Contents

List of Figures

1 Problemi introduttivi

I problemi e casi di studio che seguono derivano da esperienze reali o da esempi riportati nella letteratura scientifica; sono funzionali all'introduzione degli argomenti fondamentali affrontati nel corso, e in parte costituiscono una selezione dei problemi reali che verosimilmente sono affrontabili con le metodologie e le tecniche qui esposte.

Alcuni, in una forma anche diversa, di solito semplificata, sono poi ripresi durante le lezioni, o comunque utilizzati come spunto per l'introduzione di problematiche specifiche, o sfruttati nella sezione di esercizi.

Alcuni insiemi di dati sono inseriti, almeno in modo parziale, in queste pagine per prendere abitudine con dati e situazioni *vere* Altri problemi, evidenziati nel testo, si riferiscono invece a situazioni e a metodologie diverse (tipicamente per alcune tecniche speciali di analisi multivariata) che verranno solo accennate in questo corso.

2 Caratteristiche dei dati reali

Caratteristiche dei dati reali

Molti insiemi di dati con cui si ha che fare nella realtà hanno un difetto fondamentale: **sono veri!!!!**
Gli unici dati che si presentano *puliti* sono quelli simulati oppure quelli artificiali.
I dati veri disgraziatamente hanno tanti inconvenienti:

- non sono quasi mai completi (vi sono dati mancanti);
- difficilmente costituiscono un campione casuale semplice da una qualsiasi popolazione;
- difficilmente costituiscono una popolazione completa;
- sono spesso eterogenei (ossia provengono da diverse popolazioni);
- le osservazioni possono avere un grado di precisione delle misurazioni differente;
- qualche volta si guardano bene dal provenire esattamente da famiglie esponenziali o da universi normali;

In ogni caso non mi occuperò in questo corso dei problemi connessi con la misura delle variabili o con la quantificazione di osservazioni reali o con la costruzione di scale di misura.

Presupporrò sempre, *in tutti gli esempi di questo corso*, che le y_i e le x_{ij} (o altri simboli che userò per le variabili) siano riferiti a variabili osservate la cui misurazione e quantificazione costituisce un fatto acquisito e da non mettere in discussione.

3 Esempi agrari

In molti esperimenti agrari si vogliono mettere a confronto delle varietà di una certa coltura o pianta; oppure si vogliono confrontare dei concimi differenti o comunque confrontare tecniche diverse di produzione.

- Si vuole vedere (mediante un esperimento che conduca all'ottenimento di un campione di osservazioni) se la diversa varietà, o concime o altro fattore distintivo influenza la quantità media di raccolto per unità di area (a parità di altre condizioni)
- L'area ove si conduce l'esperimento viene suddivisa in lotti, e le diverse varietà, e/o i concimi, vengono assegnati ai vari lotti.
- E' comunque noto che se nei vari lotti vengono assegnate le stesse varietà nelle stesse condizioni, il raccolto sarà comunque diverso, anche in modo rilevante, da lotto a lotto.
- I lotti vicini avranno la tendenza ad avere livelli dei raccolti simili, e potrebbero esserci altri effetti collegati con la posizione fisica dei lotti.

- Se l'esperimento fosse condotto in un altro anno, presumibilmente il raccolto medio sarebbe sostanzialmente diverso da quello di quest'anno, anche usando la stessa varietà o concime.

Problemi statistici

Problemi statistici

- Separare l'effetto imputabile alle differenze fra i fattori controllabili, ossia le varietà (o i concimi) dagli effetti dovuti ai fattori non controllabili, ossia i diversi lotti ed altre fonti di variabilità, quali per esempio le condizioni atmosferiche.
- Stabilire qual è la varietà migliore;
- stimare la produttività media.
- Come pianificare l'esperimento se si vogliono effettuare simultaneamente i confronti fra le varietà e quelli fra i concimi? Oppure è meglio fare un esperimento per le varietà ed un altro per i concimi?
- Se esiste un concime migliore, è sempre lo stesso per tutte le varietà?

3.1 Osservazioni suddivise secondo due criteri qualitativi

Si tratta di una situazione in cui i valori osservati vengono classificati secondo due diversi aspetti.

- Nel nostro caso vengono cosiderate le produzioni unitarie di alcune varietà di frumento ottenute in 3 diverse stazioni sperimentali, le quali vengono classificate secondo la varietà (A) e secondo le stazioni (B).
- I valori osservati vengono disposti in tavole simili a quelle a doppia entrata in cui però ogni valore non rappresenta una frequenza, ma la produzione unitaria di frumento di una particolare varietà ottenuta in una particolare stazione.
- In situazioni di questo tipo si è interessati a vedere se i due fattori influenzano significativamente la produzione.

Produzioni unitarie di alcune varietà di frumento in tre stazioni sperimentali

Varietà	Stazioni sperimentali			Totali
	Ceriguola (Italia)	Schackensleben (Germania)	Skara (Svezia)	
Borsum	6,04	30,00	21,00	57,04
Dala	5,55	36,75	22,13	64,43
Gelchsheimer	9,00	48,58	20,00	77,58
Halland	8,10	38,25	21,86	68,21
Fanetzki	9,45	37,50	21,00	67,95
Marzatico	10,09	44,60	19,50	74,19
Timilia	12,80	40,50	13,90	67,20
Wagenburg	9,57	40,90	18,40	68,87
Totali	70,60	317,08	157,79	545,47

Figure 1: Esempio analisi della varianza a due vie

3.2 Osservazioni suddivise secondo due criteri qualitativi e con una variabile correlata con la risposta

Si tratta di un'analisi per il controllo statistico degli esperimenti a blocchi.

