HANDBOOK PER L’ANALISI DATI CLIMA-COVID19 con regressione lineare

Nov 2020

C. Fagarazzi

Introduzione

- L'analisi della regressione multipla è una tecnica statistica che può essere impiegata per analizzare la relazione tra una **variabile dipendente** e diverse **variabili indipendenti** (**predittori**)
- La regressione lineare multipla rappresenta un'estensione del modello di regressione lineare semplice

L'**OBIETTIVO** dell'analisi è prevedere i valori assunti da una variabile dipendente a partire dalla conoscenza di quelli osservati su più variabili indipendenti

Modello di regressione lineare multipla

Lo studio della regressione multipla consiste nel determinare una funzione che esprima nel modo migliore il **legame (in media) tra le variabili indipendenti X_1, X_2, …., X_k e la variabile dipendente Y.**

Per fare questo occorre incominciare con lo stabilire il tipo di funzione che lega la variabile dipendente a quelle indipendenti. In analogia con quanto già esposto sulla regressione semplice, ipotizziamo il tipo più semplice, quello **lineare**.

Regressione lineare multipla

Idea: Esaminare le relazione lineare fra
1 dipendente (Y) e 2 o più variabili independenti (X_i)

Modello di regressione multipla con k variabili indipendenti:

Y-intercetta

Coefficiente di regressione parziale

Errore casuale

$$Y_i = B_0 + B_1 X_1 + B_2 X_2 + K + B_k X_k + e$$

Modello lineare multiplo

I coefficienti del modello sono stimati sulla base di dati campionari

Modello di regressione multipla con k variabili indipendenti :

Stima (o valore previsto di Y

Stima dell'intercetta

Stima dei coefficienti di regressione parziale

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_2 + K + b_k X_k$$

In questo capitolo utilizzeremo sempre Excel per ottenere i parametri del modello di regressione e altre statistiche (*regression summary measures*).

PARAMETRI

- y_i ed x_1 , x_2 ….x_k sono i valori, rispettivamente, della variabile dipendente e delle k variabili indipendenti, rilevate con riferimento alla *i*-esima unità statistica;

- B_0 è la costante;

- B_1, B_2,… B_k sono i coefficienti di regressione parziale (indicano di quanto varia in media la *Y* quando *Xj* aumenta di un'unità, **a parità di valori delle altre variabili esplicative);**

- e_i è il "residuo non spiegato" relativo all'osservazione *i*-esima;

- *n* è il numero di osservazioni.

INTERPRETAZIONE

Nel modello di regressione multipla si assume che ciascun valore osservato della variabile dipendente sia esprimibile come funzione lineare dei corrispondenti valori delle variabili esplicative, più un **termine residuo** che traduce l'incapacità del modello di riprodurre con esattezza la realtà osservata.

Analisi dei residui

Una volta eseguita la regressione multipla
Potete controllare la distribuzione d residui di ciascuna variabile

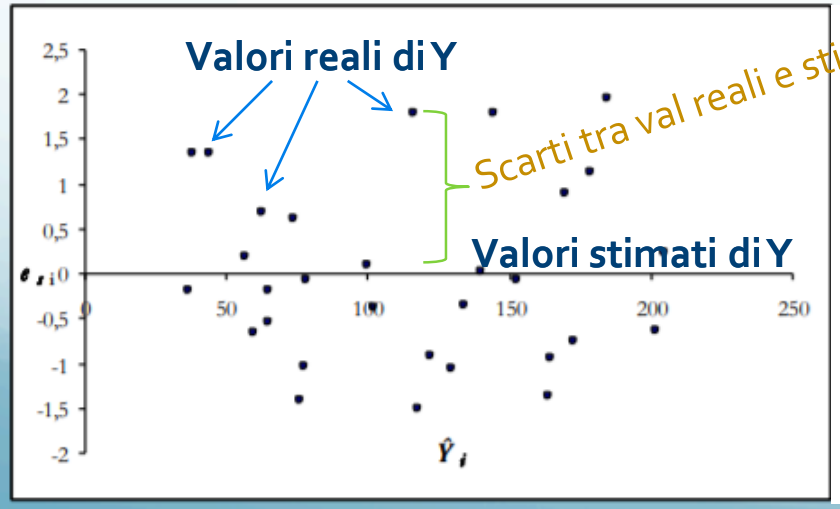
A cosa vi serve?

L'analisi grafica dei residui consente di valutare, a posteriori, se il modello ipotizzato è corretto. In tal caso, infatti, gli errori dovrebbero distribuirsi in modo normale.

Diagramma di dispersione dei residui:

in ordinata: e_{is}

in ascissa: $\hat{Y}_i$ (i valori stimati della variabile dipendente) o X_{ji}

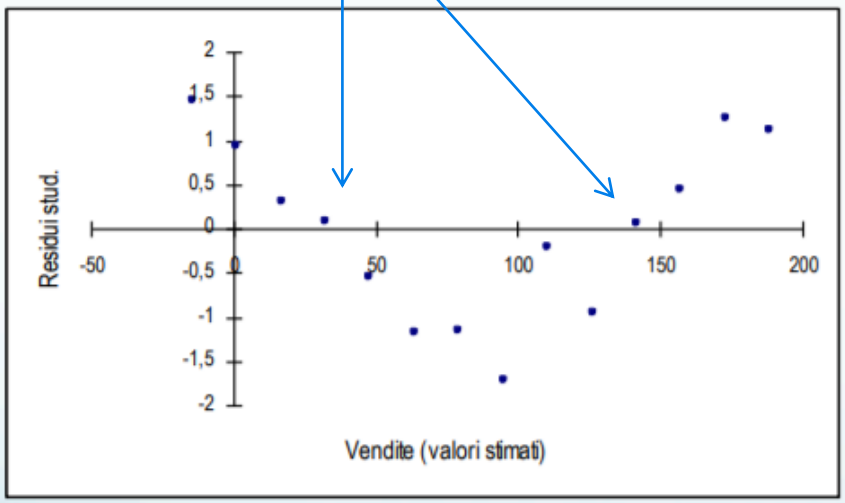


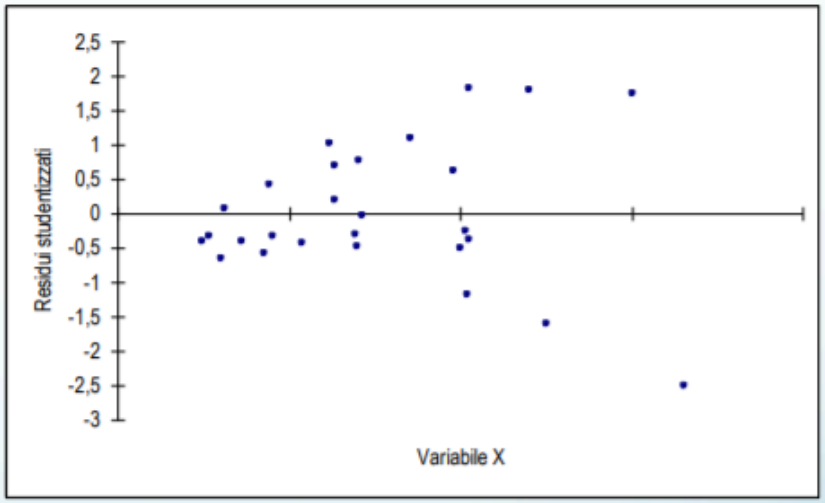
I valori degli scarti si distribuiscono in modo casuale in questo intervallo

SIGNIFICA CHE L'HP DI RELAZIONE LINEARE E' CORRETTA

Analisi dei residui

Lo scarto tra i valori reali di Y e quelli stimati non è casuale.
L'Hp di relazione lineare tra X e Y NON E' CORRETTA





