

PROGETTO STATISTICA APPLICATA

GRUPPO 10

A.A. 2021/22

PROGETTO STATISTICA APPLICATA



GRUPPO 10

A.A. 2021/22

All models are wrong, but some are useful
George E. P. Box

Indice

1	Introduzione	5
2	Analisi preliminare dei dati	6
2.1	Prestazioni del software di calcolo	7
2.2	Velocità del processore	8
2.3	Dimensione Hard-Disk	10
2.4	Numero di processi del software	12
2.5	Aging del software	14
2.6	Prestazioni della scheda audio	16
2.7	Prestazioni memoria RAM	18
2.8	Analisi di correlazione	20
2.9	Termini di interazione	21
3	Analisi di regressione	22
3.1	Regressione stepwise	22
3.2	Criteri per la scelta dei modelli	26
3.3	Diagnostica	31
4	Conclusioni	35

Capitolo 1

Introduzione

L’analisi di regressione è una tecnica impiegata per stimare la relazione funzionale tra una variabile dipendente e delle possibili variabili indipendenti, dette **predittori o regressori**. Lo scopo del presente elaborato è quello di condurre una analisi di regressione per analizzare la relazione tra le prestazioni di un software di calcolo e diversi indici prestazionali della macchina che esegue il software.

Analisi preliminare dei dati. Nella prima parte della relazione, viene condotta una analisi preliminare dei dati, che include la presentazione della distribuzione di ciascuna variabile, l’analisi di correlazione e un tentativo di regressione polinomiale per ciascun predittore. Inoltre, sono riportati i risultati relativi ad eventuali termini di interazione.

Analisi di regressione. Nella seconda parte della relazione, si presentano alcuni modelli di regressione multipla scelti seguendo un algoritmo di regressione stepwise (*forward selection*). Per ciascuno dei modelli si riportano i grafici diagnostici, impiegati per verificare le ipotesi della stima ai minimi quadrati. Infine, si presenta il modello che si ritiene più adatto ai dati.

Capitolo 2

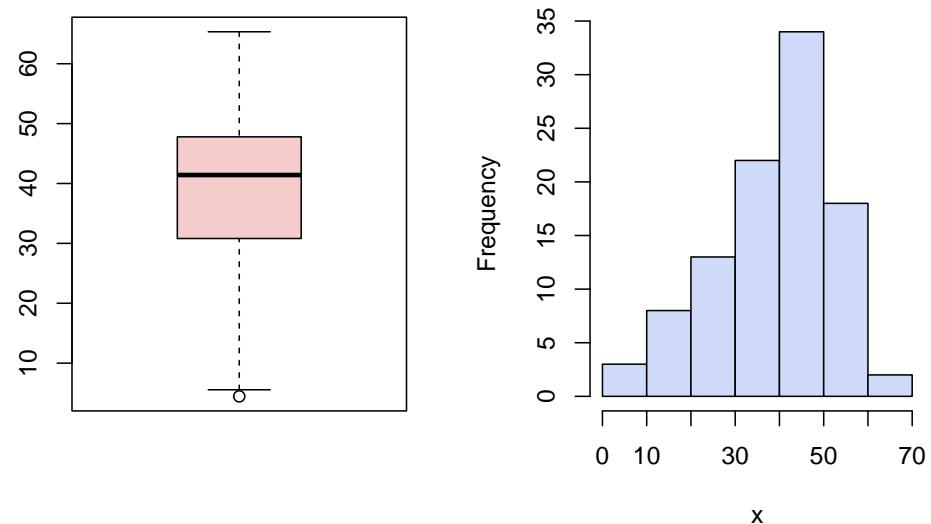
Analisi preliminare dei dati

Il dataset contiene la variabile risposta Y e sei possibili predittori X_1, \dots, X_6 .

Y	X_1	X_2	X_3	X_4	X_5	X_6
21.2095	0.1329	-1.1475	1.2257	1.6862	0.0961	-0.5140
19.8197	-1.1270	0.1785	-1.4045	1.0061	0.6818	0.0412
21.8759	-1.3342	-0.1501	-1.3210	-0.9047	-0.9003	-1.1881
65.3358	0.6699	1.1465	0.4800	-1.5387	1.5428	1.2442
46.2330	0.7900	0.9136	1.2070	0.2222	-0.9643	1.3710
54.0891	0.6467	0.5571	1.5168	-1.2268	0.5965	-0.8986

2.1 Prestazioni del software di calcolo

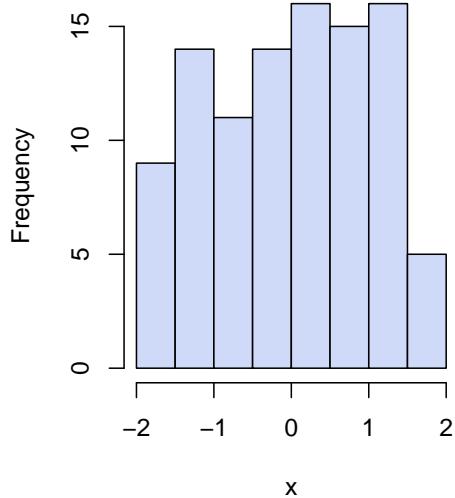
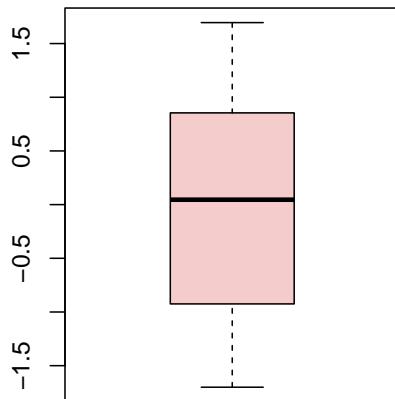
La variabile Y è un indice prestazionale relativo ad un software di calcolo.

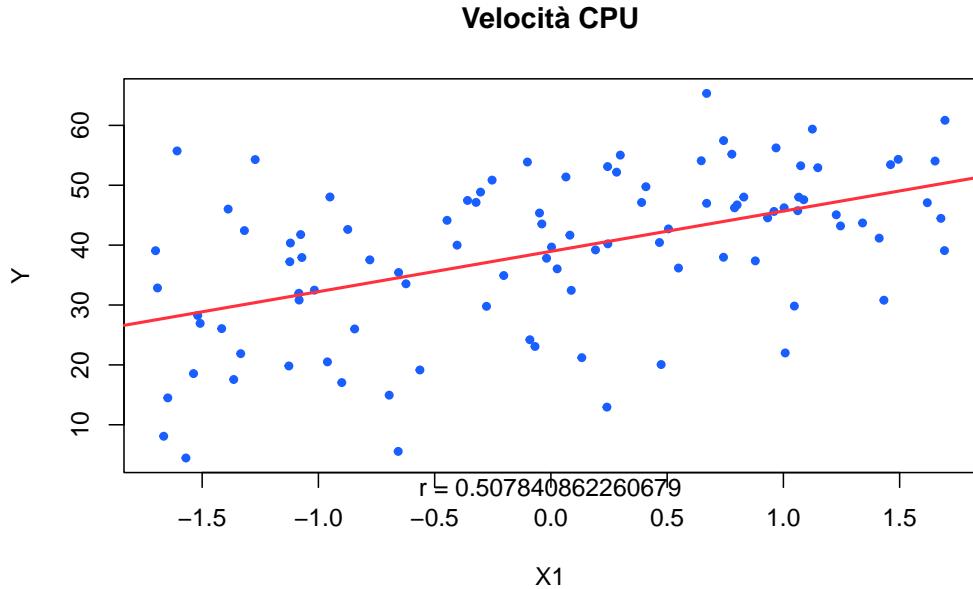


2.2 Velocità del processore

La variabile X_1 è un indice relativo alla velocità della CPU. Essendo un indice standardizzato e centrato, ha varianza unitaria e media prossima allo zero.

```
mean(dataset$x1)  
## [1] 1.335629e-17  
var(dataset$x1)  
## [1] 1
```





Esaminando lo scatter plot si può supporre una relazione lineare tra il regressore X_1 e la variabile dipendente Y . Questa ipotesi è in parte supportata dal fatto che l'indice di correlazione tra le due variabili, pari a 0.51, è significativamente maggiore di zero.

