$Y_{n+1} = 9.8 + 1.9 \times 60 \pm 21.6 = 123.8 \pm 21.6.$$

laddove (ignorando x_{n+1}) prevedevamo

$$Y_{n+1} = \bar{y} \pm sd_y = 134.4 \pm 33.8$$

L'informazione $x_{n+1} = 60$ corregge e precisa la nostra previsione.

- una previsione puntuale più bassa delle vendite del neoassunto, da 134.8 a 123.8 migliaia di sterline, perché il neoassunto ha conseguito un punteggio al test attitudinale inferiore alla media;
- una minore incertezza nella previsione delle vendite del neoassunto, nella misura in cui l'errore standard di previsione passa da 33.8 a 21.6 migliaia di sterline (errore standard di regressione) in virtù della correlazione lineare tra vendite e punteggio.

La minore incertezza nella previsione si traduce, a parità di probabilità, in un intervallo di previsione più corto...

…se si può usare una distribuzione normale per Y_{n+1}, quando si sappia che $X_{n+1} = 60$, assegnando

$$\begin{aligned} \mathbb{E}[Y_{n+1}|X_{n+1} = 60] &= 123.8 \\ sd(Y_{n+1}|X_{n+1} = 60) &= 21.6 \end{aligned}$$

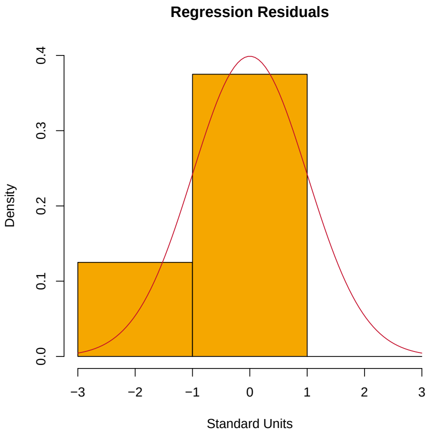
troveremo la previsione per intervallo

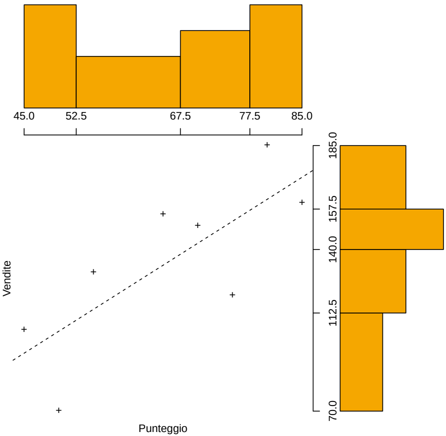
$$80.6 = 123.8 - 2 \times 21.6 \leq Y_{n+1} \leq 123.8 + 2 \times 21.6 = 167.0$$

con probabilità 95%, da confrontare con la previsione incondizionata

$$66.8 = 134.4 - 2 \times 33.8 \leq Y_{n+1} \leq 134.4 + 2 \times 33.8 = 202.0$$

di modo che la lunghezza sarà scesa da 135 a 86 migliaia di sterline.


Regression Residuals
Density
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
Standard Units


45.0
52.5
67.5
77.5
85.0
185.0
157.5
140.0
112.5
70.0
Vendite
Punteggio

In realtà ponendo $sd(Y_{n+1}|X_{n+1} = x_{n+1}) = \hat{\sigma}$ sottovalutiamo la nostra incertezza sul fenomeno oggetto di previsione:

- usando i dati non solo per stimare σ, ma anche per stimare β_0 e β_1, converrà sostituire $\hat{\sigma}$ con la sua versione "corretta"

$$\hat{\sigma}_+ = \hat{\sigma}\sqrt{\frac{n}{n-2}};$$

- inoltre i valori $\hat{\beta}_0$ e $\hat{\beta}_1$ sono affetti da incertezza e questa si rifletterà sull'assegnazione $\mathbb{E}[Y_{n+1}|X_{n+1} = x_{n+1}] = \hat{\mu}_{x_{n+1}}$, dove abbiamo introdotto la quantità $\hat{\mu}_{x_{n+1}} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} \ldots$

...in effetti $\hat{\beta}_0$ e $\hat{\beta}_1$ sono stime (corrette) con errori standard

$$\begin{aligned} se(\hat{\beta}_0) &= \frac{\sigma}{\sqrt{n}}\sqrt{1+\left(\frac{\bar{x}}{sd_x}\right)^2} \\ se(\hat{\beta}_1) &= \frac{\sigma}{sd_x\sqrt{n}} \end{aligned}$$

provenienti da stimatori[1] con correlazione $-\bar{x}\frac{se(\hat{\beta}_1)}{se(\hat{\beta}_0)}$ di modo che

$se(\hat{\mu}_{x_{n+1}}) = \frac{\sigma}{\sqrt{n}}\sqrt{1+\left(\frac{x_{n+1}-\bar{x}}{sd_x}\right)^2}$ e di conseguenza...

