4 – Regressione

Nel seguito useremo JMP per effettuare:

- La stima lineare della variabile di risposta rispetto al predittore
- Controllare la normalità e l'omoschedasticità dei residui
- Effettuare la regressione ai minimi quadrati
- In caso di non normalità dei residui, effettueremo un test di Mann-Kendall

Per la stima della pendenza e dell'intercetta della regressione è stato utilizzato uno script Python che fa uso della libreria *scipy:*

$$slope, intercept, low, up = theilslopes(y, x, alfa)$$

L'input della funzione è:

- y: variabile di risposta
- x: predittore
- $alfa$: nel nostro caso lo impostiamo sempre a 0.95, corrispondente a un livello di significatività del 95%

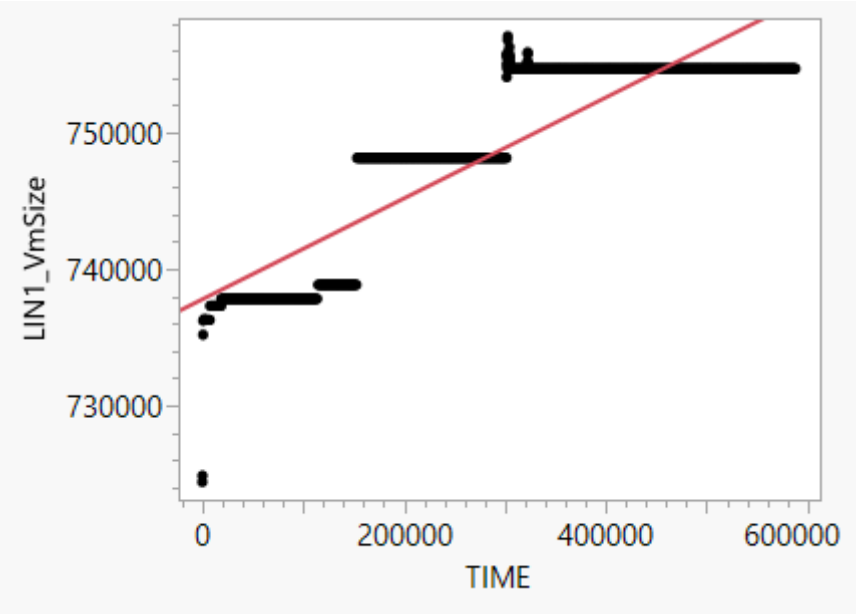
L'output della funzione è:

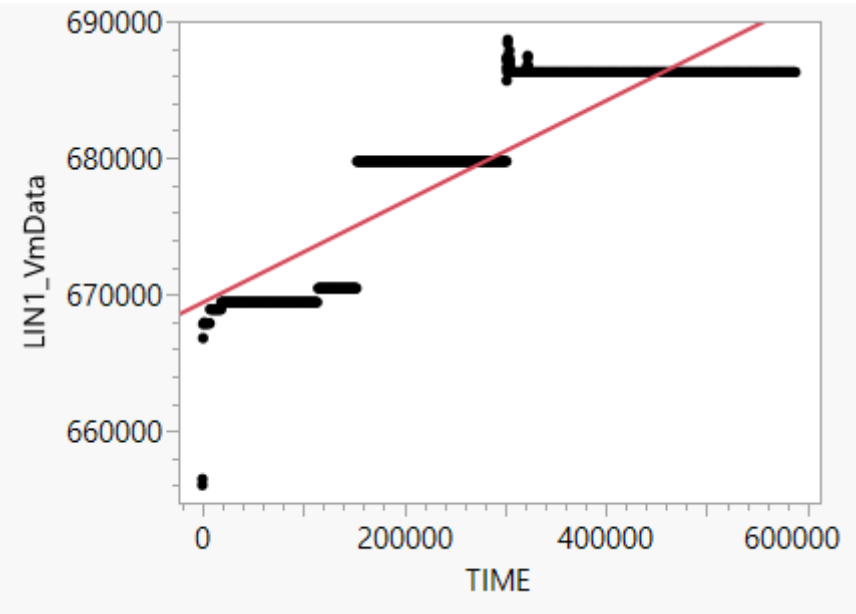
- $slope$: pendenza della regressione
- $intercept$: intercetta della retta di regressione
- low: limite inferiore dell'intervallo di confidenza della pendenza
- up: limite superiore dell'intervallo di confidenza della pendenza

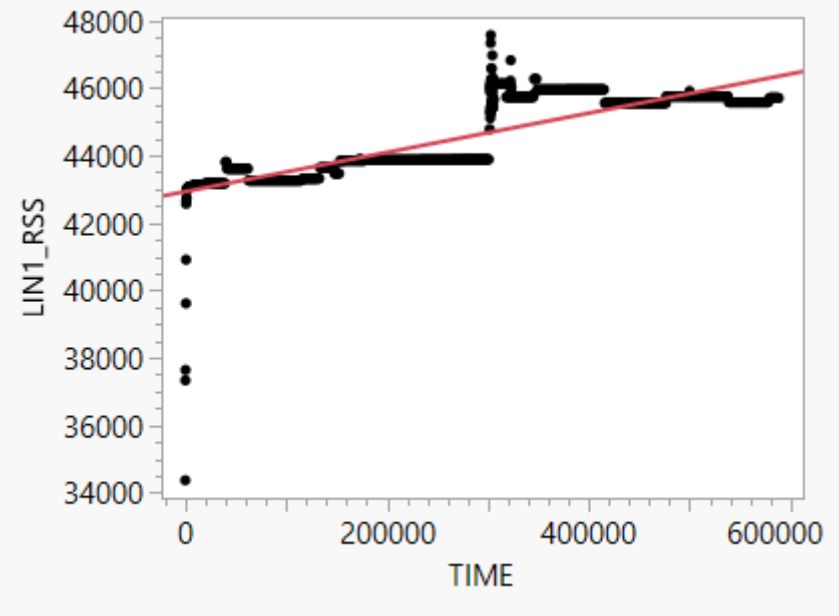
4.1 OS

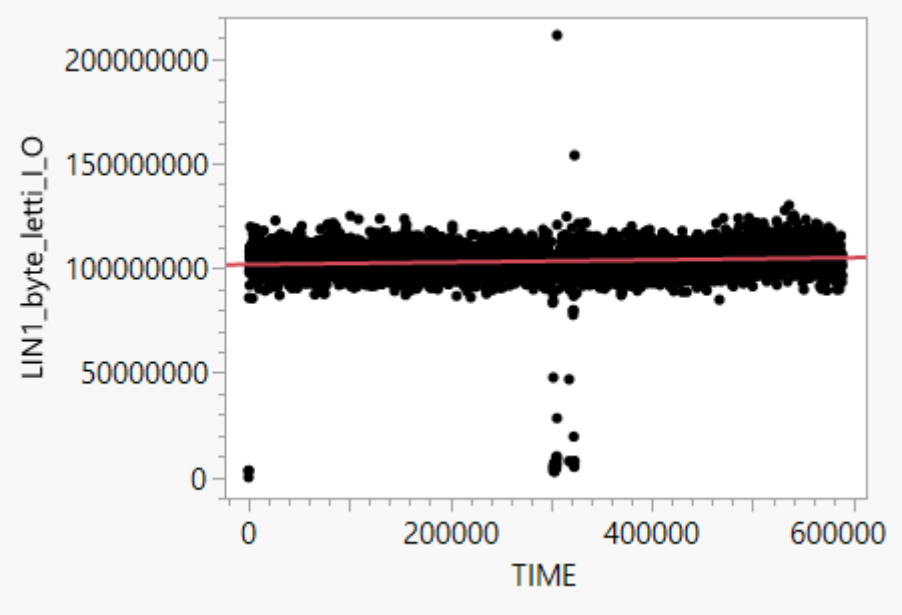
Lo scopo di questo esercizio è rilevare e stimare eventuali trend sulle 5 variabili (Size, Data, RSS, Byte letti e Byte scritti) utilizzando modelli regressivi lineari semplici, parametrici e/o non parametrici e farlo per i tre dataset os1, os2 e os3.Infine verranno confrontati i trend individuati nei tre dataset.

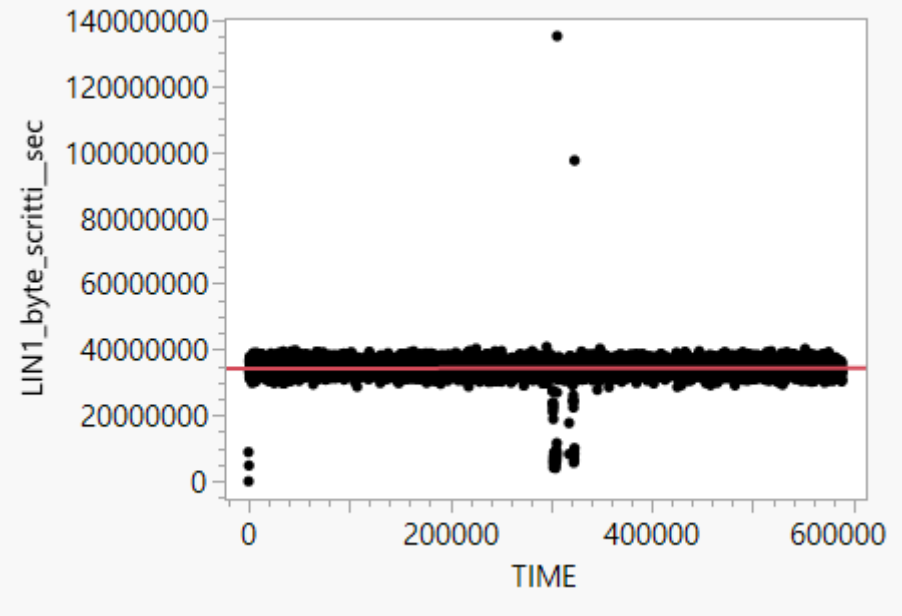
Nel dettaglio analizzeremo in prima battuta le rette di regressione delle variabili di risposta rispetto a Time, che ci serviranno per il calcolo dei residui. Le immagini vengono messe in ordine di dataset.

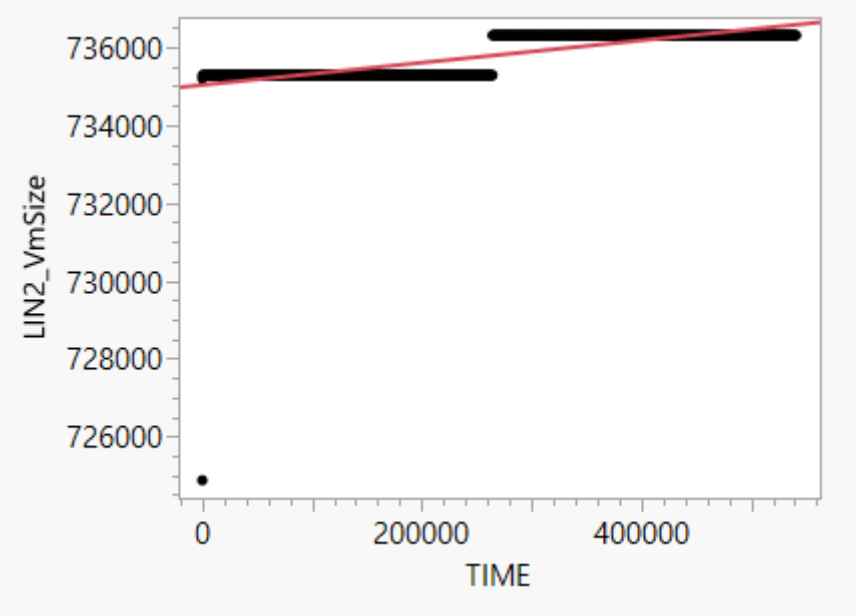


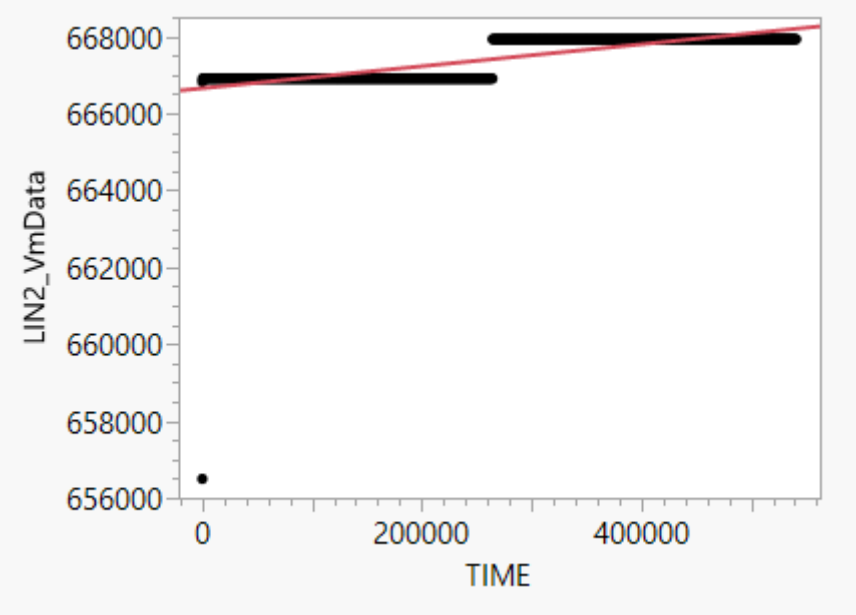


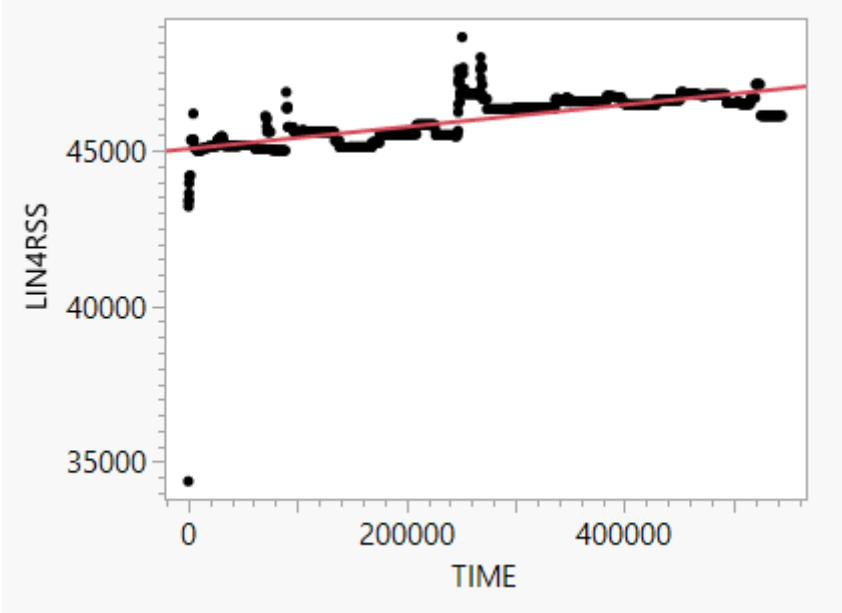

48000
46000
44000
42000
40000
38000
36000
34000
LIN1_RSS
0
200000
400000
600000
TIME

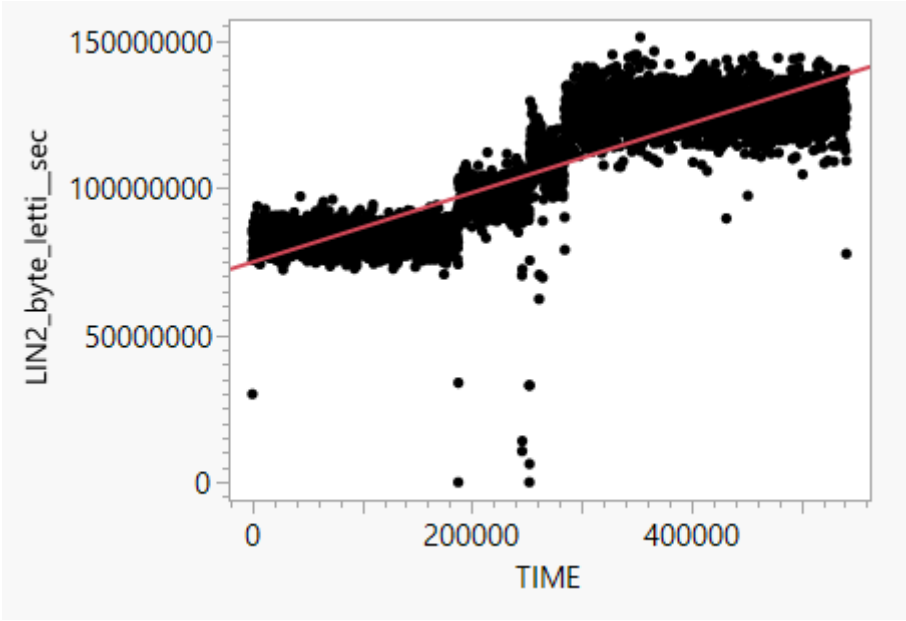

200000000
150000000
100000000
50000000
0
LIN1_byte_letti_I_O
0
200000
400000
600000
TIME

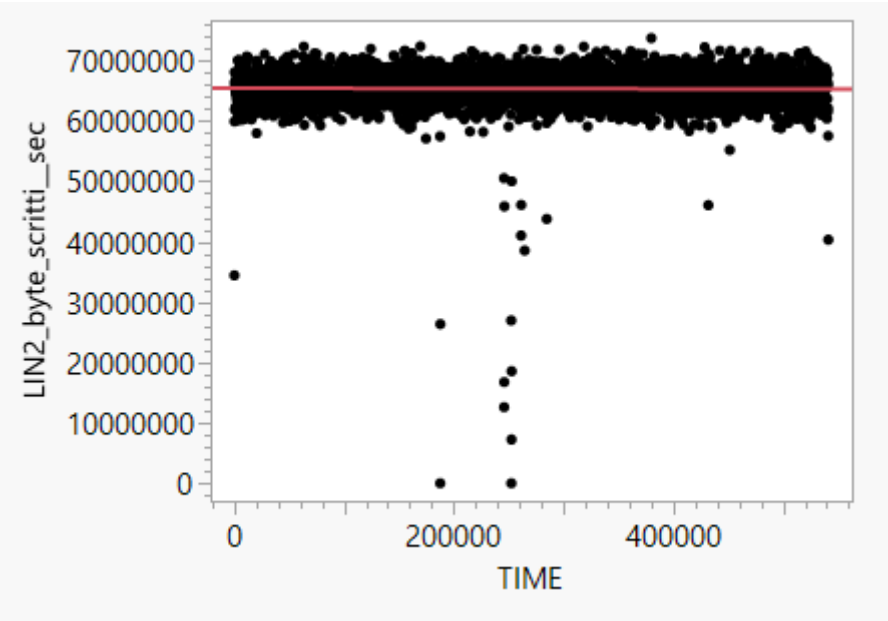

140000000
120000000
100000000
80000000
60000000
40000000
20000000
0
LIN1_byte_scritti__sec
0
200000
400000
600000
TIME

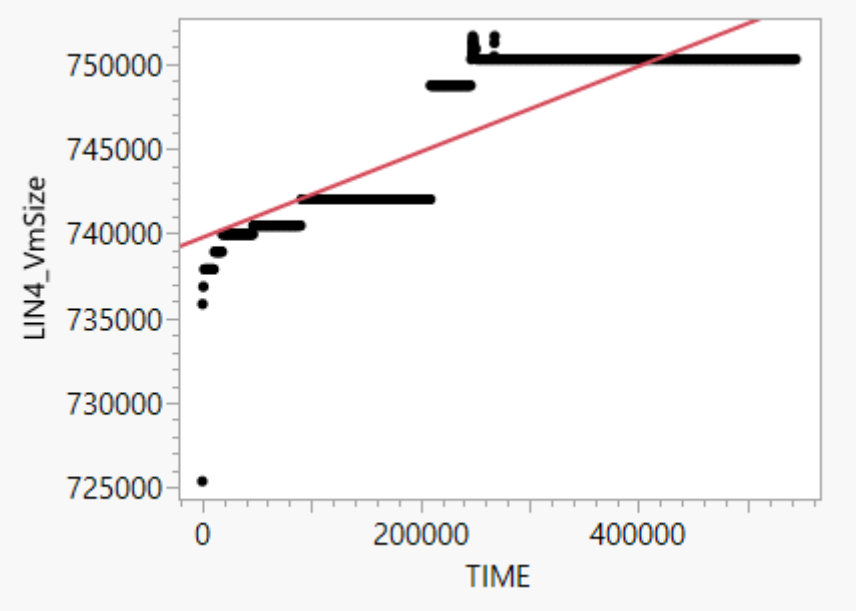

736000
734000
732000
730000
728000
726000
LIN2_VmSize
0
200000
400000
TIME

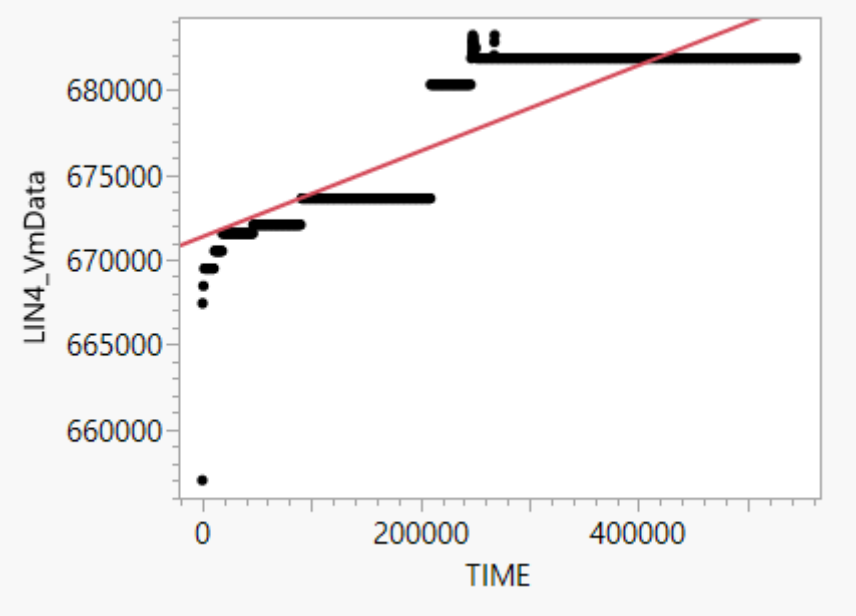

668000
666000
664000
662000
660000
658000
656000
LIN2_VmData
0
200000
400000
TIME

