Statistica Applicata anno 2021-2022

Analisi di un dataset in R

Gruppo 11

Fasi della realizzazione del progetto

La seguente relazione sarà suddivisa in 4 sezioni:

- Recupero ed eventuale pulizia dei dati;
- Analisi descrittiva dei suddetti;
- Analisi della Regressione;
- Considerazioni finali;

Introduzione

L'Analisi dei Dataset si riferisce al processo di manipolazione dei dati grezzi per scoprire informazioni utili e trarre conclusioni. Noi con questa analisi di dati, partendo da un dataset basato su vari parametri di calcolatori e tramite una ulteriore analisi di regressione, andremo a calcolare la relazione stimata tra una variabile dipendente e più variabili da noi scelte tramite l'analisi dei dati.

Sezione 1: Recupero ed eventuale pulizia del dataset

Mediante l'ambiente di sviluppo R abbiamo importato il dataset fornitoci nel file .csv e ci siamo assicurati che i dati rispettassero 2 regole principali:

- Sono caricati in un data frame, con nomi delle colonne corretti e significativi;
- Ogni colonna è del tipo corretto(nel nostro caso i valori sono tutti di tipo **numeric**);

Ora i dati sono tecnicamente corretti, successivamente dobbiamo assicurarci che essi siano consistenti;

Per verificare ciò dobbiamo passare attraverso la definizione dei vincoli che definiscono una certa variabile, i quali potrebbero decretare la modifica o l'eliminazione di alcuni dei dati raccolti:

La prima operazione praticata è stata assicurarci che la tabella non presentasse entry vuote, nel nostro caso la tabella è pienamente popolata dunque non è necessario agire ulteriormente;

Dopodichè mediante il comando *str()* abbiamo visualizzato il nostro data frame per indagare alcune discrepanze ma tutto sembra in regola: le colonne hanno nomi significativi e differenti tra di loro, contengono unicamente valori numerici;

Con la funzione *View()* abbiamo visualizzato il data frame nella sua forma tabellare per comprenderne appieno la struttura, a questo punto possiamo passare alla fase successiva:

l'analisi descrittiva;

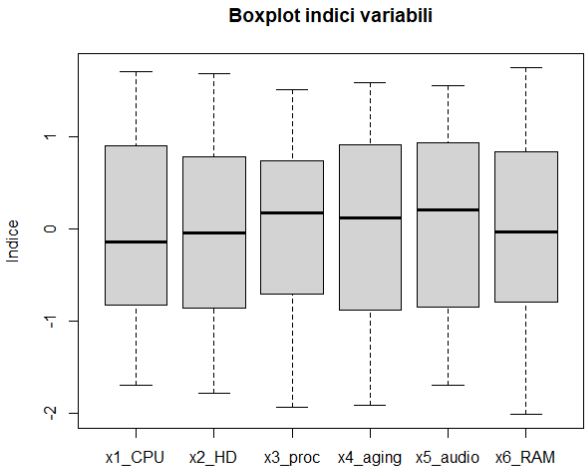
Sezione 2: Analisi di dati utilizzando gli strumenti della statistica descrittiva

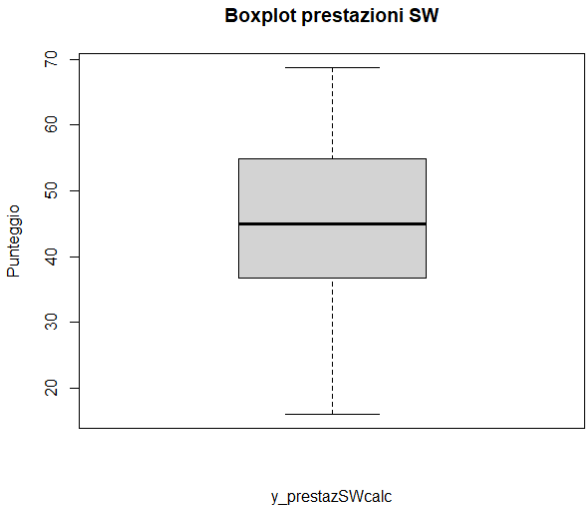
In questa fase vogliamo ottenere una prima descrizione visuale dei dati raccolti, controllare la presenza di outliers e/o anomalie, e di conseguenza rimuoverle e valutare le relazioni tra le variabili;

Mediante la funzione *summary()* andiamo a fare un resoconto del dataset mettendo in evidenza minimo, primo quantile, mediana, media, terzo quantile e massimo di ogni colonna.

x1_CPU • Min : -1.6951 • 1st Qu. : -0.8226 • Median : -.01421 • Mean : 0.0000 • 3st Qu. : 0.8783 • Max : 1.7101	**x2_HD** • Min : -1.78468 • 1st Qu. : -0.84898 • Median : -0.03711 • Mean : 0.00000 • 3st Qu. : 0.77948 • Max : 1.69267	**x3_proc** • Min : -1.9383 • 1st Qu. : -0.7027 • Median : 0.1775 • Mean : 0.0000 • 3st Qu. : 0.7357 • Max : 1.5165
x4_aging • Min : -1.9198 • 1st Qu. : -0.8829 • Median : 0.1199 • Mean : 0.0000 • 3st Qu. : 0.9029 • Max : 1.5931	**x5_audio** • Min : -1.6998 • 1st Qu. : -0.8477 • Median : 0.2097 • Mean : 0.0000 • 3st Qu. : 0.9352 • Max : 1.5610	**x6_RAM** • Min : -2.01070 • 1st Qu. : -0.77766 • Median : -0.03557 • Mean : 0.00000 • 3st Qu. : 0.82625 • Max : 1.75362
	y_presentazSwcalc • Min : 16.11 • 1st Qu. : 36.85 • Median : 44.97 • Mean : 44.96 • 3st Qu. : 54.94 • Max : 68.71	