- I dati riportati in tabella fanno riferimento al numero di barbabietole per appezzamento e produzione in tonnellate per acro, in un esperimento con disposizione degli appezzamenti a blocchi, nei quali sono stati somministrati diversi fertilizzanti.
- L'obiettivo è accertare l'effetto dei vari fertilizzanti sulla produzione delle barbabietole da zucchero.
- Poichè il numero delle piante differisce da un appezzamento all'altro, allora ci si propone di studiare l'effetto della diversa posizione e stimare le produzioni sulla base di un uguale numero di barbabietole.

Numero di barbabietole per appezzamento e produzioni (in tonnellate per acro) in un esperimento con blocchi randomizzati

Fertilizzanti	Numero e produzione	Blocchi 1	2	3	4	5	6	Somme dei trattamenti
Nessuno	Numero	183	176	291	254	225	249	1.378
	Produzione	2,45	2,25	4,38	4,35	3,42	3,27	20,12
P	Numero	356	300	301	271	288	258	1.774
	Produzione	6,71	5,44	4,92	5,23	6,74	4,74	33,78
K	Numero	224	258	244	217	192	236	1.371
	Produzione	3,22	4,14	2,32	4,42	3,28	4,00	21,38
P+*K*	Numero	329	283	308	326	318	318	1.882
	Produzione	6,34	5,44	5,22	8,00	6,96	6,96	38,92
P+*N*	Numero	371	354	352	331	290	410	2.108
	Produzione	6,48	7,11	5,88	7,54	6,61	8,86	42,48
K+*N*	Numero	230	221	237	192	247	250	1.378
	Produzione	3,74	3,24	2,82	2,15	5,19	4,13	21,23
P+*K*+*N*	Numero	322	367	400	333	314	385	2.121
	Produzione	6,10	7,68	7,37	7,83	7,75	7,39	44,12
Somme dei blocchi	Numero	2.015	1.959	2.133	1.925	1.874	2.106	12.012
	Produzione	35,00	35,30	32,91	39,52	39,95	39,35	222,03

Figure 2: esempio analisi della covarianza a due vie

4 Esempio diete suini

Recinto	Dieta	Sex	Pesoin.	accrescimento
I	A	G	48	9,94
I	B	G	48	10,00
I	C	G	48	9,75
I	C	H	48	9,11
I	B	H	39	8,51
I	A	H	38	9,52
II	B	G	32	9,24
II	C	G	28	8,66
II	A	G	32	9,48
II	C	H	37	8,50
II	A	H	35	8,21
II	B	H	38	9,25
III	C	G	33	7,63
III	A	G	35	9,32
III	B	G	41	9,34
III	B	H	46	8,43
III	C	H	42	8,90
III	A	H	41	9,32

```
IV C G 50 10,37
IV A H 48 10,56
IV B G 46 9,68
IV A G 46 10,98
IV B H 40 8,86
IV C H 42 9,51
V B G 37 9,67
V A G 32 8,82
V C G 30 8,57
V B H 40 9,20
V C H 40 8,76
V A H 43 10,42
```

```
TASSI DI ACCRESCIMENTO DI 30 ANIMALI CLASSIFICATI SECONDO
IL SESSO,  L'ALIMENTAZIONE E IL RECINTO DI PROVENIENZA
```

Si vogliono confrontare tre diete per l'alimentazione di suini. L'efficacia della dieta è misurata semplicemente dall'incremento di peso medio settimanale: interessa trovare la migliore dieta.

Gli animali sono suddivisi in 6 recinti; all'interno vi sono 6 animali (uno per ciascuna combinazione dei 2 sessi per le 3 diete))

Il peso iniziale dell'animale presumibilmente è importante.

- Il peso iniziale dell'animale è certamente importante (perchè si suppone che animali più grossi

crescano di più)

- E' presumibile che l'incremento di peso di un generico animale sia dovuto a diversi fattori più o meno controllabili, ma comunque in parte misurabili.
- che effetto ha il sesso degli animali?
- l'allocazione in un determinato recinto è importante?
- Quali diete sono migliori?
- Quali sono senz'altro da scartare?

5 dati antropometrici neonati

5.1 descrizione del problema

Rilevazione in un ospedale palermitano dei dati relativi alle nascite o ai ricoveri in un reparto di neonatologia.

