Modelli Multilineari e
Misure di adeguatezza del modello

Sommario

- Regressione multilineare
- Coefficiente di determinazione (modelli lineari e non lineari)
- Matrice di correlazione (modelli lineari)
- Matrice di correlazione asintotica (estensione a modelli non lineari)
- Analisi dei residui (modelli lineari e non lineari)

Regressioni multilineari

- Molte applicazioni di analisi della regressione coinvolgono situazioni con più di una singola variabile indipendente.
- Un modello lineare che contiene più di una variabile indipendente è detto **multilineare**
- Spesso modelli multilineari sono ottenuti a valle di linearizzazioni di modelli nonlineari.
- Continuiamo a a considerare problemi in cui si misura **una sola** variabile dipendente.

Regressioni multilineari

- Il modello dell'esperimento è:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

- Abbiamo $p=k+1$ parametri ed N condizioni sperimentali

$$\underline{\beta} = \left\| \beta_0, \beta_1 \ldots \beta_k \right\|^T$$

$$\underline{y} = \left\| y_1, y_2 \ldots y_N \right\|^T$$

- **Per poter stimare i parametri N>>p**

- Tipo di esperimento: (comunque esperimenti indipendenti)

  - ε_i:=N(0, σ^2)

Regressioni multilineari

- Il criterio della massima verosimiglianza nelle ipotesi in cui ci siamo messi porta ad uno stimatore di tipo minimi quadrati.

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

- Come al solito vogliamo minimizzare questa sommatoria al variare dei parametri da stimare:

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0$$

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k$$

Regressioni multilineari

- Con un po' di algebra si arriva ad un sistema di p equazioni in p incognite

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} &= \sum_{i=1}^{n} x_{i1} y_i \\
\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots & \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 &= \sum_{i=1}^{n} x_{ik} y_i
\end{aligned}
$$

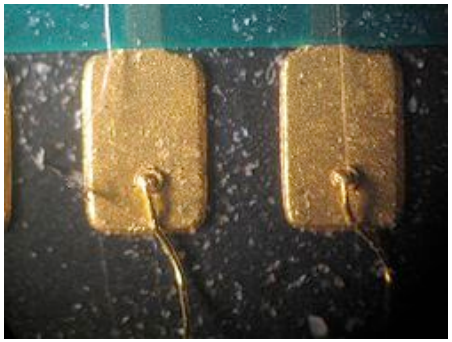
Regressioni multilineari: Esempio

- Resistenza a trazione di un wire bond per semiconduttori

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

- Modello empirico dell'esperimento:

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



Regressioni multilineari: Esempio

- Grandezze utili per i calcoli

$$n = 25, \sum_{i=1}^{25} y_i = 725.82$$

$$\sum_{i=1}^{25} x_{i1} = 206, \sum_{i=1}^{25} x_{i2} = 8{,}294$$

$$\sum_{i=1}^{25} x_{i1}^2 = 2{,}396, \sum_{i=1}^{25} x_{i2}^2 = 3{,}531{,}848$$

$$\sum_{i=1}^{25} x_{i1}x_{i2} = 77{,}177, \sum_{i=1}^{25} x_{i1}y_i = 8{,}008.37, \sum_{i=1}^{25} x_{i2}y_i = 274{,}811.31$$

Regressioni multilineari: Esempio

- Sistema di equazioni

$$
\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} &= \sum_{i=1}^{n} x_{i1}y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 &= \sum_{i=1}^{n} x_{i2}y_i
\end{aligned}
$$

- Con i dati

$$
\begin{aligned}
25\hat{\beta}_0 + 206\hat{\beta}_1 + 8294\hat{\beta}_2 &= 725.82 \\
206\hat{\beta}_0 + 2396\hat{\beta}_1 + 77{,}177\hat{\beta}_2 &= 8{,}008.37 \\
8294\hat{\beta}_0 + 77{,}177\hat{\beta}_1 + 3{,}531{,}848\hat{\beta}_2 &= 274{,}811.31
\end{aligned}
$$

- Soluzione

$$\hat{\beta}_0 = 2.26379, \quad \hat{\beta}_1 = 2.74427, \quad \hat{\beta}_2 = 0.01253$$

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

Regressioni multilineari

- Nei problemi multilineari conviene usare un approccio matriciale
- Il modello lo scriviamo così:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

X_{ij}: i parametro 1..k+1
j prova 1..n

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Regressioni multilineari

- Qualche richiamo di algebra lineare

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}, \quad \underline{\underline{X}} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{pmatrix}, \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$$

$$\underline{\underline{X}} \cdot \underline{\beta} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} \\ \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} \\ \beta_0 + \beta_1 x_{31} + \beta_2 x_{32} \\ \beta_0 + \beta_1 x_{41} + \beta_2 x_{42} \end{pmatrix}$$