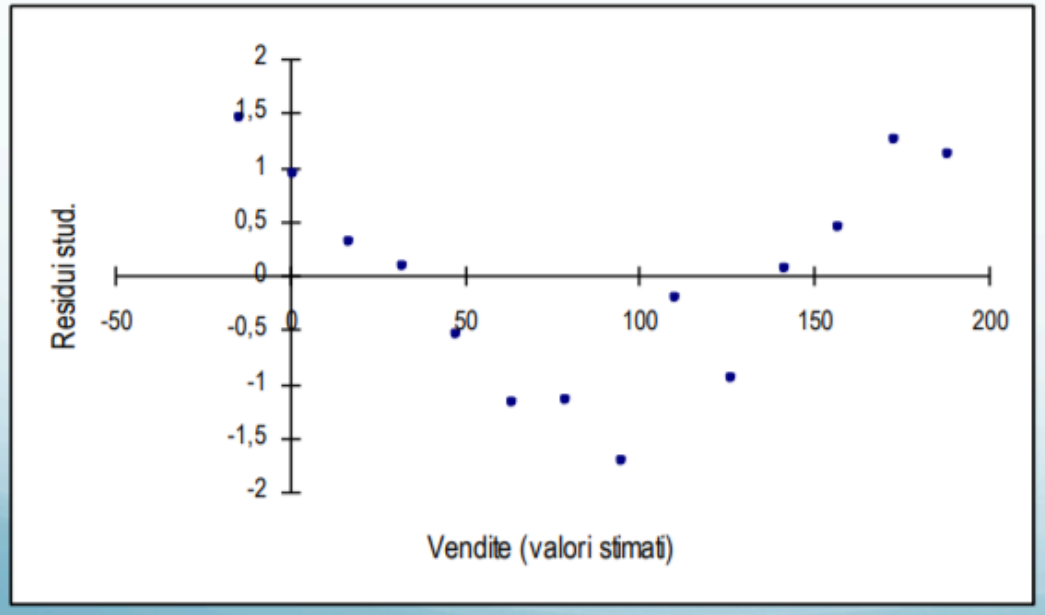
Violazione dell'ipotesi di linearità

Violazione dell'ipotesi di omoschedasticità

Violazione dell'ipotesi di **linearità**

2. dalla struttura nel diagramma di dispersione dei residui (es. crescente o decrescente)

Diagramma di dispersione dei residui

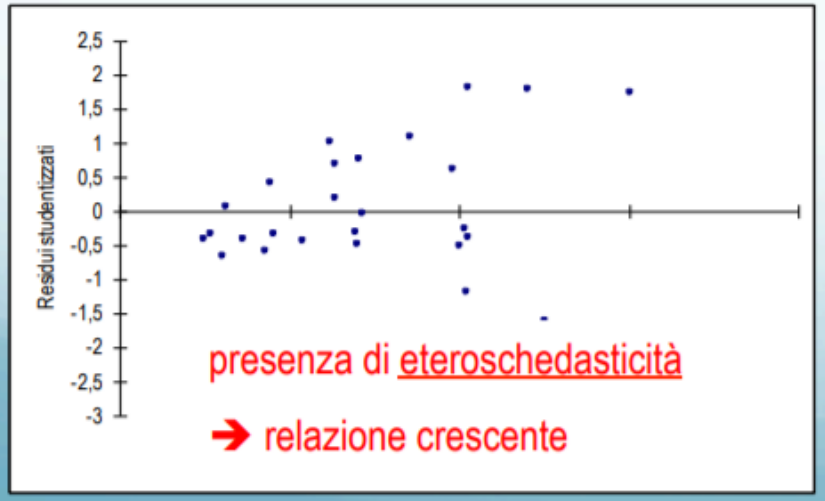


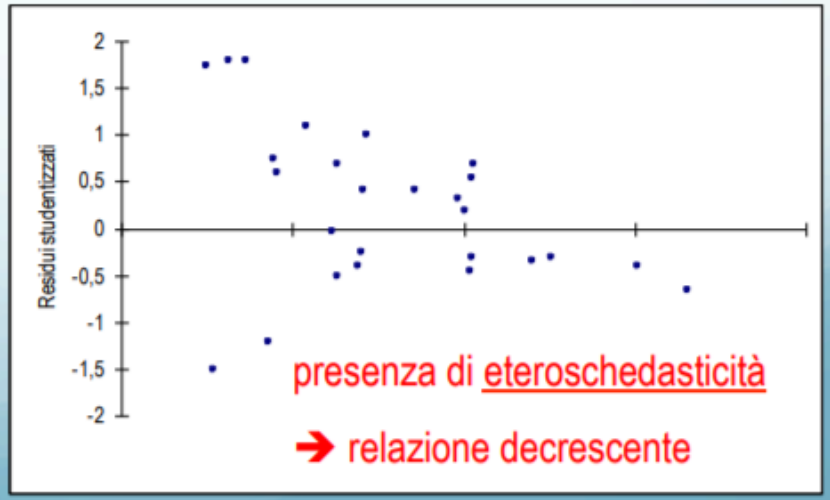
mostra non una disposizione casuale intorno allo zero ma una struttura curvilinea che indica una relazione non lineare

Violazione dell'ipotesi di **omoschedasticità**

Diagnosticata attraverso l' analisi del diagramma di dispersione dei residui:

- se la banda in cui giacciono i punti tende ad allargarsi o a restringersi la varianza degli errori tende a crescere o a decrescere al crescere della variabile esplicativa
- se invece i punti giacciono tra due parallele non si riscontra alcuna evidenza di violazione dell'assunzione



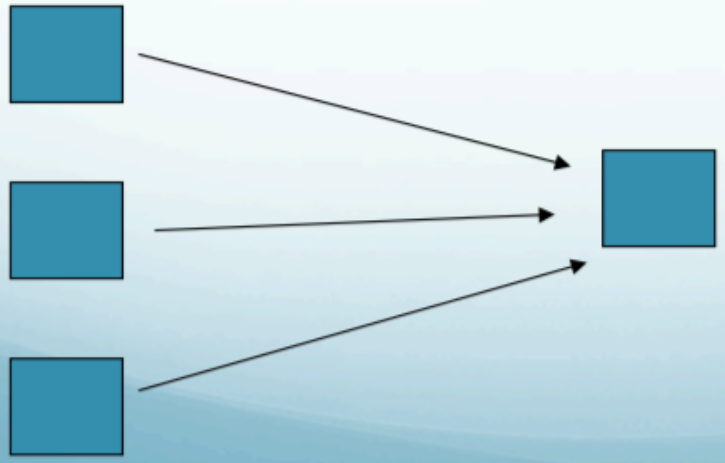


Multicollinearità

La situazione ideale per una regressione multipla dovrebbe essere: ogni X è altamente correlata con Y, ma le X non sono correlate fra loro

Test di CORRELAZIONE

	X_1	X_2	X_3
Y	.60	.50	.70
X_1		.20	.30
X_2			.20

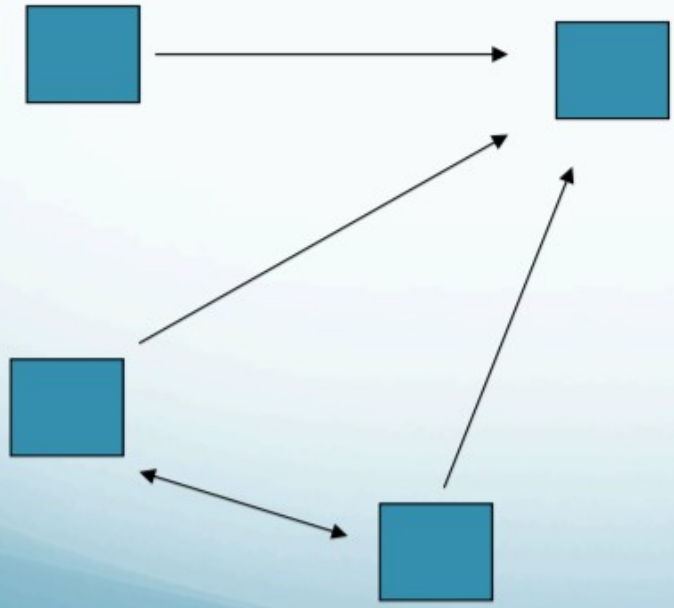


Idealmente, le correlazioni tra le X, dovrebbero essere 0

Multicollinearità

Test di CORRELAZIONE

Spesso però, due o più X sono correlate fra loro....



	X_1	X_2	X_3
Y	.60	.50	.70
X_1		**.70**	.30
X_2			.20

Quando due variabili X o più, sono tra loro correlate (moderatamente o più), parliamo di "**multicollinearità**".