Le variabili rilevate sono tutte quelle previste dalla cartella clinica da compilare per ogni parto o per ogni neonato entrato

5.2 variabili rilevate

- Una rappresentazione con una matrice di grafici è utile per avere un'idea delle relazioni a due a due fra le variabili.
- Molti software hanno la possibilità di fare tale rappresentazione direttamente, insieme con la possibilità di marcare alcuni punti particolari in tutti i grafici
- Fino ad un numero di variabili non superiori ad una decina, si tratta di una rappresentazione che fa cogliere molto delle relazioni fra le p variabili

5.3 Esempio di matrice dei dati

```
Età  fumatrici parto gestazione peso lungh. cranio
36 0 7 38 4010 535 374
30 0 0 38 4600 485 380
```

```
42 0 8 31 1450 400 298
27 0 7 34 1080 370 260
28 0 6 40 3680 515 355
38 0 7 37 2800 475 343
30 0 6 39 2990 515 336
40 0 0 39 3500 510 352
34 0 7 40 3150 490 350
33 0 7 40 2800 480 342
29 0 6 39 3550 500 340
32 0 7 38 3020 vuote vuote
vuote 0 6 38 4030 495 345
28 0 0 39 3400 500 365
32 0 7 37 2070 vuote vuote
30 0 7 38 3100 480 346
22 0 6 40 3300 510 325
37 0 7 31 1280 380 275
25 0 7 37 3260 480 345
35 0 7 34 2370 460 320
34 0 0 38 2770 470 296
21 0 7 33 1810 390 308
38 0 8 30 1370 400 280
```

```
22  0  7  40  3450  470  365
24  0  8  30  1600  390  295
22  0  7  29  1420  370  284
```

5.4 Rappresentazioni grafiche 1

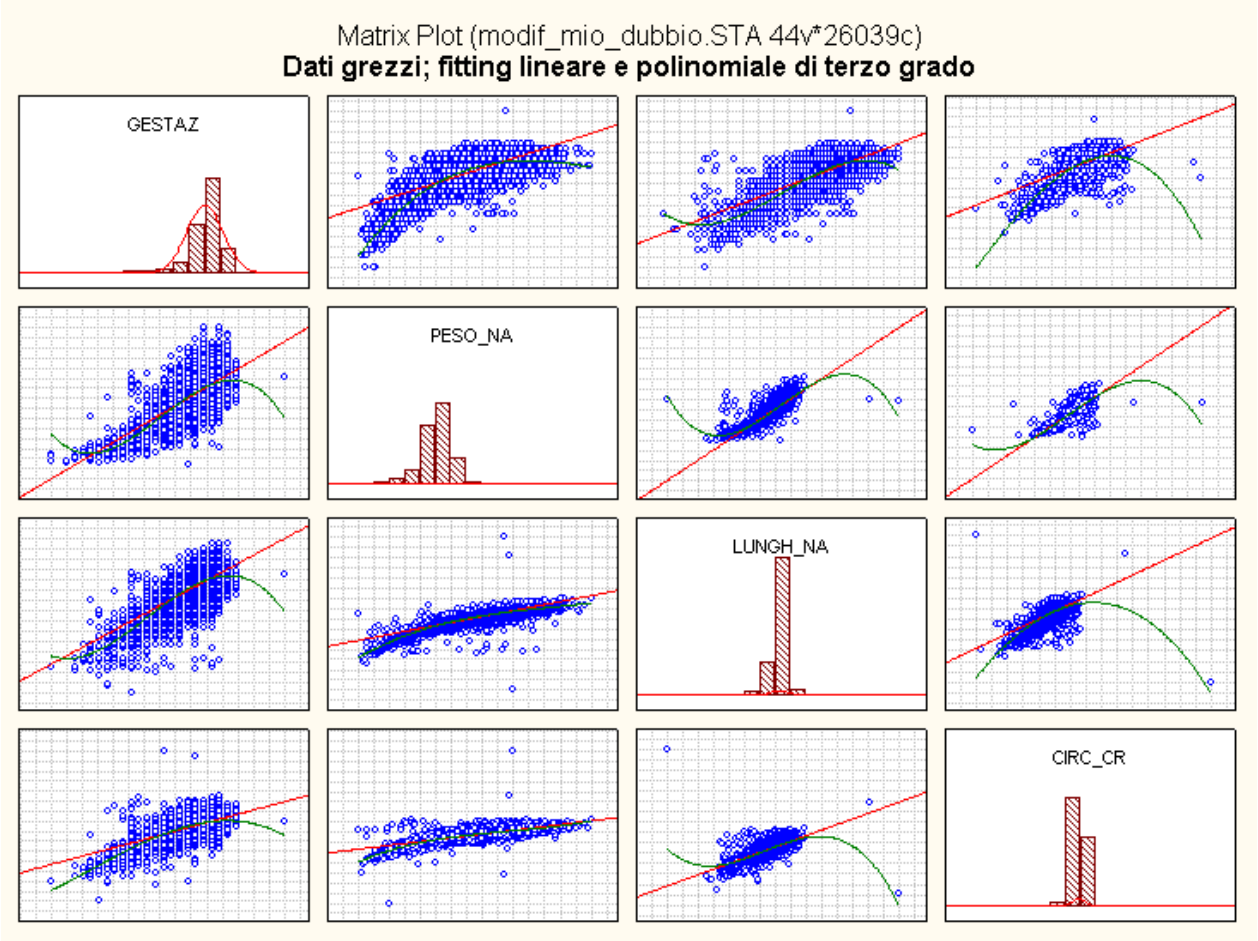

Figure 3: grafico a matrice delle tre variabili antropometriche e dell'età gestazionale: dati grezzi