Regressioni multilineari

- Qualche richiamo di algebra lineare

$$\underline{\underline{X}}^T \cdot \underline{\underline{X}} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} \\ x_{12} & x_{22} & x_{32} & x_{42} \end{pmatrix} \cdot \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{pmatrix} =$$

$$= \begin{pmatrix} 4 & x_{11}+x_{21}+x_{31}+x_{41} & x_{12}+x_{22}+x_{32}+x_{42} \\ x_{11}+x_{21}+x_{31}+x_{41} & x_{11}^2+x_{21}^2+x_{31}^2+x_{41}^2 & x_{11}x_{12}+x_{21}x_{22}+x_{31}x_{32}+x_{41}x_{42} \\ x_{12}+x_{22}+x_{32}+x_{42} & x_{11}x_{12}+x_{21}x_{22}+x_{31}x_{32}+x_{41}x_{42} & x_{12}^2+x_{22}^2+x_{32}^2+x_{42}^2 \end{pmatrix}$$

$$\underline{y}^T \cdot \underline{\underline{X}} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}^T \cdot \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ 1 & x_{31} & x_{32} \\ 1 & x_{41} & x_{42} \end{pmatrix} = \begin{pmatrix} y_1+y_2+y_3+y_4 \\ x_{11}y_1+x_{21}y_2+x_{31}y_3+x_{41}y_4 \\ x_{12}y_1+x_{22}y_2+x_{32}y_3+x_{42}y_4 \end{pmatrix}$$

Regressione multilineare

- Modello del processo:

$$g_i\left(\underline{\beta}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki}$$
$$\underline{g}\left(\underline{\beta}\right) = \underline{\underline{X}} \cdot \underline{\beta}$$

- La funzione scalare dei p parametri da minimizzare è in generale:(Forma quadratica)

$$\Phi\left(\underline{\beta}\right) = \underline{\varepsilon} \cdot \underline{\varepsilon} = \left(\underline{y} - \underline{g}\left(\underline{\beta}\right)\right)^T \cdot \left(\underline{y} - \underline{g}\left(\underline{\beta}\right)\right) = \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right)^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right)$$

- Il modello è lineare nei parametri quindi per determinare i parametri dobbiamo determinare le derivate della funzione obiettivo rispetto ai vari parametri ed uguagliarle a zero.

Regressioni multilineari

- Qualche richiamo di algebra lineare

$$\underline{y}^T \cdot \underline{\underline{X}} = \underline{y}^T \cdot \underline{\underline{X}}^T \cdot \underline{y}$$

$$\begin{aligned}
\frac{\partial \Phi\left(\underline{\beta}\right)}{\partial \underline{\beta}} &= \frac{\partial}{\partial \underline{\beta}} \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right)^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right) \\
&= -\underline{\underline{X}}^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right) + \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right)^T \cdot \left(-\underline{\underline{X}}\right) \\
&= -\underline{\underline{X}}^T \cdot \underline{y} + \underline{\underline{X}}^T \cdot \underline{\underline{X}} \cdot \underline{\beta} - \underline{y}^T \cdot \underline{\underline{X}} + \left(\underline{\underline{X}} \cdot \underline{\beta}\right)^T \cdot \underline{\underline{X}} \\
&= 2\underline{\underline{X}}^T \cdot \underline{\underline{X}} \cdot \underline{\beta} - 2\underline{\underline{X}}^T \cdot \underline{y}
\end{aligned}$$

Regressione multilineare

- In questo modo si perviene ad un sistema di p equazioni lineari in p incognite (i parametri)

$$\frac{\partial \Phi\left(\underline{\beta}\right)}{\partial \underline{\beta}} = -2\underline{\underline{X}}^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right)$$

$$\underline{\underline{X}}^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right) = \underline{0}$$

$$ovvero$$

$$\left(\underline{\underline{X}}^T \cdot \underline{\underline{X}}\right) \cdot \underline{\hat{\beta}} = \underline{\underline{X}}^T \cdot \underline{y}$$

Sistema di equazioni lineari

Matrice dei coefficienti

Vettore dei termini noti

- Perché esistano soluzioni il determinante della matrice dei coefficienti deve essere non nullo (ma dipende dalle condizioni sperimentali).

Regressione multilineare

- In definitiva, risolvendo in modo formale:

$$\hat{\beta} = \left(\underline{\underline{X}}^T \cdot \underline{\underline{X}}\right)^{-1} \cdot \left(\underline{\underline{X}}^T \cdot \underline{y}\right) \qquad (1)$$

  - Da un punto di vista pratico non conviene risolvere il sistema di equazioni lineari procedendo attraverso l'inversione della matrice caratteristica.

  - Conviene invece procedere alla soluzione del sistema di equazioni lineari con algoritmi che riducano l'onere calcolativo, per esempio con il metodo di Gauss o derivati. (vedrete un esempio alle esercitazioni)

Regressione multilineare

- Quindi il problema è di nuovo:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

- Il modello con i parametri stimati è $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

- Esercizio da fare a casa: ripetete con l'approccio matriciale l'esempio nel lucido 7

Regressione multilineare

- Come appare evidente dalla (1) lo stimatore dei parametri dipende linearmente dai risultati sperimentali per cui è una VA dello stesso tipo.

- Nelle ipotesi fatte le proprietà dello stimatore sono:

$$E\left(\underline{\hat{\beta}}\right) = \underline{\beta}$$
$$\underline{\hat{\beta}} := N\left(\underline{\beta}, \underline{\underline{V}}_{\underline{\hat{\beta}}}\right) \qquad (2)$$

$$\underline{\underline{V}}_{\underline{\hat{\beta}}} = E\left[\underbrace{\underbrace{\left(\underline{\hat{\beta}} - \underline{\beta}\right)}_{P\times 1}\underbrace{\left(\underline{\hat{\beta}} - \underline{\beta}\right)^{T}}_{1\times P}}_{P\times P}\right] \qquad (3)$$

Regressione multilineare

- Con la 1 la 2 e la 3 si ottiene:

$$\underline{\underline{V}}_{\underline{\hat{\beta}}} = \sigma^2 \left(\underline{\underline{X}}^T \cdot \underline{\underline{X}}\right)^{-1}$$