Definizione

Con il termine multicollinearità ci si riferisce alla correlazione fra le variabili indipendenti di un modello di regressione

Il suo effetto consiste nel ridurre la capacità previsiva di ogni singola variabile indipendente in modo proporzionale alla forza della sua associazione con le altre variabili indipendenti.

L' effetto della multicollinearità può interessare:

- la capacità di *spiegazione* del modello (capacità della procedura di regressione e del ricercatore di rappresentare e capire l'influenza di ciascuna variabile indipendente) sia
- la *stima* dei parametri (la sua presenza rende problematica la determinazione dei contributi individuali delle variabili indipendenti, perché i loro effetti vengono "mescolati" o confusi).

Diminuire la multicollinearità

- combinare fra loro i predittori altamente correlati (ad esempio sommandoli)
- se ci sono molti predittori altamente correlati, usare un' analisi delle componenti principali per ridurre il numero delle X e ottenere delle componenti incorrelate tra loro
- adottare come tecnica di analisi una regressione PLS

NEL NOSTRO CASO, E PLAUSIBILE CHE TEMPERATURE MIN, MED E MAX Siamo multicollineari…

Selezione delle variabili esplicative

Per la scelta di quali e quante variabili inserire nel modello bisogna giungere ad un compromesso tra il VANTAGGIO di inserire quante più variabili possibili in modo da ridurre la componente erratica e lo SVANTAGGIO dovuto all'aumento dei costi e della varianza delle stime.

Dato un insieme *q* di predittori esistono varie tecniche per selezionare il **numero ottimale di predittori** da inserire in un modello di regressione multipla:

- Usare la teoria (ricerca bibliografica)
- Metodi semi-automatici sequenziali
  - Regressione stepwise progressiva (avanti – forward)
  - Regressione stepwise a ritroso (indietro – backward)
  - Regressione stepwise convenzionale

Regressione standard

Tutte le variabili X vengono considerate assieme e tutti i coefficienti di regressione stimati contemporaneamente

- Tutte le variabili indipendenti vengono inserite nel modello
- Non si procede quindi ad alcuna selezione
- Per valutare l'importanza di ogni singolo predittore si fa riferimento al *test t*

Regressione Stepwise convenzionale

- È una combinazione delle due tecniche precedenti
- Si procede come in una regressione stepwise forward, ossia un predittore viene incluso nel modello se dà il contributo più significativo alla spiegazione della variabilità di Y.
- Aggiungendo successivamente una nuova variabile, i coefficienti di regressione delle variabili già incluse potrebbero risultare singolarmente non significativi a causa della forte correlazione con la nuova variabile.
- Pertanto, ad ogni interazione si rimettono in discussione i predittori già inseriti verificando la loro significatività attraverso il test F parziale
- Un predittore può essere rimosso nelle fasi successive se la sua capacità esplicativa viene surrogata da altri predittori.
- La regressione stepwise convenzionale (nota semplicemente come "regressione stepwise") è la più utilizzata nelle applicazioni pratiche.

RIASSUMENDO AI FINI ANALISI DATI COVID 1

- Effettuare una regressione lineare tra le singole variabili X indipendenti (**temp min, temp med, temp max, U%, PM10, ecc**) e casi COVID per selezionare tra alcune variabili multicollineari, quella che ha il maggior Indice di Determinazione Lineare R2

  Es: per selezionare tra **t min, t med e t max**

RIASSUMENDO 2

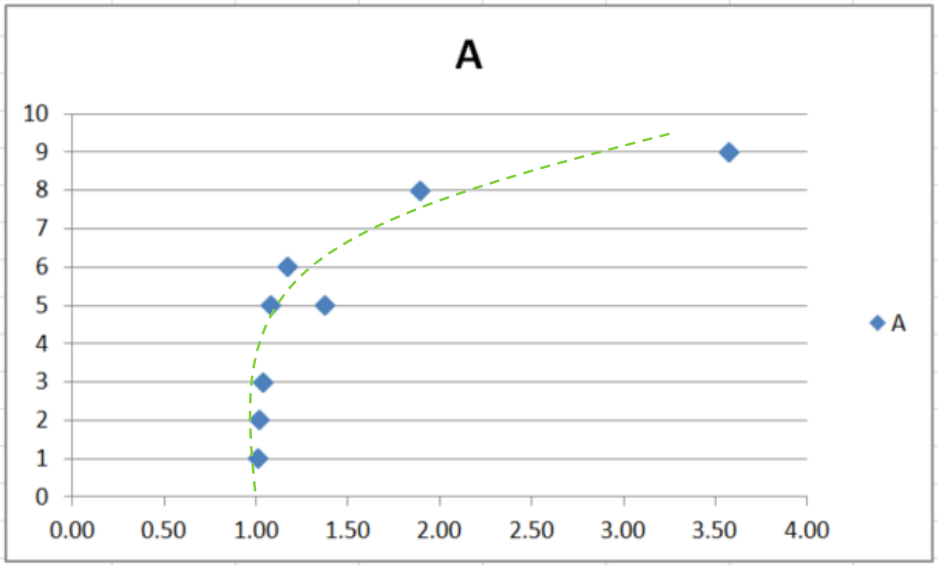
- La regressione lineare fra singole variabili X e Y serve anche per verificare se esiste relazione lineare…o non lineare. Per questa verifica, potete anche costruire dei grafici a punti per verificare

A	B	C	D	E
1	1	1.01	3	100
2	2	1.02	3.2	80
3	1	1.04	4	93
5	1	1.08	4.3	102
6	2	1.17	5	110
5	2	1.37	5.3	121
8	1	1.89	6.4	96
9	1	3.57	4	150