Per verificare se la variabile X_1 ha un effetto sul modello, si può impiegare il seguente test F:

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

La cui relativa regola decisionale è:

$$\begin{cases} F \leq F_{\alpha; \nu_1; \nu_2} & \text{si accetta } H_0 \\ F > F_{\alpha; \nu_1; \nu_2} & \text{si rifiuta } H_0 \end{cases}$$

Dove il termine $F_{\alpha; \nu_1; \nu_2}$ è il valore del quantile $\alpha = 0.05$ della distribuzione di Fisher-Snedecor con ν_1 e ν_2 gradi di libertà.

Nel caso in esame, i gradi di libertà della distribuzione sono

$$\nu_1 = df_{SQE_1} - df_{SQE_2} = (100 - 1) - 98 = 1$$

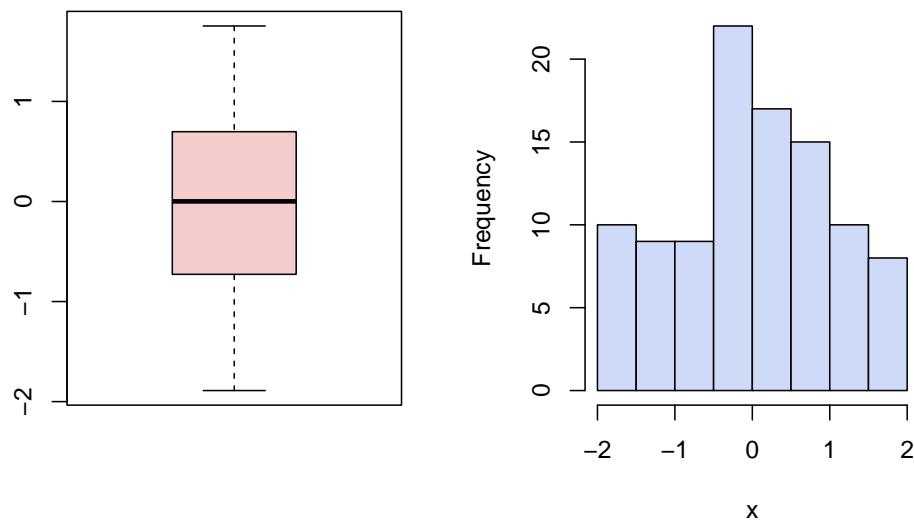
$$\nu_2 = n - p - 1 = 100 - 1 - 1 = 98$$

```
Df Sum Sq Mean Sq F value    Pr(>F)
x1          1   4481    4481   34.06 6.92e-08 ***
Residuals  98  12895     132
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1}
```

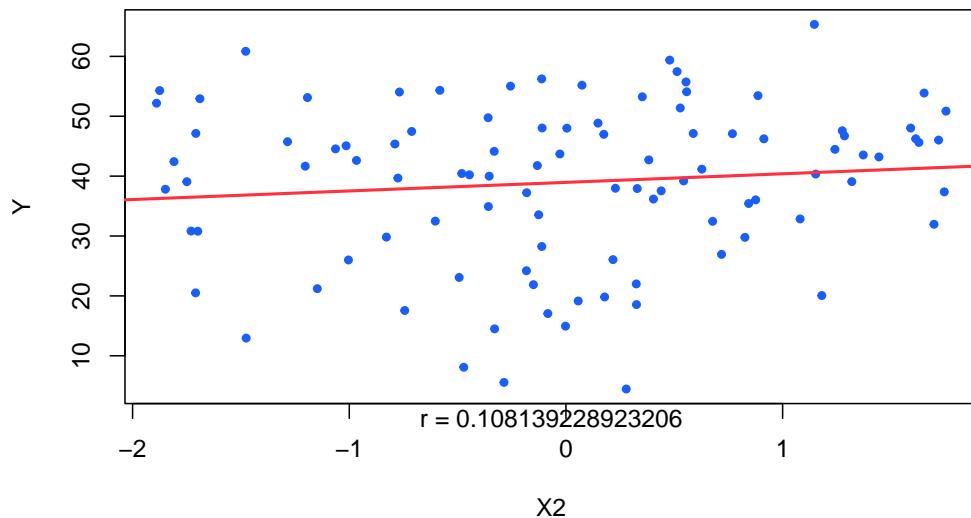
Il valore della statistica test, pari a 34.06, è maggiore del valore del quantile (0.0039523) e il *p-value* ($6.92 \cdot 10^{-8}$) ha un valore molto minore del rischio $\alpha = 0.05$. È pertanto possibile respingere l'ipotesi nulla e concludere che il predittore X_1 ha un effetto significativo sulla risposta Y .

2.3 Dimensione Hard-Disk

La variabile X_2 è un indice relativo alla dimensione dell'hard-disk.



Dimensione Hard-Disk



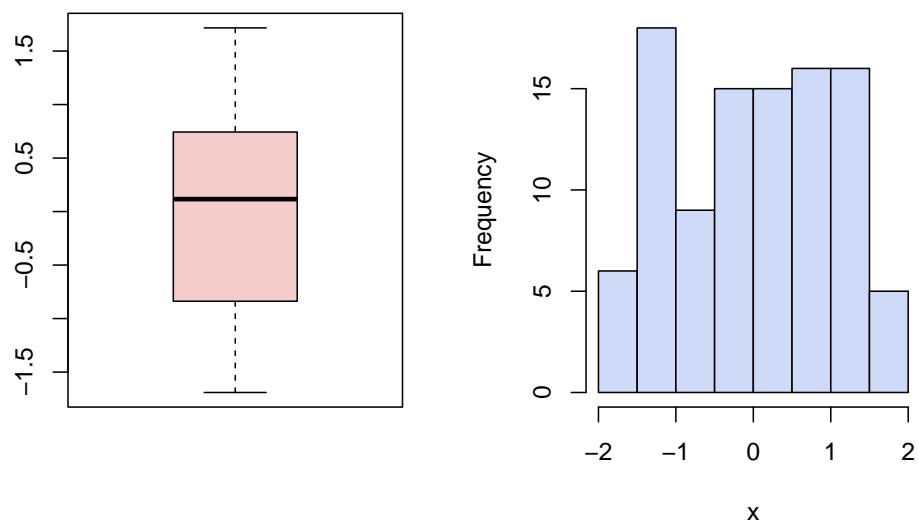
Dall'osservazione dello scatter plot non sembra essere presente una evidente relazione tra il predittore X_2 e la variabile dipendente. Inoltre, l'indice di correlazione è prossimo allo zero e il valore del *p-value* del test F non ci consente di rigettare l'ipotesi nulla, per cui non si può determinare se il predittore X_2 abbia un effetto sulla variabile dipendente.