Matrice di Covarianza

$$\Phi_{\min} = \Phi\left(\underline{\hat{\beta}}\right) = \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right)^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right)$$

- La stima basata sulla MV della varianza:

$$\hat{\sigma}^2 = \frac{\left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right)^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right)}{N} = \frac{SS_E}{N}$$

  - NB il numeratore è il minimo della funzione obiettivo. Questa espressione non viene utilizzata perché parziale in genere si usa la forma non distorta:

$$s^2 = \frac{\left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right)^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right)}{N - P} = \frac{SS_E}{N - P}$$

Regressione multilineare

RIPARAMETRIZZAZIONE

- Abbiamo già studiato questa procedura nel caso banale di una unica variabile indipendente:

$$y_i = \beta_0 + \beta_1 (x_i - \overline{x}) + \varepsilon_i$$

- Ovviamente possiamo generalizzare la riparametrizzazione ai modelli multilineari, ma non sarà in generale possibile disaccoppiare il sistema di equazioni necessarie a stimare i parametri.

- **L'unica equazione a disaccoppiarsi è quella per la valutazione dell'intercetta**

Regressione multilineare

- Il modello diventa:

$$y_i = \beta_0 + \beta_1 (x_{1i} - \bar{x}_1) + \beta_2 (x_{2i} - \bar{x}_2) + ... + \beta_k (x_{ki} - \bar{x}_k) + \varepsilon_i$$
$$b_i = \beta_i$$
$$b_0 = \beta_0 - \sum_{j=1}^{k} \beta_j \bar{x}_j$$

- In definitiva la stima consiste nella risoluzione di P-1 equazioni lineari.

- **Il vantaggio sta nel miglioramento delle proprietà della matrice caratteristica**.
  - In genere conviene riparametrizzare.

Regressione multilineare

- Che tipo di VA è lo stimatore per la varianza?

$$\chi^2_N = \frac{\left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right)^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\beta}\right)}{\sigma^2}$$

- La precedente è una forma quadratica senza doppi prodotti.

- Senza addentrarci in dettagli ma estendendo le conclusioni del caso lineare semplice possiamo partizionare la precedente chi-quadro ed ottenere:

$$\chi^2_N = Q_1 + Q_2$$

$$Q_1 = \frac{\left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right)^T \cdot \left(\underline{y} - \underline{\underline{X}} \cdot \underline{\hat{\beta}}\right)}{\sigma^2}$$

$$Q_2 = \frac{\left(\underline{\hat{\beta}} - \underline{\theta}\right)^T \cdot \underline{\underline{X}}^T \cdot \underline{\underline{X}} \cdot \left(\underline{\hat{\beta}} - \underline{\theta}\right)}{\sigma^2}$$

Regressione multilineare

- Inoltre si dimostra che

$$Q_1 = \chi^2_{N-P}, \quad Q_2 = \chi^2_P$$
$$inoltre\, le\, due\, Q\, sono\, indipendenti$$

- Lo stimatore della varianza sperimentale è indipendente dallo stimatore dei parametri

$$s^2 = \chi^2_{N-P} \frac{\sigma^2}{N-P}$$

- La Q_2 sarà utilizzata quando dovremo determinare le regioni di fiducia.

Introduzione

- Nelle precedenti sezioni si è visto come stimare i parametri di un modello matematico a partire da una campagna sperimentale
- L'errore sperimentale, inevitabilmente presente nella misura, non permette mai di trarre delle conclusioni certe e i parametri del modello sono affetti da incertezze

- Un problema essenziale è la verifica della validità del modello

  – Il modello è adeguato per descrivere i dati sperimentali a disposizione?

- Non esiste una risposta definitiva a questa domanda (l'errore sperimentale non permette di trarre delle conclusioni certe), ma esistono delle tecniche che possono essere implementate per avere utili informazioni al riguardo

Stima con i Minimi Quadrati

- Riscriviamo la SS_E valutata nel suo minimo:

$$SS_E = \sum_{i=1}^{N}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \sum_{i=1}^{N}\left(y_i - \bar{y} - \hat{\beta}_1\bar{x} - \hat{\beta}_1 x_i\right)^2 = \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 - \hat{\beta}_1\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2$$

  - Il primo termine a destra è la somma totale corretta dei quadrati SS_{TC}
  - La dipendenza dalla x della variabile dipendente determina una riduzione di SS_{TC}. Se Y non dipende da x questo termine è trascurabile e $SS_E = SS_{TC}$
  - Il secondo termine a destra quindi dipende dalla regressione

$$SS_E = SS_{TC} - SS_R$$

  - D'altra parte si può scrivere:

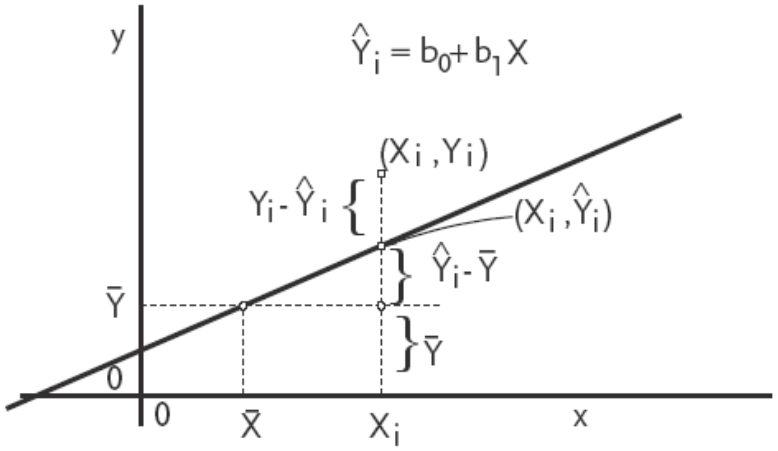
$$SS_T = \sum_{i=1}^{N} y_i^2 = n\bar{y}^2 + SS_E + SS_R = SS_M + SS_E + SS_R$$