[1] Gli stimatori ai minimi quadrati di β_0 e β_1 hanno varianza minima tra quelli corretti che sono funzioni lineari dei dati (Teorema di Gauss-Markov).

...l'errore standard di previsione, comprensivo di tutta la nostra incertezza, sarà

$$\begin{aligned} \hat{se}(Y_{n+1}|X_{n+1} = x_{n+1}) &= \sqrt{\hat{\sigma}_+^2 + \hat{se}(\hat{\mu}_{x_{n+1}})^2} \\ &= \hat{\sigma}_+ \sqrt{1 + \frac{1}{n}\left[1 + \left(\frac{x_{n+1} - \bar{x}}{sd_x}\right)^2\right]} \end{aligned}$$

tanto maggiore quanto più x_{n+1} è lontano da $\bar{x}$; si veda al riguardo la Figura 17.4.1 di Borra & Di Ciaccio (2008).

Al crescere di n l'incertezza sulla nostra previsione si riduce all'errore standard di regressione σ, ma resta il problema che la previsione non sarà affidabile se x_{n+1} non appartiene al campo di variazione dei dati: estrapolare è più azzardato che interpolare.

Nell'esempio di Rees la stima "corretta" dell'errore standard di regressione sarà

$$\hat{\sigma}_+ = 21.6 \times \sqrt{\frac{8}{6}} = 24.9$$

e quindi gli errori standard (stimati) di β_0 e β_1 saranno

$$\hat{se}_0 = \frac{24.9}{\sqrt{8}} \times \sqrt{1 + \left(\frac{65.6}{13.6}\right)^2} = 43.4$$

$$\hat{se}_1 = \frac{24.9}{13.6 \times \sqrt{8}} = 0.65$$

di modo che scriveremo (stime puntuali con errore standard)

$$\begin{aligned} \beta_0 &= 9.8 \pm 43.4 \\ \beta_1 &= 1.9 \pm 0.65 \end{aligned}$$

per esprimere la nostra incertezza sui valori dell'intercetta e del coefficiente angolare.

In particolare, concentrandoci sul parametro che ci interessa di più, per la sua interpretazione, troveremo che l'incremento nelle vendite corrispondente all'incremento di un punto nel test attitudinale potrebbe tranquillamente essere di sole mille sterline (o di ben tremila sterline).

Un intervallo di confidenza per β_1 può trovarsi mediante un opportuno coefficiente di confidenza basato sulla distribuzione t di Student con $n-2$ gradi di libertà:

$$\begin{array}{rcccl} \hat{\beta}_1 - t_\alpha \hat{se}_1 & \leq & \beta_1 & \leq & \hat{\beta}_1 + t_\alpha \hat{se}_1 \\ 1.9 - 2.45 \times 0.65 & \leq & \beta_1 & \leq & 1.9 + 2.45 \times 0.65 \\ 1.9 - 1.6 & \leq & \beta_1 & \leq & 1.9 + 1.6 \end{array}$$

se prendiamo $\alpha = 0.025$ (con $n-2=6$); concluderemo

$$0.3 \quad \leq \quad \beta_1 \quad \leq \quad 3.5$$

al livello di confidenza 95%.

La verifica dell'ipotesi di irrilevanza del test attitudinale

$$H_0 : \quad \beta_1 = 0$$

può effettuarsi mediante la statistica test (osservata)

$$t = \frac{\hat{\beta}_1}{\hat{se}_1} = \frac{1.9}{0.65} = 2.92$$

da confrontarsi con un opportuno valore critico della distribuzione t di Student con $n-2$ gradi di libertà: se prendiamo $\alpha = 0.025$ (con $n-2=6$) questo sarà $t_\alpha = 2.45$ e concluderemo che il test attitudinale contribuisce significativamente a spiegare le vendite; la stessa conclusione si ottiene osservando che il valore zero non appartiene all'intervallo di confidenza.