C	A
1.01	1
1.02	2
1.04	3
1.08	5
1.17	6
1.37	5
1.89	8

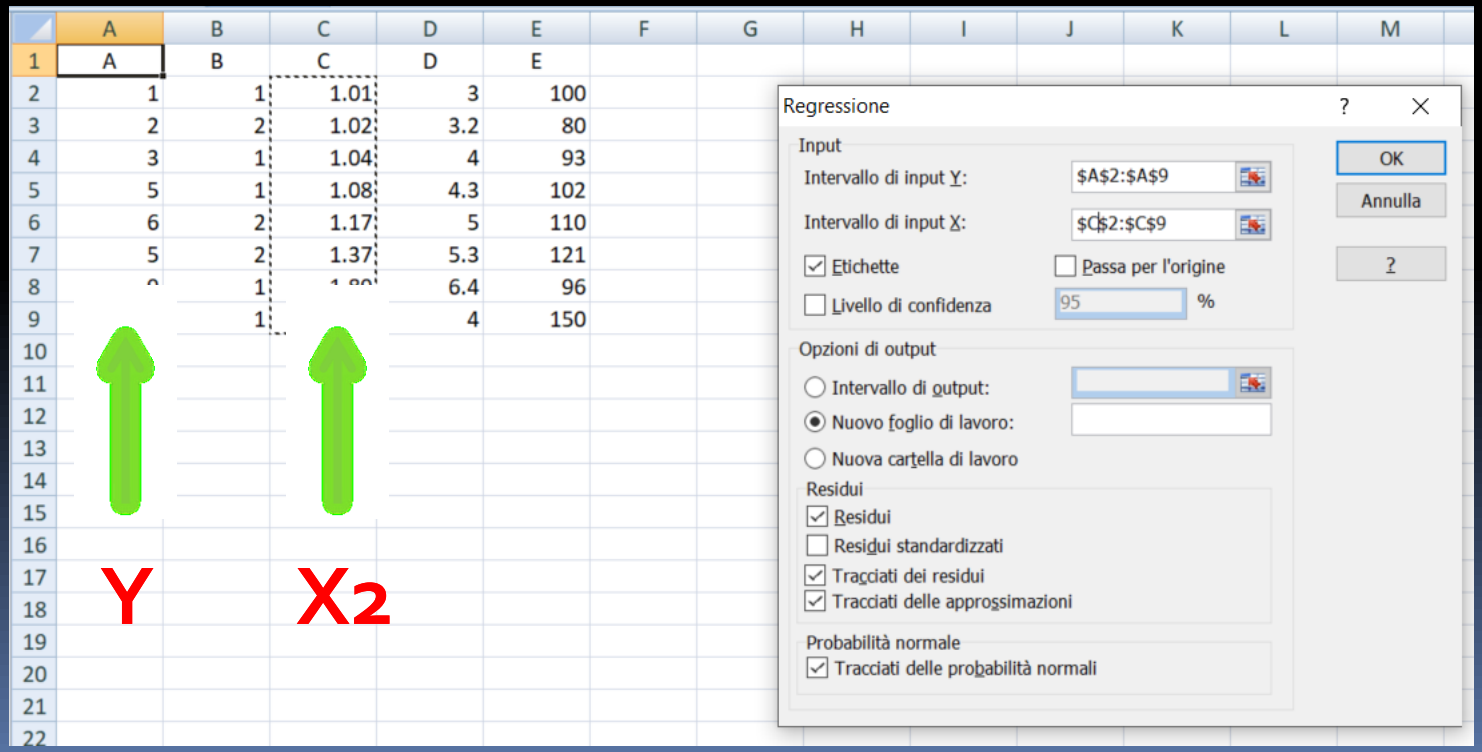
X2 Y

Si vede che esiste una relazione Non lineare



RIASSUMENDO 3

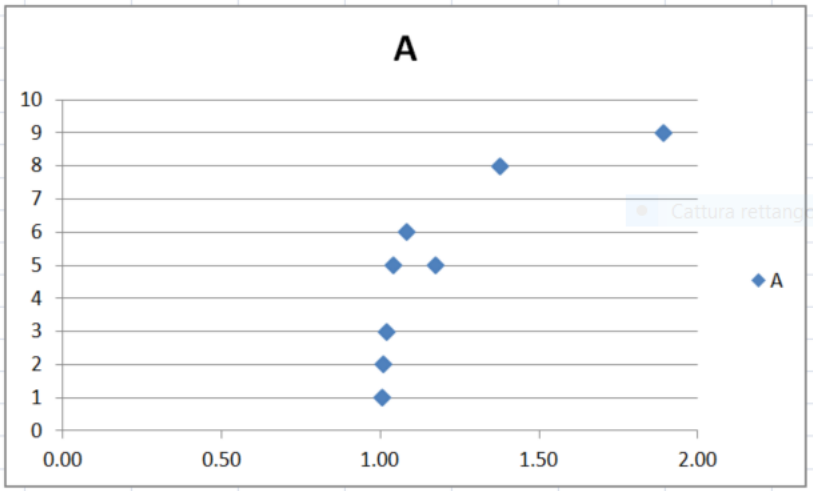
- Le singole regressioni lineari servono quindi sia a selezionare quella con maggiore R2 tra variabili multicollineari (es. tmin, t med, tmax) ma anche a verificare esistenza di relazione lineare o quadratica o esponenziale ecc. (quest'ultimo aspetto lo verificate attraverso i "tracciati dei residui")



RIASSUMENDO steep by steep con Excel

Vostro archivio dati

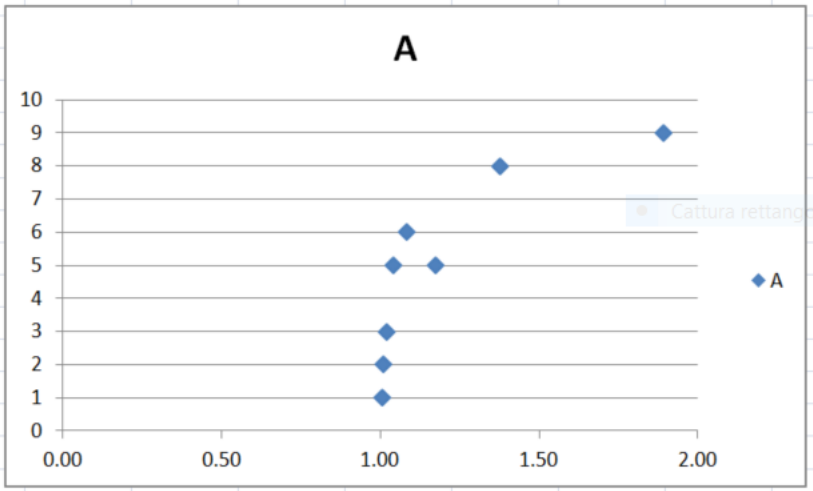
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**A (y)**	B (X1)	C (X2)	D (X3)	E (X4)		**A (y)**	C (X2)	D (X3)	E (X4)		**A (y)**	C (X2)	D (X3)	
2	**1**	1	1.01	3	100		**1**	1.01	3	100		**1**	1.01	3	
3	**2**	2	1.02	3.2	80		**2**	1.02	3.2	80		**2**	1.02	3.2	
4	**3**	1	1.04	4	93		**3**	1.04	4	93		**3**	1.04	4	
5	**5**	1	1.08	4.3	102		**5**	1.08	4.3	102		**5**	1.08	4.3	
6	**6**	2	1.17	5	110		**6**	1.17	5	110		**6**	1.17	5	
7	**5**	2	1.37	5.3	121		**5**	1.37	5.3	121		**5**	1.37	5.3	
8	**8**	1	1.89	6.4	96		**8**	1.89	6.4	96		**8**	1.89	6.4	
9	**9**	1	3.57	4	150		**9**	3.57	4	150		**9**	3.57	4	
10															
11															
12															
13															
14															
15															
16	**A**	C	**A**	C^0.5	**A**										
17	**1**	1.01	**1**	1.00	**1**										
18	**2**	1.02	**2**	1.01	**2**										
19	**3**	1.04	**3**	1.02	**3**										
20	**5**	1.08	**5**	1.04	**5**										
21	**6**	1.17	**6**	1.08	**6**										
22	**5**	1.37	**5**	1.17	**5**										
23	**8**	1.89	**8**	1.37	**8**										
24	**9**	3.57	**9**	1.89	**9**										
25															
26															



RIASSUMENDO steep by steep con Excel

Vostro archivio dati

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	**A (y)**	B (X1)	C (X2)	D (X3)	E (X4)		**A (y)**	C (X2)	D (X3)	E (X4)		**A (y)**	C (X2)	D (X3)	
2	**1**	1	1.01	3	100		**1**	1.01	3	100		**1**	1.01	3	
3	**2**	2	1.02	3.2	80		**2**	1.02	3.2	80		**2**	1.02	3.2	
4	**3**	1	1.04	4	93		**3**	1.04	4	93		**3**	1.04	4	
5	**5**	1	1.08	4.3	102		**5**	1.08	4.3	102		**5**	1.08	4.3	
6	**6**	2	1.17	5	110		**6**	1.17	5	110		**6**	1.17	5	
7	**5**	2	1.37	5.3	121		**5**	1.37	5.3	121		**5**	1.37	5.3	
8	**8**	1	1.89	6.4	96		**8**	1.89	6.4	96		**8**	1.89	6.4	
9	**9**	1	3.57	4	150		**9**	3.57	4	150		**9**	3.57	4	
10															
11															
12															
13															
14															
15															
16	**A**	C	**A**	C^0.5	**A**										
17	**1**	1.01	**1**	1.00	**1**										
18	**2**	1.02	**2**	1.01	**2**										
19	**3**	1.04	**3**	1.02	**3**										
20	**5**	1.08	**5**	1.04	**5**										
21	**6**	1.17	**6**	1.08	**6**										
22	**5**	1.37	**5**	1.17	**5**										
23	**8**	1.89	**8**	1.37	**8**										
24	**9**	3.57	**9**	1.89	**9**										
25															
26															