Stima con i Minimi Quadrati

- La regressione ripartisce la somma dei quadrati, SS_T, in tre termini:
  1. La somma dei quadrati dovuti alla media SS_M;
  2. La somma degli scarti quadratici dovuti agli errori SS_E (deviazione dalla linea di regressione);
  3. La somma degli scarti quadratici dovuti alla regressione SS_R.

- In altro modo si può dire che ciascun y_i è composto da tre parti:

$$y_i = \bar{y} + (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$



Coefficiente di determinazione

- Si definisce coefficiente di determinazione il rapporto

$$R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SS_R}{S_{TC}} = 1 - \frac{SS_E}{S_{TC}} \qquad 0 \le R^2 \le 1$$

  - S_{TC} è una misura della variabilità in y senza considerare la variabilità della variabile indipendente x
  - SS_E è una misura della variabilità rimanente dopo che x è stata considerata
  - R^2 è la porzione di variazione descritta dalla regressione lineare.
  - Il secondo termine a destra paragona la varianza non spiegata dal modello con la varianza totale dei dati.

Coefficiente di determinazione

- R^2 e una statistica che da informazioni sulla bontà del fit di un modello.
- Tale coefficiente da una misura di quanto bene la linea di regressione approssima i dati sperimentali.
- Un valore unitario indicherebbe che la regressione passa perfettamente tra i dati. **(ATTENZIONE!)**
- Possiamo avere valori di R^2 al di fuori dell'intervallo 0,1. Questo può accadere quando la regressione non è lineare.
- Se facciamo crescere il numero di parametri R^2 può crescere.

Coefficiente di determinazione

- Cautele quando si interpreta R^2

- R^2 non fornisce informazioni su se:
  - Le variabili indipendenti considerate siano la vera causa della variazione della variabile dipendente
  - Esista una distorsione dovuta alle variabili indipendenti omesse
  - Il modello sia corretto
  - Siano state scelte le migliori variabili indipendenti.
  - Il modello possa essere migliorato trasformando le variabili indipendenti

Coefficiente di determinazione aggiustato

- È possibile normalizzare il coefficiente R^2 in funzione del numero p di parametri e del numero n di prove sperimentali:

$$SS_E = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

$$SS_{TC} = \sum_{i=1}^{N} \left(y_i - \bar{y} \right)^2$$

$$R_C^2 = 1 - \frac{SS_E}{df_E} \frac{df_{TC}}{SS_{TC}} = 1 + \left(R^2 - 1 \right) \frac{n-1}{n-p}$$

- Tale espressione risulta più significativa del coefficiente di determinazione classico in quanto tiene conto anche del numero di parametri presente nel modello.
  - Al crescere del numero dei parametri possiamo avvicinarci sempre meglio ai dati anche se ciò può non essere significativo dal punto di vista del modello

- Utile soprattutto nel caso di regressioni multiple.

Matrice di correlazione

- Nell'ipotesi di regressione multilineare:

$$\underline{y} = \underline{\underline{X}} \cdot \underline{\beta}$$

- La matrice di covarianza è:

$$\underline{\underline{V}}_{\underline{\hat{\beta}}} = \sigma^2 \left(\underline{\underline{X}}^T \cdot \underline{\underline{X}} \right)^{-1}$$

- È possibile valutare una sua versione "normalizzata", ovvero la matrice di correlazione:

$$\begin{cases} c_{ij} = \dfrac{\operatorname{cov}\left(B_i B_j\right)}{\sigma_{B1}\sigma_{B2}} & i \neq j \\ c_{ij} = 1 & i = j \end{cases}$$

Matrice di correlazione

- Se i termini fuori diagonale della matrice di correlazione c_{ij} sono molto prossimi al valore unitario:

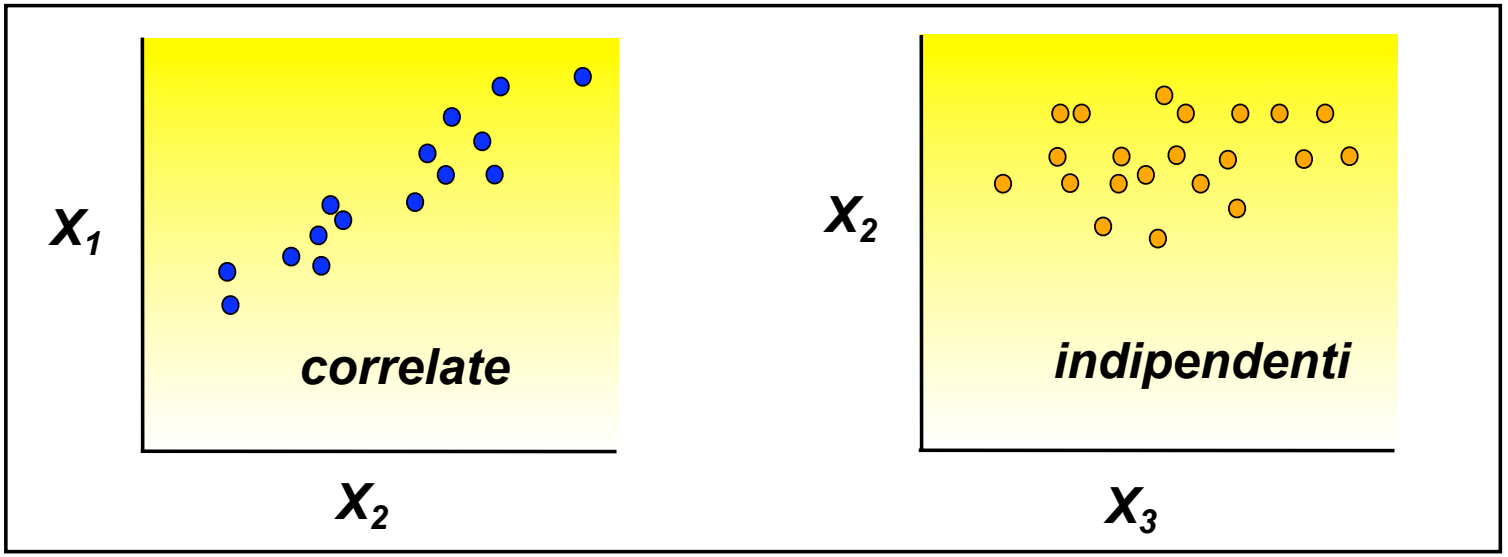
$$\left|c_{ij}\right| \approx 1 \qquad\qquad i \neq j$$

  - Le stime dei parametri non sono molto affidabili:
  - Il modello sperimentale è "troppo complicato" per descrivere la campagna sperimentale

- Questo può essere legato a:
  - Un modello difettoso che contempla la dipendenza da un numero eccessivo di variabili
  - La campagna sperimentale è stata progettata male

Multicollinearità

- Può accadere che nel caso di regressione multilineare variabili indipendenti siano correlate e pertanto non indipendenti.



- Tale eventualità può avere effetti disastrosi sulla stima dei parametri

Multicollinearità

- Richiamando relazioni precedentemente introdotte:

$$\underline{\underline{V}}_{\underline{\hat{\beta}}} = \sigma^2 \left(\underline{\underline{X}}^T \cdot \underline{\underline{X}} \right)^{-1}$$

- L'eventualità di una dipendenza lineare tra le variabili dipendenti, di fatto, rende la matrice delle prove sperimentali vicina alla condizione di singolarità.

- Gli elementi della matrice inversa di X^TX possono assumere valori enormi

- La varianza dello stimatore può essere drammaticamente amplificata rispetto alla varianza sperimentale

- Come individuarla:
  - Valori elevati di R^2 ma intervalli di fiducia molto ampi per i singoli coefficienti di regressione.
  - Correlazioni tra le variabili regressore

Quantificazione della multicollinearità

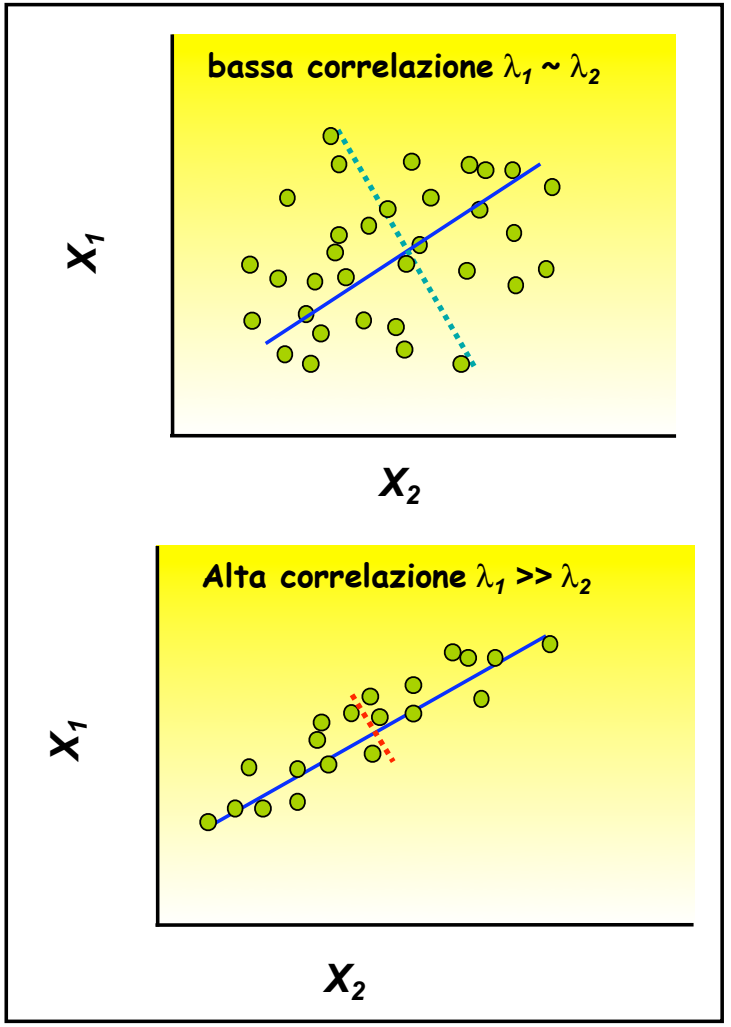
- Autovalori: se tutti i k autovalori della matrice delle condizioni sperimentali sono approssimativamente uguali, la multicollinearità è bassa.