L'errore standard di $\hat{\mu}_{x_{n+1}} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1} = 9.8 + 1.9 \times 60 = 123.8$ può stimarsi come

$$\hat{se}(\hat{\mu}_{x_{n+1}}) = \frac{24.9}{\sqrt{8}} \sqrt{1 + \left(\frac{60 - 65.6}{13.6}\right)^2} = 9.5$$

e quindi per il parametro $\mu_{x_{n+1}} = \beta_0 + \beta_1 x_{n+1}$ (per le vendite medie dei venditori con punteggio x_{n+1}) avremo la stima con errore standard

$$\mu_{x_{n+1}} = 123.8 \pm 9.5$$

da cui eventualmente ricaveremo un intervallo di confidenza usando la distribuzione t di Student con $n - 2 = 6$ gradi di libertà.

Infine, per quanto riguarda la previsione delle vendite di un venditore neoassunto con punteggio $x_{n+1} = 60$, troveremo

$$\hat{se}(Y_{n+1}|X_{n+1} = 60) = \sqrt{(24.9)^2 + (9.5)^2} = 26.7$$

da confrontare con $\hat{se}(Y_{n+1}) = 38.3$ (se non conosciamo x_{n+1}); usando la distribuzione t di Student con $n - 2 = 6$ gradi di libertà per $\frac{Y_{n+1} - \mathbb{E}[Y_{n+1}|X_{n+1}=x_{n+1}]}{\hat{se}(Y_{n+1}|X_{n+1}=x_{n+1})}$, prevederemo quindi

$$58.4 = 123.8 - 2.45 \times 26.7 \leq Y_{n+1} \leq 123.8 + 2.45 \times 26.7 = 189.2$$

con probabilità 95% (lunghezza 131 da confrontare con 181).

Abbiamo visto che il coefficiente di determinazione $R^2 = r^2_{xy}$ è la proporzione di varianza totale spiegata dalla retta di regressione...

...in quanto tale è una misura della bontà di adattamento[2] della retta di regressione ai dati (quindi della sua utilità) se sussistono le condizioni per stimarla (ai minimi quadrati):

- linearità della relazione statistica

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

ovvero correlazione lineare tra Y e X;

[2] Al riguardo si veda anche la Sezione 17.3 di Borra & Di Ciaccio (2008) sulla tavola ANOVA e sul test F.

- omoschedasticità
$$sd(\epsilon_i) = \sigma$$
ovvero stessa deviazione standard per tutti i termini di errore;
- indipendenza stocastica dei termini d'errore $\epsilon_1, \ldots, \epsilon_n$ (in caso contrario SQM non è la funzione di perdita più adatta);
- normalità dei termini di errore
$$\epsilon_i \sim \mathcal{N}(0, \sigma)$$
che permette di interpretare i minimi quadrati in termini di "massima verosimiglianza" (in caso contrario non è detto che i minimi quadrati siano l'approccio migliore) e di usare una distribuzione t di Student a fini inferenziali (stima e previsione).

L'analisi dei residui permette di controllare se sussistono o meno le condizioni per stimare la retta di regressione (ai minimi quadrati).

Borra e Di Ciaccio (2008, Sezione 17.5) discutono questo argomento, mostrando fra l'altro grafici dei residui che segnalano:

- relazioni statistiche non lineari;
- eteroschedasticità (termini di errore con deviazioni standard diverse fra loro);
- termini di errore fra loro dipendenti (correlati);
- termini di errore non normali.

I due grafici seguenti, relativi all'esempio dei venditori, non mostrano evidenza contro linearità, omoschedasticità e indipendenza (si suppongano i venditori in ordine di assunzione)...

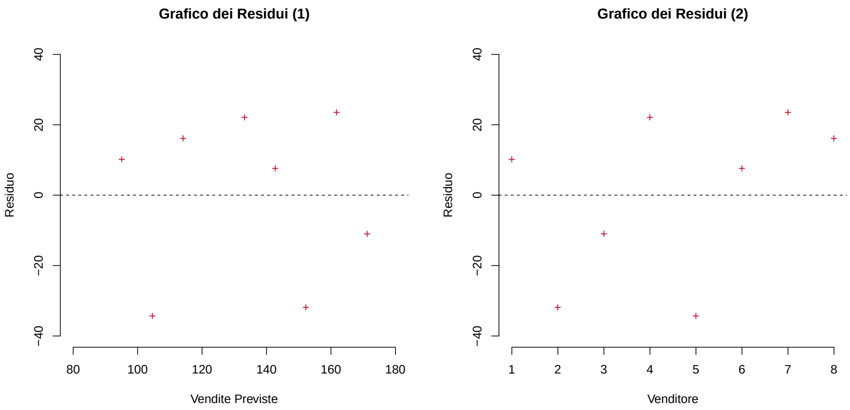


Grafico dei Residui (1)
Residuo
40
20
0
−20
−40
80
100
120
140
160
180
Vendite Previste
Grafico dei Residui (2)
Residuo
40
20
0
−20
−40
1
2
3
4
5
6
7
8
Venditore