5.5 Rappresentazioni grafiche 2

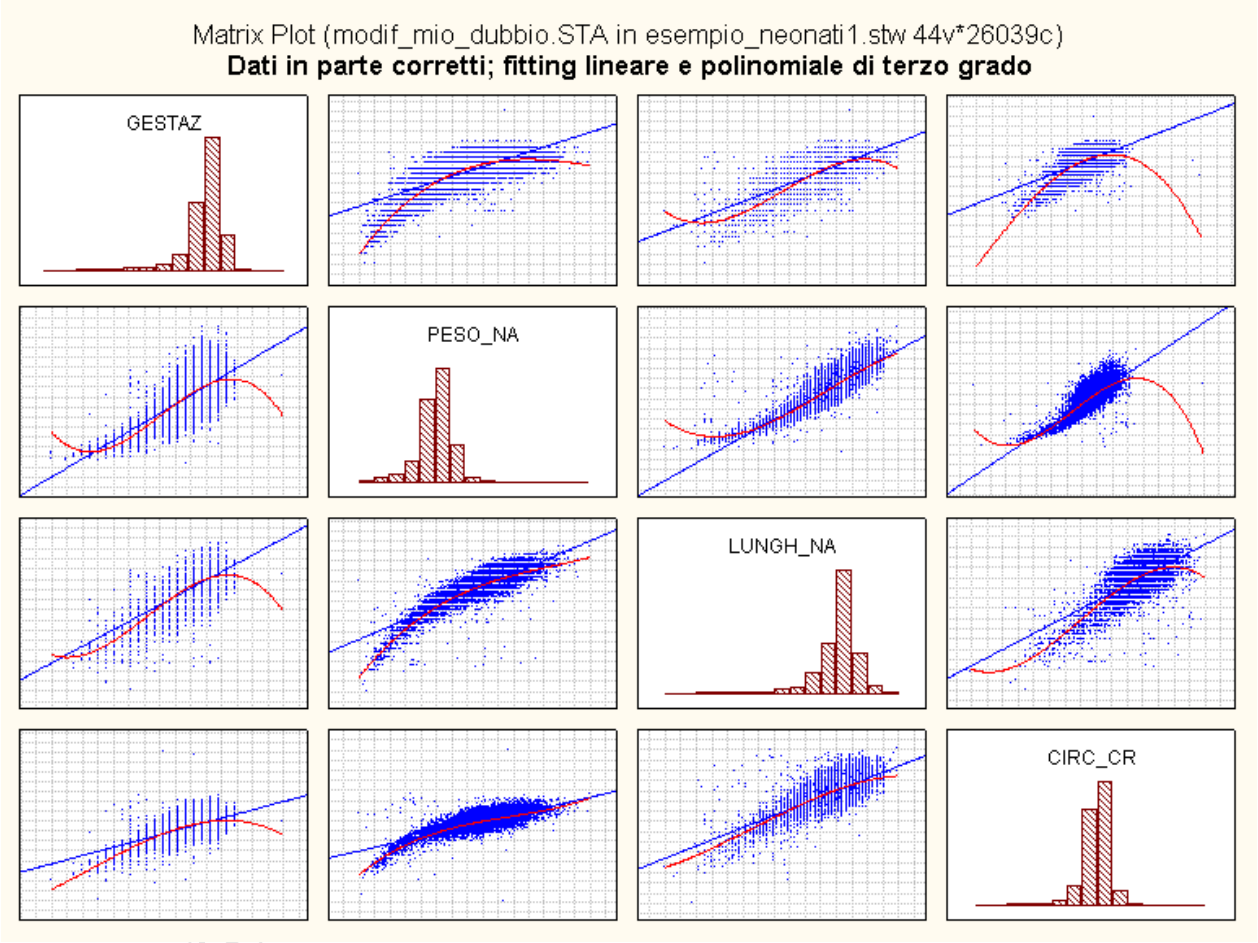

Figure 4: grafico a matrice delle tre variabili antropometriche e dell'età gestazionale: dati corretti (almeno in parte)

5.6 Problematiche statistiche (solo alcune!)

- nelle rappresentazioni grafiche si sono riportate solo alcune variabili a titolo di esempio: non si è tenuto conto di molte variabili che influiscono su queste, quali il tipo di parto, l'età della madre, etc.
- E' possibile costruire degli standard di peso, altezza e circonferenza cranica alla nascita in funzione dell'età gestazionale?
- la relazione fra peso ed età gestazionale è lineare, o è meglio espressa da un polinomio? e di che grado?
- Che incidenza hanno i diversi tipi di parto?
- Esistono patologie più frequenti in funzione di alcuni fattori?
- Lo status materno (essere fumatrice, tipo di parto, età della madre, etc. influenza le caratteristiche del neonato?)
- Etc. etc. ...