- Condizionamento prossimo a 1 indica bassa multicollinearità

- Matrice di correlazione: Se vi sono dei termini fuori diagonale per cui

$$\left|c_{ij}\right| \approx 1$$

può esistere una forte correlazione tra le variabili regressore x_i e x_j.



Come migliorare le stime dei parametri

- Si è visto come una campagna sperimentale condotta in modo poco attento possa avere delle conseguenze disastrose sulla stima dei parametri
- Da un punto di vista intuitivo, la scelta delle condizioni sperimentali deve essere dettata dall'esigenza di rendere le prove sperimentali quanto più possibile linearmente indipendenti
- Da un punto di vista rigoroso, la matrice X^TX deve essere quanto più possibile lontana dalle condizioni di singolarità
- Una possibile politica può essere la ricerca delle condizioni sperimentali per cui il determinante sia massimo
- Tale filosofia è alla base delle cosiddetti Progettazioni di Campagne Sperimentali D-ottimali (D-Optimal Design, dove D sta per determinante)

Matrice di correlazione per modelli non lineari

- Nel caso di modelli non lineari non è possibile effettuare dei test rigorosi, dato che la maggior parte delle variabili non sono assimilabili a VA normali.
- Si deve far ricorso a delle approssimazioni
- È possibile solo esprimere giudizi qualitativi

Analisi dei residui

- Si definisce residuo alla prova i-esima: $e_i = y_i - \hat{y}_i$

- Il residuo rappresenta la componente dell'osservazione sperimentale che il modello non è in grado di descrivere
- Il residuo (idealmente) descrive la parte aleatoria dell'esperimento

$$e_i := N(0,\sigma)$$

- essendo σ la deviazione standard dell'errore sperimentale

$$y_i = \hat{y}_i + e_i$$

Osservazione sperimentale — **Parte deterministica** — **Parte aleatoria**

- Si può facilmente verificare che:

$$\frac{\sum e_i^2}{n-2} = \frac{SS_E}{n-2} = MS_E = s^2$$

Analisi dei residui

- È possibile anche introdurre il concetto di residuo standardizzato:

$$d_i = \frac{e_i}{\sqrt{MS_E}}$$

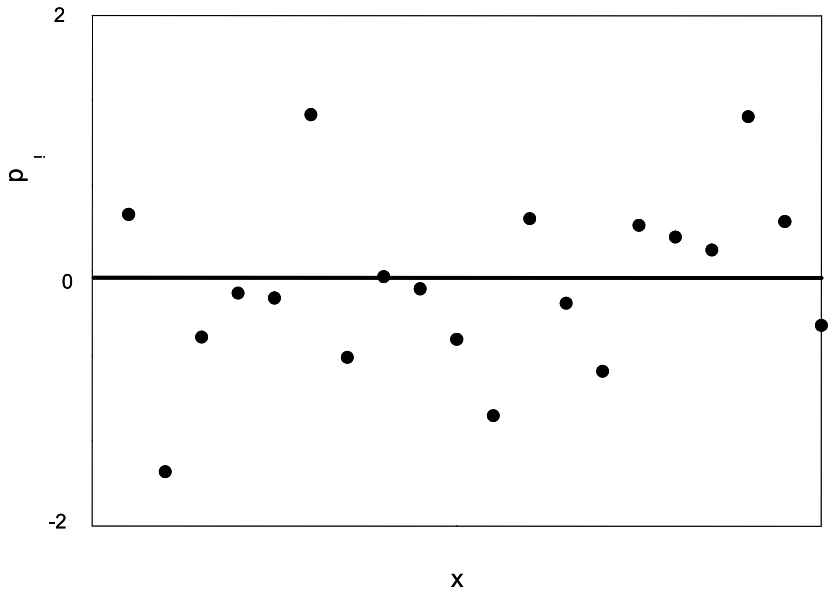
- I residui standardizzati hanno media 0 e varianza più o meno unitaria.

- L'analisi dei residui è un'analisi di tipo grafico
- È possibile rappresentare graficamente l'errore di
  - rispetto al valore previsto dal modello corrispondente $\hat{y}_i$
  - rispetto alla variabile regressore xi
  - non si rappresenta graficamente rispetto all'osservazione yi

- Se il modello descrive esattamente le osservazioni, i residui si dovrebbero "comportare" come genuini numeri casuali
- Può essere utile per la determinazione di inadeguatezze del modello.

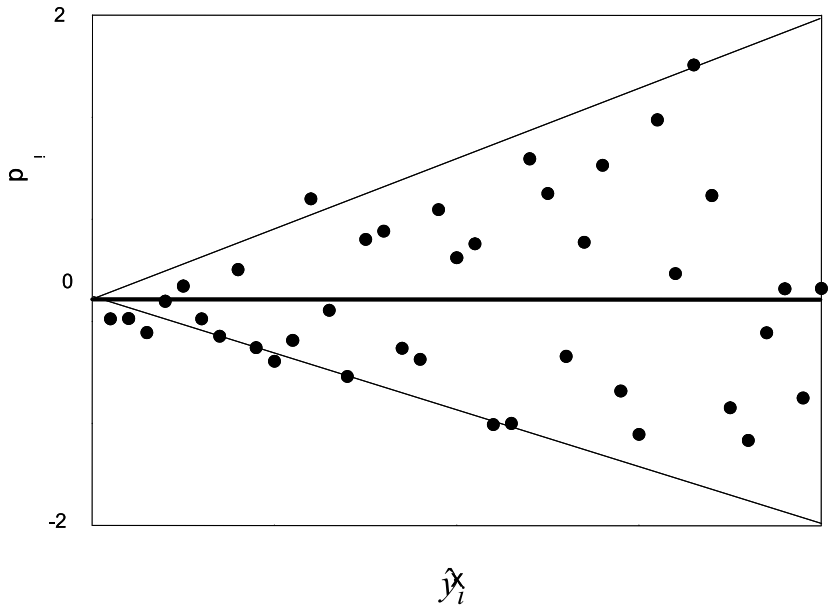
Analisi dei residui

- Se nel diagramma i residui sono contenuti in una banda orizzontale, senza la presenza di una struttura, allora non appaiono evidenti difetti nel modello



Analisi dei residui – Varianza non uniforme

- Situazioni in cui l'analisi dei residui rivela un comportamento anomalo.