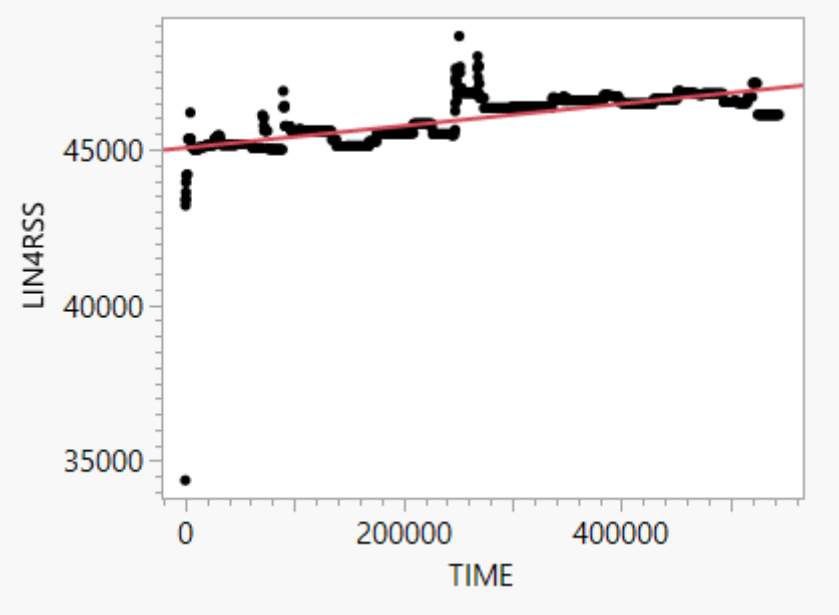

45000
40000
35000
LIN4RSS
0
200000
400000
TIME

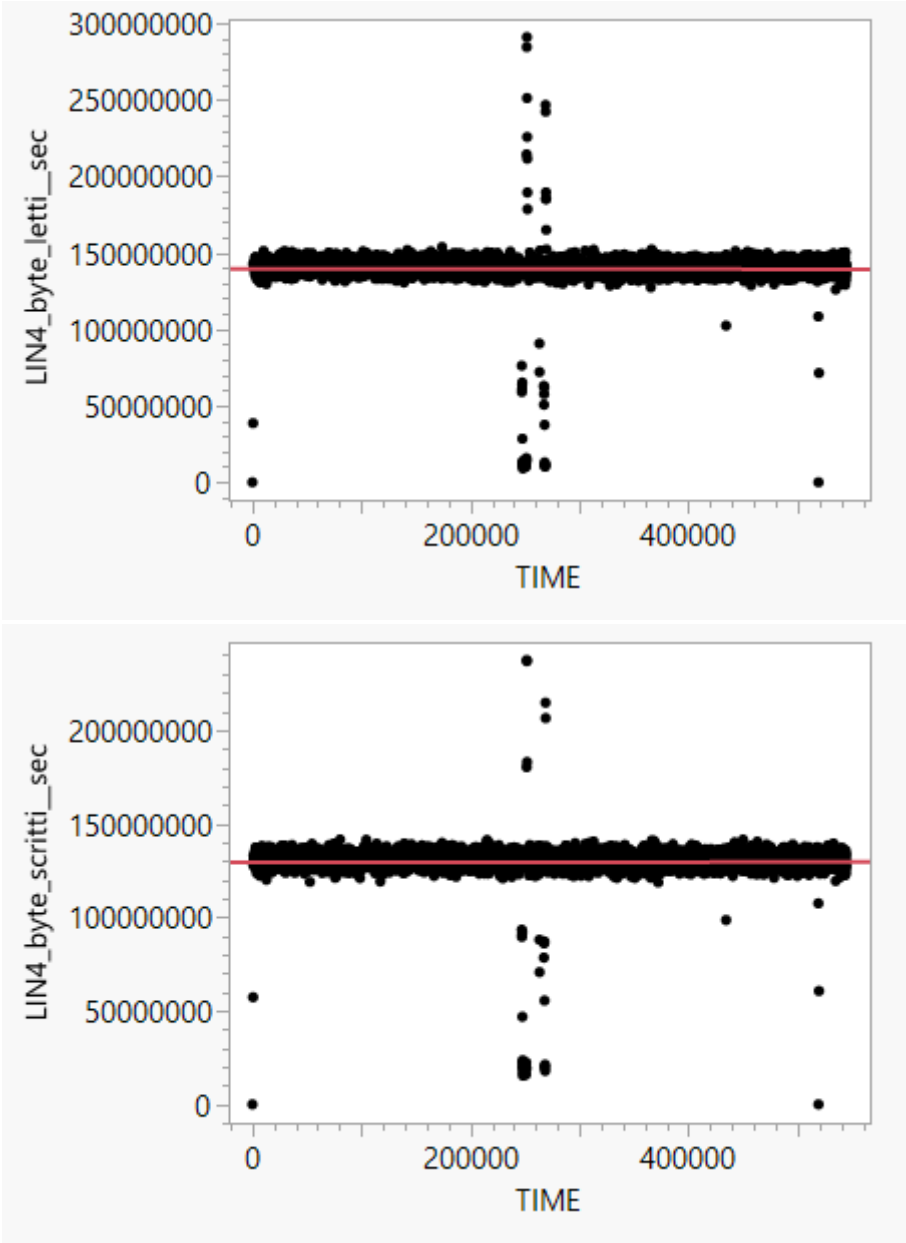

LIN2_byte_letti__sec
150000000
100000000
50000000
0
0
200000
400000
TIME


LIN2_byte_scritti__sec
70000000
60000000
50000000
40000000
30000000
20000000
10000000
0
0
200000
400000
TIME

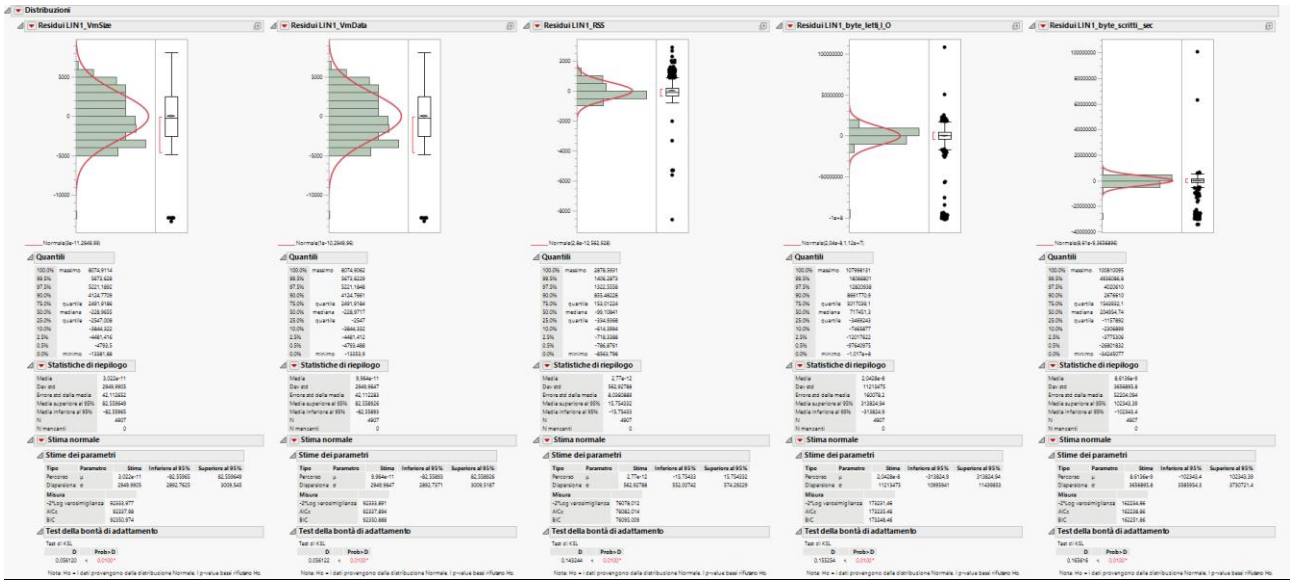

LIN4_VmSize
750000
745000
740000
735000
730000
725000
0
200000
400000
TIME

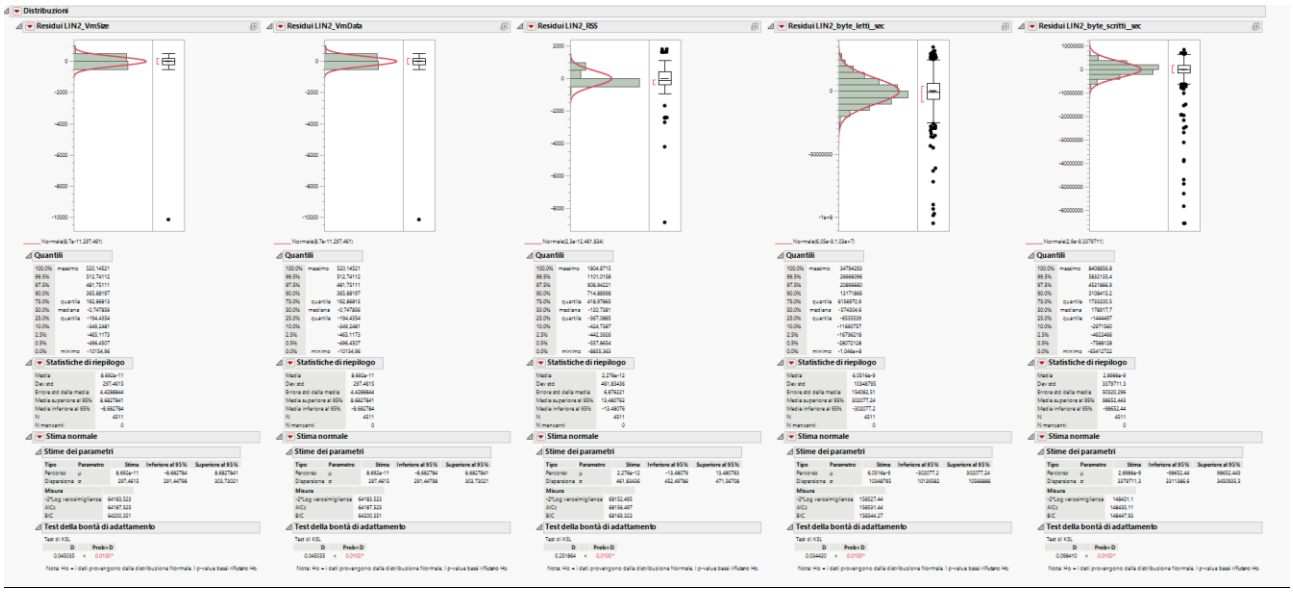

LIN4_VmData
680000
675000
670000
665000
660000
0
200000
400000
TIME

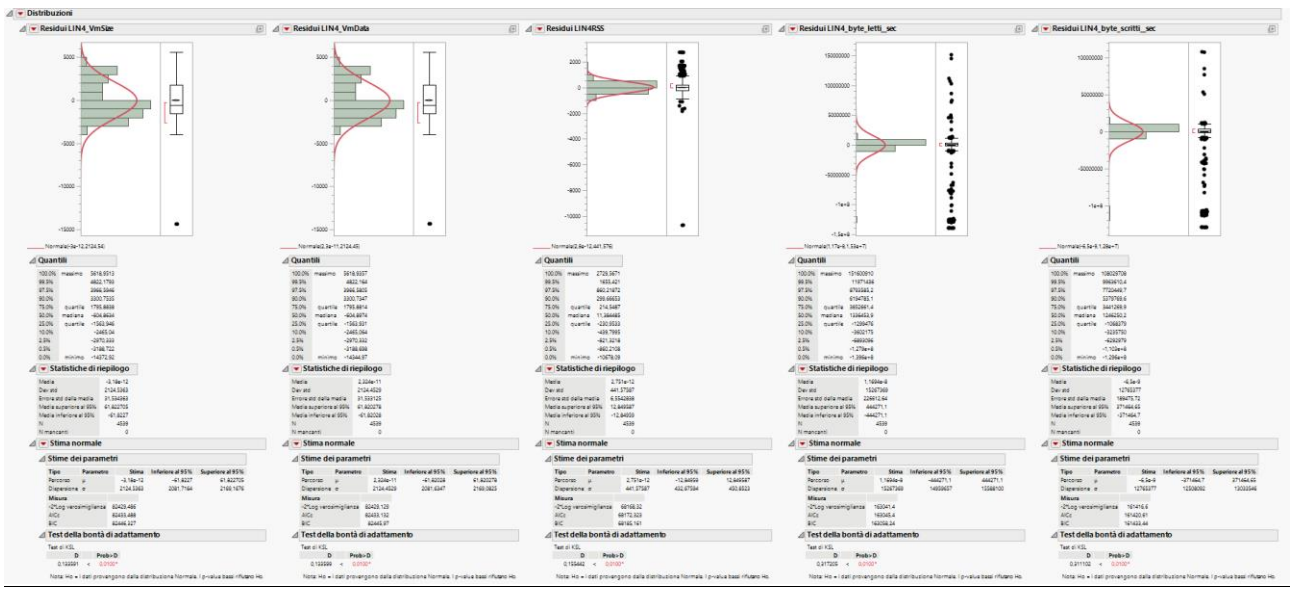

LIN4RSS
45000
40000
35000
0
200000
400000
TIME



Successivamente analizzeremo la distribuzione dei residui che rappresentano gli errori rispetto alla predizione del modello di regressione.







Dalle immagini è evidente la non-normalità dei dati, viene ritenuto inutile quindi testare la loro omoschedasticità, per cui si procede direttamente ad effettuare un test di regressione lineare non parametrico: il test di Mann-Kendall.

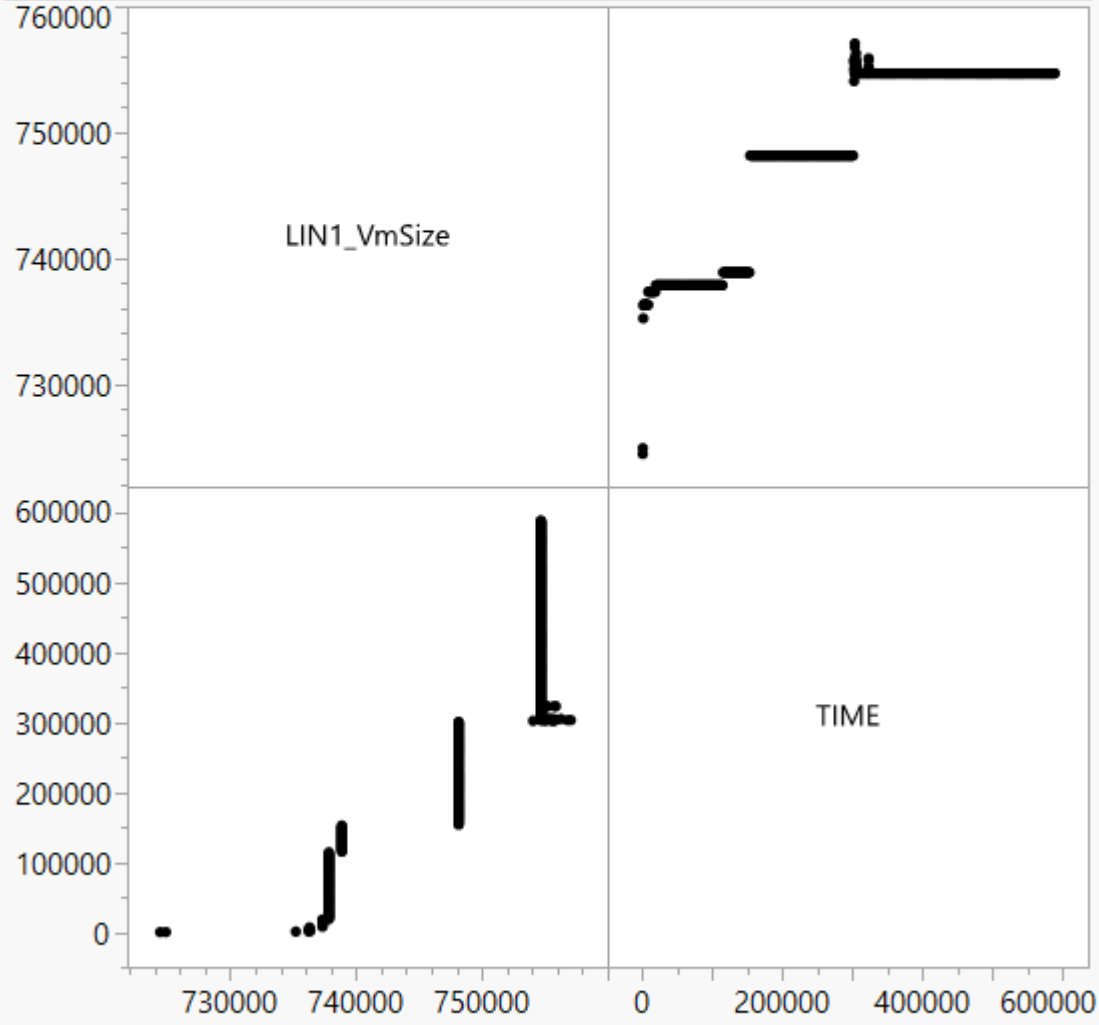
Multivariato

Correlazioni

	LIN1_VmSize	TIME
LIN1_VmSize	1,0000	0,9052
TIME	0,9052	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8-,6-,4-,2 0 ,2 ,4 ,6 ,8
TIME	LIN1_VmSize	0,8076	<,0001*	

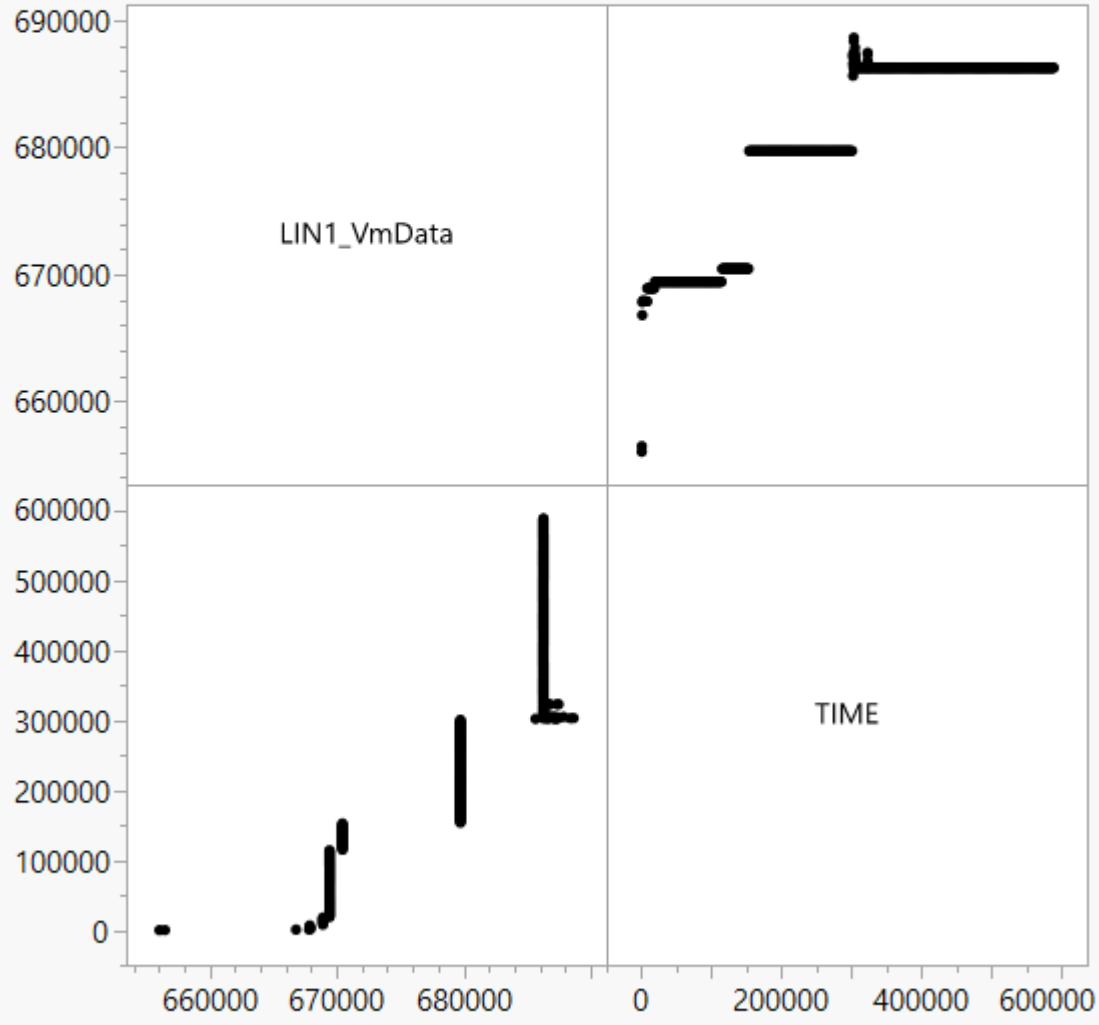
Multivariato

Correlazioni

	LIN1_VmData	TIME
LIN1_VmData	1,0000	0,9052
TIME	0,9052	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
TIME	LIN1_VmData	0,8076	<,0001*	