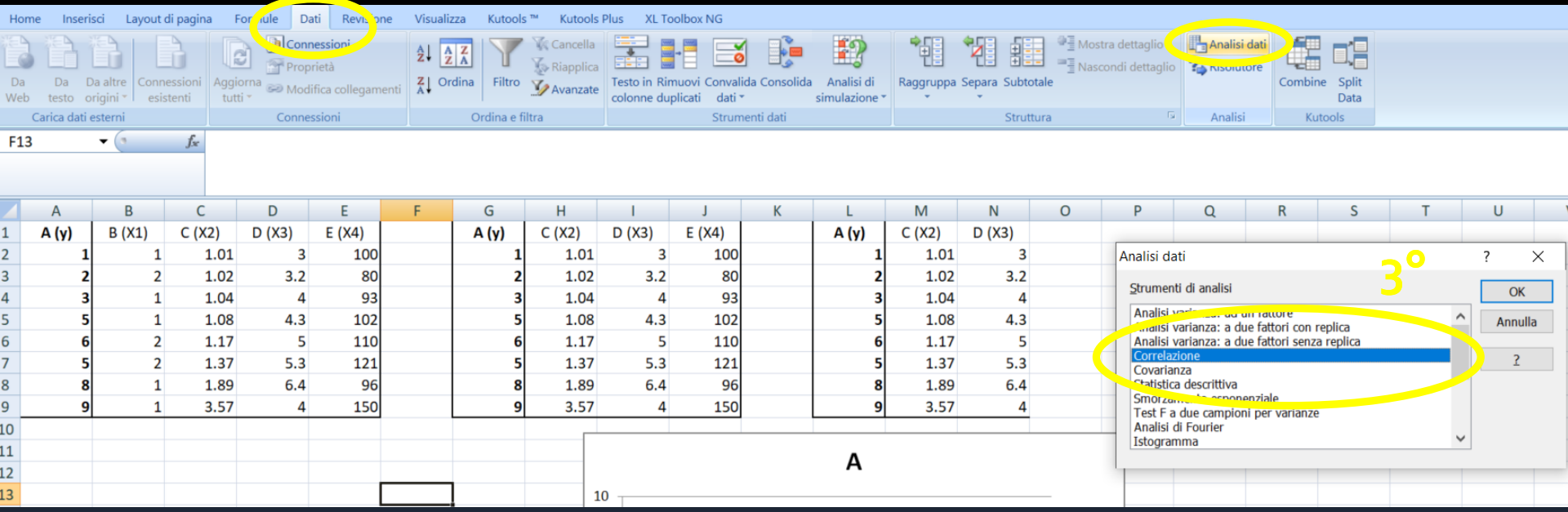


RIASSUMENDO steep by steep con Excel

ANALI DI CORRELAZIONE PER VERIFICARE SE ESISTE MULTICOLLINERAIRA TRA VARIABILI X

1°

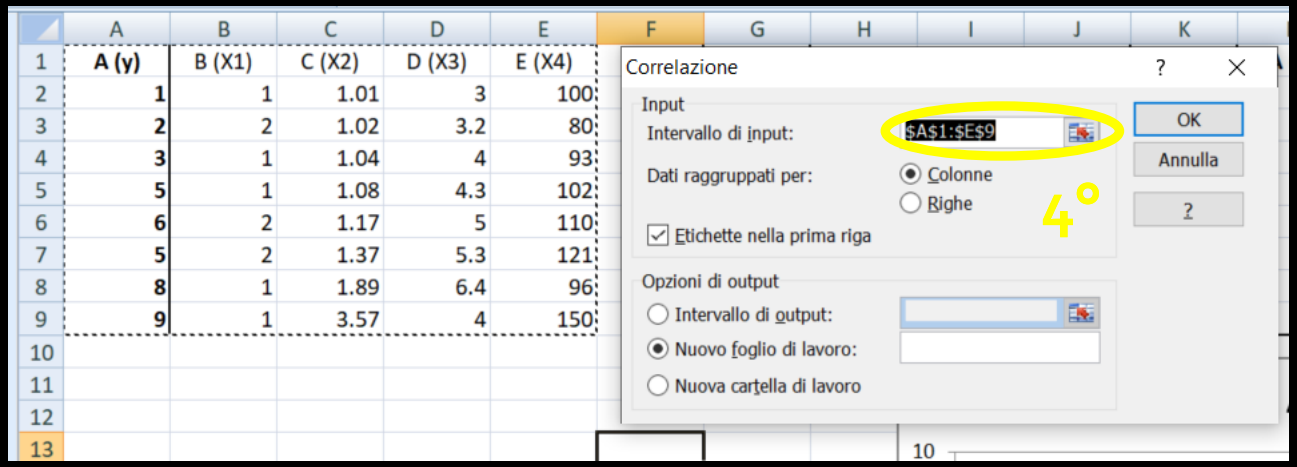
2°



3°

RIASSUMENDO steep by steep con Excel

ANALI DI CORRELAZIONE PER VERIFICARE SE ESISTE MULTICOLLINERAIRA TRA VARIABILI X



RIASSUMENDO steep by steep con Excel

OUTPUT ANALI DI CORRELAZIONE

CON INDICAZIONI SU INTERPRETAZIONE DEI RISULTATI…VEDI FILE XLS

Calcoliamo la matrice di correlazione tra le variabili in gioco.

	A	B	C	D	E
A	1				
B	-0.1602	1			
C	0.78096	-0.3117	1		
D	0.67875	0.07344	0.16254	1	
E	0.66753	-0.11022	0.81646	0.13986	1

Variance Inflation Factors

VIF D-E	1.01995
VIF C-E	2.999
VIF B-E	1.0123
VIF C-D	1.02714
VIF B-D	1.00542
VIF B-C	1.10761
VIF A-B	1.02634
VIF A-C	2.56341
VIF A-D	1.85425
VIF A-D	1.80374

1 SI VERIFICA QUALI VARIABILI X (B, C, D, E) sono maggiormente correlate con la variabile Y da spiegare (A)

Come regola generale `e bene che entrino nel modello le variabili maggiormente correlate con la variabile da spiegare e le meno incorrelate tra loro.

IN QUESTO CASO VARIABILI C, D, E hanno alta correlazione, mentre la variabile B non spiega il fenomeno e può essere esclusa

2 SI VERIFICA SE TRA LE VARIABILI CON ALTA CORRELAZIONE RISPETTO A Y (A) ALCUNE HANNO MULTICOLLINEARITA'

Nella matrice di correlazione il coeff di correlazione C-E e 0,816461…più elevato dei coeff A-C (0,780958) e A-E (0,667531)

Inoltre il coeff VIF di C-E è pari a 2,999 molto elevato anche se min di 5

DOBBIAMO QUINDI ESCLUDERE UNA DELLE DUE VARIABILI, QUALE?

Escluderemo la variabile E perché è quella meno correlata rispetto a Y (A)….prima verificheremo se R2 corretto di C,D,E è maggiore di R2 corretto di C,D. Se così fosse lasciamo anche E

3 IL MODELLO DI REGRESSIONE MULTIPLA SARA' REALIZZATO INSERENDO PROGRESSIVAMENTE LE VARIABILI: C, D, E

Nel caso specifico, dopo l'inserimento di ogni variabile dovremo controllare che il coeff R2 corretto della regressione con C e D sia maggiore di C

Poi controllare che il coeff R2 corretto di C, D e E sia maggiore di C e D. In quest'ulitmo caso dovremo stare particolarmente attenti perché esiste una multicollinerarità tra C ed E.