Box Plot





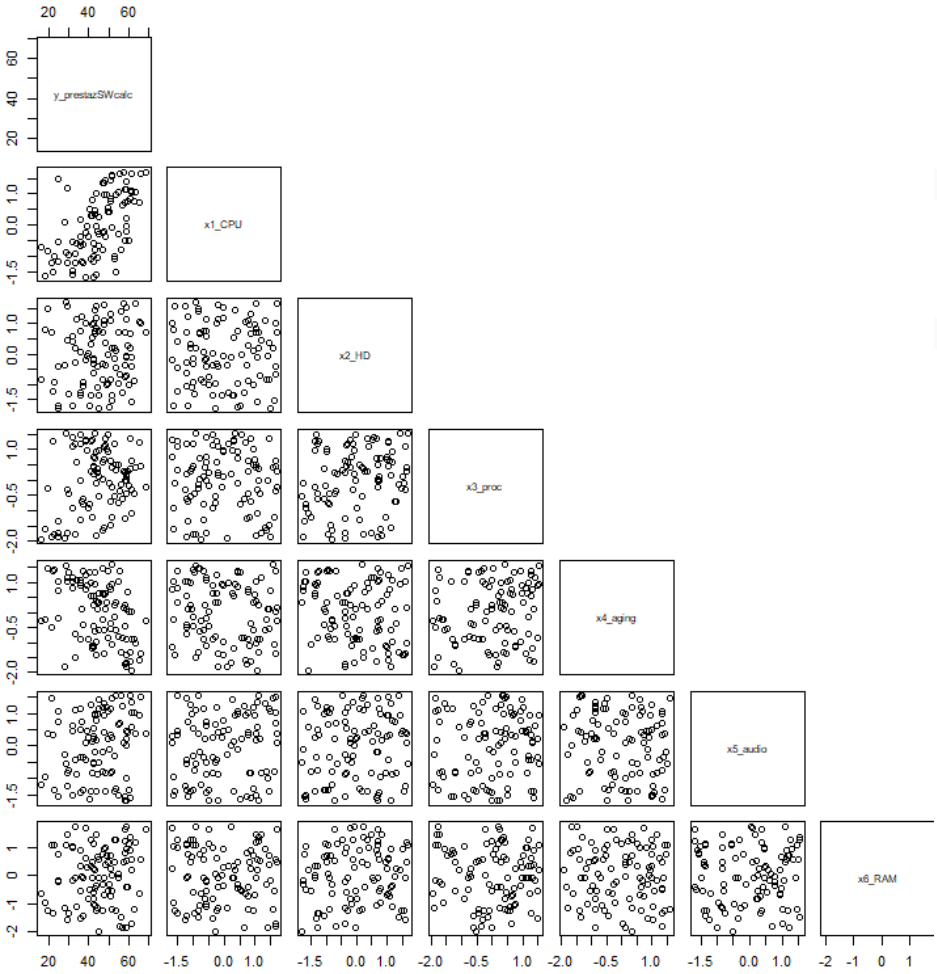
E' evidente che i box plot non presentino outliers dunque si può procedere nell'analisi.

Analisi grafica degli outliers

Mediante la funzione *boxplot()* cerchiamo gli outliers graficamente per ogni colonna.

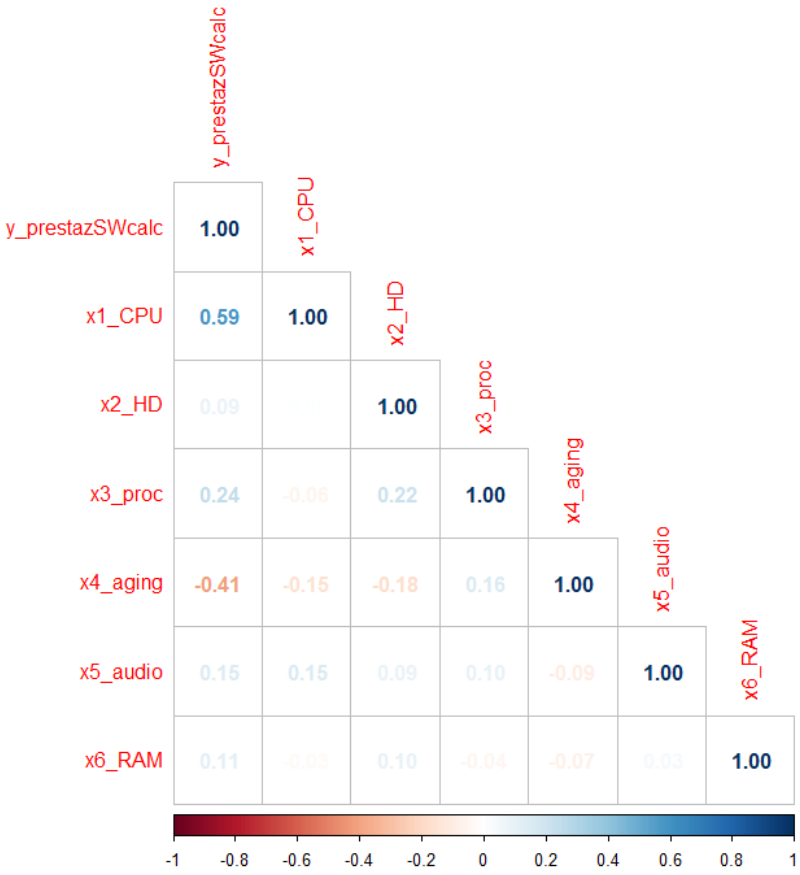
Si evidenzia, come visto in precedenza mediante il box plot che gli outliers sono nulli;