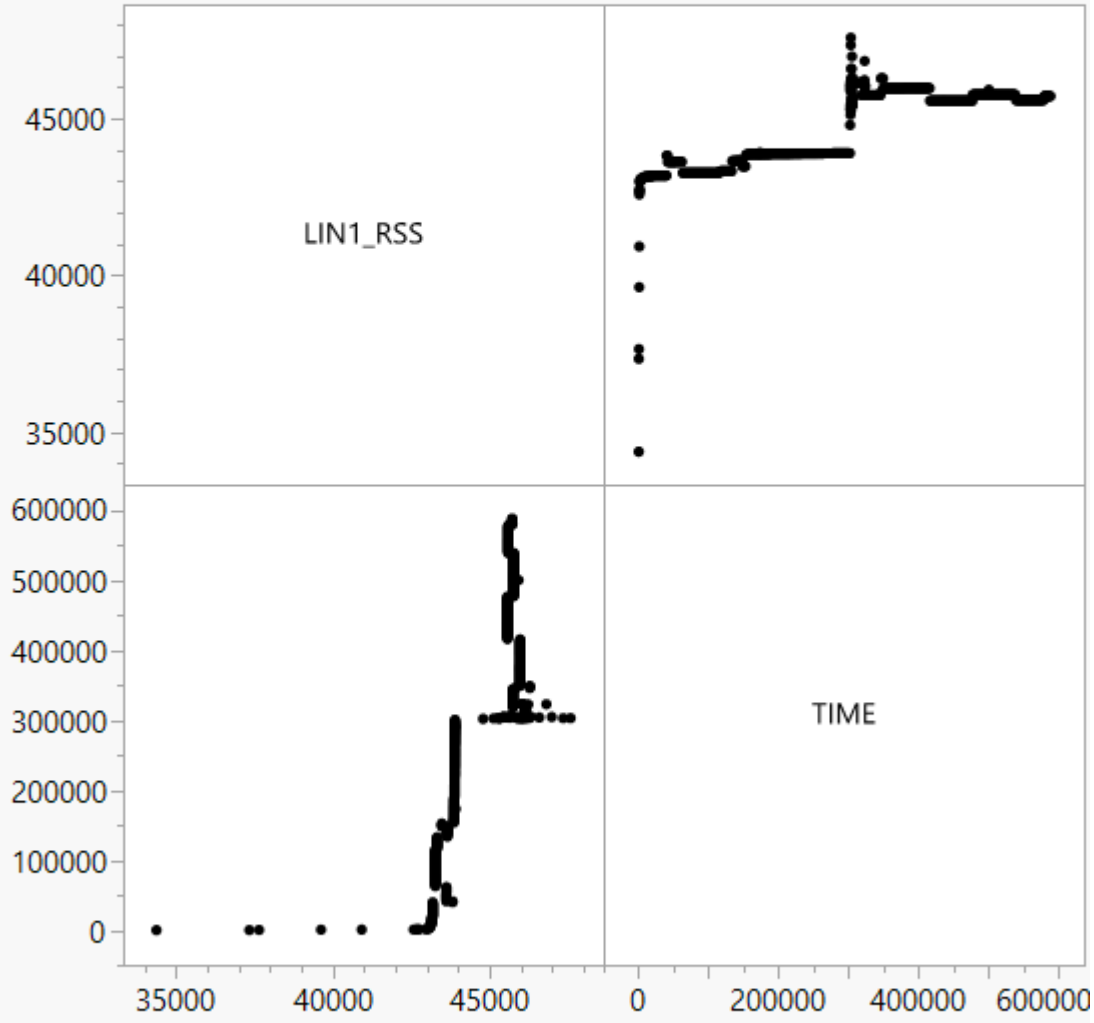
Multivariato

Correlazioni

	LIN1_RSS	TIME
LIN1_RSS	1,0000	0,8691
TIME	0,8691	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
TIME	LIN1_RSS	0,6827	<,0001*	

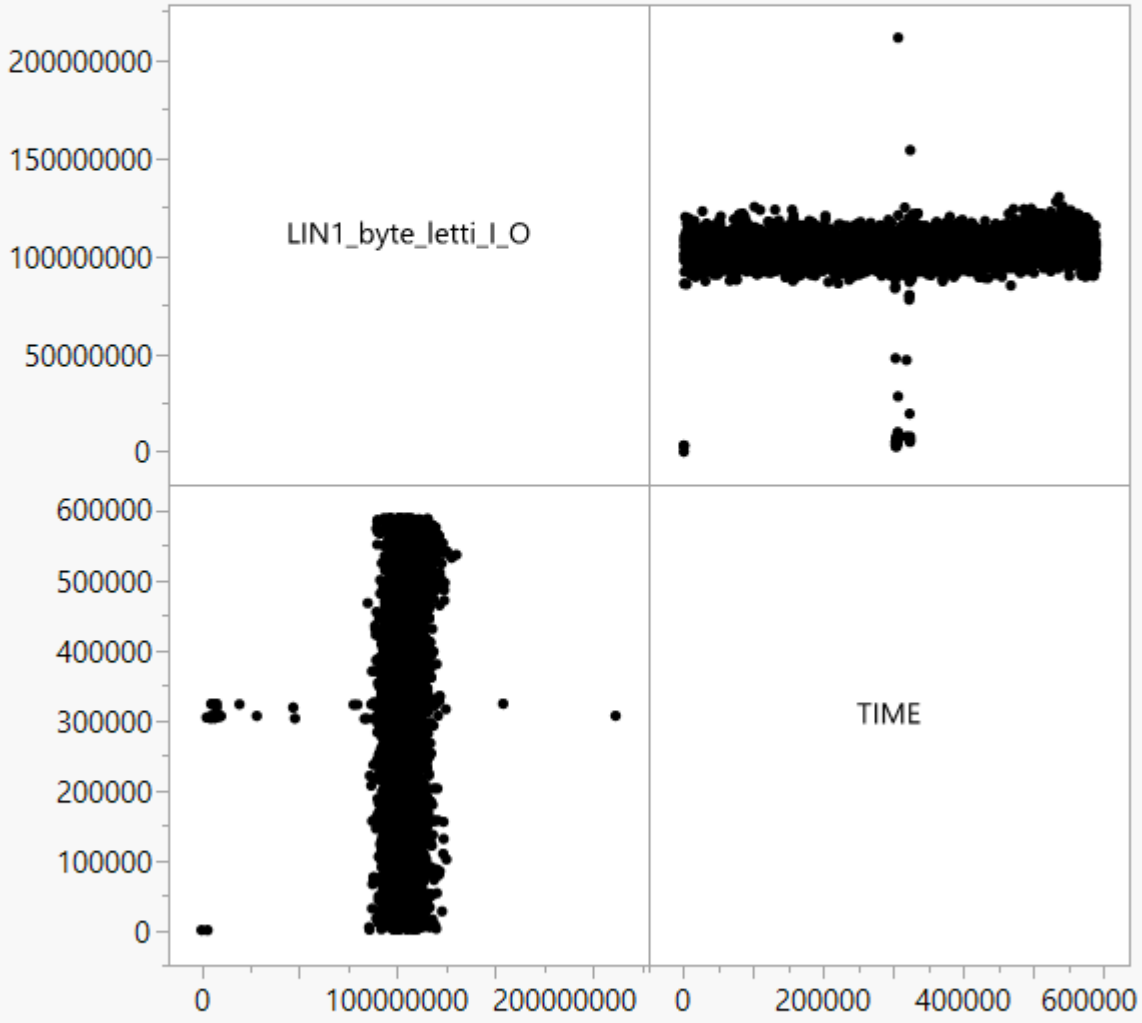
Multivariato

Correlazioni

	LIN1_byte_letti_I_O	TIME
LIN1_byte_letti_I_O	1,0000	0,0834
TIME	0,0834	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8-,6-,4-,2 0 ,2 ,4 ,6 ,8
TIME	LIN1_byte_letti_I_O	0,0823	<,0001*	

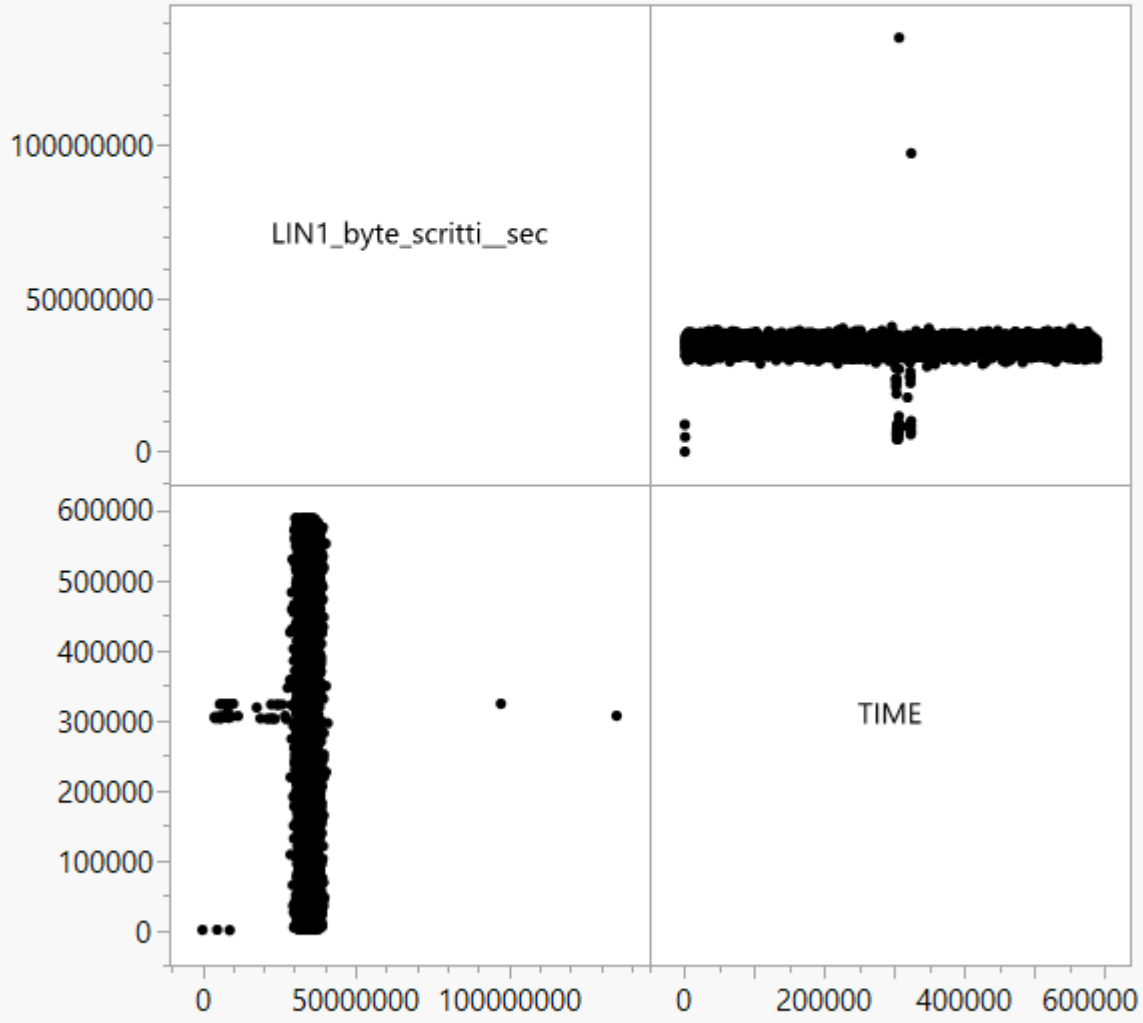
Multivariato

Correlazioni

	LIN1_byte_scritti__sec	TIME
LIN1_byte_scritti__sec	1,0000	0,0073
TIME	0,0073	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
TIME	LIN1_byte_scritti__sec	-0,0017	0,8555	

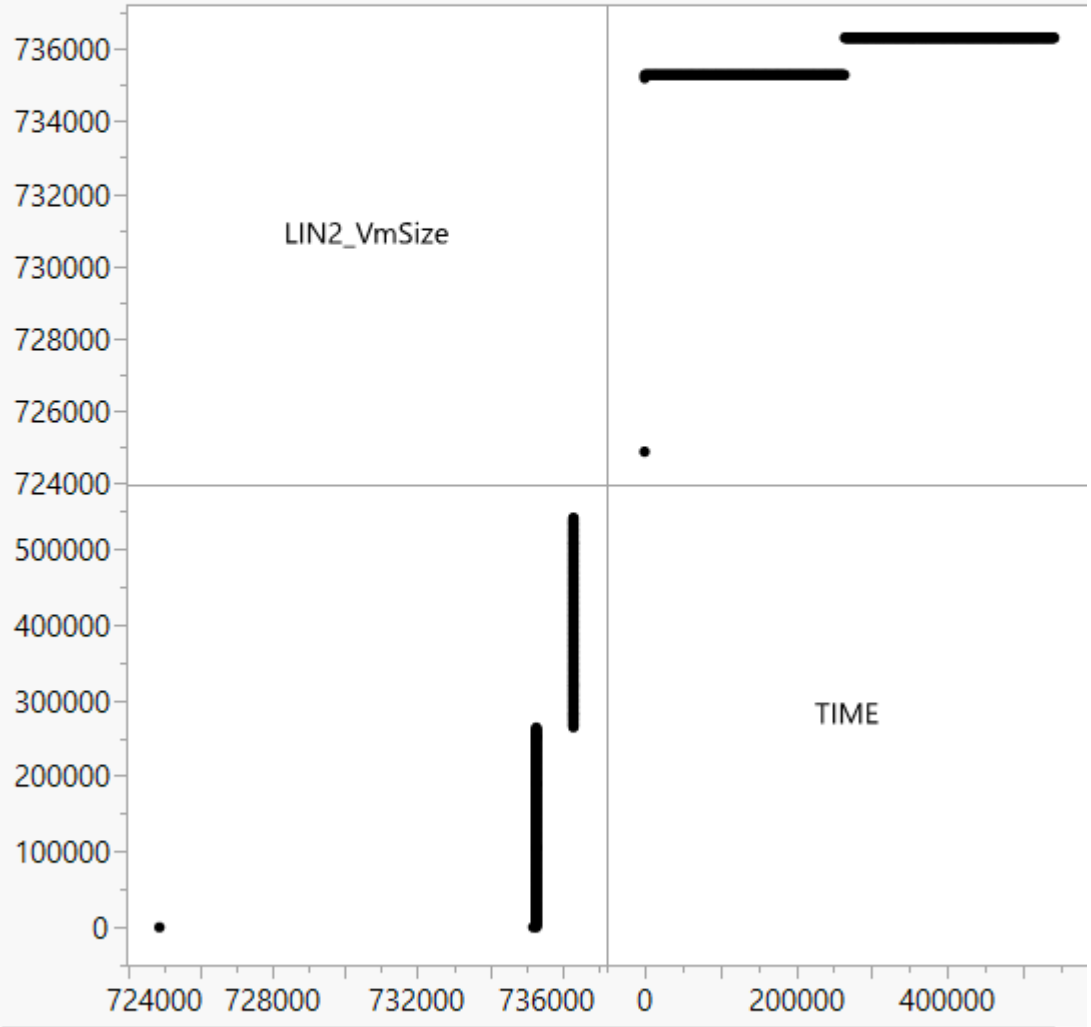
Multivariato

Correlazioni

	LIN2_VmSize	TIME
LIN2_VmSize	1,0000	0,8327
TIME	0,8327	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8-,6-,4-,2 0 ,2 ,4 ,6 ,8
TIME	LIN2_VmSize	0,7065	<,0001*	

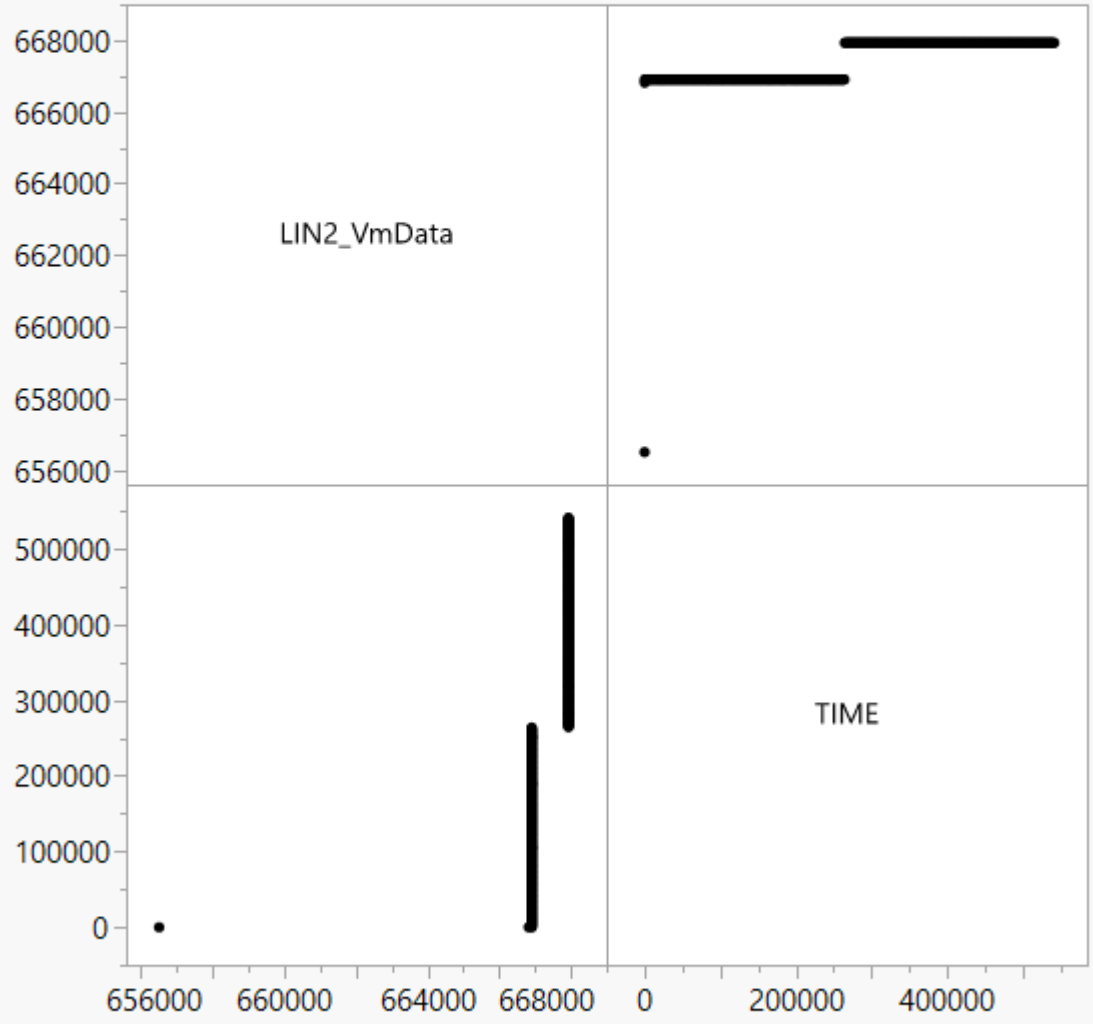
Multivariato

Correlazioni

	LIN2_VmData	TIME
LIN2_VmData	1,0000	0,8327
TIME	0,8327	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8-,6-,4-,2 0 ,2 ,4 ,6 ,8
TIME	LIN2_VmData	0,7065	<,0001*	

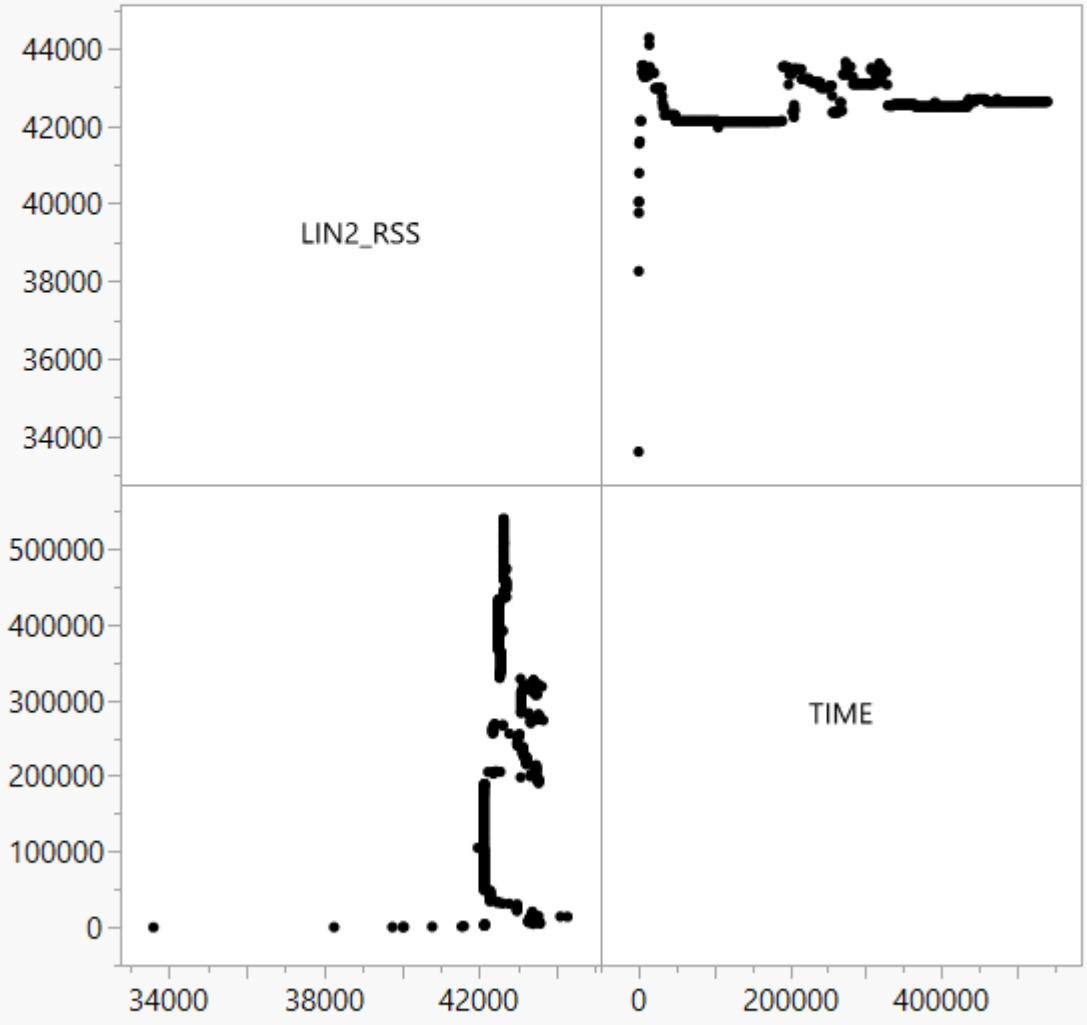
Multivariato

Correlazioni

	LIN2_RSS	TIME
LIN2_RSS	1,0000	0,1840
TIME	0,1840	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8-,6-,4-,2 0 ,2 ,4 ,6 ,8
TIME	LIN2_RSS	0,1788	<,0001*	