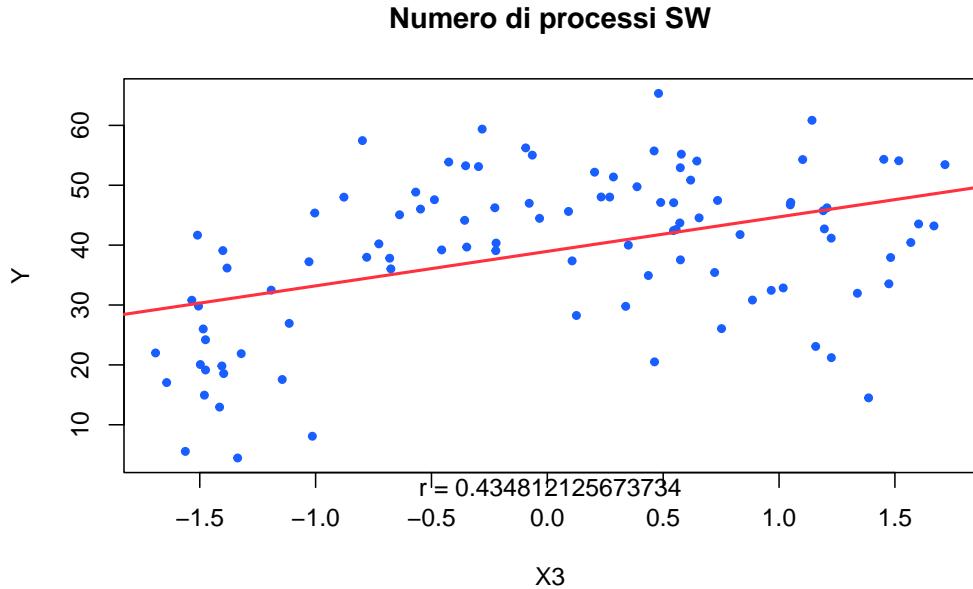
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x2	1	203	203.2	1.16	0.284
Residuals	98	17173	175.2		

Ulteriori considerazioni sono riportate nella sezione sull'analisi di regressione.

2.4 Numero di processi del software

La variabile X_3 è un indice relativo al numero di processi del software.





Per questo predittore, l'indice di correlazione, pari a 0.4348121, è tale da far supporre una correlazione lineare. Tuttavia, osservando l'andamento curvilineo dei punti dello scatter plot, possiamo ipotizzare una relazione polinomiale tra la variabile dipendente e il predittore.

Si riporta l'analisi del seguente modello di regressione polinomiale

$$Y = \beta_0 + \beta_1 X_3 + \beta_2 X_3^2 + \beta_3 X_3^3 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	45.788	1.565	29.255	< 2e-16 ***
x3	-1.123	2.734	-0.411	0.6821
I(x3^2)	-6.508	1.174	-5.544	2.59e-07 ***
I(x3^3)	3.441	1.413	2.435	0.0168 *
<hr/>				
Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.'
	0.1 '	0.1 '	0.1 '	1

Residual standard error: 10.33 on 96 degrees of freedom
 Multiple R-squared: 0.4108, Adjusted R-squared: 0.3924

F-statistic: 22.31 on 3 and 96 DF, p-value: 4.8e-11

I coefficienti sono stati stimati attraverso il metodo dei minimi quadrati. Il predittore di primo grado deve essere escluso dal modello in quanto l'intervallo di confidenza del parametro β_1 contiene lo zero.

lower	upper
-6.549602	4.303483

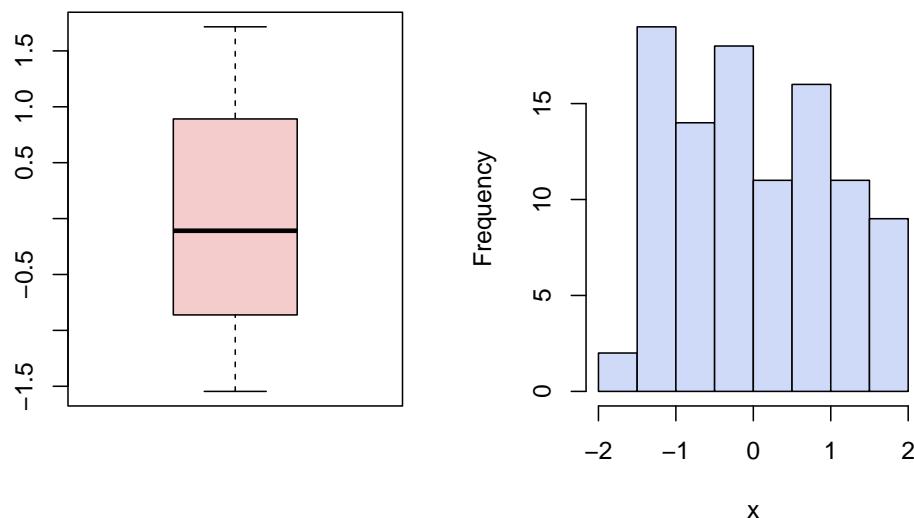
Lo stimatore impiegato per determinare l'intervallo di confidenza è

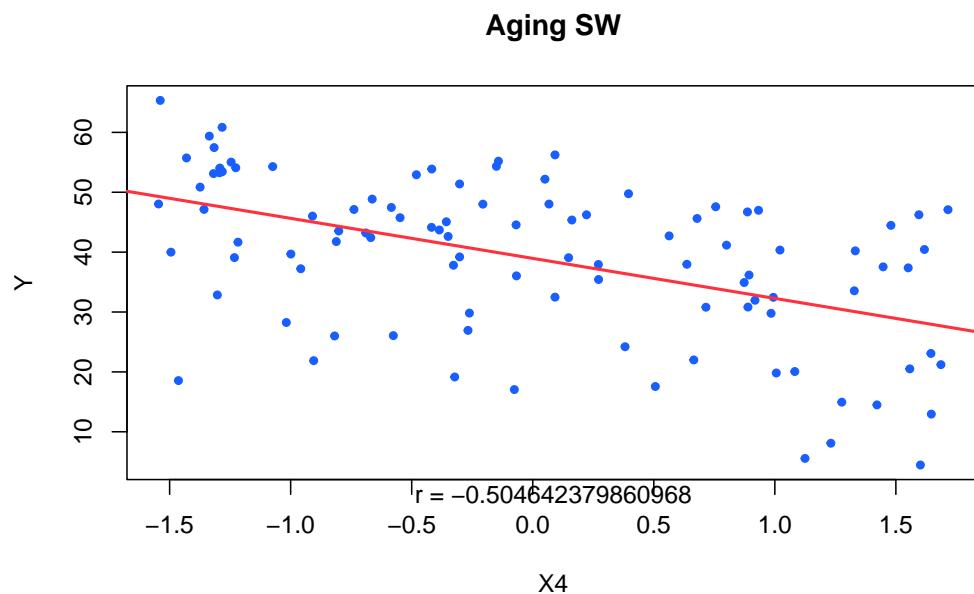
$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{var\hat{\beta}_j}} \sim T(n - p - 1)$$

Si conferma quindi la relazione polinomiale di terzo grado, in cui è assente il solo termine lineare.

2.5 Aging del software

La variabile X_4 è un indice relativo all'aging del software.





Il coefficiente di correlazione tra il predittore e la variabile dipendente, in questo caso, è negativo (-0.5046), ma comunque sufficientemente lontano da zero.