Analisi degli scatter plot



E’ a questo punto facile notare che vi sono delle correlazioni tra alcune variabili, in particolare possiamo notare correlazione lineare diretta sicuramente tra prestazioni e CPU e correlazione lineare inversa tra prestazioni ed aging, coerentemente a quanto ci si aspetterebbe da un sistema reale.

Analizziamo ora nel dettaglio la correlazione tra le variabili



Il grafico soprastante mette in evidenza quanto rilevato precedentemente mediante lo scatter plot riportando un coefficiente di correlazione forte di 0.59 tra prestazioni e cpu e di -0.41 tra prestazioni ed aging, è inoltre possibile notare che vi è correlazione, sebbene debole di 0.24, tra processo e prestazioni.

Tabelle di Frequenza

CPU

Classi	Centro di Classe	Frequenza	Frequenza Relativa	Frequenza Cumolativa	Frequenza Relativa Cumolativa
-1.6951≥ i >-1.2695	-1.4823	8	0.08	8	0.08
-1.2695≥ i >-0.8439	-1.0567	16	0.16	24	0.24
-0.8439≥ i >-0.4183	-0.6311	16	0.16	40	0.40
-0.4182≥ i > 0.0073	-0.2054	14	0.14	54	0.54
0.0073≥ i > 0.4329	0.2201	9	0.09	63	0.63
0.4329≥ i > 0.8585	0.6457	12	0.12	75	0.75
0.8585≥ i > 1.2841	1.0713	12	0.12	87	0.87
1.2841≥ i ≥1.7200	1.5020	13	0.13	100	1.00

HDD

Classi	Centro di Classe	Frequenza	Frequenza Relativa	Frequenza Cumolativa	Frequenza Relativa Cumolativa
-1.78468≥ i >-1.35008	-1.56738	11	0.11	11	0.11
-1.35008≥ i >-0.91558	-1.13283	12	0.12	23	0.23
-0.91558≥ i >-0.48088	-0.69823	9	0.09	32	0.32
-0.48088≥ i >-0.04628	-0.26358	18	0.18	50	0.50
-0.04628≥ i > 0.38832	0.17102	11	0.11	61	0.61
0.38832≥ i > 0.82292	0.60562	15	0.15	76	0.76
0.82292≥ i > 1.25752	1.04022	13	0.13	89	0.89
1.25752≥ i ≥ 1.6927	1.47511	11	0.11	100	1.00

PROCESSI

Classi	Centro di Classe	Frequenza	Frequenza Relativa	Frequenza Cumolativa	Frequenza Relativa Cumolativa
-1.9384≥ i >-1.5065	-1.72245	12	0.12	12	0.12
-1.5065≥ i >-1.0747	-1.2906	7	0.07	19	0.19
-1.0747≥ i >-0.6429	-0.8638	7	0.07	26	0.26
-0.6429≥ i >-0.2111	-0.4320	13	0.13	39	0.39
-0.2111≥ i > 0.2207	0.0048	13	0.13	52	0.52
0.2207≥ i > 0.6525	0.4366	19	0.19	71	0.71
0.6525≥ i > 1.0843	0.8684	11	0.11	82	0.82
1.0843≥ i ≥1.5170	1.3006	18	0.18	100	1.00

AGING

Classi	Centro di Classe	Frequenza	Frequenza Relativa	Frequenza Cumolativa	Frequenza Relativa Cumolativa
-1.9198≥ i >-1.4807	-1.7002	8	0.08	8	0.08
-1.4807≥ i >-1.0416	-1.2611	11	0.11	19	0.19
-1.0416≥ i >-0.6025	-0.8220	13	0.13	32	0.32
-0.6025≥ i >-0.1634	-0.3829	14	0.14	46	0.46
-0.1634≥ i > 0.2757	0.0561	10	0.10	56	0.56
0.2757≥ i > 0.7148	0.4952	13	0.13	69	0.69
0.7148≥ i > 1.1539	0.9343	16	0.16	85	0.85
1.1539≥ i ≥1.5940	1.3739	15	0.15	100	1.00