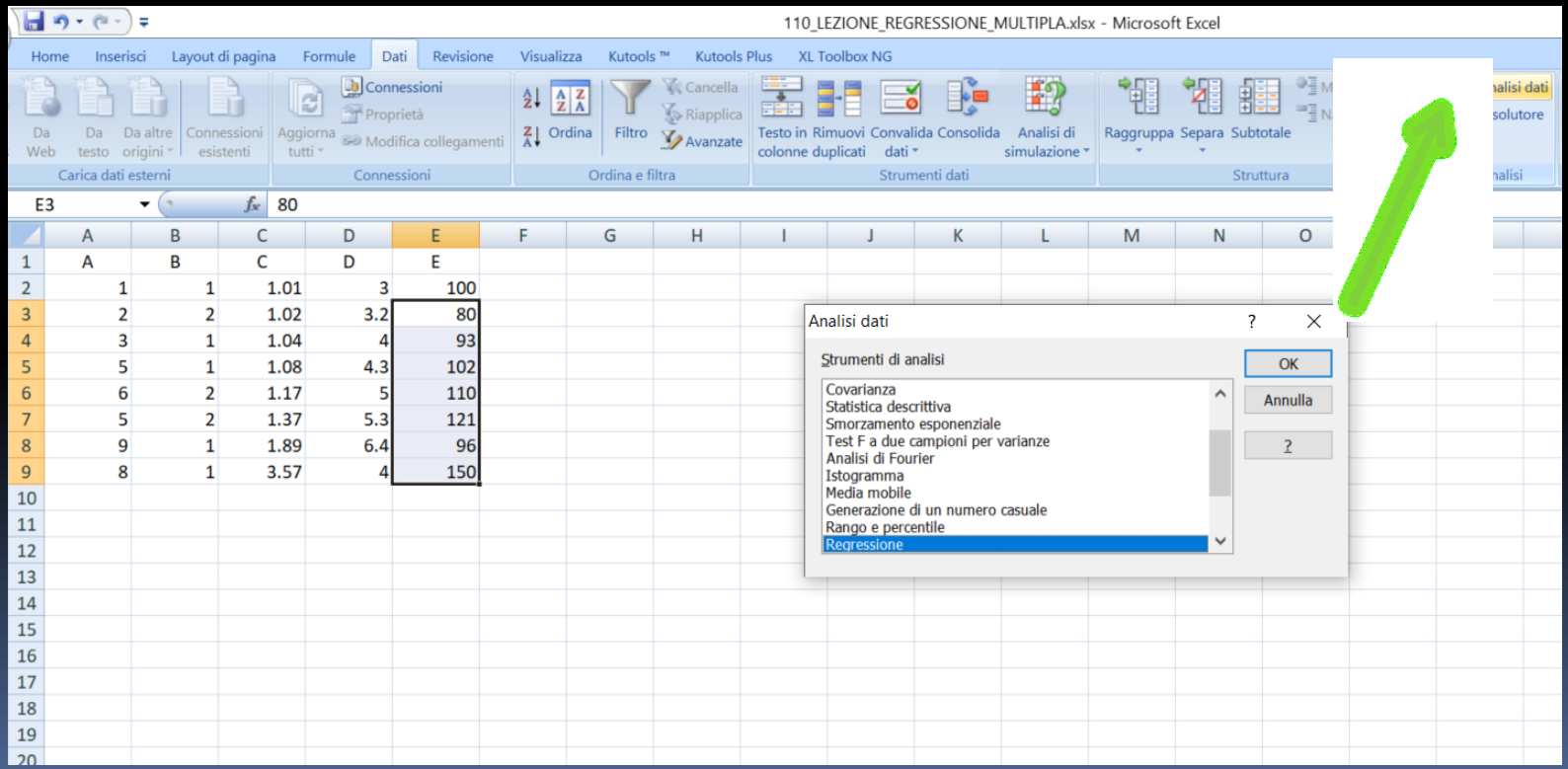
- Esame della matrice dei coefficienti di correlazione
  1. Abbiamo multicollinearità se la correlazione fra coppie di variabili *X* è più elevata di quella con la variabile *Y*
  2. Possiamo verificarla con esame dell Variance Inflation Factor (VIF)

  Se $VIF_i > 5$ al primo test la multicollinearità è presente

  Se $VIF_i > 10$ al secondo test la multicollinearità è presente
- 3. Pochi rimedi
  - –Utilizzare nuovi dati
  - –Eliminare una delle variabili X correlate

RIASSUMENDO steep by steep con Excel

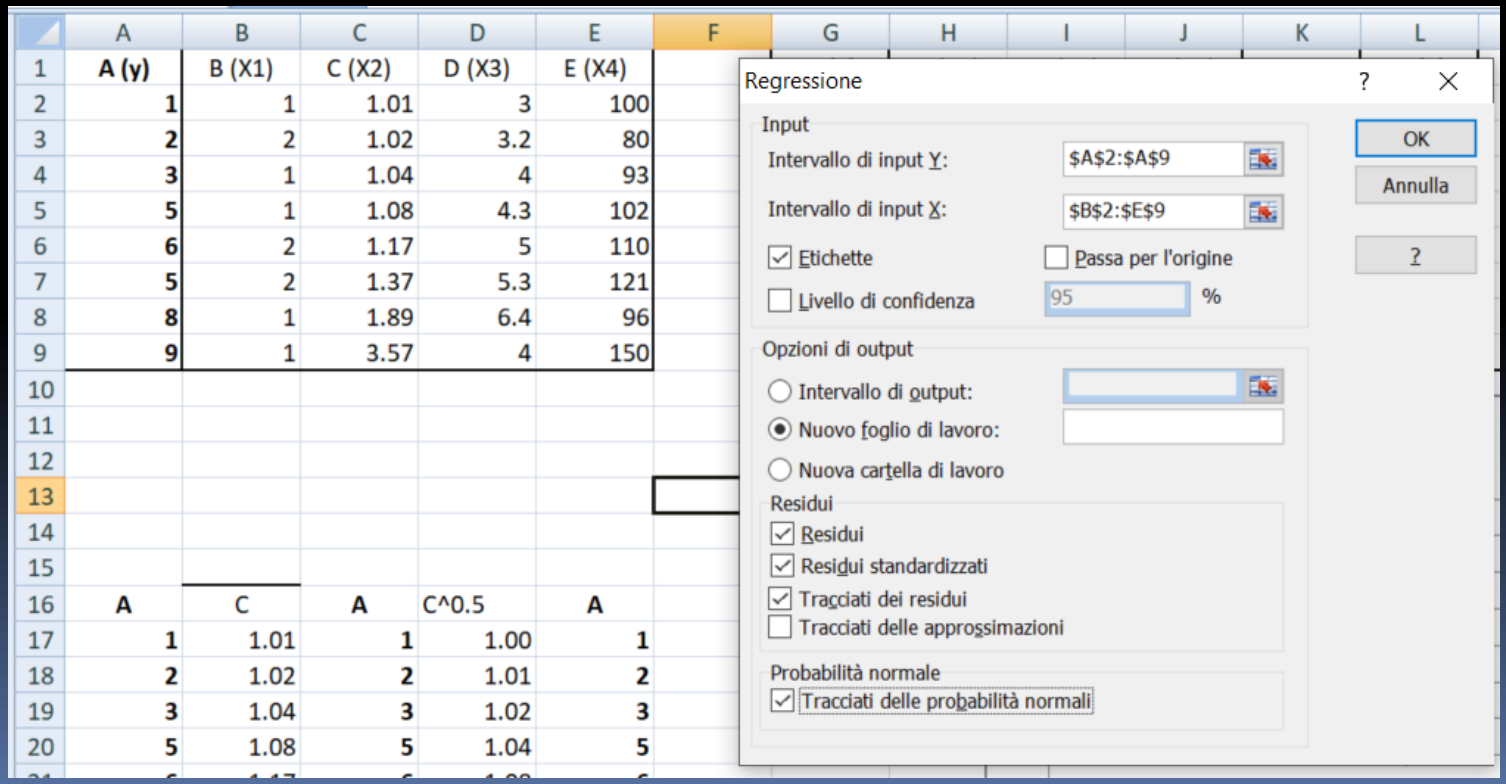
- Realizzare regressione mulipla con excel, vi ricordo, menu DATI/REGRESSIONE



	A	B	C	D	E
1	A	B	C	D	E
2	1	1	1.01	3	100
3	2	2	1.02	3.2	80
4	3	1	1.04	4	93
5	5	1	1.08	4.3	102
6	6	2	1.17	5	110
7	5	2	1.37	5.3	121
8	9	1	1.89	6.4	96
9	8	1	3.57	4	150

RIASSUMENDO steep by steep con Excel

- Definire Intervallo Y (casi COVID)..ideale sarebbe n. casi ogni 1.000 tamponi
- Definire intervallo X (VARIABILI X1, X2, X3 X4)
- Spuntare gli output come in figura, serviranno tutte queste informazioni



	A	B	C	D	E
1	A (y)	B (X1)	C (X2)	D (X3)	E (X4)
2	1	1	1.01	3	100
3	2	2	1.02	3.2	80
4	3	1	1.04	4	93
5	5	1	1.08	4.3	102
6	6	2	1.17	5	110
7	5	2	1.37	5.3	121
8	8	1	1.89	6.4	96
9	9	1	3.57	4	150
10					
11					
12					
13					
14					
15					
16	A	C	A	C^0.5	A
17	1	1.01	1	1.00	1
18	2	1.02	2	1.01	2
19	3	1.04	3	1.02	3
20	5	1.08	5	1.04	5

RIASSUMENDO steep by steep con Excel

OUTPUT REGRESSIONE MULTIPLA CON TUTTE LE VARIABILI
CON INDICAZIONI SU INTERPRETAZIONE DEI RISULTATI…VEDI FILE XLS

OUTPUT RIEPILOGO

tistica della regressione	
R multiplo	0.9623
R al quadr	0.92602
R al quadr	0.77807
Errore sta	1.18111
Osservazi	7