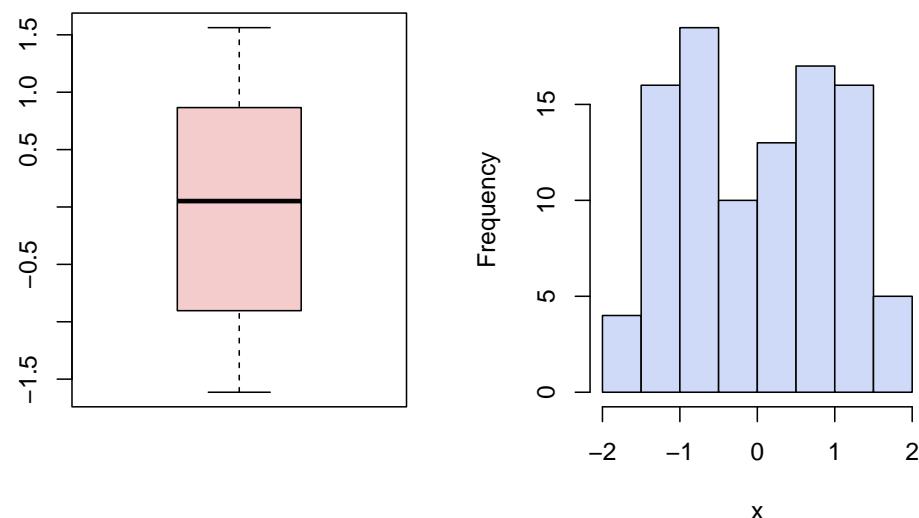
Il risultato del test F conferma l'ipotesi alternativa, secondo cui il predittore X4 ha un effetto sulla variabile di risposta.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x4	1	4425	4425	33.48	8.62e-08	***
Residuals	98	12951	132			

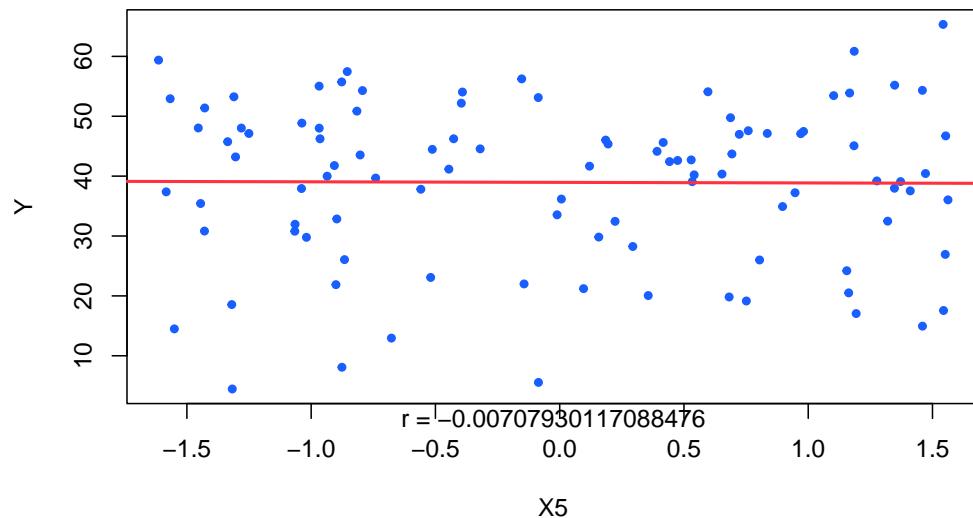
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'
					0.05	'. '
					0.1	' '
					1	

2.6 Prestazioni della scheda audio

La variabile X_5 è un indice prestazionale relativo alla scheda audio



Prestazioni scheda audio



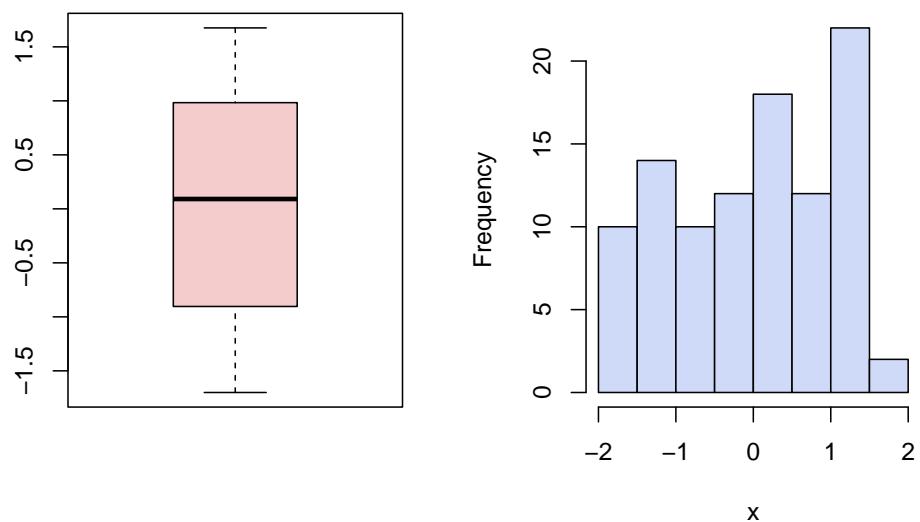
Il coefficiente di correlazione è quasi nullo, il che farebbe supporre l'assenza di una relazione tra le due variabili. Inoltre, il valore del *p-value* del test F non ci consente di rigettare l'ipotesi nulla, per cui non si può determinare se il predittore X_5 abbia un effetto sulla variabile dipendente.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x5	1	1	0.87	0.005	0.944
Residuals	98	17375	177.30		

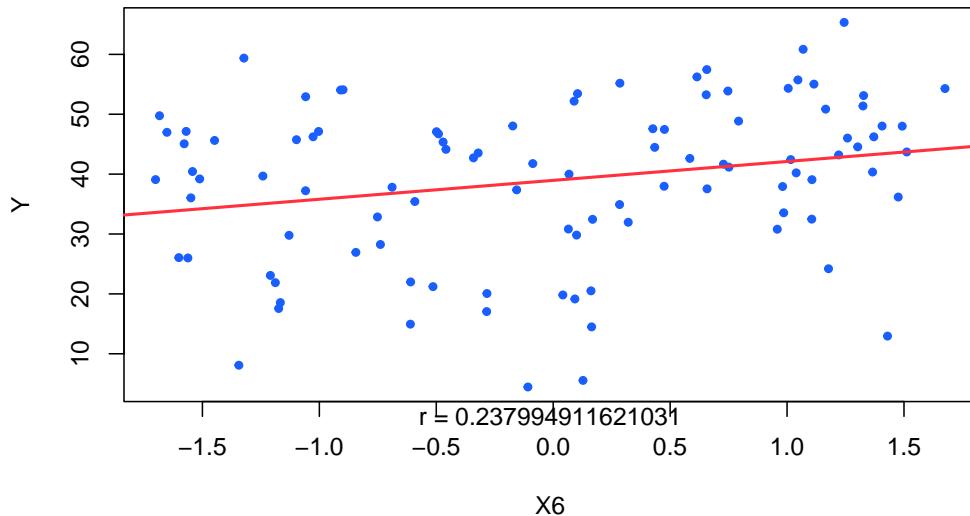
Ulteriori considerazioni sono riportate nella sezione sull'analisi di regressione.