Un paio di avvertenze finali (prima di passare a un'applicazione):

- La stima ai minimi quadrati della retta di regressione è piuttosto sensibile alla presenza di valori anomali; questi possono essere individuati ispezionando visivamente il diagramma di dispersione (Borra & Di Ciaccio, 2008, Sezione 17.6).
- L'evidenza di un'associazione (lineare) tra due caratteri statistici non implica l'esistenza di una relazione causale tra gli stessi; si veda al riguardo la Sezione 17.7 di Borra & Di Ciaccio (2008).

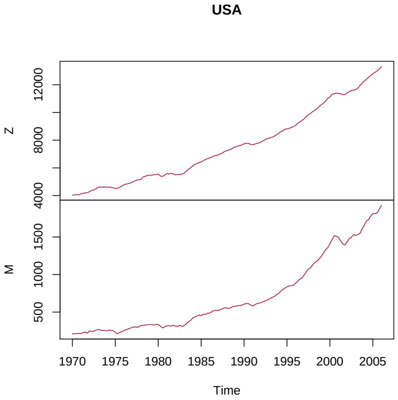
Vediamo ora un'applicazione in ambito macroeconomico, su dati relativi agli Stati Uniti d'America (cortesia del prof. G. Bonifati).

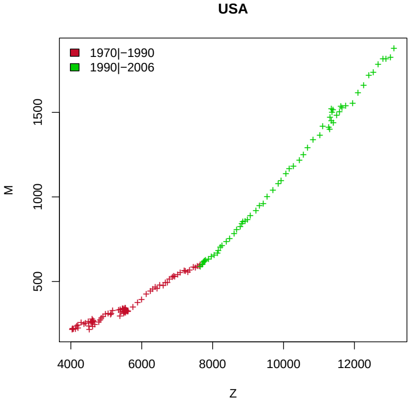
La variabili di interesse sono

$$\begin{aligned} M &= \text{“importazioni”} \\ Z &= \text{“spesa complessiva”} \end{aligned}$$

dove la spesa complessiva (PIL) è il totale di “consumi”, “investimenti”, “spesa pubblica” ed “esportazioni nette”: $Z = C + I + G + (X - M)$; si noti che M è parte di $C + I + G$ e dunque, nel calcolo di Z, si elide.

È disponibile la serie storica di M e Z (cadenza quadrimestrale) dal primo quadrimestre del 1970 al primo quadrimestre del 2006; M e Z sono misurate in “miliardi di dollari attualizzati al 2000”...


USA
Z
M
12000
8000
4000
1500
1000
500
1970
1975
1980
1985
1990
1995
2000
2005
Time


USA
1970|−1990
1990|−2006
M
1500
1000
500
4000
6000
8000
10000
12000
Z

Restringendo l'attenzione a un'opportuna finestra temporale, si può ipotizzare una relazione statistica lineare tra le importazioni M e la spesa complessiva Z:

$$M = \beta_0 + \beta_1 Z + \epsilon.$$

Il coefficiente β_1 è detto propensione marginale all'importazione (di un dato paese, qui gli USA, in un dato momento storico, qui ancora da determinare); essa indica

l'aumento delle importazioni corrispondente a un aumento unitario (1 miliardo di dollari) nella spesa complessiva (PIL)