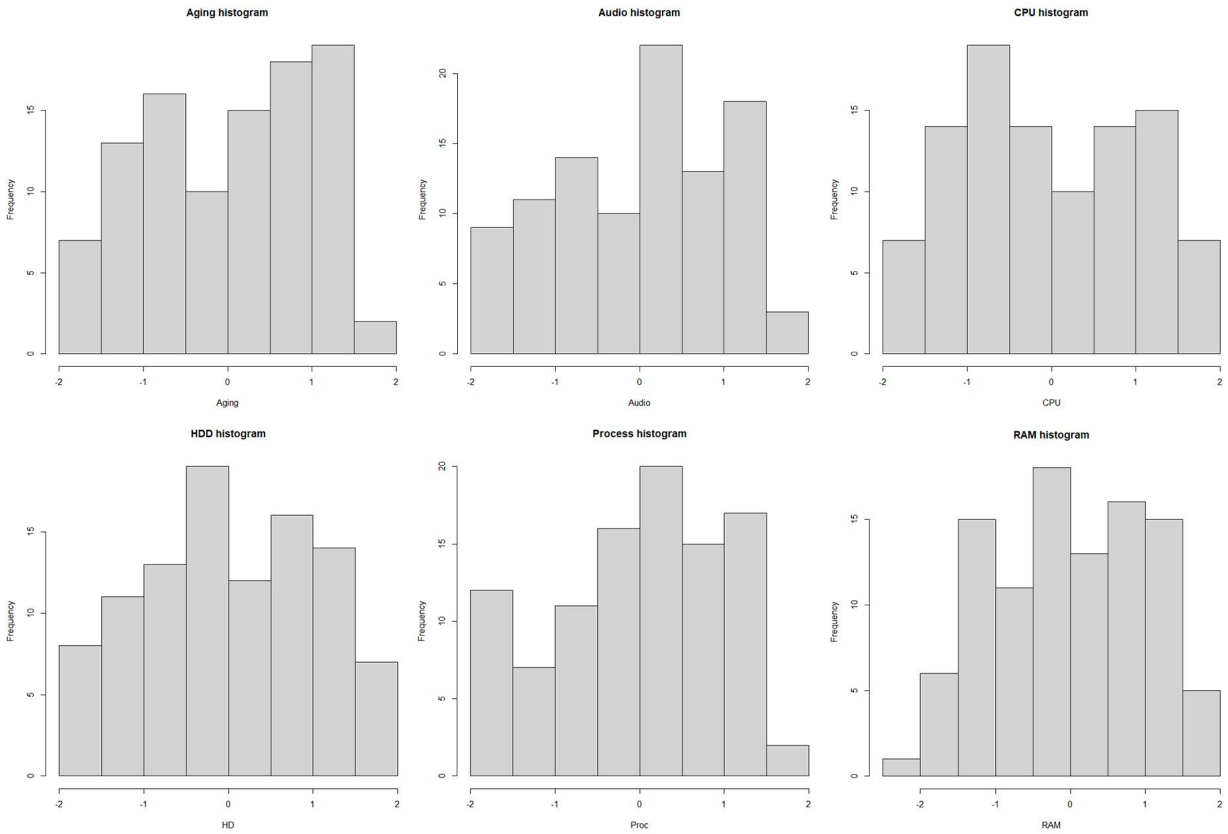
AUDIO

Classi	Centro di Classe	Frequenza	Frequenza Relativa	Frequenza Cumolativa	Frequenza Relativa Cumolativa
-1.6999≥ i >-1.2922	-1.4960	18	0.18	18	0.18
-1.2922≥ i >-0.8846	-1.0884	5	0.05	23	0.23
-0.8846≥ i >-0.4770	-0.6808	12	0.12	35	0.35
-0.4770≥ i >-0.0694	-0.2732	9	0.09	44	0.44
-0.0694≥ i > 0.3382	0.1344	12	0.12	56	0.56
0.3382≥ i > 0.7458	0.5420	15	0.15	71	0.71
0.7458≥ i > 1.1534	0.9496	15	0.15	86	0.86
1.1534≥ i ≥1.5620	1.3577	14	0.14	100	1.00

RAM

Classi	Centro di Classe	Frequenza	Frequenza Relativa	Frequenza Cumolativa	Frequenza Relativa Cumolativa
-2.01071≥ i >-1.54021	-1.77546	7	0.07	8	0.08
-1.54021≥ i >-1.06971	-1.30496	13	0.13	20	0.20
-1.06971≥ i >-0.59921	-0.83446	10	0.10	30	0.30
-0.59921≥ i >-0.12871	-0.36396	13	0.13	43	0.43
-0.12871≥ i > 0.34179	0.10654	14	0.14	57	0.57
0.34179≥ i > 0.81229	0.57704	18	0.18	75	0.75
0.81229≥ i > 1.28279	1.04754	16	0.16	91	0.91
1.28279≥ i ≥1.75363	1.51821	9	0.09	100	1.00

Analisi degli istogrammi



Dagli istogrammi ottenuti e dai test di Shapiro svolti sulle variabili indipendenti osservate è evidente che tutte le colonne non sono distribuite in accordo a una normale.

Possiamo dunque procedere in seguito con la regressione lineare, ipotizzando che le variabili di maggior influenza siano sicuramente cpu, processo ed aging.

Sezione 3: Analisi in regressione

In un primo momento dobbiamo capire quali variabili utilizzare per la regressione, dunque facciamo un'analisi del modello lineare relativo al dataset.