2.7 Prestazioni memoria RAM

La variabile X_6 è un indice relativo alle prestazioni della memoria RAM



Prestazioni RAM



Il coefficiente di correlazione tra il predittore X_6 e la variabile di risposta è molto basso, il che porterebbe ad escludere una correlazione lineare tra le due variabili. Tuttavia, dall'analisi di regressione polinomiale, sembrerebbe essere presente una relazione quadratica.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	35.819	1.960	18.275	< 2e-16 ***
x6	3.526	1.290	2.734	0.00744 **
I(x6^2)	3.170	1.507	2.104	0.03797 *
<hr/>				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				

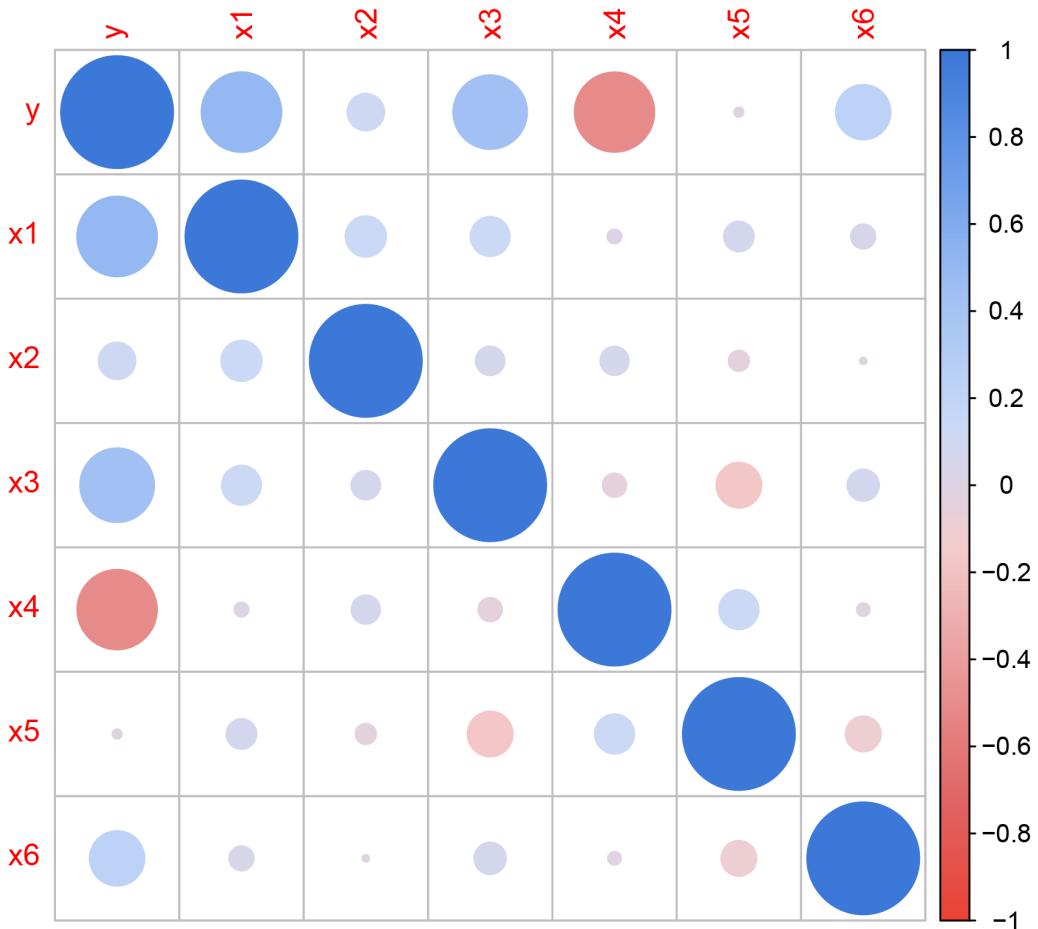
Residual standard error: 12.71 on 97 degrees of freedom

Multiple R-squared: 0.09781, Adjusted R-squared: 0.07921

F-statistic: 5.258 on 2 and 97 DF, p-value: 0.00679

2.8 Analisi di correlazione

Si riporta di seguito un grafico riassuntivo della matrice di correlazione.



Poichè l'indice di correlazione per tutte le coppie di variabili indipendenti è molto basso, si è portati ad escludere la presenza di una **collinearità** tra i predittori, ossia di una situazione in cui un predittore può essere ottenuto come combinazione lineare di un altro regressore, rendendo di fatto ridondante l'inserimento di una delle due variabili nel modello.

2.9 Termini di interazione

La ricerca dei termini di interazione da inserire nel modello di regressione multipla è stata eseguita con un approccio *greedy*, ossia tentando diverse combinazioni tra i termini di interazione. Dall'analisi di regressione effettuata su tutti i possibili termini di interazione, emerge che solo i termini di interazione che coinvolgono due o tre variabili sono significativi. Restringendo l'analisi solo a questi termini prodotto, i termini composti dal prodotto a tre variabili hanno perso significatività a confronto con i termini che coinvolgono due o tre variabili. Infine, dalla regressione con termini prodotto a due variabili, gli unici termini che forniscono un effetto significativo alla variabile risposta sono $X1 \cdot X6$, $X3 \cdot X5$, $X3 \cdot X4$, $X4 \cdot X6$, di cui si riportano gli intervalli di confidenza dei parametri stimati.

	lower	upper
x1:x6	-3.9660920	-0.88632240
x2:x4	0.2689505	3.21462122
x3:x5	0.3075373	3.53616113
x4:x6	-3.0115855	-0.06732781

Inoltre, sulla base dei test F, concludiamo che dei quattro possibili predittori solo $X1 \cdot X6$ e $X3 \cdot X5$ potrebbero avere un effetto significativo sulla variabile.

```
[1] Test F predittore X1:X6
      Df Sum Sq Mean Sq F value Pr(>F)
x1:x6       1    827   826.5   4.895 0.0293 *
Residuals  98  16549   168.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

[1] Test F predittore X3:X5
      Df Sum Sq Mean Sq F value Pr(>F)
x3:x5       1    719   718.6   4.228 0.0424 *
Residuals  98  16657   170.0
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Capitolo 3

Analisi di regressione

3.1 Regressione stepwise

Per determinare il modello di regressione multipla adatto ai dati forniti si impiega un algoritmo di regressione stepwise. In particolare, per scegliere quale combinazione di regressori vada inclusa nel modello si procederà usando il criterio del p-value, ossia:

1. Partendo dal *null model*, si aggiunge un predittore al modello;
2. Se il test T sul predittore ha un p-value maggiore del rischio di prima specie $\alpha = 0.05$, il predittore viene rimosso dal modello;
3. Si continuano ad aggiungere predittori, eventualmente rimuovendo variabili che nei successivi test T non hanno un p-value sufficientemente basso.

Il primo modello, di cui si riportano i coefficienti, è il seguente

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

(Intercept)	x1
38.957310	6.727985

Si aggiunge il predittore X_2 al modello

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	38.9573	1.1516	33.828	< 2e-16 ***
x1	6.6546	1.1675	5.700	1.29e-07 ***
x2	0.5588	1.1675	0.479	0.633

Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Residual standard error: 11.52 on 97 degrees of freedom

Multiple R-squared: 0.2597, Adjusted R-squared: 0.2444

F-statistic: 17.01 on 2 and 97 DF, p-value: 4.652e-07

Il valore del *p-value* è maggiore del rischio α , per cui si rimuove $X2$ dal modello. Il successivo modello da testare è il seguente, a cui sono stati aggiunti i termini polinomiali del predittore $X3$.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3^2 + \beta_3 X_3^3 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	45.5900	1.2671	35.979	< 2e-16 ***
x1	5.9218	0.8575	6.906	5.41e-10 ***
I(x3^2)	-6.4175	0.9581	-6.698	1.43e-09 ***
I(x3^3)	2.4787	0.4468	5.547	2.56e-07 ***

Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Residual standard error: 8.448 on 96 degrees of freedom

Multiple R-squared: 0.6057, Adjusted R-squared: 0.5933

F-statistic: 49.15 on 3 and 96 DF, p-value: < 2.2e-16

Tutti i termini hanno un *p-value* inferiore al rischio di prima specie. Si aggiunge il predittore $X4$ al modello, ottenendo così il modello

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3^2 + \beta_3 X_3^3 + \beta_4 X_4 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	44.8061	0.9145	48.994	< 2e-16 ***
x1	6.0777	0.6166	9.857	3.34e-16 ***