5.7 Alcuni problemi e spunti

Alcune delle relazioni fra variabili sono tipicamente non lineari:

Figure 5: esempio delle nascite: relazione fra altezza e peso dei nati

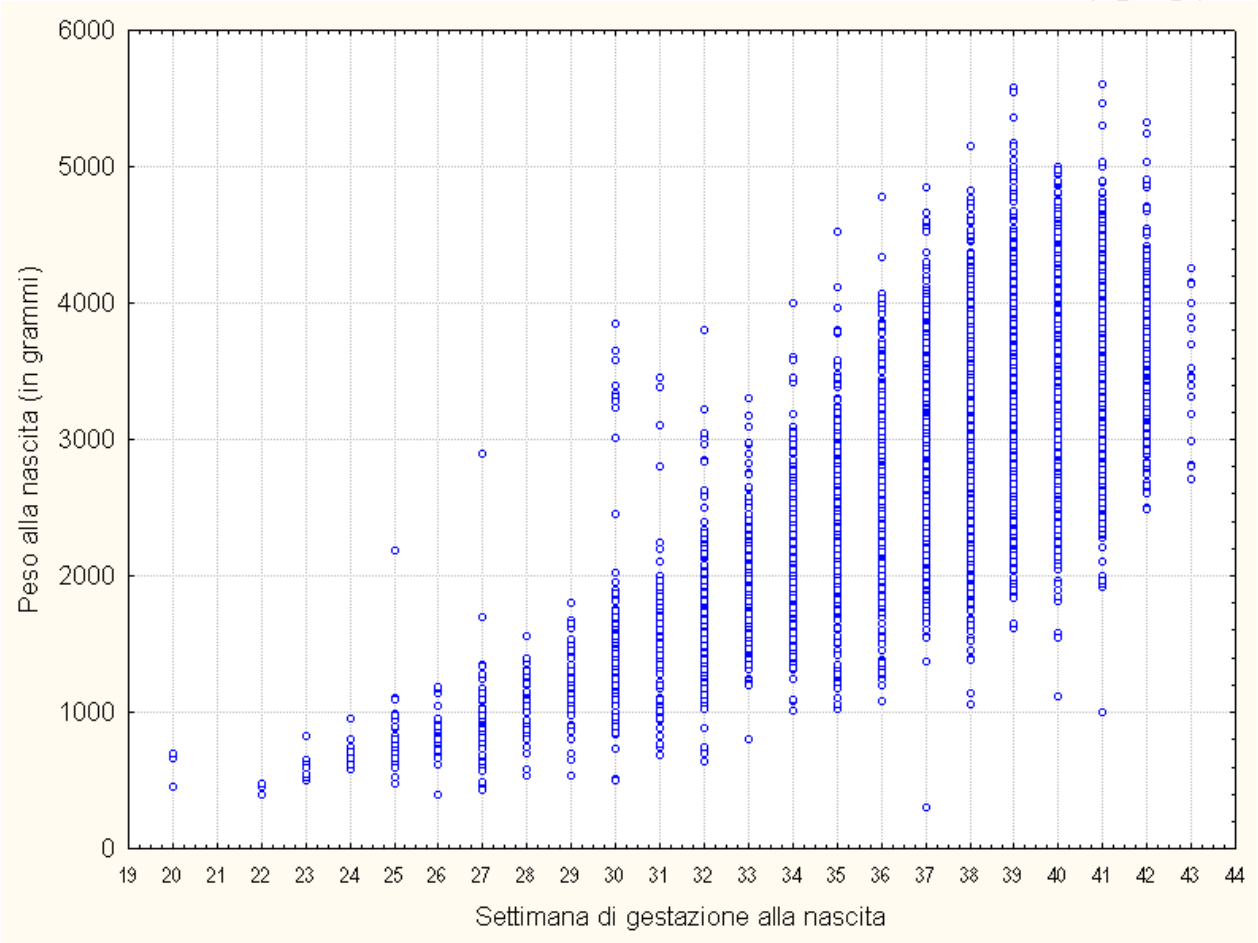

Figure 6: esempio delle nascite: relazione fra peso dei nati e durata della gestazione