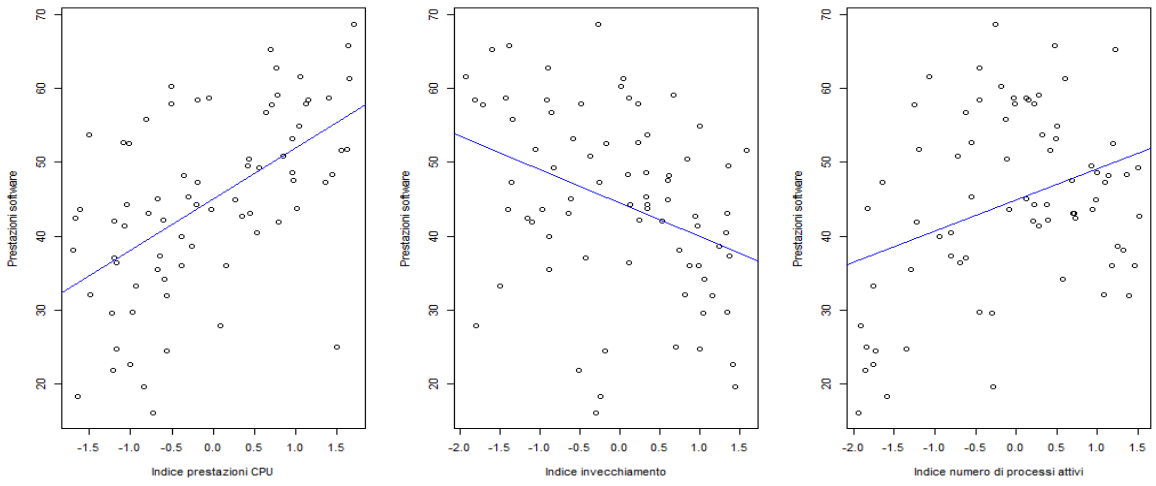
Suddividiamo il dataset in due parti, identificando un 75% dei dati come train set ed il restante 25% come test set, così da poter successivamente testare il modello e vedere quanto i valori calcolati si discostano dai valori veri.

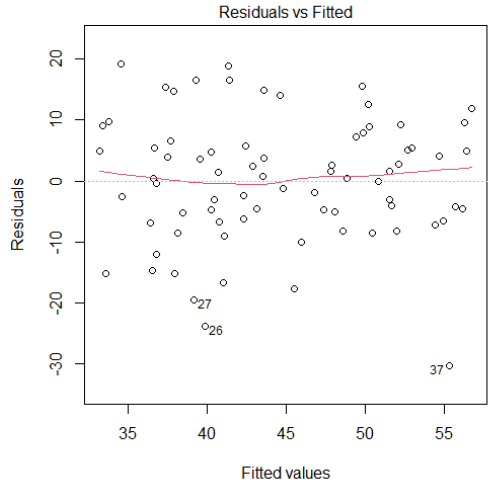
Analisi dei singoli modelli lineari monomiali

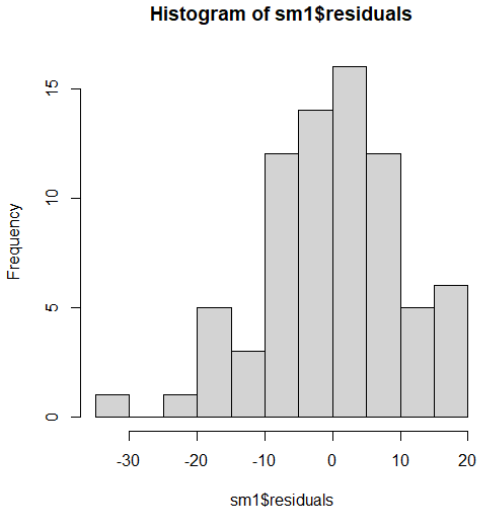
Tramite le funzioni lm (linear model) e summary analizziamo le relazioni numeriche tra le singole variabili indipendenti e l'output, considerando il relativo valore assunto dal p-value e scartandole dal modello se quest'ultimo risulta essere superiore al livello di rischio prefissato pari a 0.05.

Dall'analisi, come già predetto nei precedenti test, notiamo che le variabili con il p-value conforme agli standard sono proprio quelle relative alle prestazioni della CPU utilizzata, al numero di processi attivi ed all'invecchiamento del sistema.

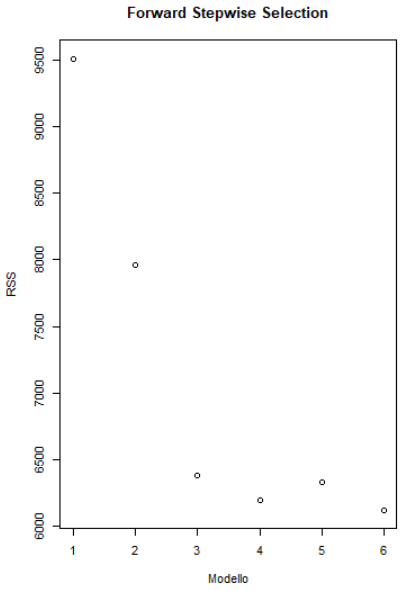
Inoltre, attraverso l'indice Multiple R-squared, notiamo che la variabile di maggior risalto per l'aderenza al modello reale è x1, ovvero l'indice di performance della CPU, con un valore pari a 0.3124, seguita poi da x4 (l'indice di aging) con un R^2 pari a 0.1293, ed infine abbiamo la variabile x3 col suo valore di riferimento pari a 0.1205.