```

I(x3^2)      -5.6451     0.6934   -8.141 1.52e-12 ***
I(x3^3)      2.3070     0.3217    7.171 1.61e-10 ***
x4          -5.8706     0.6161   -9.529 1.67e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 6.073 on 95 degrees of freedom
 Multiple R-squared: 0.7984, Adjusted R-squared: 0.7899
 F-statistic: 94.05 on 4 and 95 DF, p-value: < 2.2e-16

	lower	upper
(Intercept)	42.990531	46.621613
x1	4.853634	7.301857
I(x3^2)	-7.021774	-4.268493
I(x3^3)	1.668365	2.945656
x4	-7.093652	-4.647614

Anche per questo modello, il *p-value* di tutti i predittori è minore della probabilità di rischio α . Si aggiunge quindi il predittore X_5 al modello

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3^2 + \beta_3 X_3^3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	44.9272	0.8994	49.950	< 2e-16 ***
x1	5.9617	0.6077	9.811	4.64e-16 ***
I(x3^2)	-5.7582	0.6827	-8.434	3.88e-13 ***
I(x3^3)	2.3888	0.3181	7.510	3.35e-11 ***
x4	-6.0087	0.6081	-9.881	3.29e-16 ***
x5	1.3125	0.6120	2.144	0.0346 *

Signif. codes:	0 '***'	0.001 '**'	0.01 '*'	0.05 '.' 0.1 ' ' 1

Residual standard error: 5.961 on 94 degrees of freedom
 Multiple R-squared: 0.8078, Adjusted R-squared: 0.7976
 F-statistic: 79.01 on 5 and 94 DF, p-value: < 2.2e-16

	lower	upper
(Intercept)	43.14135163	46.713100
x1	4.75520144	7.168247

I(x3^2)	-7.11371463	-4.402676
I(x3^3)	1.75729719	3.020355
x4	-7.21618156	-4.801285
x5	0.09724295	2.527714

Sebbene in questo caso il *p-value* sia prossimo al livello di rischio, non è presente sufficiente evidenza per escludere il predittore dal modello. Si aggiunge il predittore X_6 , il suo quadrato e si esamina il modello

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3^2 + \beta_3 X_3^3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_6^2 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	44.6155	1.0838	41.167	< 2e-16 ***
x1	5.8154	0.5555	10.469	< 2e-16 ***
I(x3^2)	-5.7337	0.6226	-9.209	1.05e-14 ***
I(x3^3)	2.3232	0.2891	8.036	3.04e-12 ***
x4	-5.9689	0.5610	-10.639	< 2e-16 ***
x5	1.5407	0.5597	2.753	0.00712 **
x6	2.5963	0.5536	4.690	9.47e-06 ***
I(x6^2)	0.2829	0.6608	0.428	0.66960
<hr/>				
Signif. codes:	0	'***'	0.001	'**'
	0.01	'*'	0.05	'. '
	0.1	' '	1	

Residual standard error: 5.411 on 92 degrees of freedom

Multiple R-squared: 0.845, Adjusted R-squared: 0.8332

F-statistic: 71.63 on 7 and 92 DF, p-value: < 2.2e-16

Sebbene dall'analisi di regressione polinomiale il termine X_6^2 mostrava un effetto significativo sulla variabile di risposta, il *p-value* del test T per il modello appena esaminato è maggiore del rischio di prima specie, per cui il regressore viene rimosso dal modello. Nello *step* successivo, si aggiunge al modello il predittore $X_1 : X_6$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3^2 + \beta_3 X_3^3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_1 \cdot X_6 + \varepsilon$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
--	----------	------------	---------	----------

```

(Intercept) 44.6421    0.7490  59.605 < 2e-16 ***
x1          5.8433     0.5046  11.579 < 2e-16 ***
I(x3^2)    -5.3854     0.5730 -9.399 4.20e-15 ***
I(x3^3)    2.2342     0.2650  8.432 4.52e-13 ***
x4          -6.1328    0.5053 -12.138 < 2e-16 ***
x5          1.3506     0.5123  2.636  0.00984 **
x6          2.7491     0.5026  5.469 3.86e-07 ***
x1:x6      -2.1690    0.5056 -4.290 4.41e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.945 on 92 degrees of freedom
Multiple R-squared: 0.8705, Adjusted R-squared: 0.8607
F-statistic: 88.38 on 7 and 92 DF, p-value: < 2.2e-16

```

Per brevità, non sono state riportate la successiva aggiunta e rimozione del predittore $X_3 \cdot X_5$. L'ultimo modello esaminato è l'unico modello dotato di un indice di determinazione alto e predittori con *p-value* sufficientemente bassi. Prima di prendere una decisione, tuttavia, bisogna verificarne la bontà di adattamento.

3.2 Criteri per la scelta dei modelli

Esistono diversi criteri per decidere quale modello di regressione fornisca il miglior *fit* dei dati, ciascuno dei quali caratterizzato da un indice.

L'indice di determinazione, indicato con R^2 , è un indice che misura quanto il modello scelto riesce a spiegare la variabilità totale dei dati. È così definito:

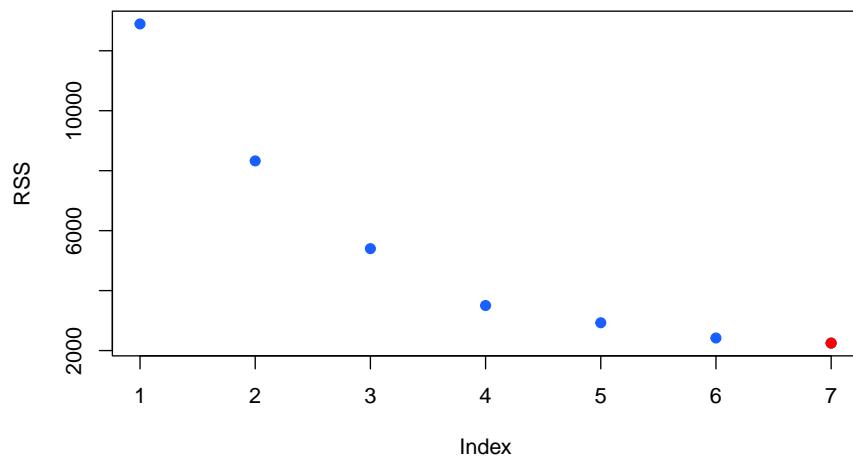
$$R^2 = 1 - \frac{SQE}{SQTOT}$$

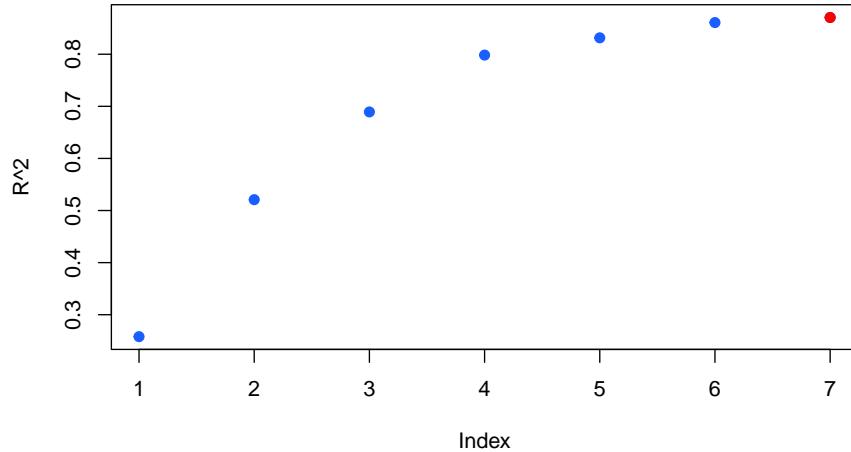
Dove il termine *SQE*, indicato in inglese con *RSS* (Residual Sum of Squares), è la somma delle differenze tra i valori della variabile Y osservati e i valori calcolati a partire dal modello di regressione.