- La varianza dei residui varia con la stima di y (eteroschedasticità): sarebbe adeguata una stima pesata dei parametri.

Analisi dei residui – Struttura nei residui

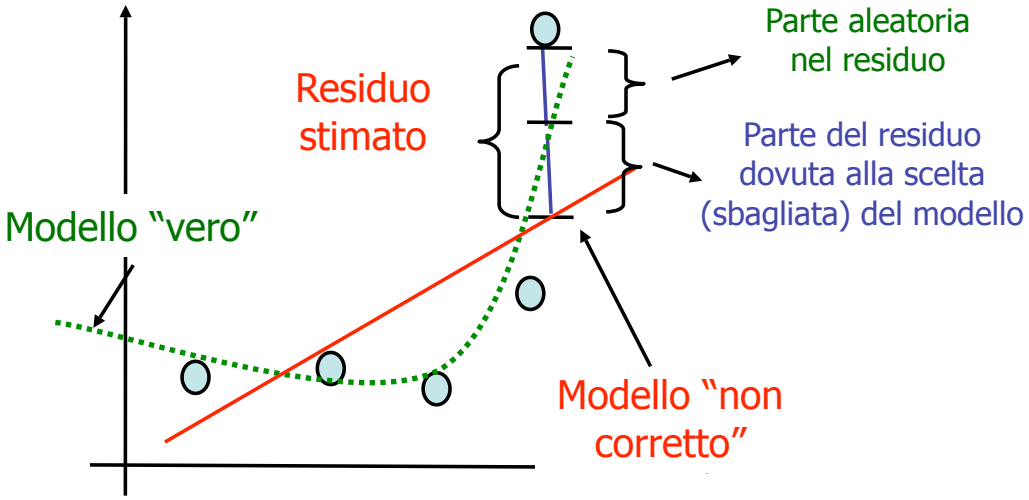
- Se il modello supposto per la descrizione dei dati non è "corretto", nel residuo cade anche una parte deterministica che il modello non riesce a descrivere

$$y_i = \hat{y}_i + e_i(\mathbf{x}) + \varepsilon_i$$

Osservazione sperimentale

Parte deterministica catturata dal modello

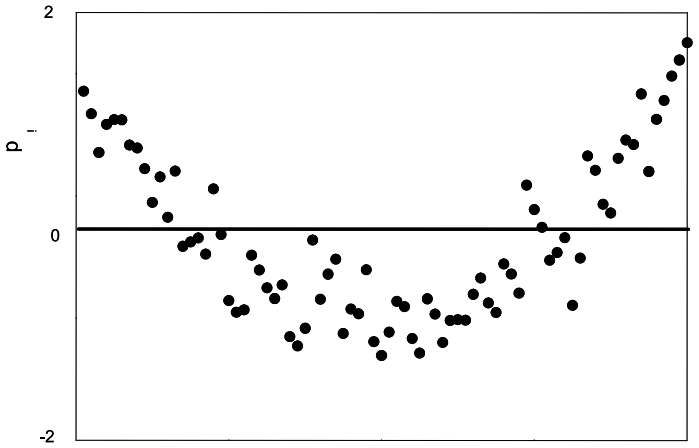
Residuo: non è più un numero casuale dato che in esso cade il determinismo che il modello "sbagliato" non può catturare



In questo caso il valore del residuo "dipende" dalla variabile indipendente

Analisi dei residui – Struttura nei residui

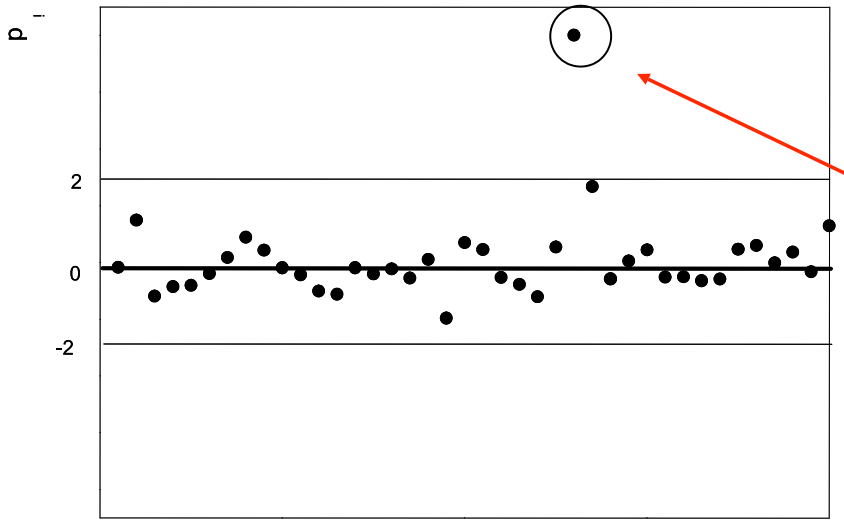
- Nel caso rappresentato in figura, si evince la presenza di una struttura nei residui al variare della predizione di y
- Questo scenario è in conflitto con l'assunzione iniziale (di di natura puramente casuale dell'osservazione)