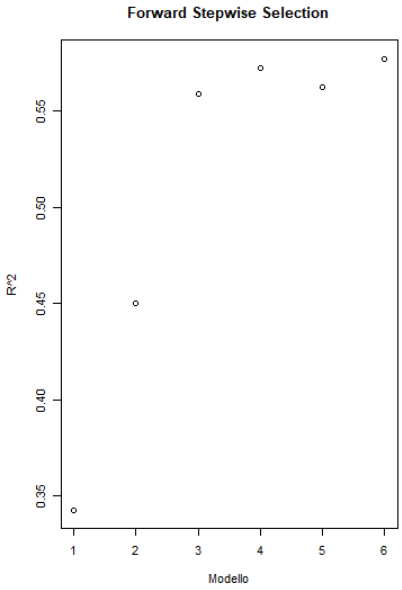


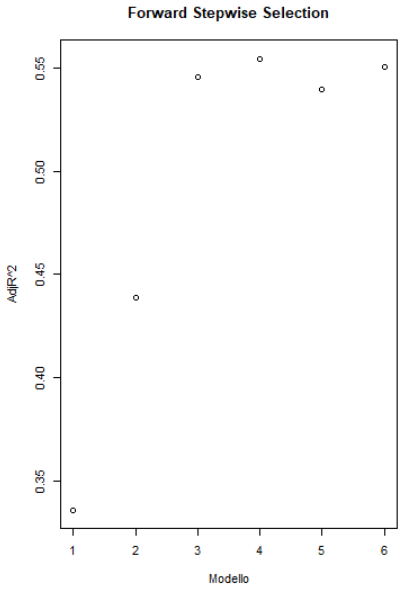




Per avere un'ulteriore conferma sulle successive variabili da aggiungere al modello al fine di avviare la regressione multipla lineare abbiamo utilizzato la funzione step, per analizzare i valori dell'AIC (Akaike Information Criterion) e dell'RSS (Residual Sum of Squares), la quale ci ha fornito un possibile set di variabili da utilizzare in base al numero di elementi da considerare.







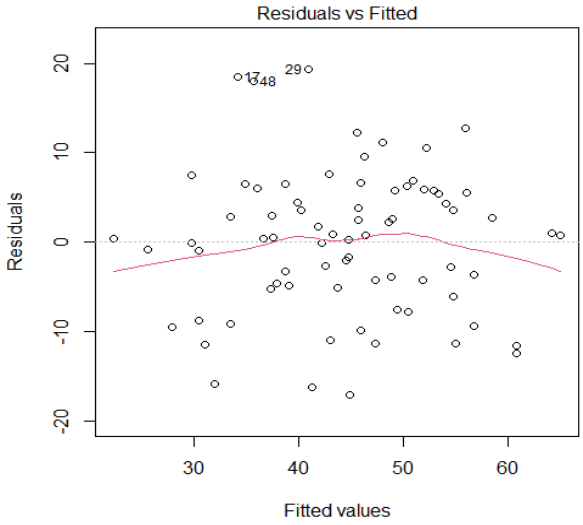
Analisi dei modelli lineari multipli

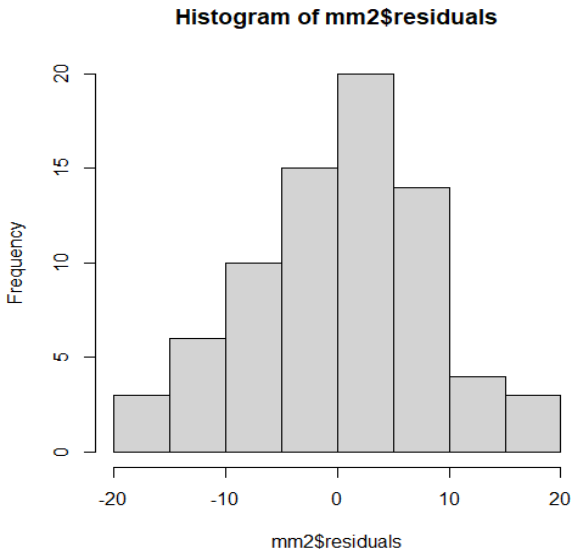
Abbiamo poi svolto una regressione lineare multipla con la funzione lm prendendo in considerazione le prime due variabili il cui R^2 risultasse più alto dalle analisi dei modelli lineari (quindi x1 e x4) e ne abbiamo ottenuto i risultati attraverso la funzione summary. Queste ultime hanno un ruolo molto importante nella definizione del modello finale, essendo il loro p-value minore rispetto al livello di rischio alpha pari a 0.05 ed incrementando l'R^2 portandolo a 0.4041.

Per essere certi che il miglioramento dell'indice R^2 corretto (Adjusted R^2) fosse significativo, abbiamo usato la funzione anova prendendo in considerazione il precedente modello migliore, in base a tale indice, e l'abbiamo confrontato con l'ultimo miglior modello analizzato senza variabili il cui p-value risultasse superiore a 0.05. Il valore del p-value risultante è anch'esso inferiore rispetto al livello di rischio, quindi possiamo rigettare l'ipotesi nulla (secondo cui il miglioramento non è significativo) e stabilire che il modello è statisticamente rilevante nella definizione del modello finale.