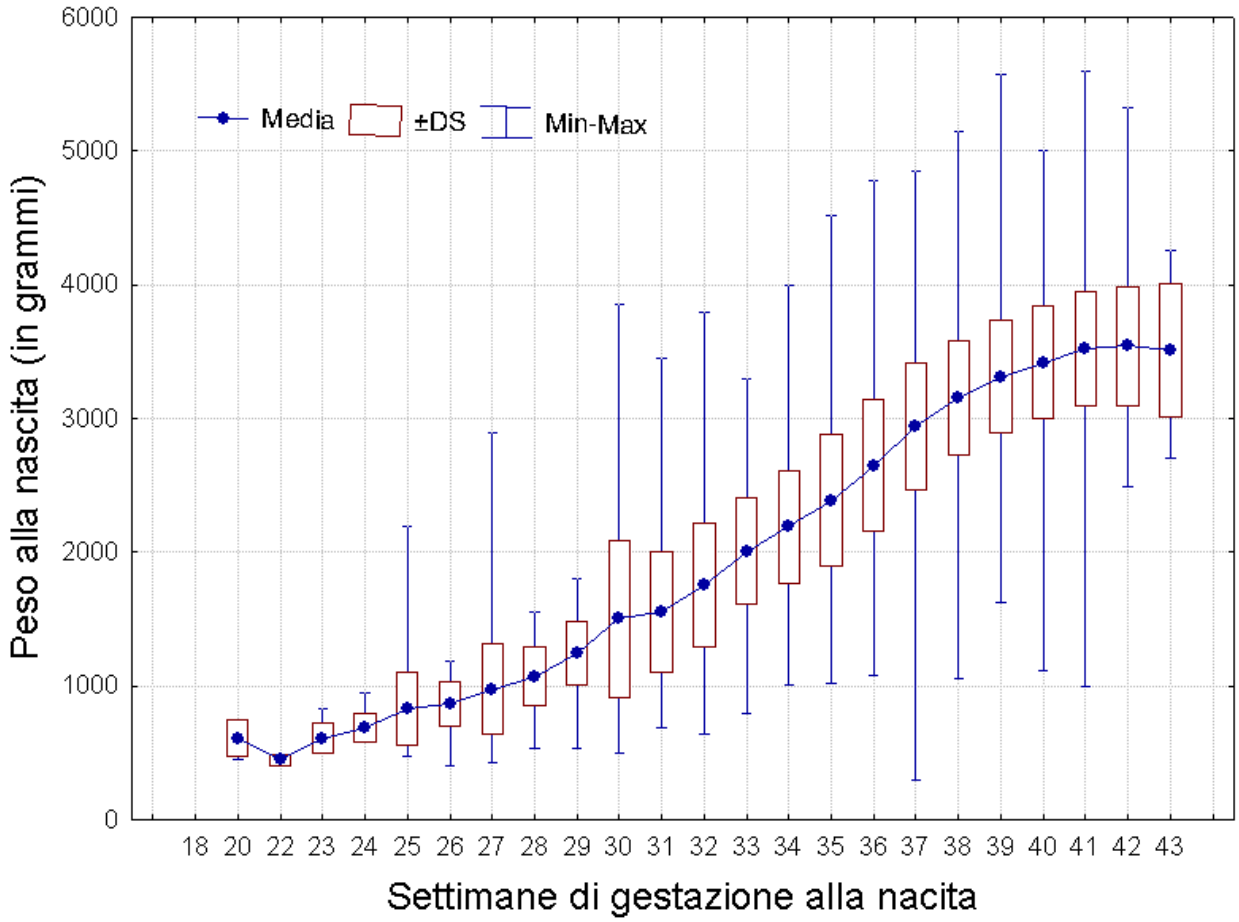

Figure 7: esempio delle nascite: relazione fra peso dei nati e durata della gestazione

Figure 8: esempio delle nascite: relazione fra peso dei nati e durata della gestazione

Figure 9: esempio delle nascite: relazione fra peso dei nati e durata della gestazione

Non si confonda la curva ottenuta dalle medie dei pesi dei nati in funzione della durata della gestazione (che è una curva di regressione)con la curva di crescita (intrauterina) del peso per un singolo neonato in funzione della settimana di gravidanza (che è una curva di crescita individuale)

regressione

5.8 Una relazione non lineare: regressione non parametrica

Nella figura è riportata la relazione (per il dataset delle nascite) fra peso all'ingresso e peso all'uscita (solo a titolo di esempio e trascurando tutte le altre variabili che consentirebbero di selezionare meglio i casi)

- E' evidente che la relazione fra le due variabili non è lineare.
- Infatti, come si vede anche dal grafico, si sa che il neonato comunque non esce dal reparto se non ha raggiunto un certo peso (parte sinistra del grafico); (nel grafico sono riportati tutti i casi a prescindere dal numero di giorni di permanenza e a prescindere dal numero di settimane di gestazione)
- E' improbabile che sia utile adattare un'unica relazione di regressione: è meglio procedere per via esplorativa

Figure 10: esempio delle nascite: relazione fra peso all'ingresso e peso all'uscita

regressione non parametrica

Si può cercare di stimare una relazione di regressione non parametrica come si vede dal grafico (è irrilevante qual'è la tecnica particolare usata) da un certo punto in poi, la relazione può essere considerata senz'altro lineare.

E' ovvio che un'unica relazione lineare sarebbe del tutto insoddisfacente

6 dati antropometrici

6.1 descrizione del problema

In un'indagine antropometrica, si esamina un grosso campione di ragazzi a cavallo dell'età puberale, su ciascuno dei quali vengono rilevati diversi caratteri antropometrici.

6.2 Esempio di matrice dei dati

	5 HABITUS	6 ALTEZZA	7 PESOKG	8 TORACECM	9 CRANIOCM	10 BISACROM	11 BITROCAN	12 SPANCM	G
1	L	143	34	71	53	31	23	145	A
2	L	159	44	74	52	36	26	161	A
3	N	162	48	74	53	36	26	162	A
4	N	177	56	85	58	38	30		A
5	B	148	53	86	55	33	27	146	M
6	L	163	47	80	56	35	28	162	A
7	L	154	38	71	53	34	24	150	A
8	N	143	37	69	56	29	23	146	A
9	B	154	61	96	56	36	28	159	G
10	N	159	57	86	55	37	29	165	M
11	N	143	40	74	57	33	23	146	A
12	N	152	43	73	57	33	26	151	A
13	L	172	54	85	58	38	29	172	A
14	L	168	55	77	57	36	28	166	A
15	L	145	35	70	55	32	23	147	A
16	N	150	36	67	56	33	23	152	A