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*nificatività F*
Regressio	4	34.9243	8.73107	6.25878	0.14248
Residuo	2	2.79002	1.39501		
Totale	6	37.7143			

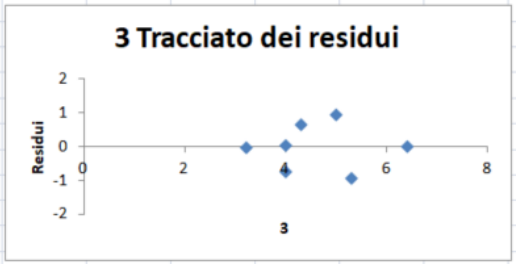
	Coefficient	*ore standa*	*Stat t*	*di signific*	*feriore 95*	*periore 95*	*eriore 95.0*	*eriore 95.0%*
Intercetta	-4.14844	3.62063	-1.14578	0.37049	-19.7267	11.4299	-19.7267	11.4299
1	-0.44638	1.05787	-0.42196	0.71409	-4.99804	4.10528	-4.99804	4.10528
1.01	1.59172	1.04376	1.52499	0.26676	-2.89922	6.08267	-2.89922	6.08267
3	1.18578	0.46258	2.56342	0.12441	-0.80453	3.17609	-0.80453	3.17609
100	0.0207	0.03918	0.52829	0.65006	-0.14787	0.18927	-0.14787	0.18927

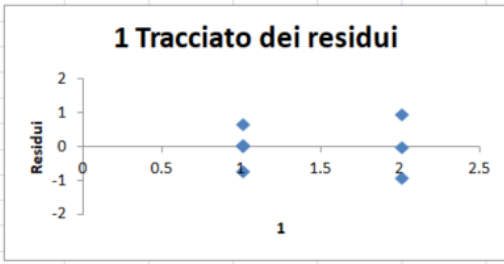
OUTPUT RESIDUI

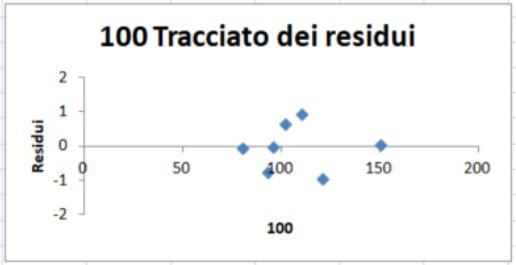
sservazion	*Previsto 1*	*Residui*
1	2.03281	-0.03281
2	3.72952	-0.72952
3	4.33878	0.66122
4	5.03084	0.96916
5	5.93635	-0.93635
6	7.99022	0.00978
7	8.94149	0.05851

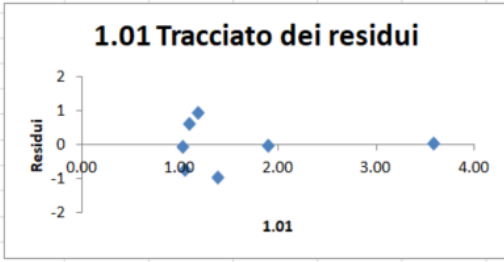
OUTPUT DATI

Percentile	*1*
7.14286	2
21.4286	3
35.7143	5
50	5
64.2857	6
78.5714	8
92.8571	9









4 IL MODELLO DI REGRESSIONE MULTIPLA CON TUTTE LE VARIABILI PRESENTA UN R2 corretto di 0,778066. F è pari a 6,25878 e la significatività (P-value) di F è 0,142483, quindi rifiutiamo l'ipotesi nulla con un livello di significatività $\alpha = 0.05$, la regressione non è quindi rappresentativa con una probabilità del 95%.

Si procede quindi ad eliminare la Variabile B come indicato da matrice correlazione 02. e verifichiamo R2 corretto e F

RIASSUMENDO st by st con Excel

OUTPUT RIEPILOGO

tistica della regressione	
R multiplc	0.95887
R al quadr	0.91944
R al quadr	0.83887
Errore sta	1.00638
Osservazi	7

ANALISI VARIANZA

	gdl	*SQ*	*MQ*	*F*	*nificatività F*
Regressio	3	34.6759	11.5586	11.4125	0.03787
Residuo	3	3.0384	1.0128		
Totale	6	37.7143			

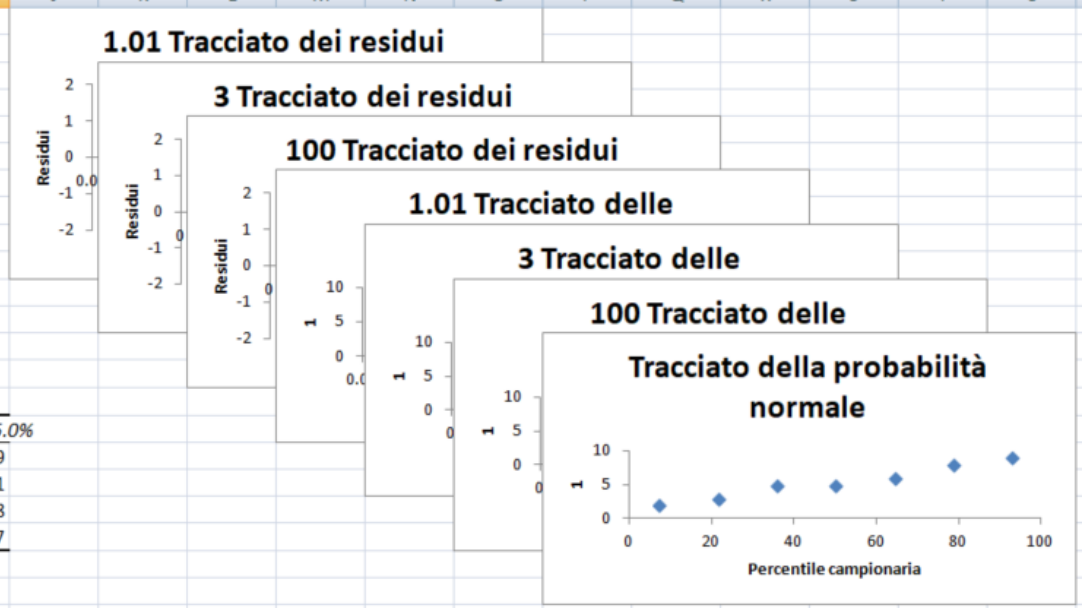
	Coefficient	*ore stand*	*Stat t*	*di signific*	*feriore 95*	*periore 95*	*eriore 95.0*	*eriore 95.0%*
Intercetta	-4.6146	2.93791	-1.5707	0.21428	-13.964	4.73519	-13.964	4.73519
1.01	1.81229	0.76979	2.35425	0.09992	-0.6375	4.26211	-0.6375	4.26211
3	1.20689	0.39184	3.08009	0.05413	-0.0401	2.45388	-0.0401	2.45388
100	0.01492	0.03128	0.47712	0.66588	-0.0846	0.11447	-0.0846	0.11447

OUTPUT RESIDUI

sservazion	*Previsto 1*	*Residui*	*idui standard*
1	2.29017	-0.2902	-0.4078
2	3.48686	-0.4869	-0.6842
3	4.05982	0.94018	1.32118
4	5.18664	0.81336	1.14297
5	6.07962	-1.0796	-1.5171
6	7.96835	0.03165	0.04448
7	8.92853	0.07147	0.10043

OUTPUT DATI

Percentile	*1*
7.14286	2
21.4286	3
35.7143	5
50	5
64.2857	6
78.5714	8
92.8571	9



OUTPUT REGRESSIONE
CON INDICAZIONI SU SIGNIFICATIVITA'
RISULTATI E EVENTUALE
MULTICOLLINEARITA' ...VEDI FILE XLS

5

IL MODELLO DI REGRESSIONE MULTIPLA CON LE VARIABILI C, D, E PRESENTA UN R2 corretto di 0,838873 (superiore a regr inclusa B). F è pari a 11,41254 e la significatività (P-value) di F è 0,037868, quindi accettiamo l'ipotesi nulla con un livello di significatività α = 0.05, la regressione è quindi rappresentativa con una probabilità del 95%.