Abbiamo poi svolto le stesse analisi introducendo l'ultima variabile il cui p-value risultasse concorde con il limite imposto ed abbiamo notato un aumento dell'indice di permeabilità dei dati col modello reale pari a 0.5791. Da una conseguente analisi anova, l'aumento risulta essere statisticamente significativo.

Terminata l'analisi delle variabili da aggiungere al modello lineare multiplo, abbiamo svolto le stesse analisi sulla linearità dei residui e sull'ipotesi di una loro distribuzione normale rispetto ai modelli lineari migliori basandosi sull'R^2.



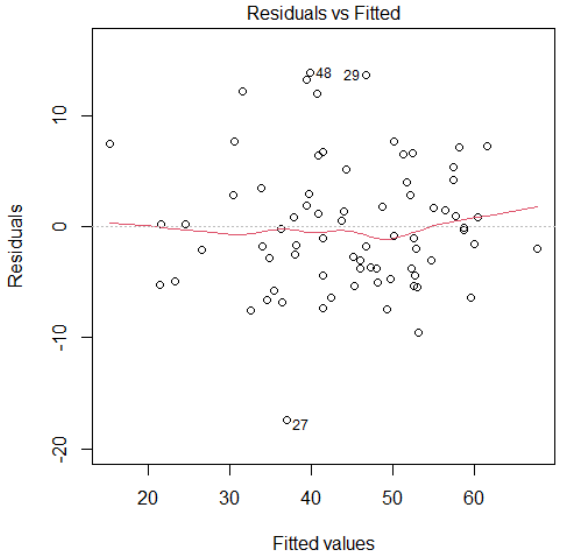


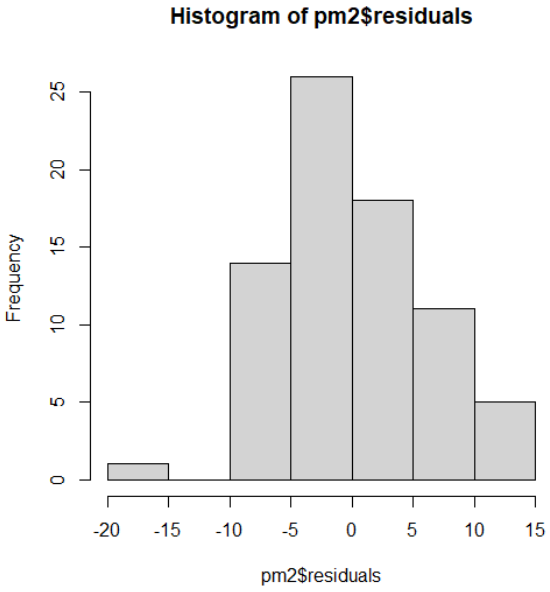
Analisi dei modelli lineari polinomiali multipli

Infine, abbiamo svolto una regressione polinomiale multipla confrontando il modello ottenuto dal prodotto delle variabili a due a due, di tutte e tre le variabili ed il quadrato delle singole, notando che il p-value di $x1^2$, $x4^2$, x1*x4, x1*x3, x4*x3 e di x1*x3*x4 è troppo elevato, quindi le abbiamo eliminate ed abbiamo ottenuto il modello migliore in base all'R^2 pari a 0.7733, confermato anche dall'analisi dell'anova fatta tra l'ultimo modello ottenuto ed il modello a regressione lineare multipla con l'indice più alto e con tutte le variabili con p-value basso.

Si identifica come modello migliore poiché, da una successiva analisi del modello, prendendo anche in considerazione l'elevazione al cubo della variabile x3, abbiamo notato che l'R^2 ha avuto un leggero aumento pari a 0.7744, ma il suo p-value risulta estremamente alto, quindi non possiamo tenerlo in considerazione.

Trovato dunque il modello migliore tramite regressione polinomiale multipla, abbiamo svolto le analisi di linearità e normalità dei residui, come per gli altri modelli analizzati.





Calcolo numerico outliers

Terminati i seguenti test, abbiamo analizzato, tramite la funzione augment(), la possibile presenza di valori tali per cui potessero essere presenti eventuali outliers nel modello ma, come già evidenziato dai boxplot delle variabili notando l'assenza di punti distaccati dalle estremità inferiori e superiori, non ne risulta l'effettiva presenza, essendo i valori assoluti dei massimi e dei minimi dei residui standardizzati tutti minori di 3.