ed è l'oggetto di interesse nell'analisi che segue.

```
> load("datiUSAbyBonifati.rda")
> datiUSAantichi <- as.data.frame(datiUSA)[time(datiUSA) < 1990,]
> regLin1 <- lm(M~Z, data = datiUSAantichi)
> summary(regLin1)

Call:
lm(formula = M ~ Z, data = datiUSAantichi)

Residuals:
    Min      1Q  Median      3Q     Max
-45.039 -14.161   5.066  16.517  37.535

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.604e+02  1.289e+01  -20.20   <2e-16 ***
Z            1.105e-01  2.279e-03   48.47   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.33 on 78 degrees of freedom
Multiple R-squared: 0.9679, Adjusted R-squared: 0.9675
F-statistic:  2349 on 1 and 78 DF,  p-value: < 2.2e-16
```

```
> load("datiUSAbyBonifati.rda")
> datiUSArecenti<-as.data.frame(datiUSA)[time(datiUSA) >= 1990,]
> regLin2 <- lm(M~Z, data = datiUSArecenti)
> summary(regLin2)

Call:
lm(formula = M ~ Z, data = datiUSArecenti)

Residuals:
    Min      1Q  Median      3Q     Max
-40.051 -13.701  -4.225  14.691  70.703

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.264e+03  1.549e+01  -81.65   <2e-16 ***
Z            2.387e-01  1.512e-03  157.89   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.04 on 63 degrees of freedom
Multiple R-squared: 0.9975, Adjusted R-squared: 0.9974
F-statistic: 2.493e+04 on 1 and 63 DF,  p-value: < 2.2e-16
```

Pertanto, se nel ventennio 1970-1990 la propensione marginale all'importazione degli USA era pari a

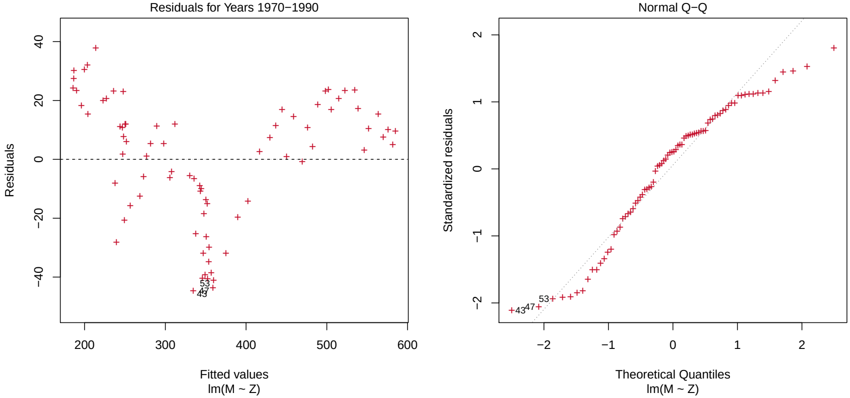
$$11.05 \pm 0.23$$

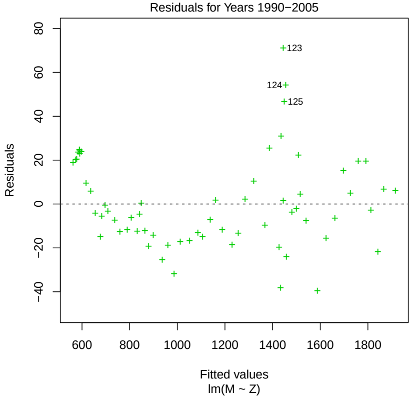
punti percentuali, nel periodo dal 1990 al 2005 essa si attesta al livello di

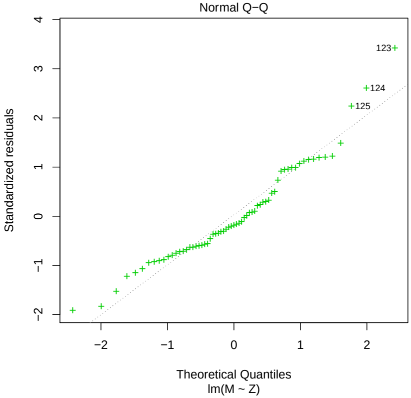
$$23.87 \pm 0.15$$

punti percentuali (circa il doppio del livello precedente).

I valori di R^2 (96.79% nel primo caso, 99.75% nel secondo) indicano un ottimo adattamento locale della retta di regressione (modello lineare).


Residuals for Years 1970−1990
Residuals
Fitted values
lm(M ~ Z)
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(M ~ Z)
53
47
43


Residuals for Years 1990−2005
Residuals
Fitted values
lm(M ~ Z)
123
124
125


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(M ~ Z)
123
124
125

- BORRA, S. & DI CIACCIO, A. (2008).
  Statistica: Metodologie per le Scienze Economiche e Sociali (Seconda Edizione).
  McGraw-Hill, Milano.
- REES, D. G. (2001).
  Essential Statistics.
  CRC Press, Boca Raton.