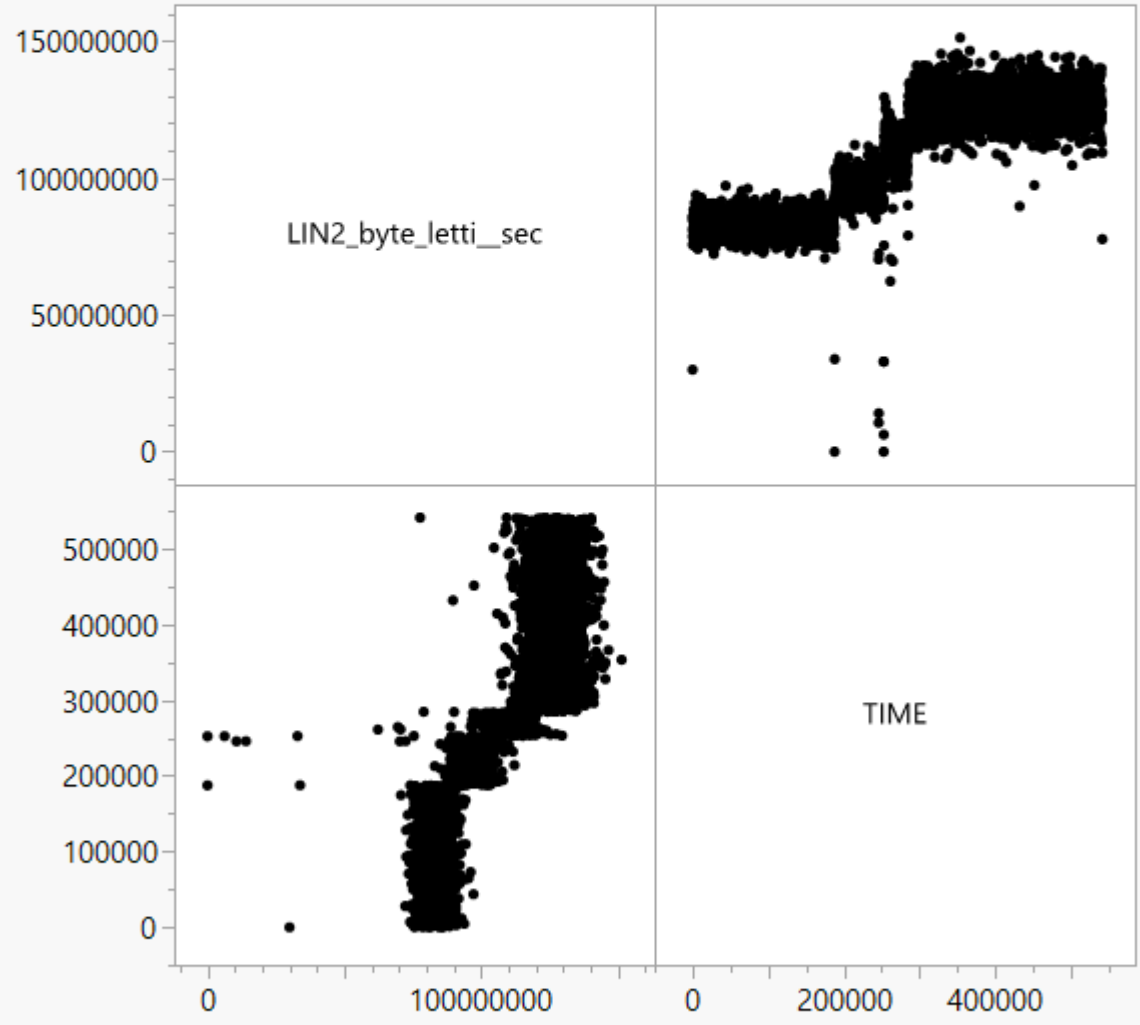
Multivariato

Correlazioni

	LIN2_byte_letti__sec	TIME
LIN2_byte_letti__sec	1,0000	0,8720
TIME	0,8720	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8-,6-,4-,2 0 ,2 ,4 ,6 ,8
TIME	LIN2_byte_letti__sec	0,6259	<,0001*	

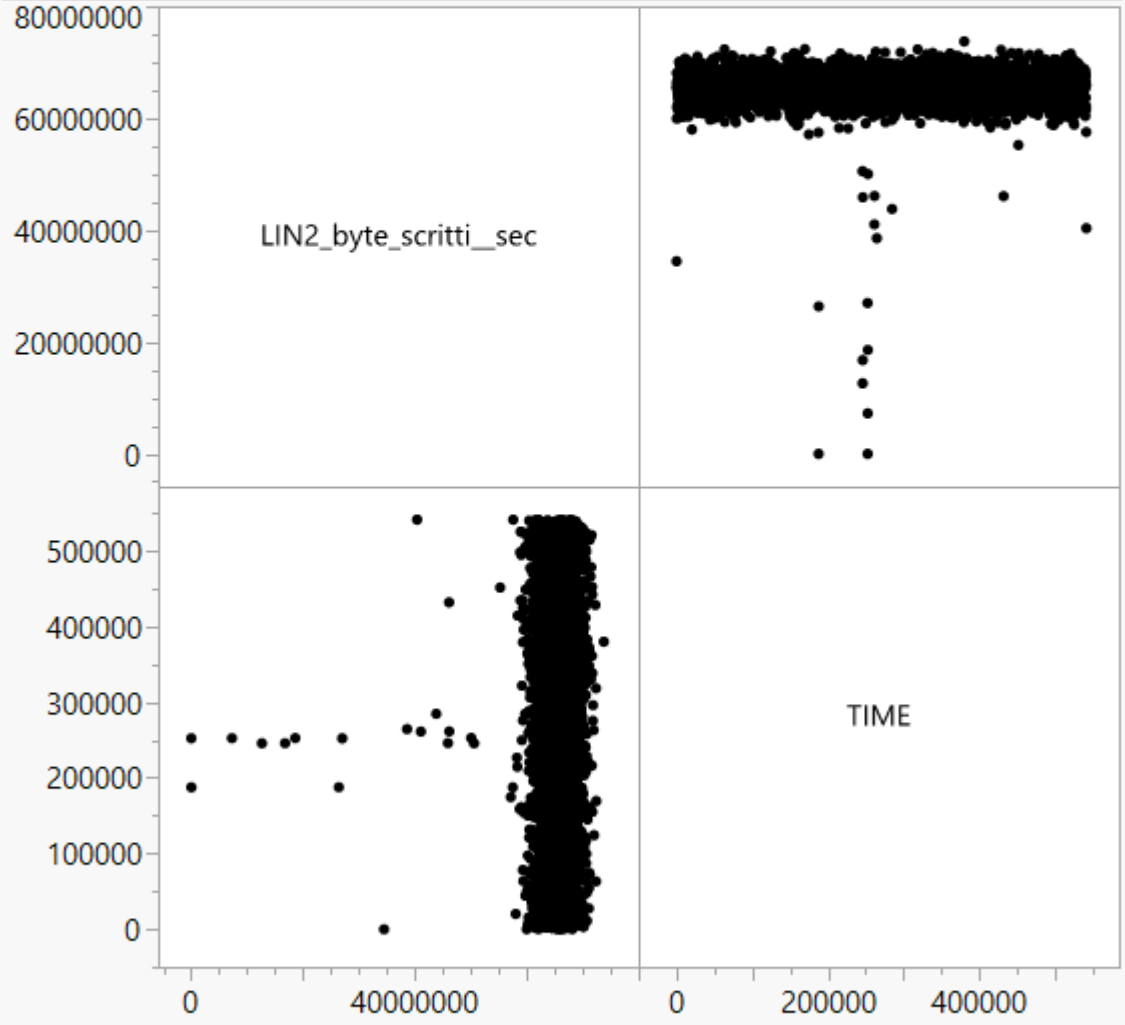
Multivariato

Correlazioni

	LIN2_byte_scritti__sec	TIME
LIN2_byte_scritti__sec	1,0000	-0,0154
TIME	-0,0154	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
TIME	LIN2_byte_scritti__sec	-0,0210	0,0342*	

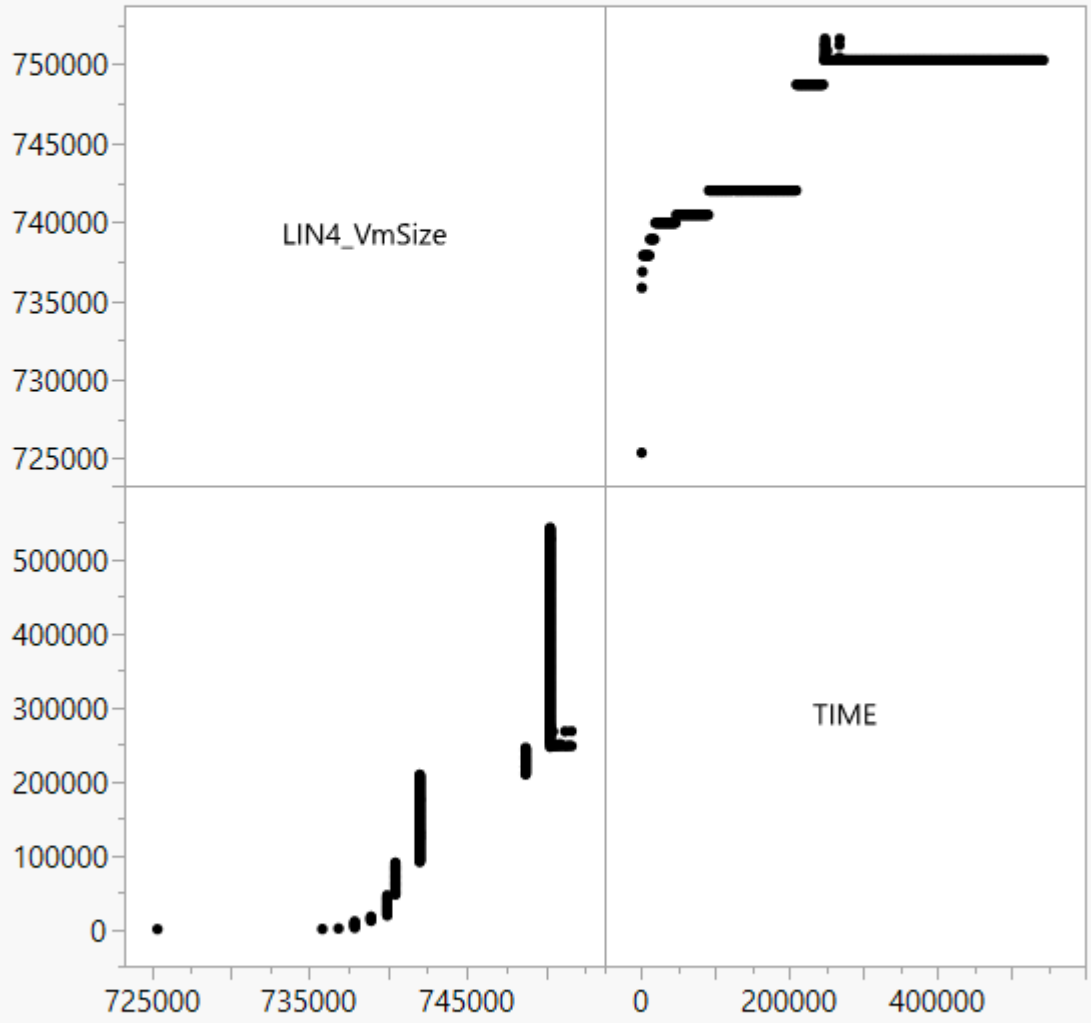
Multivariato

Correlazioni

	LIN4_VmSize	TIME
LIN4_VmSize	1,0000	0,8817
TIME	0,8817	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8-,6-,4-,2 0 ,2 ,4 ,6 ,8
TIME	LIN4_VmSize	0,7832	<,0001*	

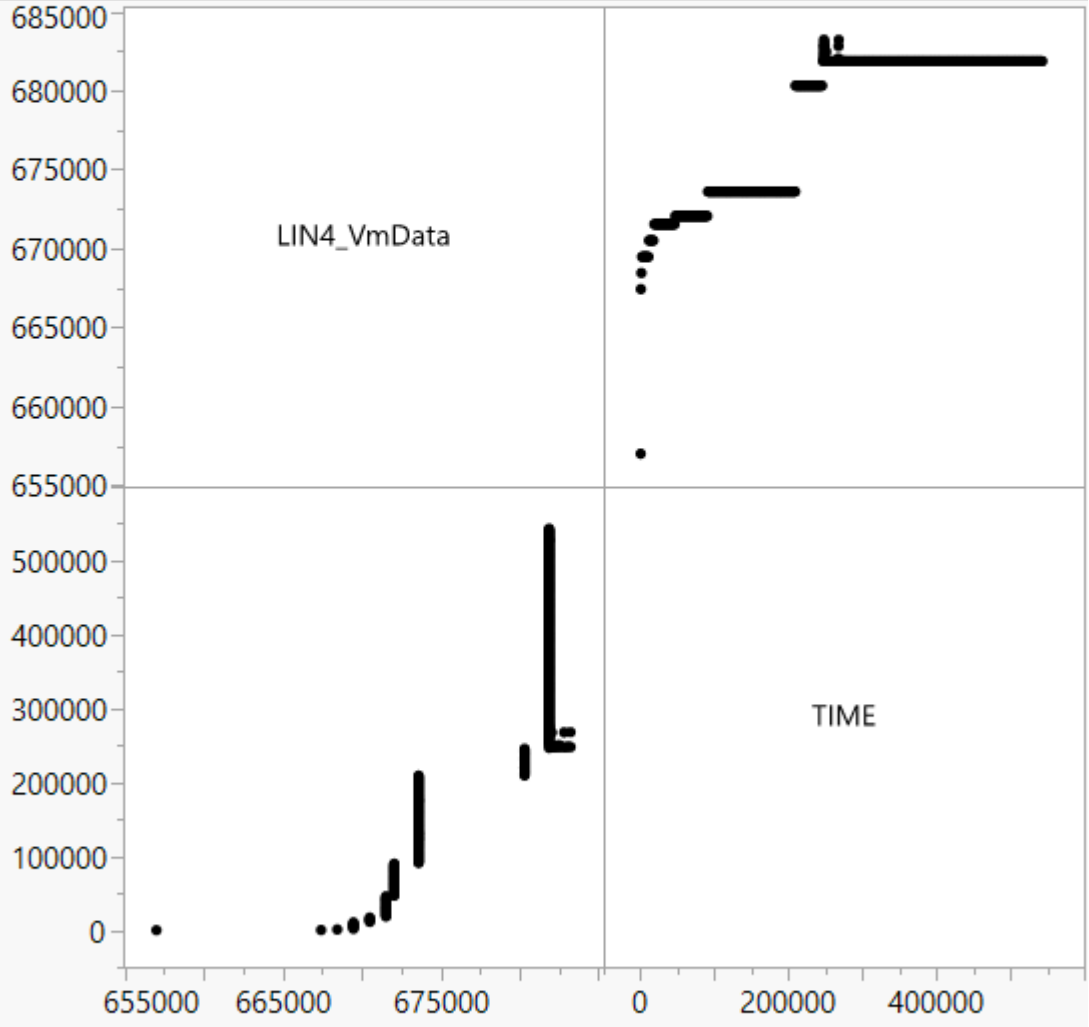
Multivariato

Correlazioni

	LIN4_VmData	TIME
LIN4_VmData	1,0000	0,8817
TIME	0,8817	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
TIME	LIN4_VmData	0,7832	<,0001*	

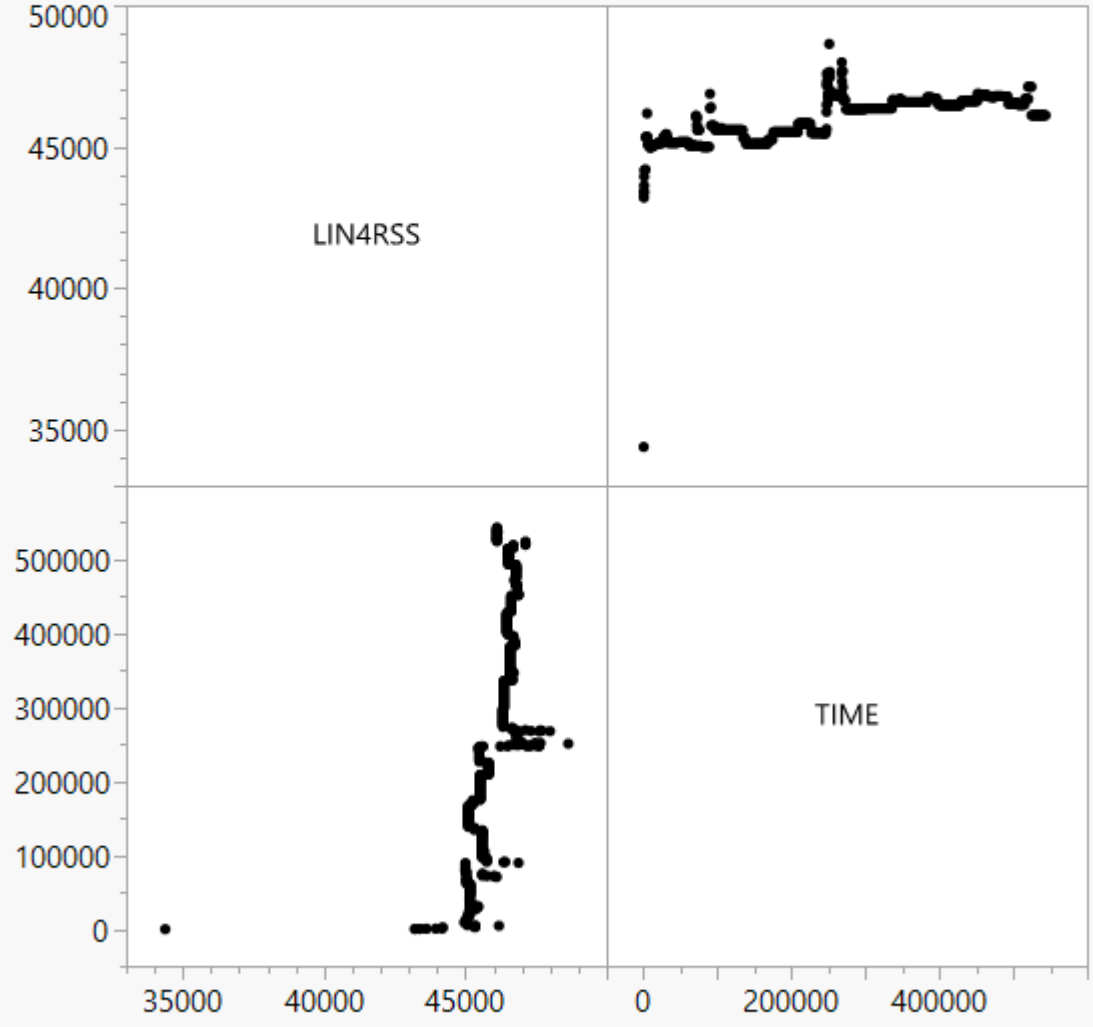
Multivariato

Correlazioni

	LIN4RSS	TIME
LIN4RSS	1,0000	0,7826
TIME	0,7826	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
TIME	LIN4RSS	0,5938	<,0001*	

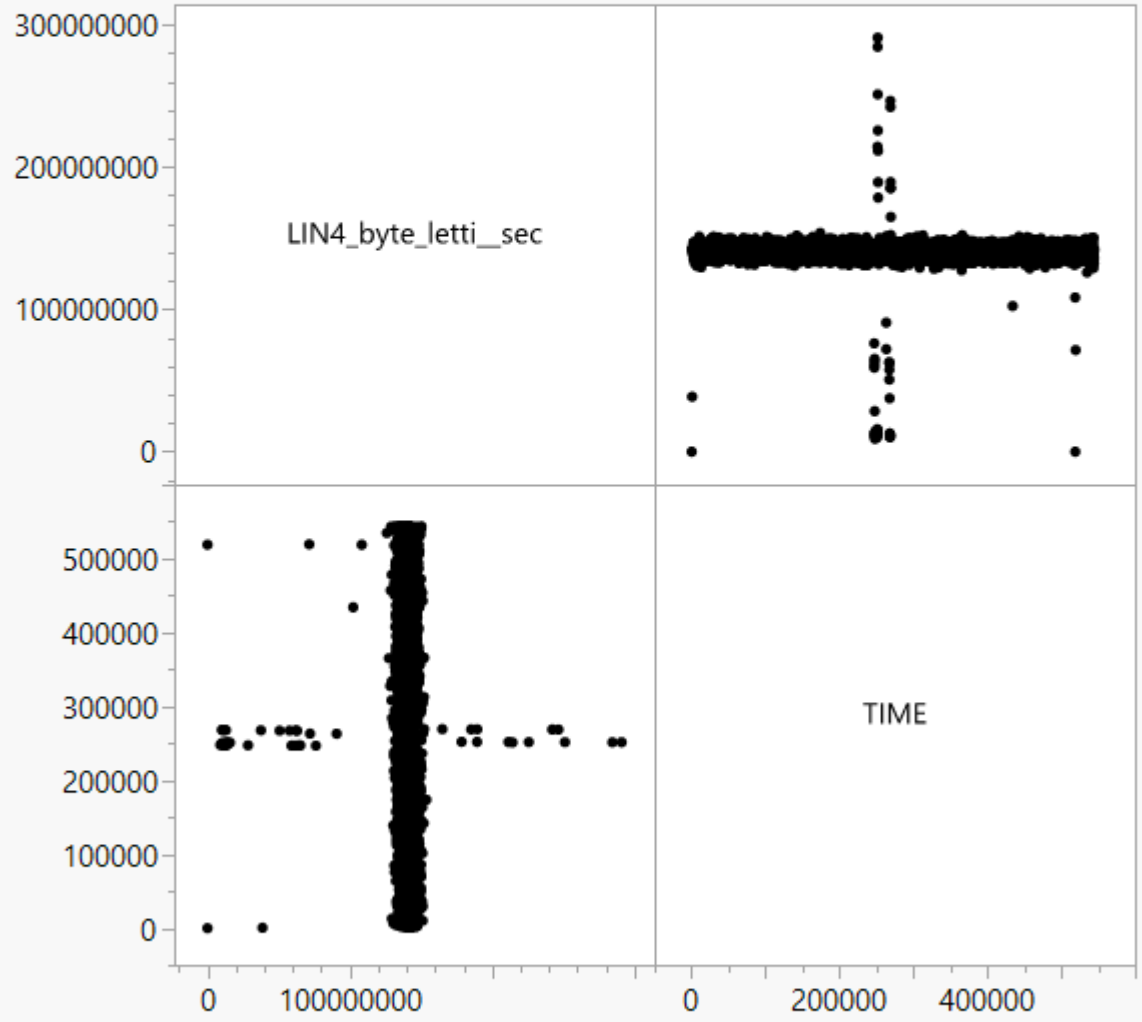
Multivariato

Correlazioni

	LIN4_byte_letti__sec	TIME
LIN4_byte_letti__sec	1,0000	-0,0071
TIME	-0,0071	1,0000