Test sul modello finale

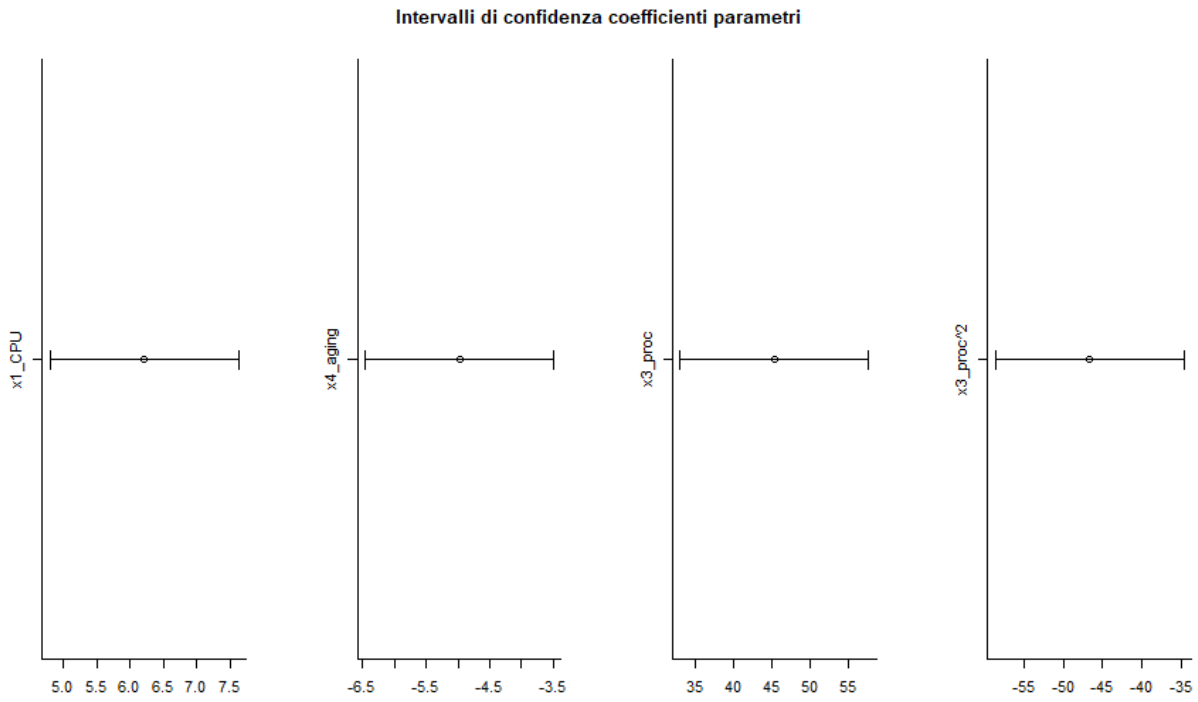
Le nostre ultime analisi sul modello finale sono state svolte attraverso la predizione dei valori possibili del test set ed osservandone i gradi di accuratezza tra i calcoli del modello e i valori reali.

Abbiamo ottenuto i seguenti valori:

- SQE (Variabilità delle unità sperimentali) = 3139.093
- MSQE (Stimatore non distorto della varianza) = 33.0431
- S (Deviazione campionaria) = 5.748311
- MSE (Mean Squared Error) = 27.51494

Intervalli di confidenza dei parametri

Infine, abbiamo ottenuto gli intervalli di confidenza sui parametri del modello tramite la funzione confint, considerandone il 2.5° e 97.5° percentile.



Sezione 4: Considerazioni finali

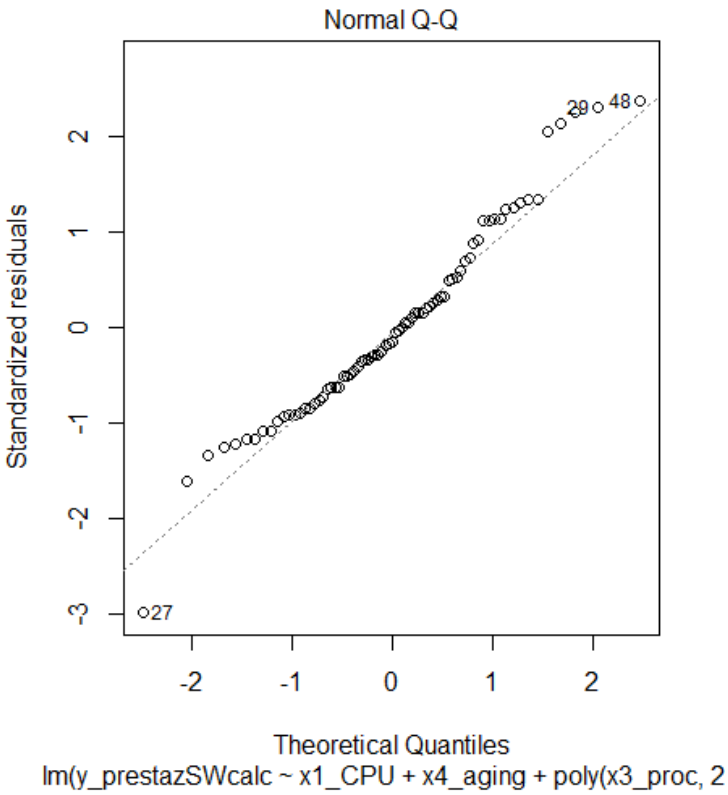
Il miglior modello finale trovato in base all'R^2 ed ai p-value adeguati delle variabili indipendenti utilizzate ha la seguente forma:

Prestazioni del software di calcolo = 6.2181 * (Indice prestazioni CPU) +

- 4.9773 * (Indice invecchiamento PC) +

+ 45.3142 * (numero di processi attivi) +

- 46.6915 * (numero di processi attivi)^2



Fine Relazione

Gruppo11

- Capobianco Federica
- D’Angelo Antonio
- Ferrigno Antonio