Pertanto, possiamo rifiutare H0 (nessuna relazione) in favore di H1 (esiste una relazione). Ciò significa che il modello di regressione multipla che è stato proposto non è una mera costruzione teorica, ma effettivamente esiste ed è statisticamente significativo. C'è evidenza che almeno una variabile indipendente influenza significativamente Y !!!

Pur essendo gia' accettabile si procede ad eliminare la Variabile E poiché multicollinerare con la variabile C ...come indicato da matrice correlazione 02. e verifichiamo R2 corretto e F

RIASSUMENDO st by st con Excel

OUTPUT RIEPILOGO

tistica della regressione	
R multiplo	0.95568
R al quadr	0.91332
R al quadr	0.86998
Errore sta	0.90401
Osservazi	7

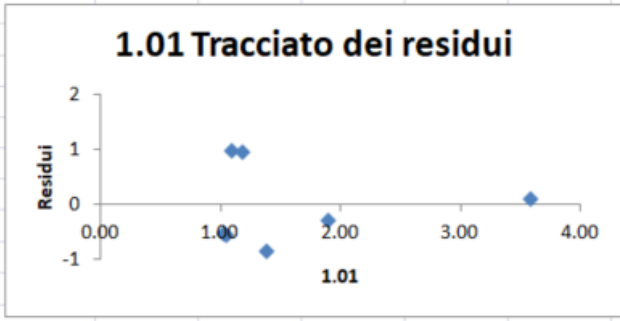
ANALISI VARIANZA

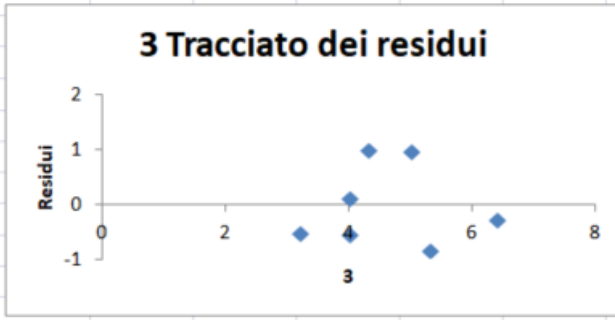
	gdl	*SQ*	*MQ*	*F*	*nificatività F*
Regressio	2	34.4453	17.2227	21.0742	0.00751
Residuo	4	3.26896	0.81724		
Totale	6	37.7143			

	Coefficient	*ore standa*	*Stat t*	*di significe*	*feriore 95%*	*periore 95*	*eriore 95.0*	*eriore 95.0%*
Intercetta	-3.55916	1.7368	-2.04926	0.10978	-8.3813	1.26298	-8.3813	1.26298
1.01	2.112	0.3997	5.28398	0.00615	1.00226	3.22174	1.00226	3.22174
3	1.22217	0.3508	3.48396	0.02526	0.2482	2.19615	0.2482	2.19615

OUTPUT RESIDUI

sservazion	*Previsto 1*	*Residui*	*idui standard*
1	2.50624	-0.50624	-0.68585
2	3.52728	-0.52728	-0.71436
3	3.98317	1.01683	1.37758
4	5.02819	0.97181	1.3166
5	5.82223	-0.82223	-1.11394
6	8.2554	-0.2554	-0.34601
7	8.87749	0.12251	0.16598





OUTPUT REGRESSIONE
CON INDICAZIONI SU SIGNIFICATIVITA'
RISULTATI E EVENTUALE
MULTICOLLINEARITA' …VEDI FILE XLS

6

IL MODELLO DI REGRESSIONE MULTIPLA CON LE VARIABILI C, D PRESENTA UN R2 corretto di 0,8699 (superiore a regr inclusa B e E). F è pari a 21,07418 e la significatività (P-value) di F è 0,007513, quindi accettiamo l'ipotesi nulla con un livello di significatività α = 0.01, la regressione è quindi rappresentativa con una probabilità del 99%.

La regressione multipla con le sole variabili C e D è erisultato molto più rappresentativo che con 4 varibili (R2 corretto 0,8699) con una probilità molto più elevata (99%).

Pertanto, possiamo rifiutare H0 (nessuna relazione) in favore di H1 (esiste una relazione). Ciò significa che il modello di regressione multipla che è stato proposto non è una mera costruzione teorica, ma effettivamente esiste ed è statisticamente significativo. C'è evidenza che le due variabili indipendenti influenzano significativamente Y !!!

RIASSUMENDO st by st con Excel

OUTPUT RIEPILOGO

Statistica della regressione	
R multiplo	0.95567936
R al quadrat	0.91332303
R al quadrat	0.86998455
Errore stand	0.90401327
Osservazior	7

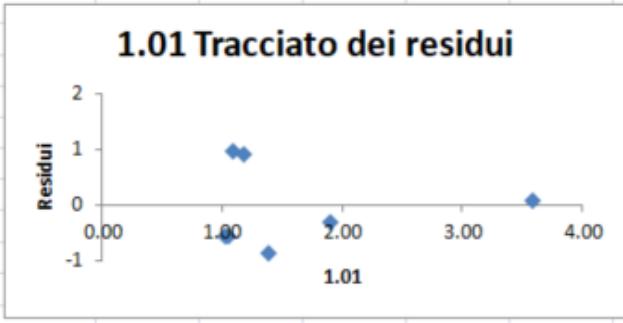
ANALISI VARIANZA

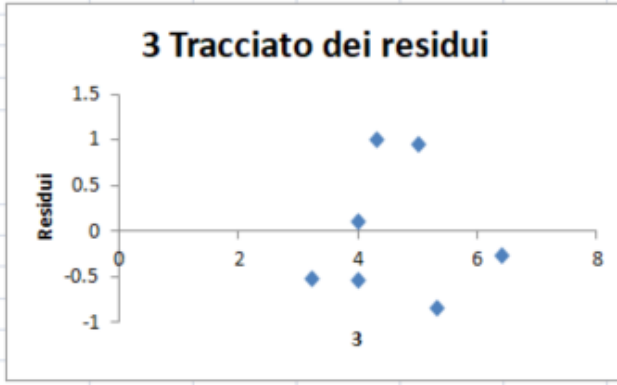
	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	2	34.4453258	17.2226629	21.0741804	0.0075129
Residuo	4	3.26895995	0.81723999		
Totale	6	37.7142857			

	Coefficienti	*Errore standard*	*Stat t*	*Valore di significatività*	*Inferiore 95%*	*Superiore 95%*	*Inferiore 95.0%*	*Superiore 95.0%*
Intercetta	-3.5591633	1.73680373	-2.0492605	0.10978244	-8.3813035	1.2629769	-8.3813035	1.2629769
1.01	2.11199825	0.39969843	**5.283979**	**0.006154**	1.0022575	3.22173901	1.0022575	3.22173901
3	1.22217339	0.35080021	**3.483959**	**0.025261**	0.24819586	2.19615093	0.24819586	2.19615093

OUTPUT RESIDUI

Osservazione	*Previsto 1*	*Residui*	*Residui standard*
1	2.50624096	-0.506241	-0.6858484
2	3.52728411	-0.5272841	-0.7143573
3	3.98317375	1.01682625	1.37758237
4	5.0281877	0.9718123	1.31659808
5	5.82222798	-0.822228	-1.1139433
6	8.25539857	-0.2553986	-0.3460105
7	8.87748693	0.12251307	0.16597903





OUTPUT REGRESSIONE
CON INDICAZIONI SU TEST STATISTICO t
E SIGNIFICATIVITA' DI OGNI VARIABILE
...VEDI FILE XLS

7 Le singole variabili esplicative C e D sono significative?

Inferenza riguardo al coefficiente di regressione parziale:t Test

Il valore del test statistico t per ogni variabile cade nella zona di rifiuto (p-values < 0,05) **C = 0,006154 D = 0,025261 quindi sono entrambe significative**

Per la variabile C il t osservato = 5,283979. Così, H0 deve essere inequivocabilmente rifiutata in favore di H1; in questo caso, si può affermare che la variabile C influenza significativmente la variabile dipendente Y. Per D, t osservato = 3,483959. Così, Ho deve essere rifiutata in favore di H1; in questo caso, si può ritenere che la variabile D ha una influenza significativa sulla variabile dipendente Y.