Figure 11: Esempio di matrice di di dati

6.3 Le variabili (solo alcune)

Variable	Descriptive Statistics (PUBTRAS_totale_corretto.STA)								
	Valid N	Mean	Minimum	Maximum	Std.Dev.	Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis
PRT	1519	285.83	1	844	241.52	0.828	0.063	-0.600	0.125
ANNO	1519	81.89	81	83	0.99	0.219	0.063	-1.955	0.125
DATA	1519	8112.97	7773	8553	356.03	0.223	0.063	-1.941	0.125
DATANASC	1512	3448.65	391	7079	584.53	0.061	0.063	1.545	0.125
HABITUS	1518								
ALTEZZA	1519	151.85	126	183	10.09	0.319	0.063	-0.348	0.125
PESOKG	1519	45.04	21	100	10.70	0.966	0.063	1.438	0.125
TORACECM	1517	75.67	57	104	7.77	0.719	0.063	0.479	0.125
CRANIOCM	1517	54.75	43	63	1.66	0.003	0.063	1.952	0.125
BISACROM	1516	34.51	23	55	2.99	0.557	0.063	1.438	0.125
BITROCAN	1516	26.36	20	38	2.78	0.439	0.063	0.360	0.125
SPANCM	1446	153.45	16	311	13.19	-0.254	0.064	28.010	0.128
GRASSO	1518								
CARNAG	1518								
OCCHI	1518								
CAPELLI	1518								
Fase_puberale	1519	1.95	0	5	1.17	0.818	0.063	-0.442	0.125

Figure 12: Indici sintetici descrittivi per lacune variabili

6.4 Rappresentazioni grafiche

Rappresentazione delle sole 7 variabili antropometriche:

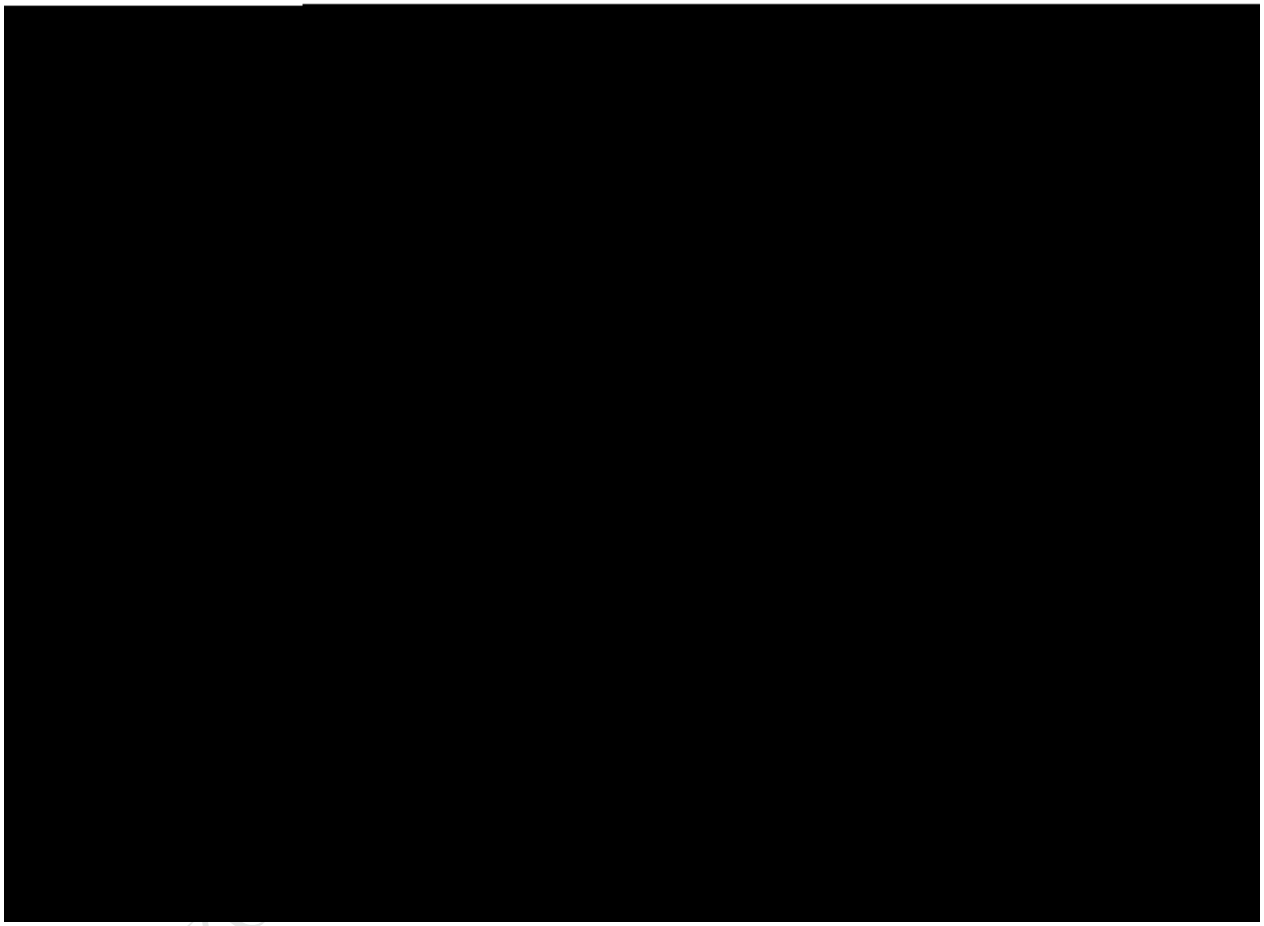

Figure 13: grafico a matrice delle coppie di variabili:dati grezzi

Alcune osservazioni, come si vede dal grafico a matrice, presentano valori dubbi per alcune delle variabili, presumibilmente dovuti ad errori di trascrizione; sono stati considerati come dati mancanti ed è stato rappresentato di nuovo il grafico