- Nel residuo vi è una parte deterministica che non è stata catturata completamente dal modello
- È necessario estendere e/o modificare il modello

Analisi dei residui – Punti isolati

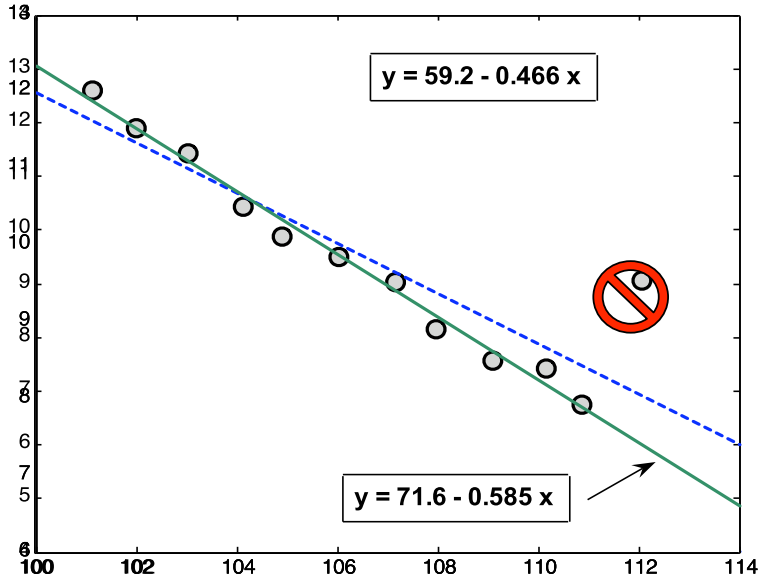
- L'analisi dei residui può aiutare a individuare punti sperimentali che siano frutto di una misura palesemente errata.



- I residui normalizzati devono comunque essere compresi più o meno nella banda [-2,2]

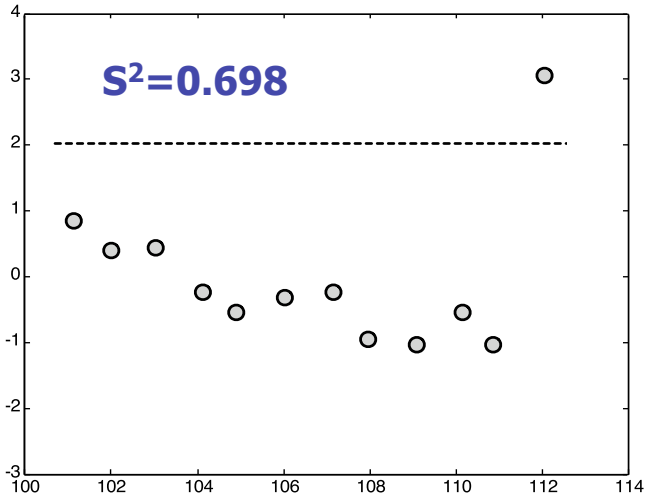
Analisi dei residui – Punti isolati

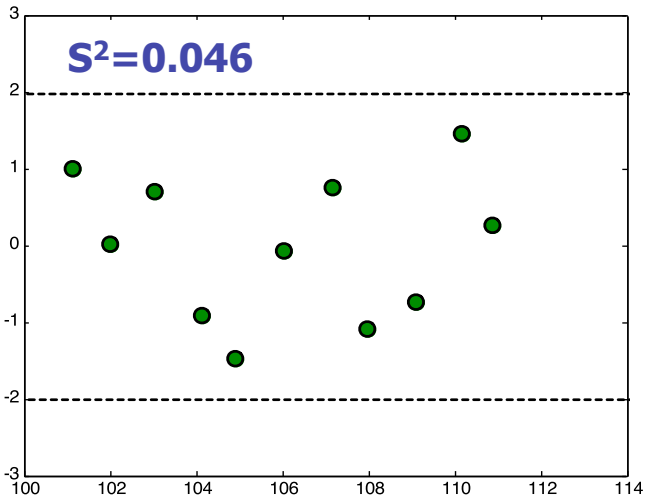
- Attenzione: la presenza di punti isolati può influenzare drammaticamente l'interpretazione dei residui



Analisi dei residui – Punti isolati

- I residui standardizzati nel primo caso presentano una struttura molto evidente: questo non è dovuto alla mancata efficienza del modello ma alla presenza del punto isolato che perturba significativamente la stima dei parametri della regressione
- Rimuovendo il punto isolato, la struttura dei residui migliora significativamente. Da osservare come la varianza sia diminuita di un ordine di grandezza





Sommario

- Generalizzazione della regressione lineare al caso multilineare
- È possibile implementare tecniche per confermare se il modello scelto per descrivere il processo in esame sia adeguato o meno
- Analisi di tipo quantitativo (determinazione scalari e/o matrici)
  - Coefficiente di determinazione
  - Matrice di correlazione
- Analisi di tipo qualitativo (via grafica)
  - Matrice di correlazione
- Non esiste un metodo che sia univocamente riconosciuto come il più efficiente
  - Conviene eseguire quante più analisi possibile e confrontarne i risultati