Vi sono 3 valori mancanti. Le correlazioni sono stimate per metodo REML.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
TIME	LIN4_byte_letti__sec	-0,0506	<,0001*	

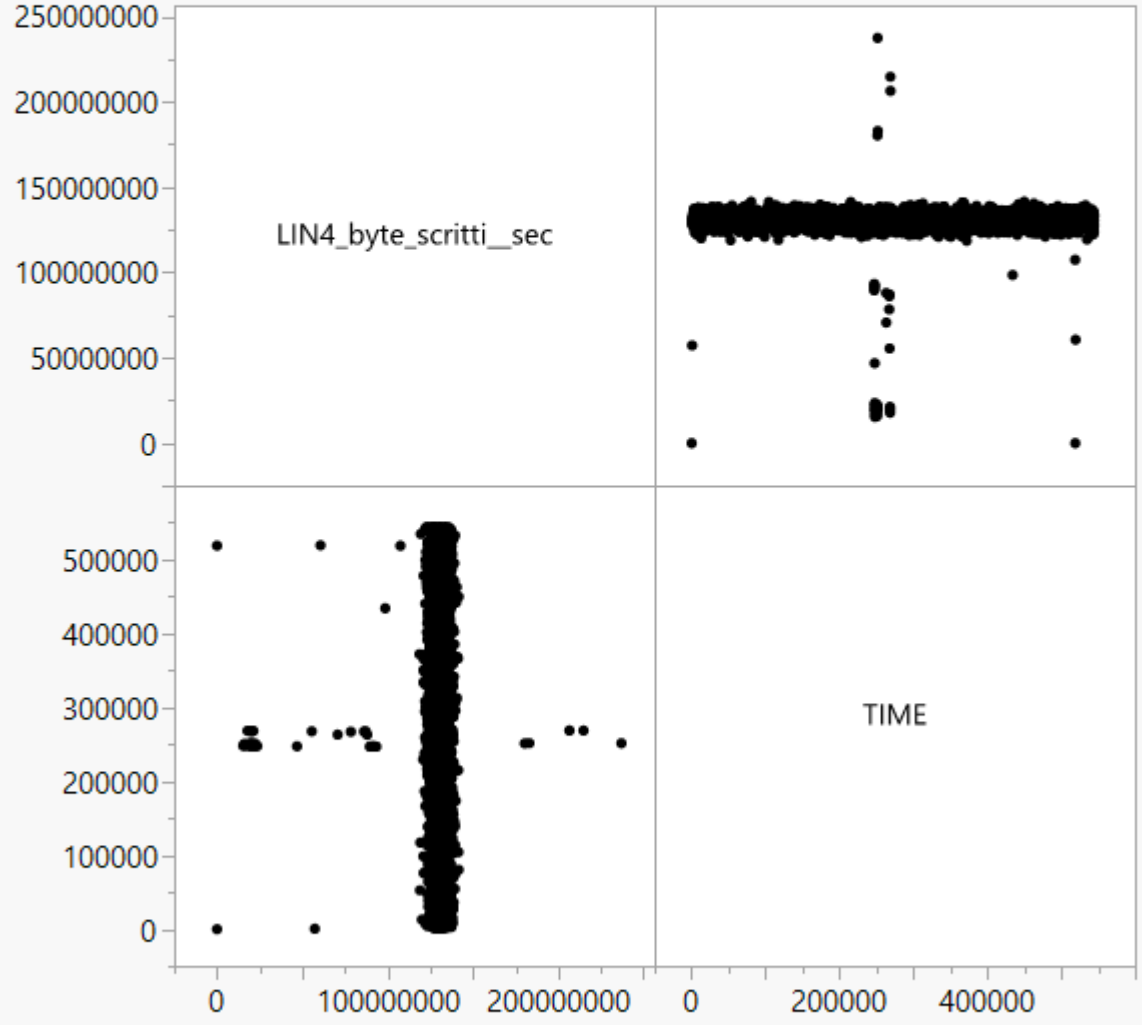
Multivariato

Correlazioni

	LIN4_byte_scritti__sec	TIME
LIN4_byte_scritti__sec	1,0000	0,0056
TIME	0,0056	1,0000

Vi sono 3 valori mancanti. Le correlazioni sono stimate per metodo REML.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
TIME	LIN4_byte_scritti__sec	-0,0171	0,0841	

Questo test è utilizzato per determinare se una serie temporale ha un trend monotonamente crescente o decrescente quando i dati non sono normali né lineari, a patto però che si abbia assenza di autocorrelazione. Questo metodo fornisce due tipi di informazione:

• Il Kendall Tau, che misura la monotonia dei dati. Il τ varia tra-1 e 1; un valore prossimo a ±1 indica una forte associazione tra le due variabili, che sarà positiva o negativa (correlazione o correlazione inversa) a seconda del segno, mentre un valore prossimo allo 0 indica assenza di relazione tra le variabili.

• La significatività, che rappresenta la soglia entro cui l'ipotesi di assenza di trend è accettata: il trend è statisticamente significativo se il p-value è minore di 0.05.

Solo in due casi è risultato che il trend non è statisticamente significativo, in particolare (OS1 – Byte scritti e OS3 – Byte scritti).

Una volta individuata la presenza di trend abbiamo utilizzato uno stimatore di Theil-Sen per stimare la pendenza della retta di regressione. Abbiamo implementato uno script python per fare questo e i risultati ottenuti sono i seguenti:

OS	Metric	Slope	Interval_Lower	Interval_Upper	Intercept
os1	LIN_VmSize	0.031	0.030	0.032	738854.3947841726
os1	LIN_VmData	0.031	0.030	0.032	670450.3947841726
os1	LIN_RSS	0.005	0.0049	0.0051	42428.80701754386
os1	LIN_byte_letti__sec	4.737	3.673	5.804	102763229.503125
os1	LIN_byte_scritti__sec	-0.030	-0.360	0.298	/
os2	LIN_VmSize	0.002	0.0	0.0021	735790.3761488455
os2	LIN_VmData	0.002	0.0	0.0021	667414.3761488455
os2	LIN_RSS	0.001	0.0005	0.0007	42384.85453267744
os2	LIN_byte_letti__sec	114.553	112.73	116.37	77733384.23784722
os2	LIN_byte_scritti__sec	-0.488	-0.939	-0.036	65693615.57629428
os3	LIN_VmSize	0.021	0.021	0.022	744374.599
os3	LIN_VmData	0.021	0.021	0.022	675970.599
os3	LIN_RSS	0.003	0.0033	0.0035	45386.471
os3	LIN_byte_letti__sec	-1.891	-2.618	-1.167	141291080.771
os3	LIN_byte_scritti__sec	-0.562	-1.198	0.076	/

In presenza delle variabili in cui non era stato individuato un trend, abbiamo a conferma di tale test che l'intervallo di confidenza del coefficiente angolare della retta di regressione comprende lo 0. In corrispondenza dei valori nan vuol dire che il valore non era rappresentabile dal calcolatore, in quanto molto alto.

I dataset analizzati sono relativi all'occupazione di memoria in una macchina virtuale: nel caso di os2 tutte le variabili presentano trend, mentre per os1 e os3 solo le prime 4. È possibile dunque confrontare i trend ottenuti nei diversi dataset rispetto alla stessa variabile, in particolare:

• Per VMSize e VMData si ha che in tutti e tre i casi la pendenza è positiva, anche se piuttosto bassa. In generale è possibile affermare che queste variabili crescano con il passare del tempo, per cui si nota un utilizzo crescente della memoria in generale.

• Analogo è il caso per VMRSS, per cui è possibile fare le stesse considerazioni e dedurre che si ha un trend crescente anche per l'utilizzo dello spazio di archiviazione.

• Per quanto riguarda Byte_letti si nota che il trend cambia a seconda del dataset: si osserva infatti un trend positivo per os2 (con pendenza anche piuttosto elevata) e negativo invece per os3; impossibile

invece fare confronti sulla variabile Byte_scritti_sec in quanto nel caso di os2 non presenta trend statisticamente significativi, e non è possibile dunque individuarne pendenza e intercetta.

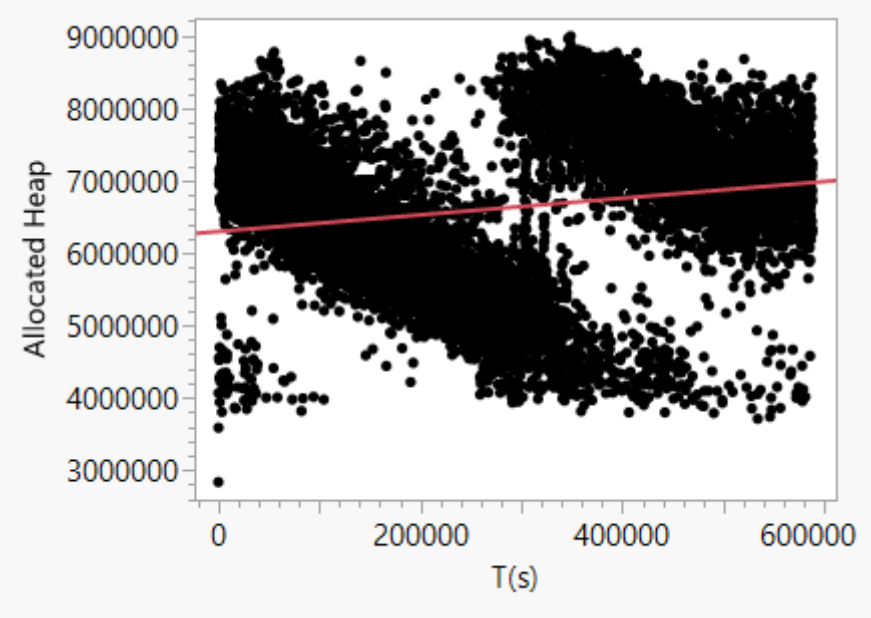
4.2 VMRes

Supponendo di avere un limite massimo alla memoria heap di 1 GByte. Rilevare un eventuale trend di consumo dello heap nell'esperimento in figura. Se rilevato il trend, stimare il tempo in cui lo heap satura (failure prediction).

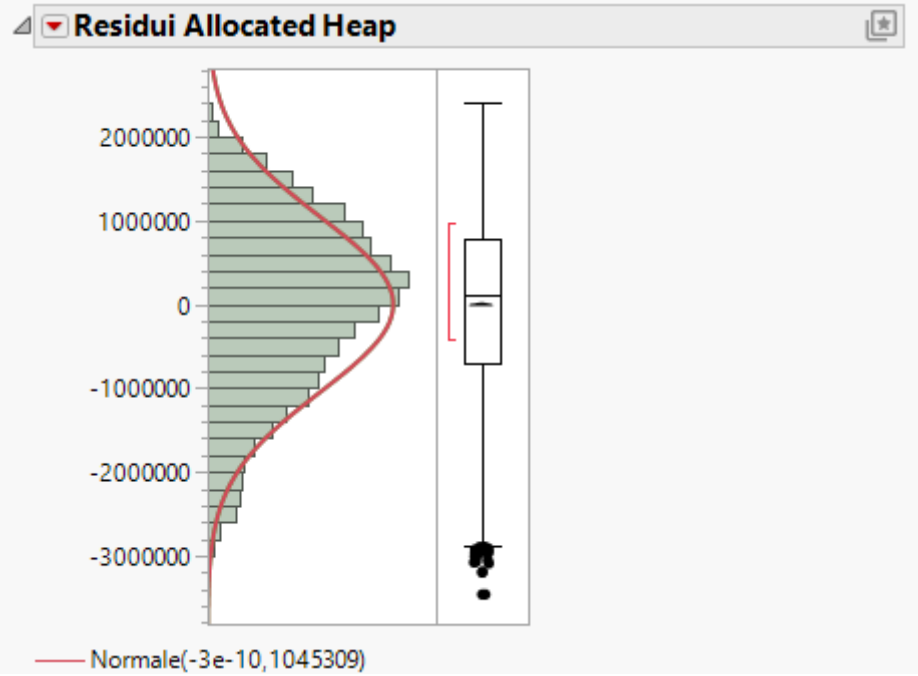
Ci sono stati forniti due workload per questo homework, entrambi con lo stesso obiettivo.

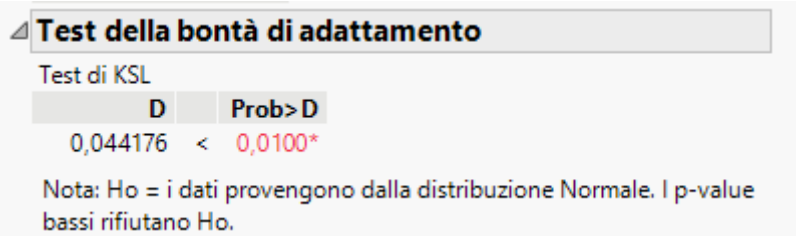
Per quanto riguarda VMRes1:

Prima di tutto abbiamo fatto la stima lineare dello heap allocato

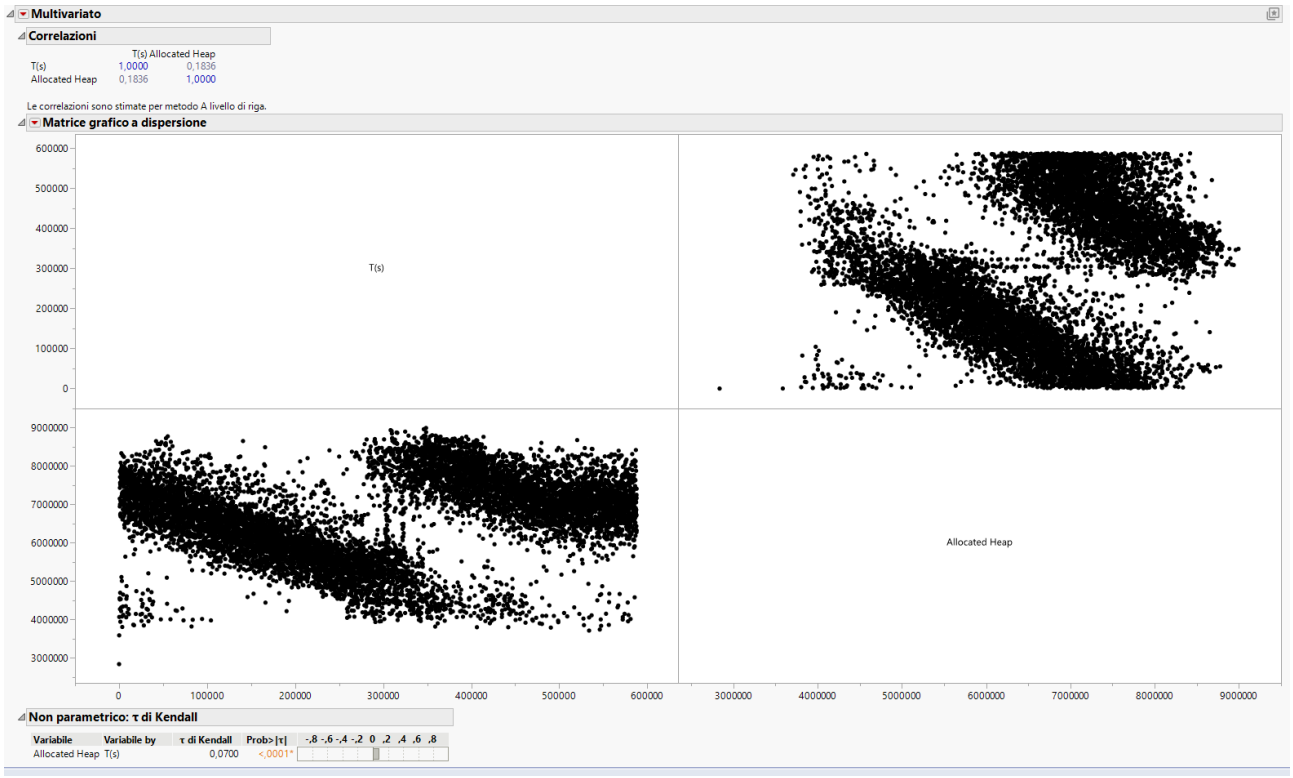


Calcolando successivamente i residui abbiamo poi analizzato la loro distribuzione, con il test di Shapiro-Wilk:





Il test ha rigettato l'ipotesi nulla ritenendo i residui con una distribuzione non normale. A valle del risultato di tale test è stato condotto il test non parametrico di Kendall per individuare la presenza di trend:



Il test non rigetta l'ipotesi di un trend e quindi possiamo stimare la retta di regressione con Theil-Sen:

Slope	Interval_Lower	Interval_Upper	Intercept
0.6911224682945296	0.563777264562484	0.8178240089963452	6562275.461669506

Secondo questa retta siamo poi riusciti a stimare il tempo (s) per saturare 1GB di memoria heap. Abbiamo fissato la dimensione sulle y e abbiamo visto entro quanto tempo si raggiungeva; l'output è il seguente:

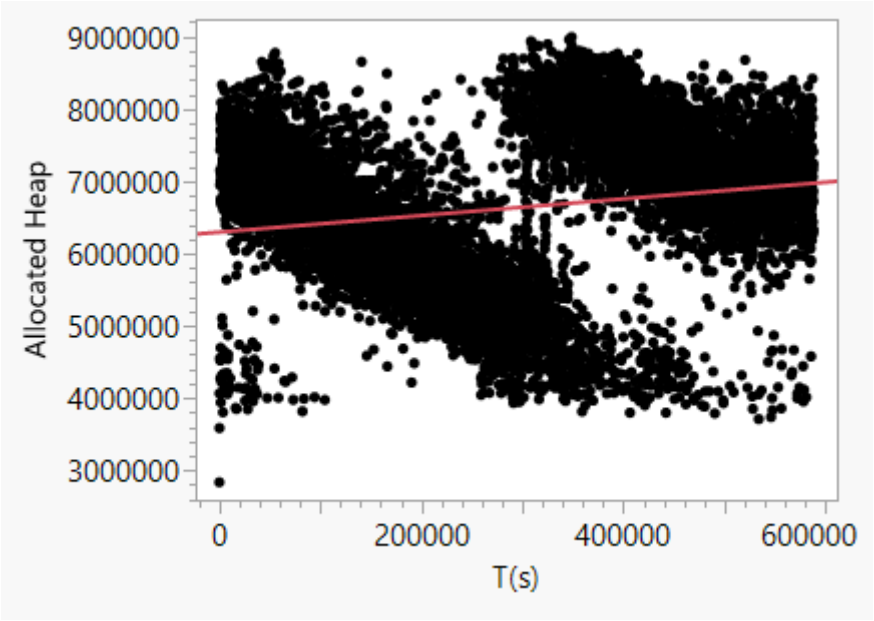
$$Allocated\ Heap = m + bT(s) = 6562275.46 + (0.69 \pm 0.13)\mathrm{T}(s)$$

Si pone Allocated Heap = 1 GB = 1073741824 e si calcola la formula inversa per trovare il corrispettivo $T(s)$:

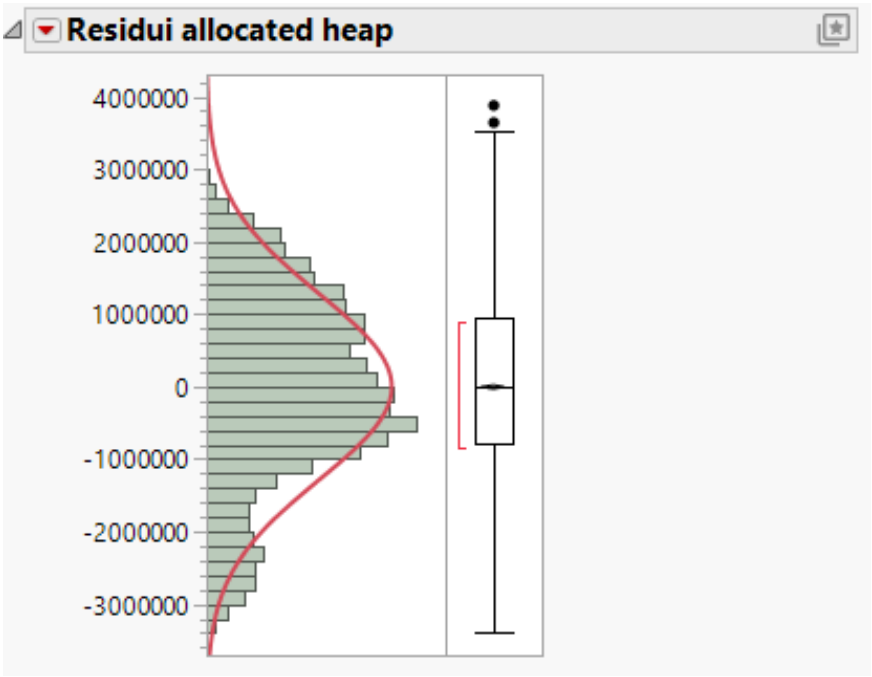
$$T_{failure}(s) = \frac{1073741824 - 6562275.46}{(0.69 \pm 0.13)} = (1546637027 \pm 277369603)s = \ (49 \ \pm \ 8.8)anni$$

Per quanto riguarda VMRes2:

Prima di tutto abbiamo fatto la stima lineare dello heap allocato



Calcolando successivamente i residui abbiamo poi analizzato la loro distribuzione, con il test di Shapiro-Wilk:



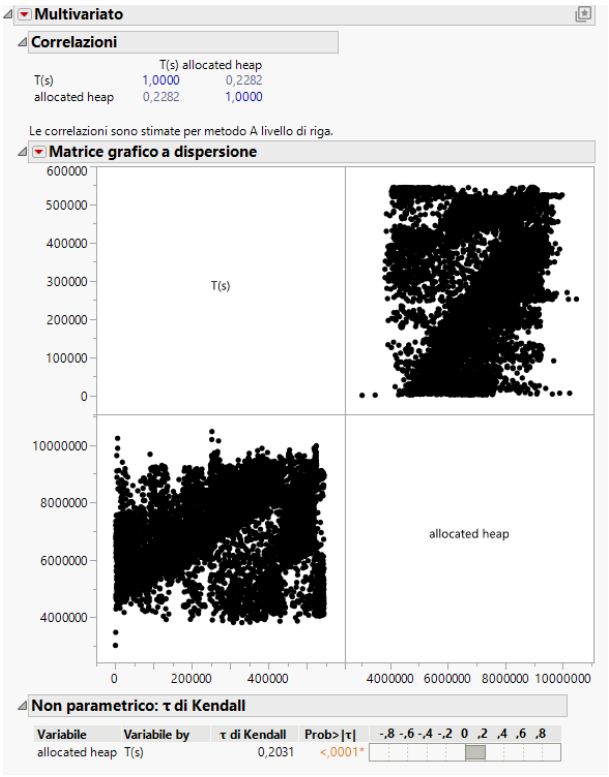
Test della bontà di adattamento

Test di KSL

D		Prob>D
0,030198	<	0,0100*

Nota: Ho = i dati provengono dalla distribuzione Normale. I p-value bassi rifiutano Ho.