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

Maggiore è l'indice di determinazione R^2 , migliore è la bontà di adattamento del modello. Questa osservazione è equivalente ad affermare che minore è l' RSS , migliore è la bontà di adattamento del modello.

Di seguito si riporta la serie dei valori di RSS ed R^2 per i sottoinsiemi del modello ottenuto tramite regressione stepwise, in modo da evidenziare quali predittori forniscano un contributo significativo alla bontà di adattamento del modello.





I modelli indicati dai grafici non corrispondono ai modelli descritti prima, ma riguardano modelli ottenuti tramite un algoritmo di selezione del miglior sottoinsieme di predittori, di cui si riporta la legenda

Index	x1	x4	x5	x6	I(x3^2)	I(x3^3)	x1:x6
1	(1)	"*	"	"	"	" "	" "
2	(1)	"*	"*	"	"	" "	" "
3	(1)	"*	"*	"	"	" "	" "
4	(1)	"*	"*	"	"	"*	" "
5	(1)	"*	"*	"	"*	"*	" "
6	(1)	"*	"*	"	"*	"*	"*
7	(1)	"*	"*	"*	"*	"*	"*

Come atteso, il modello che contiene tutti i predittori è il modello che massimizza l'indice R^2 . Il coefficiente di determinazione aumenta, anche se leggermente, con l'introduzione di nuovi predittori. Per questo motivo sono stati introdotti altri indici che assegnano una penalità che aumenta con il numero di coefficienti, come l'AIC, il BIC e il Cp. L'indice AIC (*Akaike Information Criterion*) è un indice generalmente definito come

$$AIC = -2 \ln L + 2k$$

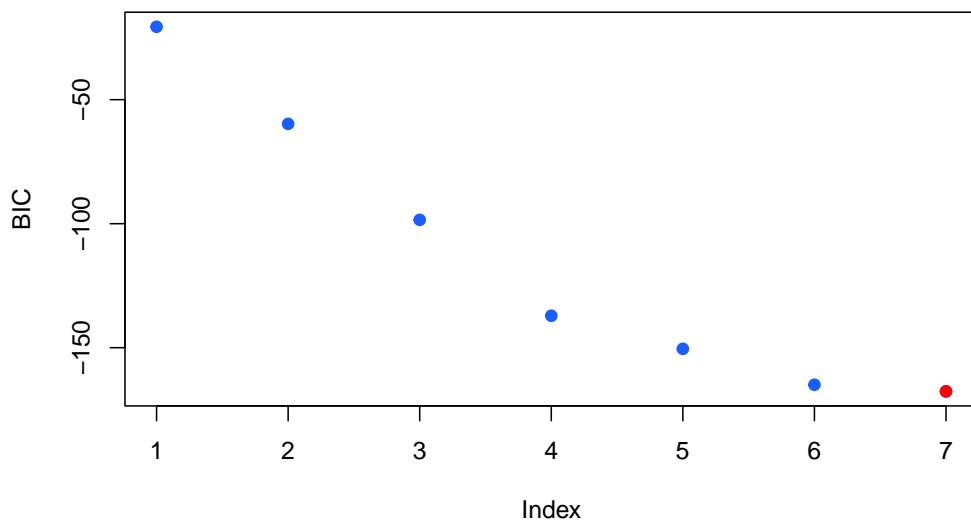
Dove k è il numero di parametri del modello e L è il massimo della funzione di verosimiglianza del modello stimato. Minore è il valore dell'AIC, migliore è la bontà di adattamento.

L'indice BIC (*Bayesian Information Criterion*), invece, è un indice definito come

$$BIC = -2 \ln L + k \ln n$$

Dove k è il numero di parametri del modello, n il numero di osservazioni e L il massimo della funzione di verosimiglianza. L'indice BIC introduce una penalizzazione maggiore rispetto all'indice AIC. Minore è il valore di questo indice, migliore è la bontà di adattamento.

Di seguito si riporta la serie dei valori BIC associati a ciascun sottoinsieme di predittori presi dal modello. Sorprendentemente, il modello con più predittori è il modello ad avere sia il miglior R^2 che il BIC minore.



Anche se entrambi i criteri selezionano il modello con tutti i predittori come il migliore, lo scarto tra gli indici R^2 e BIC del modello con tutti i predittori e del modello che non comprende il predittore X_5 è molto basso. Per l'indice R^2 lo scarto è di 0.0097793, mentre per l'indice BIC lo scarto è di -2.6772217.

Si ritiene opportuno confrontare i due modelli attraverso l'uso di grafici diagnostici.

3.3 Diagnostica

I grafici diagnostici sono dei grafici che elaborano i residui di un modello di regressione per:

- verificare le ipotesi della stima ai minimi quadrati
- fornire informazioni su eventuali situazioni patologiche

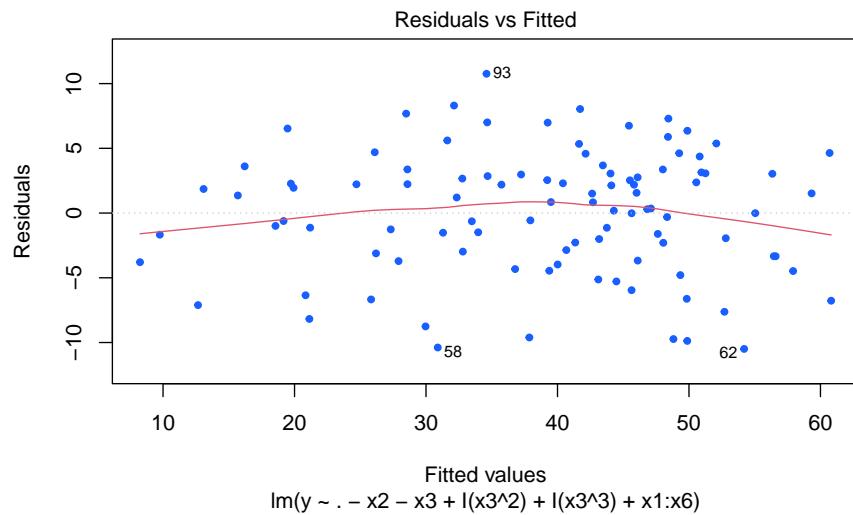
Per determinare le stime dei parametri del modello di regressione, si usa in genere il metodo dei minimi quadrati, che consiste nel trovare i parametri che minimizzano il quadrato dei residui. Dato il modello di regressione lineare

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p+1} X_{p+1} + \varepsilon$$