Il test ha rigettato l'ipotesi nulla ritenendo i residui con una distribuzione non normale. A valle del risultato di tale test è stato condotto il test non parametrico di Kendall per individuare la presenza di trend:



Il test non rigetta l'ipotesi di un trend e quindi possiamo stimare la retta di regressione con Theil-Sen:

Slope	Interval_Lower	Interval_Upper	Intercept
2.9030932760062105	2.714759535655058	3.0923529411764705	6041585.556431387

Secondo questa retta siamo poi riusciti a stimare il tempo (s) per saturare 1GB di memoria heap. Abbiamo fissato la dimensione sulle y e abbiamo visto entro quanto tempo si raggiungeva; l'output è il seguente:

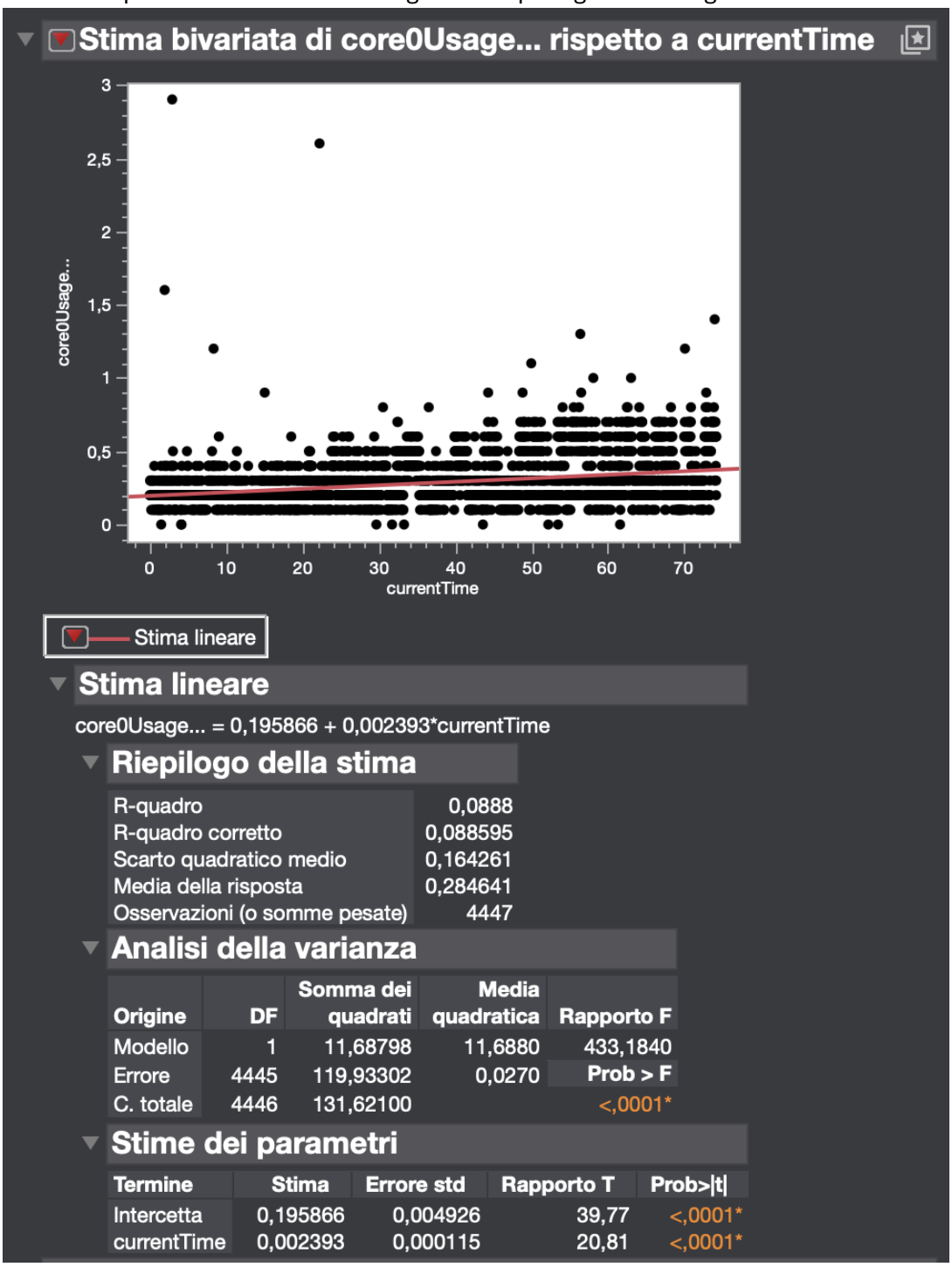
$$T_{failure}(s) = \frac{1073741824 - 6041585.55}{(2.9 \pm 0.19)} = (368172496 \pm 24225634)s \; = \; (11.7 \; \pm \; 0.8)anni$$

VMRes1 ci mette di più a saturare 1GB di memoria heap. Infatti la differenza di circa **38 anni** indica che **VMRes1 è molto più lento rispetto a VMRes2** per saturare 1GB di memoria heap.

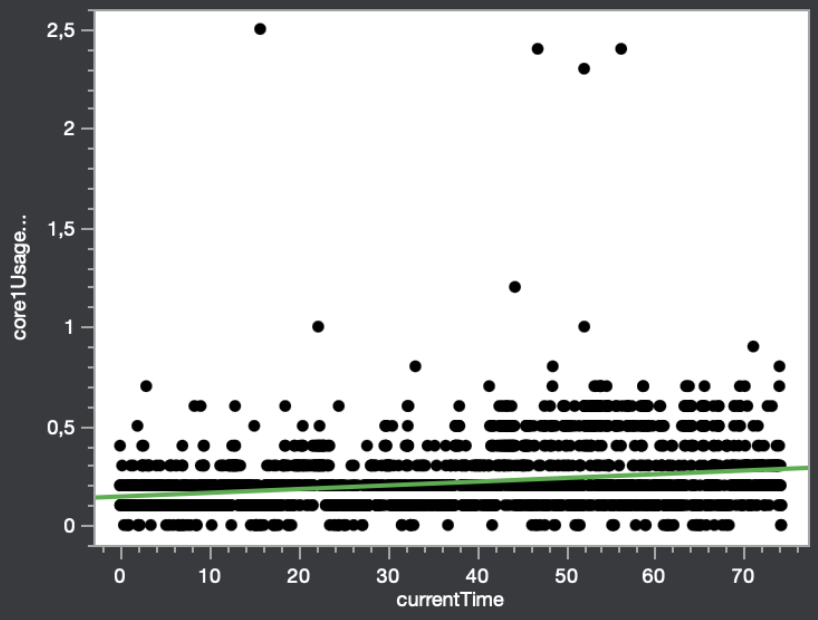
4.3 Cloud

Il seguente Workload di basso livello è relativo all'addestramento di un algoritmo per il riconoscimento dei pedoni nell'ambito di "Autonomous driving".

1) Le CPU sono state utilizzate unformemente (efficienza dello scheduling)?
   Tracciamo prima di tutto le rette di regressione per ogni coreUsage:



Stima bivariata di core1Usage... rispetto a currentTime



Stima lineare

Stima lineare

core1Usage... = 0,1437895 + 0,001871*currentTime

Riepilogo della stima

R-quadro	0,069305
R-quadro corretto	0,069096
Scarto quadratico medio	0,146922
Media della risposta	0,2132
Osservazioni (o somme pesate)	4447

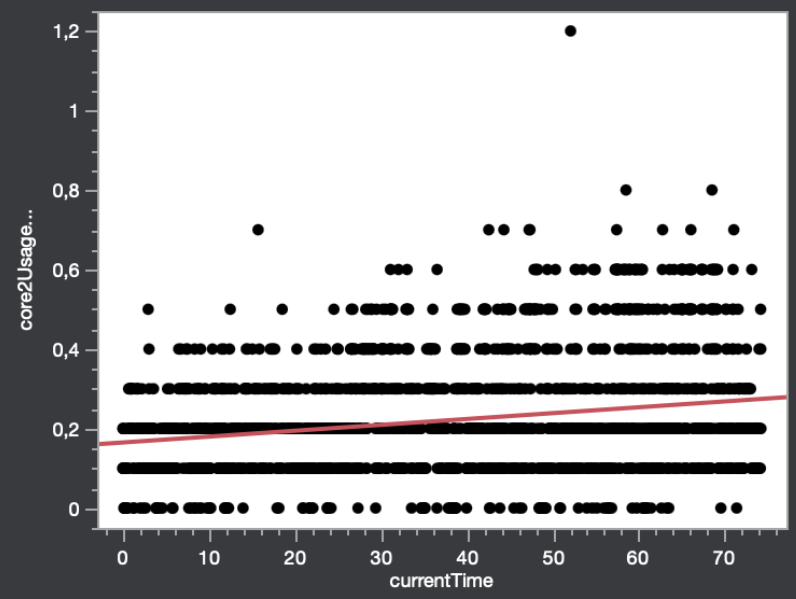
Analisi della varianza

Origine	DF	Somma dei quadrati	Media quadratica	Rapporto F
Modello	1	7,14502	7,14502	331,0012
Errore	4445	95,95015	0,02159	Prob > F
C. totale	4446	103,09517		<,0001*

Stime dei parametri

Termine	Stima	Errore std	Rapporto T	Prob>|t|
Intercetta	0,1437895	0,004406	32,64	<,0001*
currentTime	0,001871	0,000103	18,19	<,0001*

Stima bivariata di core2Usage... rispetto a currentTime



Stima lineare

Stima lineare

core2Usage... = 0,1649518 + 0,0014758*currentTime

Riepilogo della stima

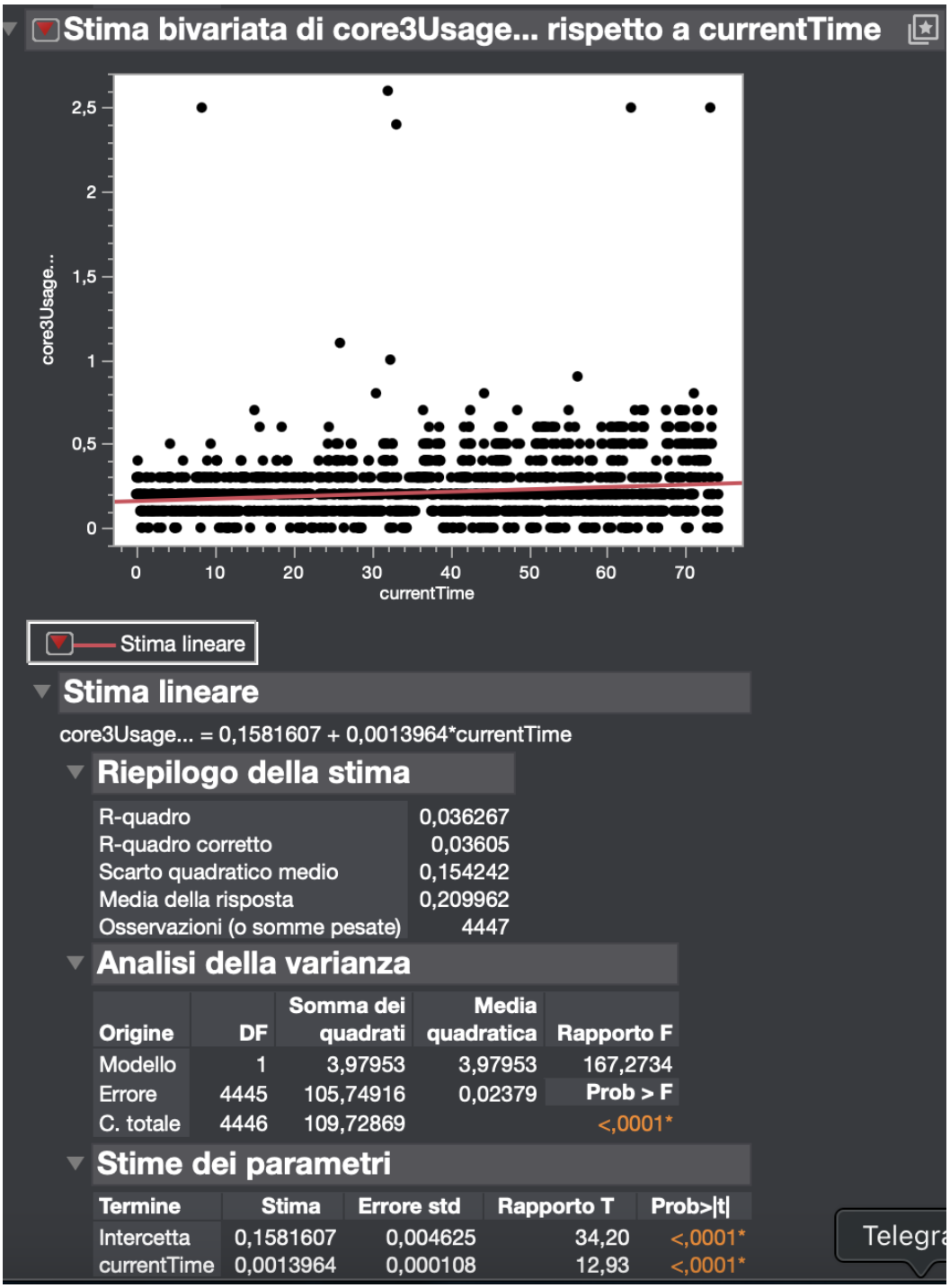
R-quadro	0,056071
R-quadro corretto	0,055859
Scarto quadratico medio	0,129748
Media della risposta	0,219699
Osservazioni (o somme pesate)	4447

Analisi della varianza

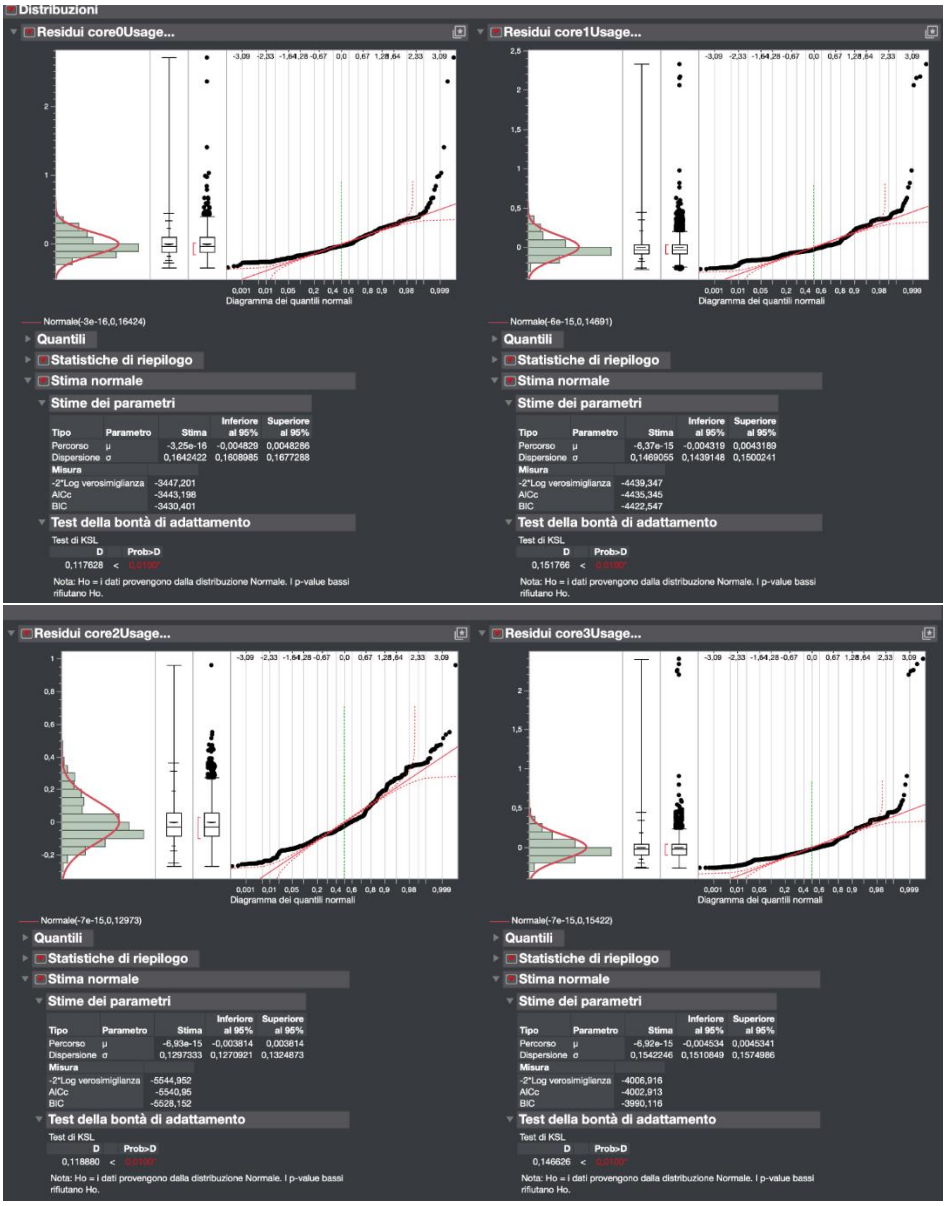
Origine	DF	Somma dei quadrati	Media quadratica	Rapporto F
Modello	1	4,445018	4,44502	264,0421
Errore	4445	74,829378	0,01683	Prob > F
C. totale	4446	79,274396		<,0001*

Stime dei parametri

Termine	Stima	Errore std	Rapporto T	Prob>|t|
Intercetta	0,1649518	0,003891	42,40	<,0001*
currentTime	0,0014758	9,082e-5	16,25	<,0001*



Poi andiamo a verificare la normalità dei residui:



L'ipotesi nulla del test di Shapiro-Wilk viene rigettata; quindi, possiamo concludere che le distribuzioni dei residui non sono normali. Non serve quindi andare a verificare l'omoschedasticità. Andremo direttamente ad effettuare il test di Kendall per analizzare l'eventuale presenza di una tendenza:

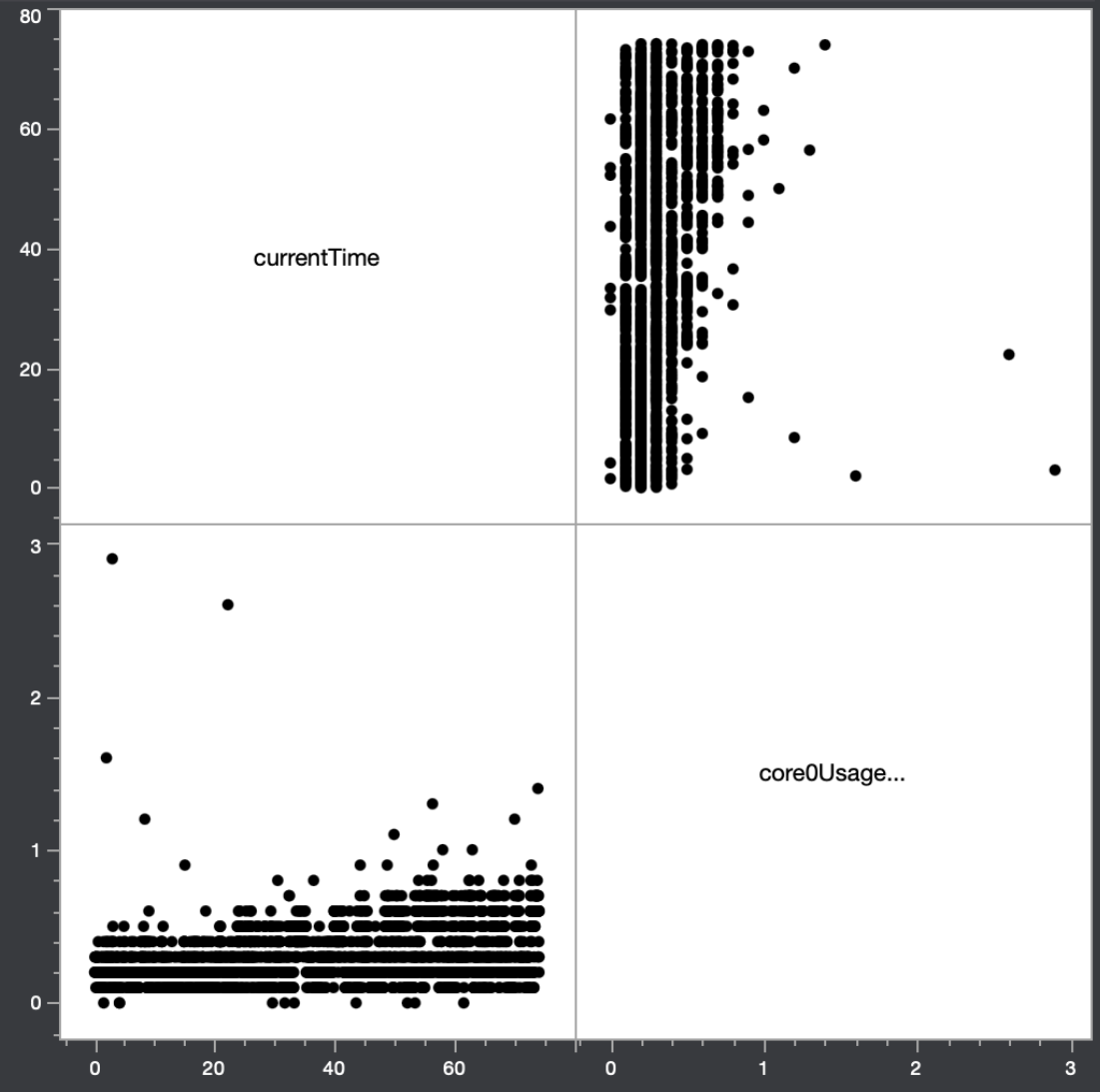
Multivariato

Correlazioni

	currentTime	core0Usage...
currentTime	1,0000	0,2980
core0Usage...	0,2980	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
core0Usage...	currentTime	0,2000	<,0001*	

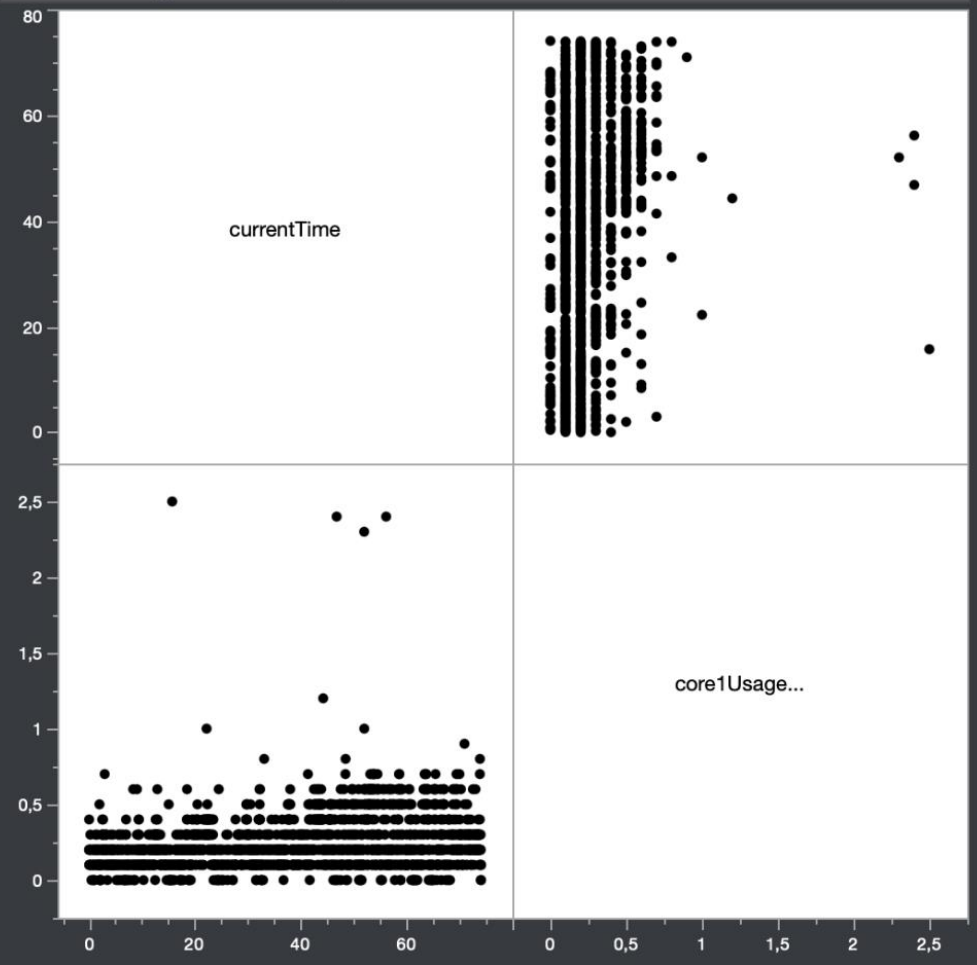
Multivariato

Correlazioni

	currentTime	core1Usage...
currentTime	1,0000	0,2633
core1Usage...	0,2633	1,0000

Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
core1Usage...	currentTime	0,2114	<,0001*	

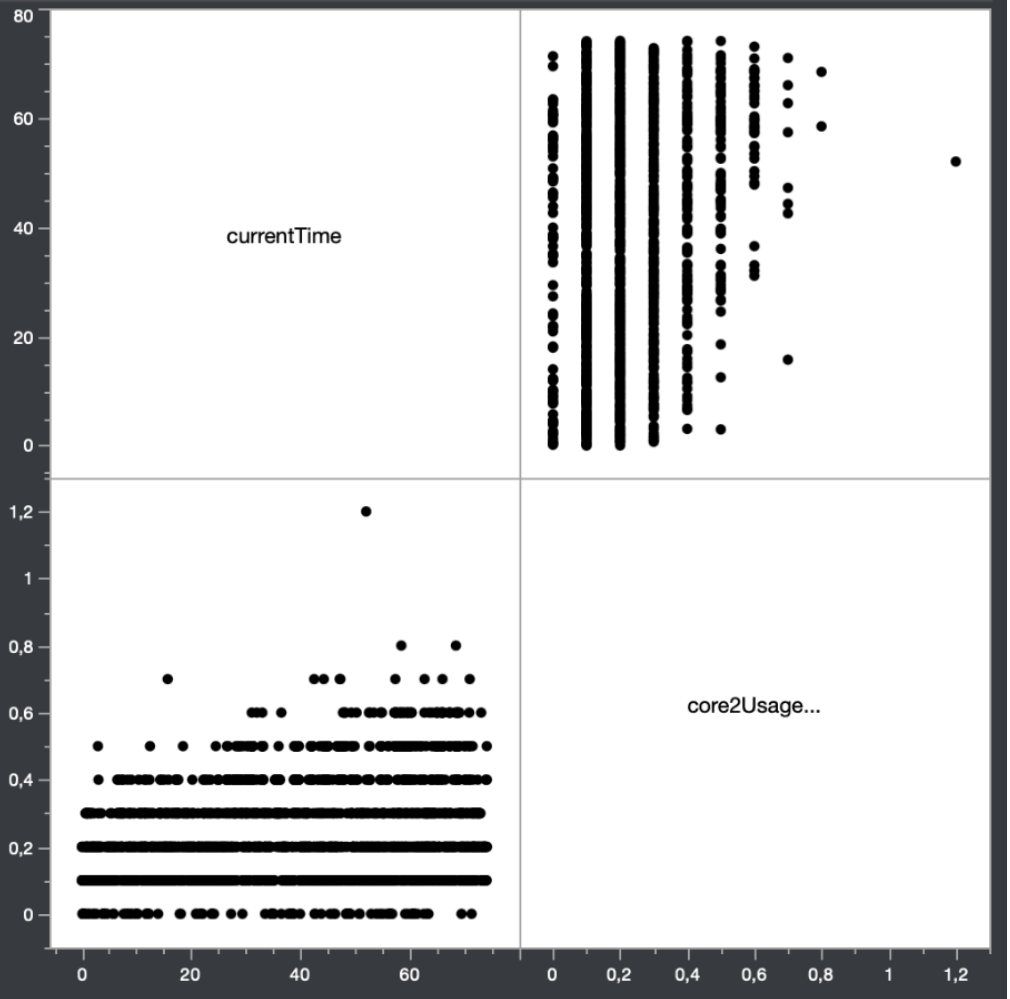
Multivariato

Correlazioni

	currentTime	core2Usage...
currentTime	1,0000	0,2368
core2Usage...	0,2368	1,0000

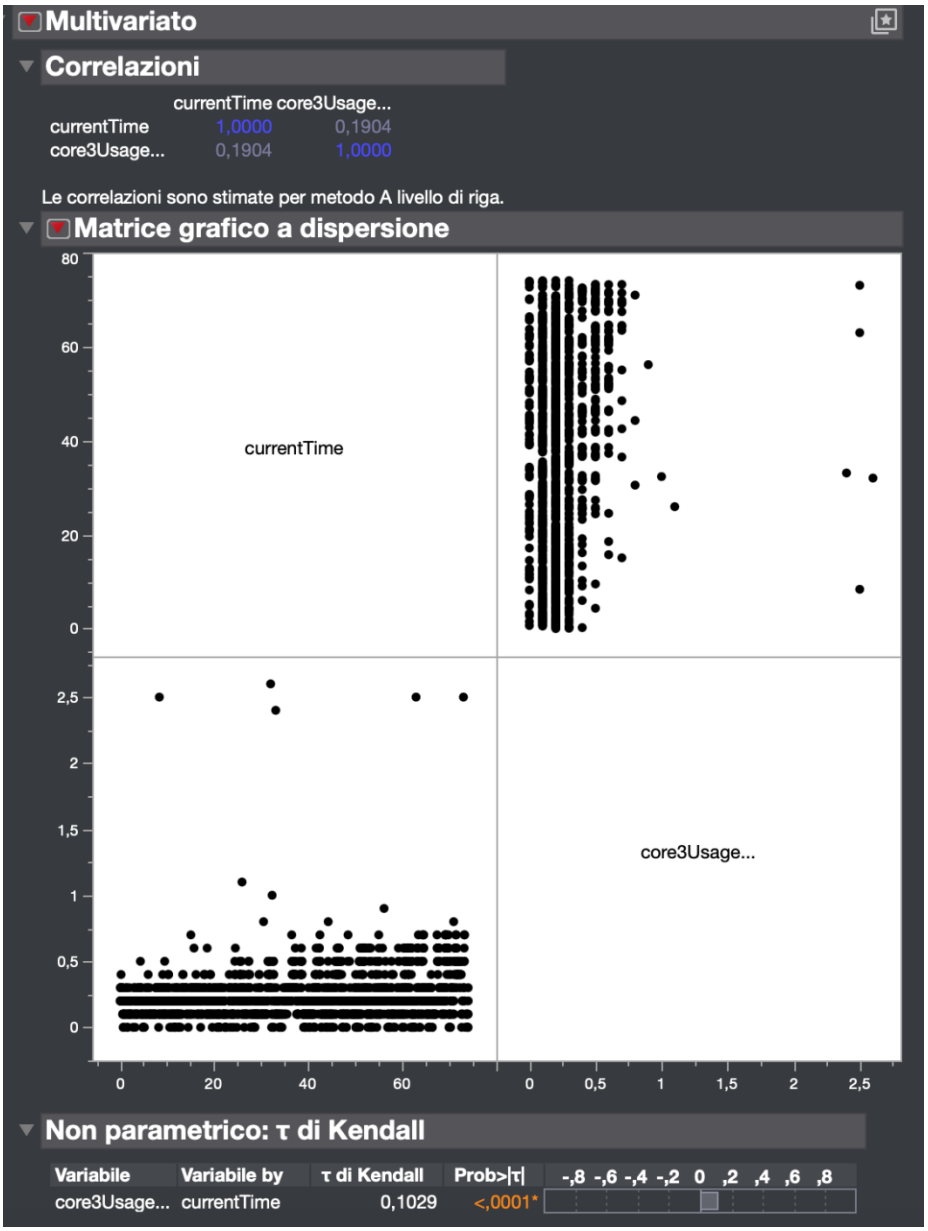
Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
core2Usage...	currentTime	0,1539	<,0001*	



Il test di Kendall rigetta l'ipotesi nulla e ci segnala la presenza di un trend in tutti e 4 i core. Per verificare l'efficienza dello scheduling andiamo a stimare i parametri di regressione con gli stimatori di Theil-Sen:

Metric	Slope	Interval_Lower	Interval_Upper	Intercept
Core0Usage	0.0	0.0	0.0	0.2
Core1Usage	0.0	0.0	0.0	0.2
Core2Usage	0.0	0.0	0.0	0.2
Core3Usage	0.0	0.0	0.0	0.2

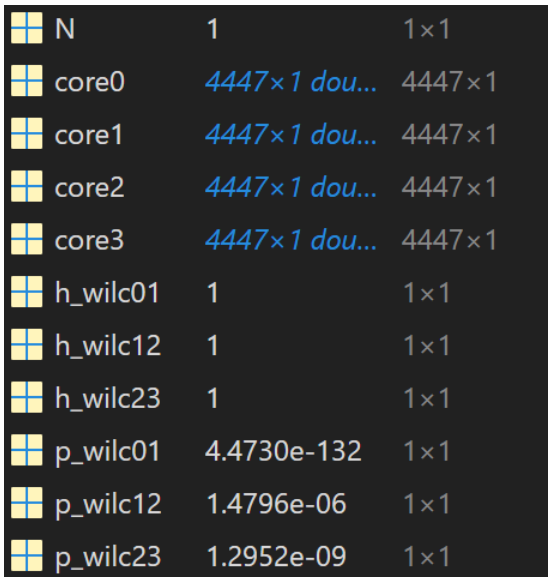
Ci accorgiamo che lo slope della retta è 0, questo è strano in quanto contraddice il test di Man-Kendall eseguito su JMP. Da un test visivo si nota che la pendenza delle rette di regressione è molto bassa quindi questo potrebbe essere dovuto ad un errore nella procedura di Theil-Sen quando lavora con valori molto piccoli compresi fra 0 ed 1.
Motivo per cui abbiamo utilizzato un ulteriore script Python sfruttando la libreria *sklearn*, in particolare la libreria *linear_model*. In questo modo abbiamo ripetuto le procedure della regressione ai minimi quadrati, specificando un livello di confidenza del 95%. Di seguito i risultati:

Metric	Slope	Interval_Lower	Interval_Upper
Core0Usage	0.0239	0.0216	0.0261
Core1Usage	0.0187	0.0166	0.0207
Core2Usage	0.0147	0.0129	0.0165
Core3Usage	0.0139	0.0118	0.0160

Viene quindi rilevato un trend positivo nell'utilizzo di tutti e 4 i core. Dal valore delle pendenze delle rette di regressione questo non sembra essere un trend rilevante per nessuno dei 4 core. Possiamo verificarlo andando a testare l'ipotesi che le 4 distribuzioni dell'utilizzo dei core non sono significativamente diverse quindi testando a coppie le distribuzioni di:

- core0 e core1
- core1 e core2
- core2 e core3

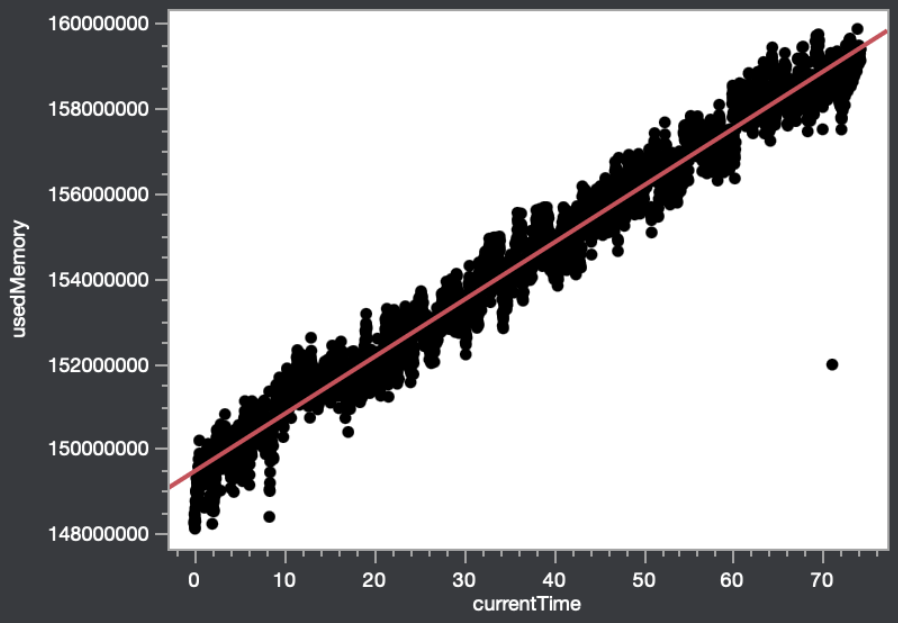
Non possiamo assumere che gli esperimenti siano IID, motivo per cui non è possibile applicare il Teorema del Limite Centrale nonostante la grande mole di dati in nostro possesso. Abbiamo quindi optato per un test non parametrico di Wilcoxon utilizzando Matlab proprio come fatto nel capitolo 3 di questa tesina. Di seguito presentiamo i risultati:



Vediamo quindi che le distribuzioni dei dati sono significativamente diverse. Per quanto piccoli e confrontabili siano gli slope possiamo però concludere che core0 sembra essere il core più utilizzato e quindi possiamo concludere che lo scheduling non risulta essere efficiente.

2) C'è un trend positivo nel consumo di memoria?
Per verificare la presenza di un trend nel consumo di memoria i passaggi sono analoghi a quelli precedenti:

Stima bivariata di usedMemory rispetto a currentTime



Stima lineare

Stima lineare

usedMemory = 149460531 + 134281,64*currentTime

Riepilogo della stima

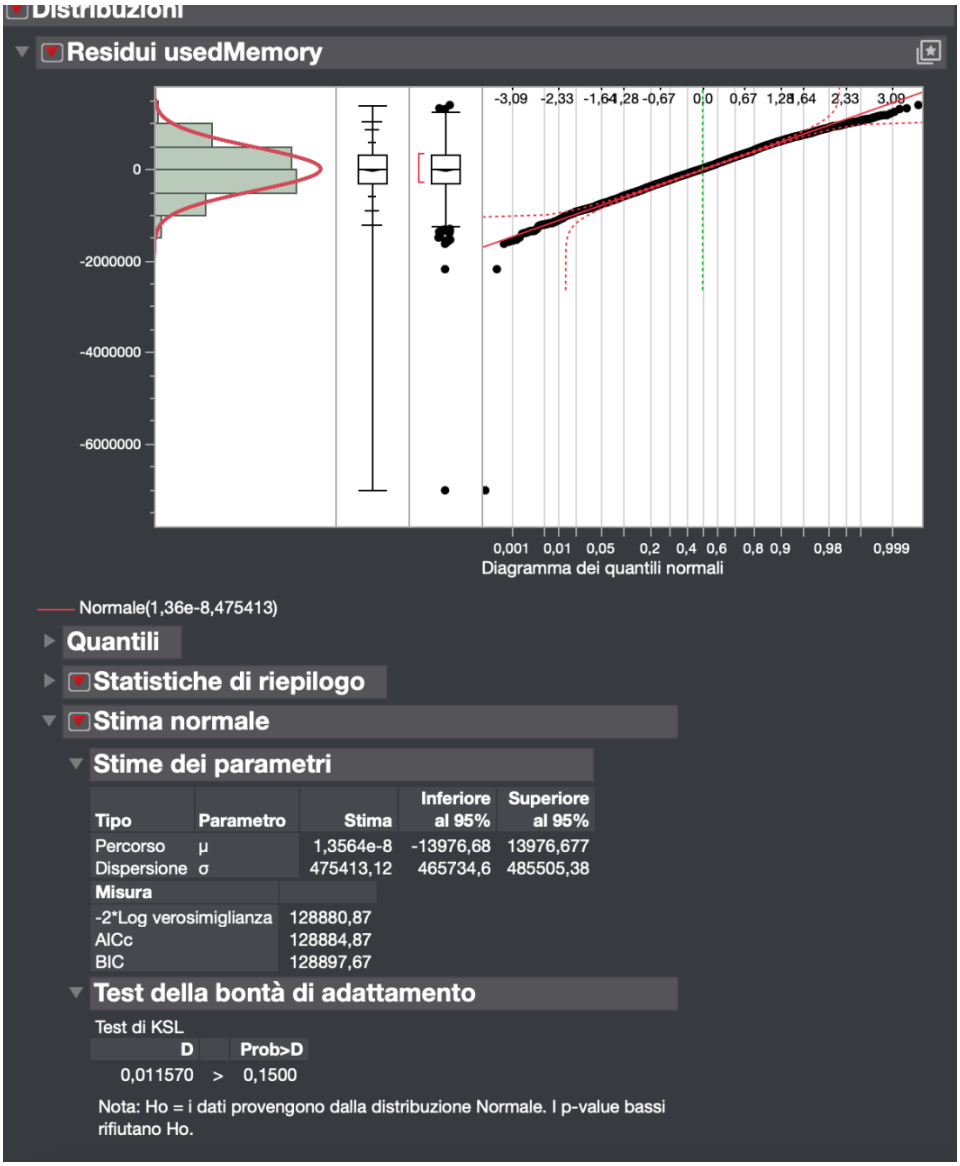
R-quadro	0,973421
R-quadro corretto	0,973415
Scarto quadratico medio	475466,6
Media della risposta	1,544e+8
Osservazioni (o somme pesate)	4447

Analisi della varianza

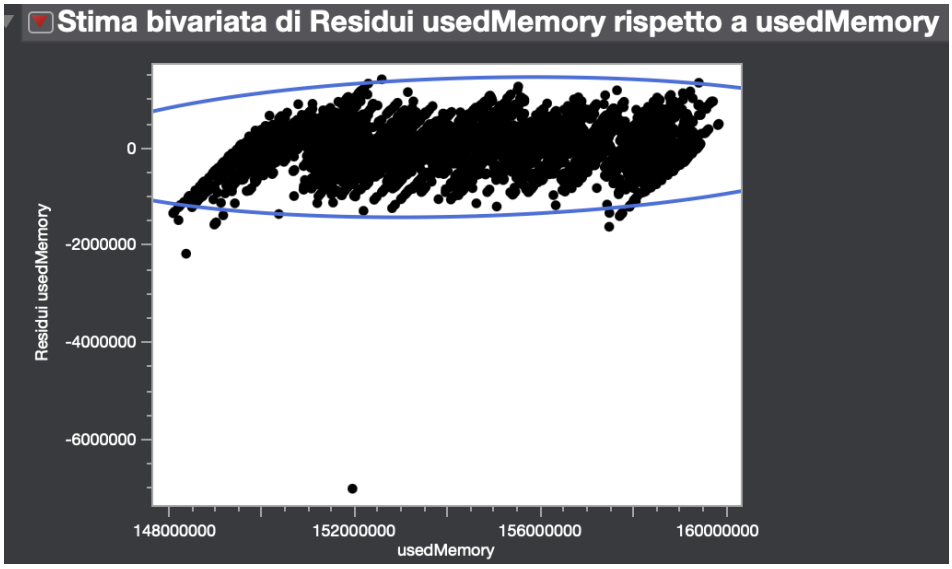
Origine	DF	Somma dei quadrati	Media quadratica	Rapporto F
Modello	1	3,6802e+16	3,68e+16	162791,5
Errore	4445	1,0049e+15	2,261e+11	Prob > F
C. totale	4446	3,7807e+16		<,0001*

Stime dei parametri

Termine	Stima	Errore std	Rapporto T	Prob>|t|
Intercetta	149460531	14257,33	10483	<,0001*
currentTime	134281,64	332,8134	403,47	<,0001*



La distribuzione dei residui è normale e andiamo quindi a verificare l'omoschedasticità:



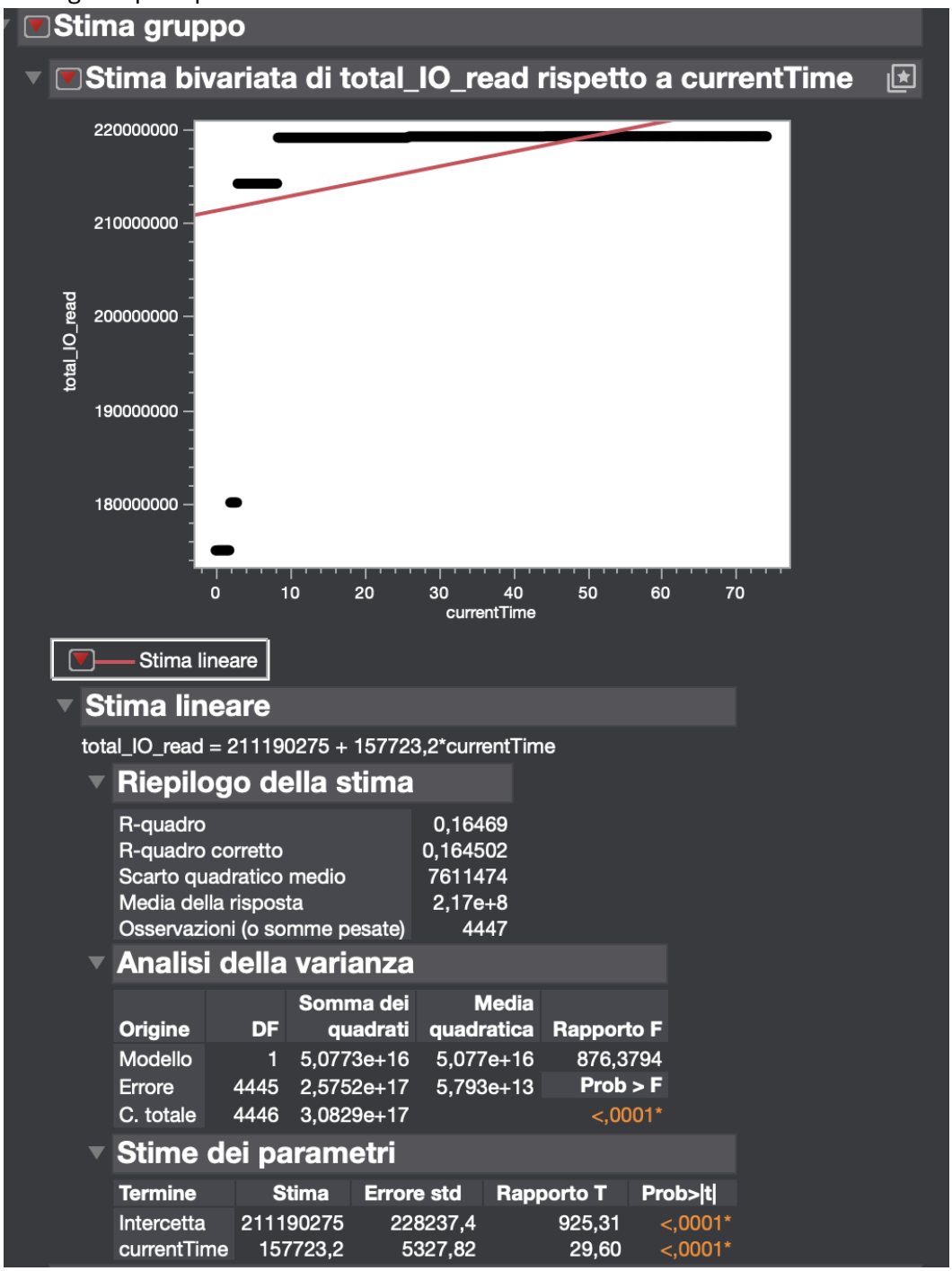
Sono compresi tutti entro una certa fascia e possiamo concludere che sono omoschedastici.

La presenza di trend l'abbiamo verificata attraverso un test visivo della retta di regressione, che dà evidenza di un trend positivo, di cui abbiamo trovato i parametri attraverso Theil-Sen:

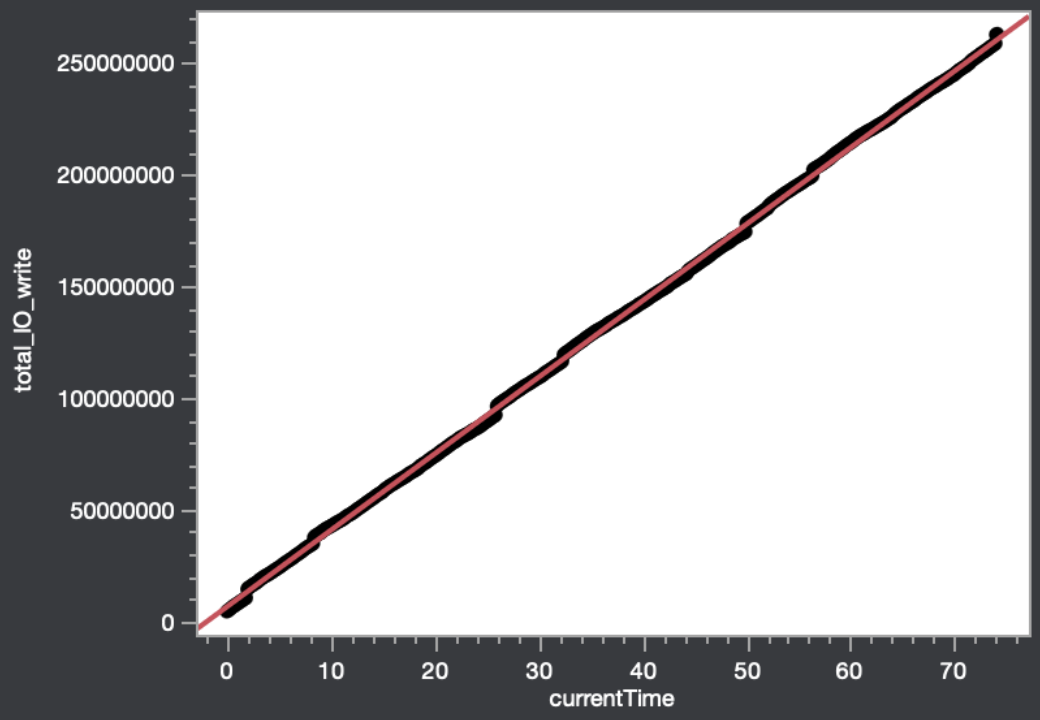
Metric	Slope	Interval_Lower	Interval_Upper	Intercept
usedMemory	134223.19452028826	133560.06322584432	134886.37241686913	149534143.56207412

Come possiamo notare il coefficiente angolare della retta di regressione è positivo, confermando ciò che abbiamo detto in precedenza.

3) C'è un trend sull'operazioni di I/O di read e write?
Per verificare la presenza di un trend nel nelle operazioni I/O di lettura e scrittura i passaggi sono analoghi a quelli precedenti



Stima bivariata di total_IO_write rispetto a currentTime



Stima lineare

Stima lineare

total_IO_write = 6346804,5 + 3430493,6*currentTime

Riepilogo della stima

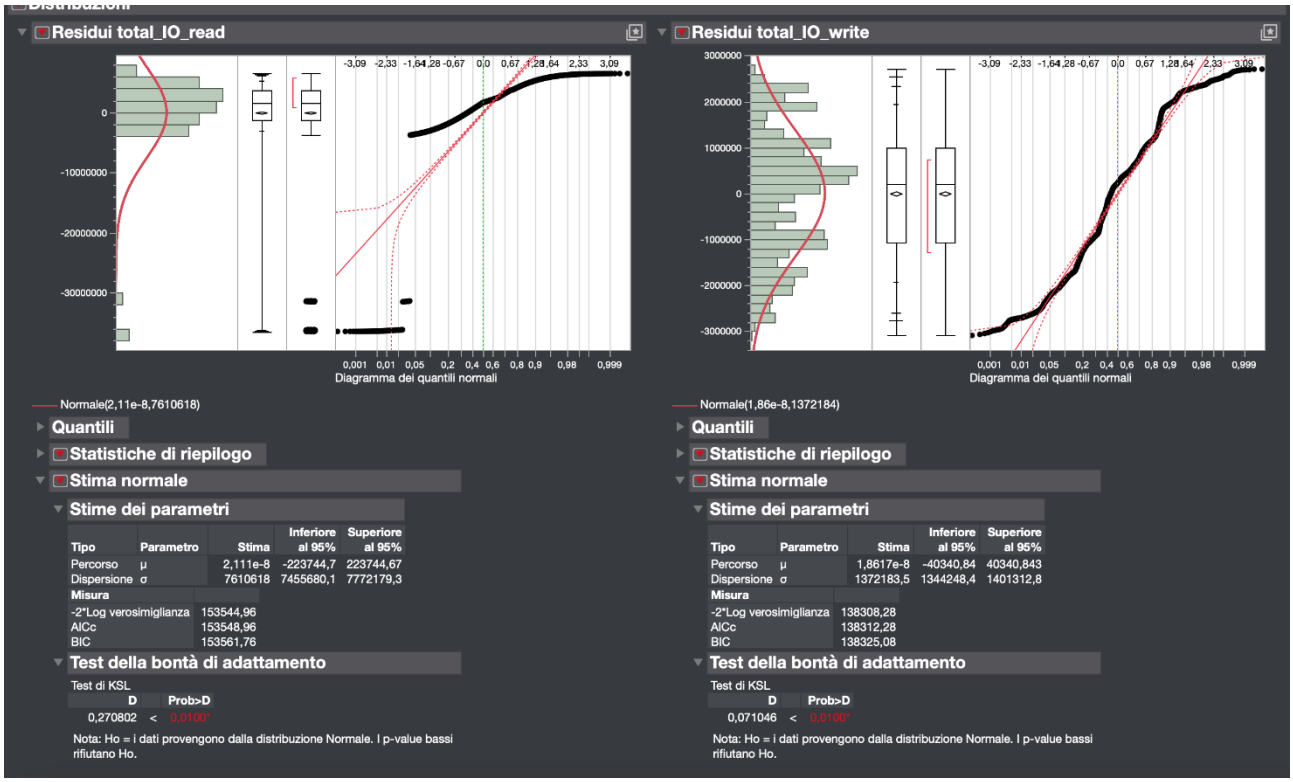
R-quadro	0,999652
R-quadro corretto	0,999652
Scarto quadratico medio	1372338
Media della risposta	1,336e+8
Osservazioni (o somme pesate)	4447

Analisi della varianza

Origine	DF	Somma dei quadrati	Media quadratica	Rapporto F
Modello	1	2,4019e+19	2,402e+19	12753508
Errore	4445	8,3713e+15	1,883e+12	Prob > F
C. totale	4446	2,4027e+19		<,0001*

Stime dei parametri

Termine	Stima	Errore std	Rapporto T	Prob>|t|
Intercetta	6346804,5	41150,88	154,23	<,0001*
currentTime	3430493,6	960,5982	3571,2	<,0001*



La distribuzione dei residui non è normale e quindi andiamo a effettuare il test non parametrico di Kendall per verificare la presenza di un trend:

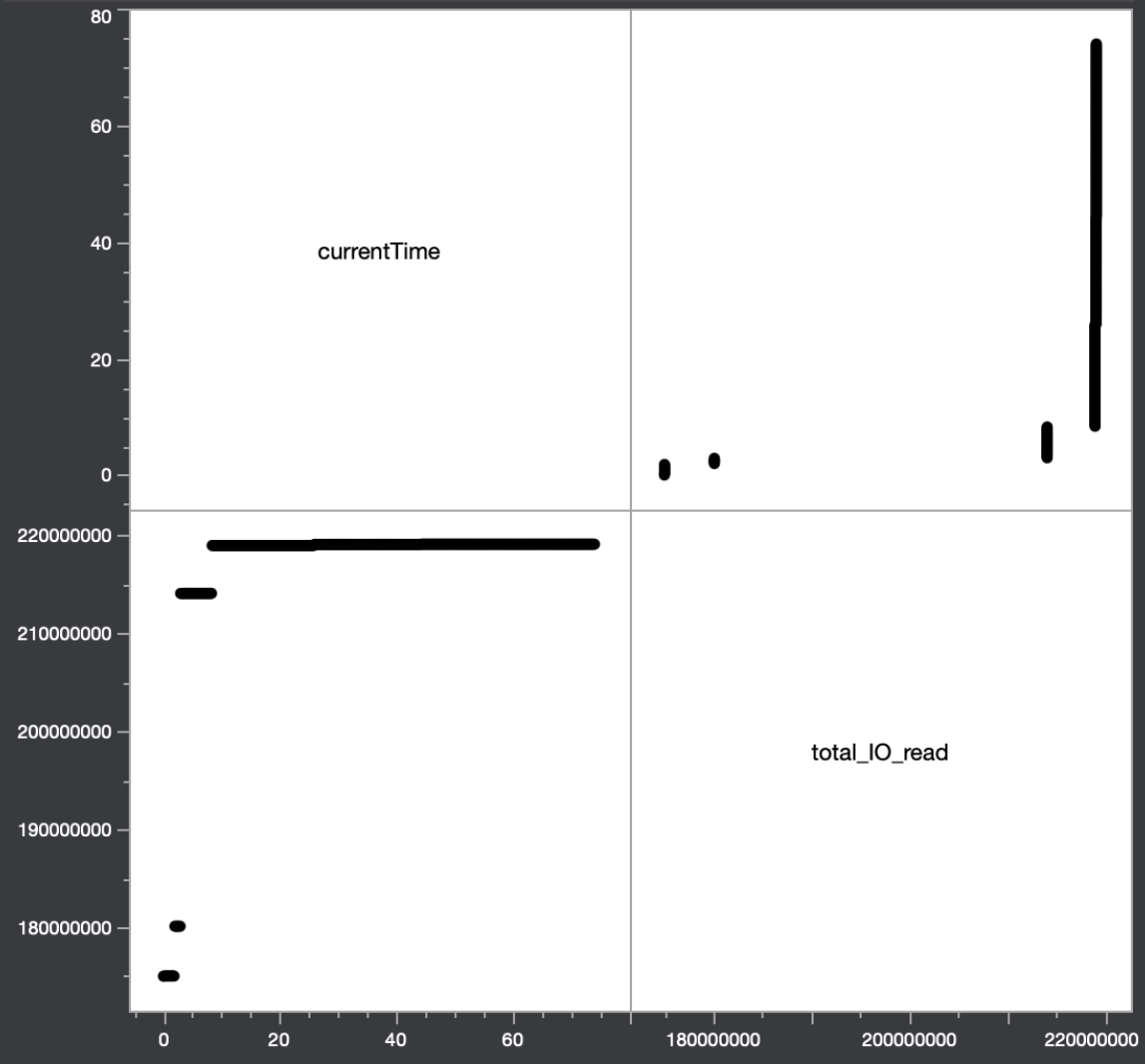
Multivariato

Correlazioni

	currentTime	total_IO_read
currentTime	1,0000	0,4058
total_IO_read	0,4058	1,0000

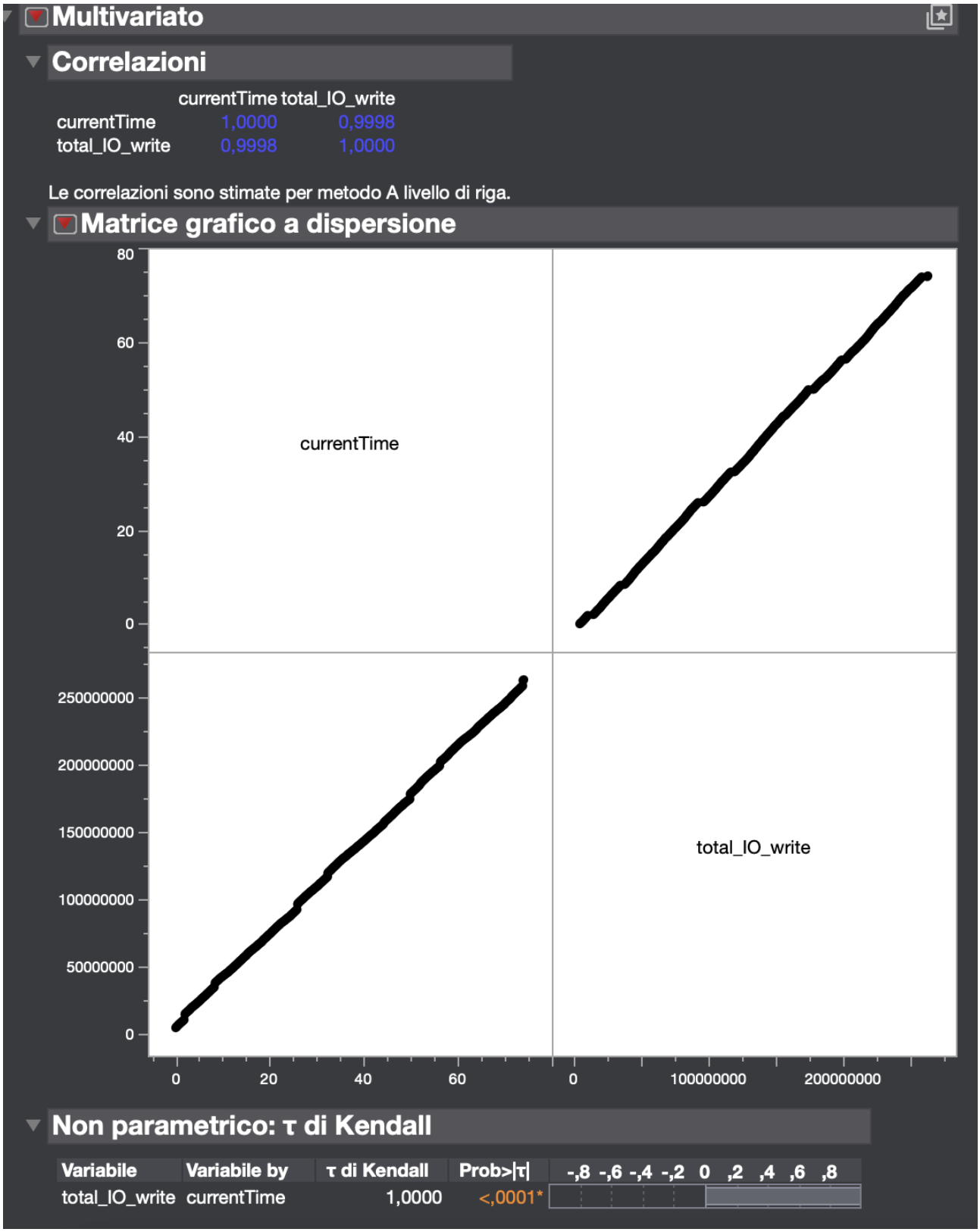
Le correlazioni sono stimate per metodo A livello di riga.

Matrice grafico a dispersione



Non parametrico: τ di Kendall

Variabile	Variabile by	τ di Kendall	Prob>|τ|	-,8 -,6 -,4 -,2 0 ,2 ,4 ,6 ,8
total_IO_read	currentTime	0,8460	<,0001*	



Sia per le letture che per le scritture abbiamo un trend, quindi analizziamo i parametri di regressione attraverso Theil-Sen:

Metric	Slope	Interval_Lower	Interval_Upper	Intercept
Total_IO_read	2412.14726148131	2314.8040901164964	2500.5263692419003	219026036.81401965
Total_IO_write	3424919.414559959	3422796.882874008	3426857.88403178	7353274.635074779

Entrambe le operazioni di I/O hanno un trend positivo con pendenza delle scritture piuttosto elevata.