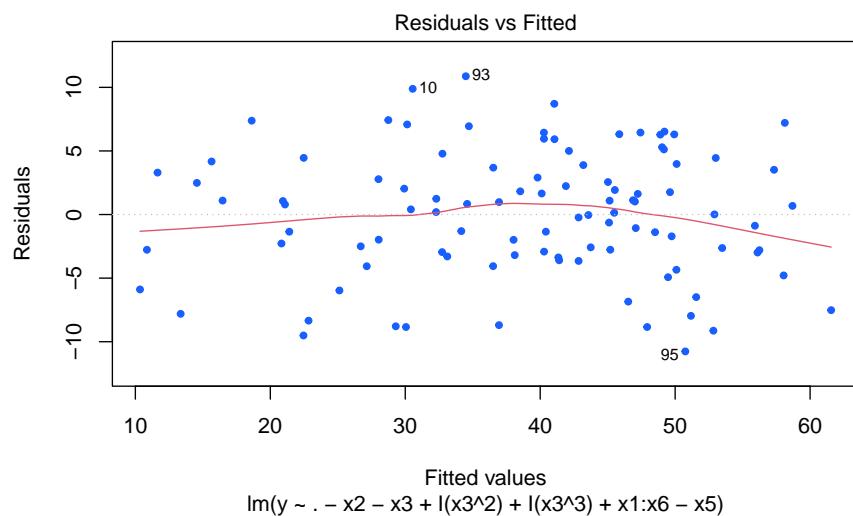
Per applicare il metodo dei minimi quadrati, si assume per ipotesi che il termine di errore ε sia distribuito come una normale a media nulla e varianza σ^2 .

Essendo la media campionaria dei residui uno stimatore della media di ε , il grafico dei residui, che mette a confronto i residui con i *fitted values* di un modello di regressione, viene impiegato per verificare che la media del termine ε è nulla. Inoltre, si potrebbe anche verificare che i residui siano affetti da *eteroschedasticità*, ossia siano composti da distribuzioni con diverse varianze. Nel grafico dei residui, si suppone *eteroschedasticità* se la distribuzione dei residui non è omogenea rispetto all'asse delle ascisse.

Si riporta il grafico dei residui del modello a sette predittori.



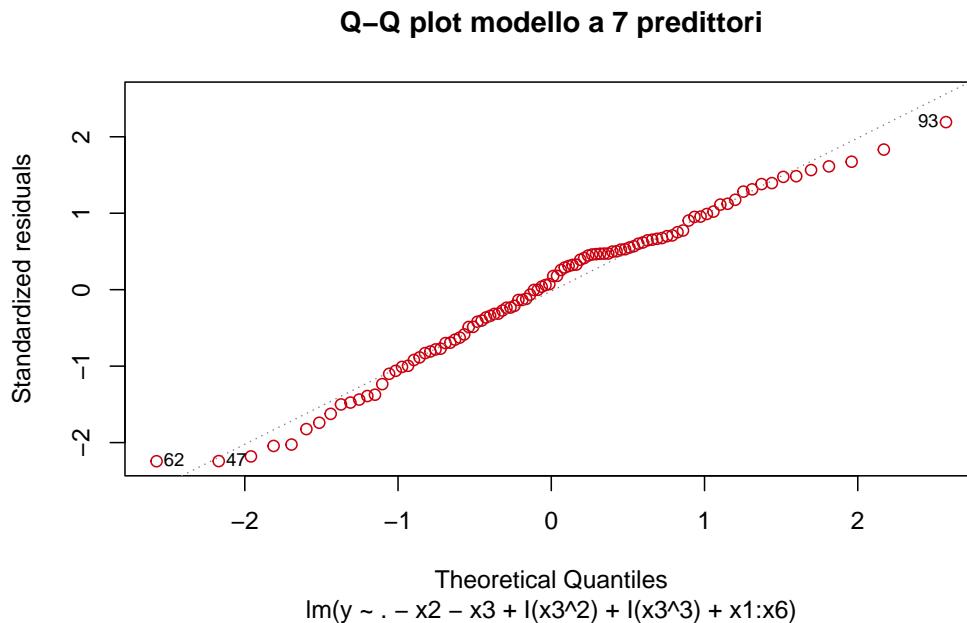
Si riporta il grafico dei residui del modello senza il predittore X_5



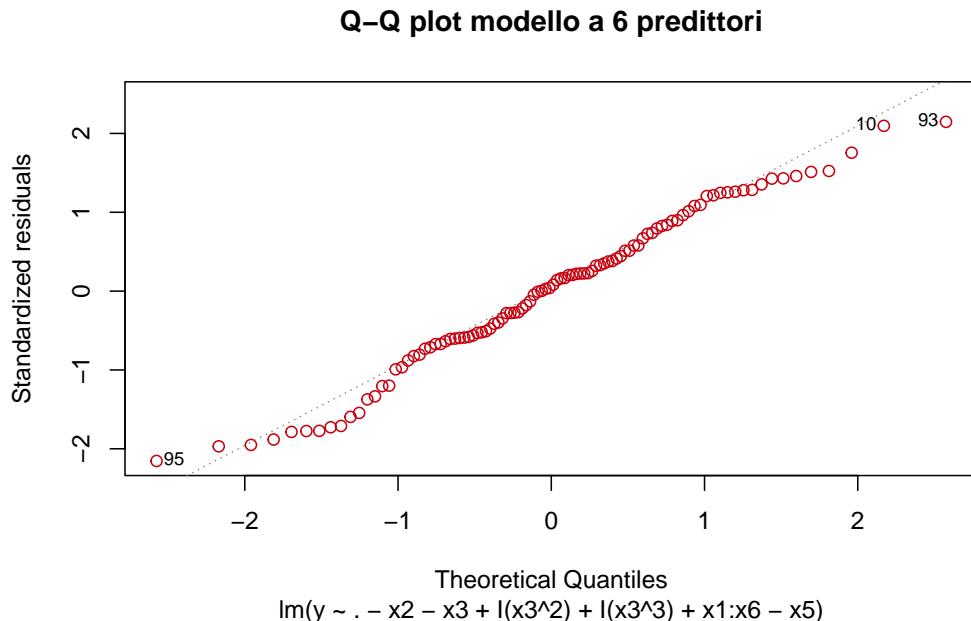
Per entrambi i grafici si può osservare che i residui sono distribuiti in maniera omogenea rispetto all'asse delle ascisse. Non essendoci differenze evidenti tra i due grafici, non si è certi dell'effetto del predittore X_5 .

Il grafico Quantile-Quantile, detto *Q-Q plot* è un grafico diagnostico impiegato per confrontare i quantili di due distribuzioni. Il Q-Q plot può essere impiegato per determinare se l'ipotesi che il termine di errore seguia una distribuzione normale è verificata. Si confrontano quindi i quantili di una distribuzione normale standard con i residui, naturalmente standardizzati. Se i punti del grafico sono disposti come la bisettrice del primo e terzo quadrante, allora si può ritenere valida l'ipotesi che il termine ε sia distribuito come una distribuzione normale.

Si riporta il Q-Q plot del modello con sette predittori.



Si riporta il Q-Q plot del modello senza il predittore X_5



Anche per questo grafico diagnostico, la differenza tra i due modelli è minima. Tra i due modelli, quello con il plot QQ più vicino ad una retta è il modello con sette predittori.

Dopo questa analisi, è confermato che l'introduzione del predittore X_5 migliora, seppur, di poco la bontà di adattamento del modello.

Capitolo 4

Conclusioni

In conclusione, si ritiene che il modello più adatto ai dati forniti sia

$$Y = 44.64 + 5.84X_1 - 5.38X_3^2 + 2.23X_3^3 - 6.13X_4 + \\ + 1.35X_5 + 2.74X_6 - 2.16X_1 \cdot X_6 + \varepsilon$$

Nonostante si sia potuto riscontrare che l'introduzione del predittore X_5 non comporta un vantaggio netto, si preferisce includere quest'ultimo nel modello in quanto non si ha sufficiente evidenza per affermare altrimenti.