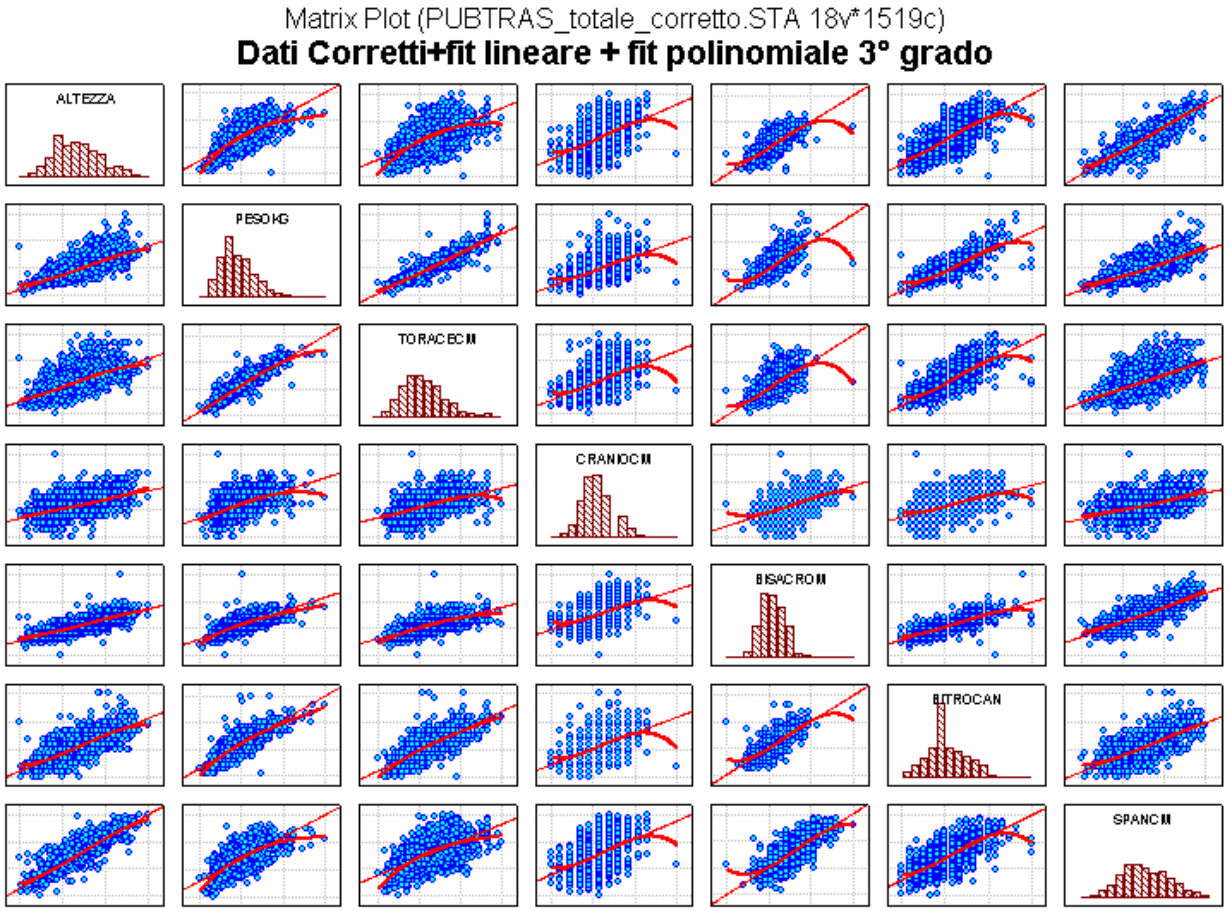

Figure 14: grafico a matrice delle coppie di variabili

Variable	Correlations (PUBTRAS_totale_corretto.STA)							
	ANNO	DATA	DATANASC	HABITUS	ALTEZZA	PESOKG	TORACECM	CRANIOCM
ANNO	1.00	1.00	0.61	0.06	-0.02	-0.03	-0.01	0.11
DATA	1.00	1.00	0.61	0.06	-0.01	-0.03	-0.01	0.11
DATANASC	0.61	0.61	1.00	-0.01	-0.48	-0.35	-0.28	-0.11
HABITUS	0.06	0.06	-0.01	1.00	0.09	0.09	0.10	-0.00
ALTEZZA	-0.02	-0.01	-0.48	0.09	**1.00**	**0.72**	**0.58**	**0.46**
PESOKG	-0.03	-0.03	-0.35	0.09	**0.72**	**1.00**	**0.91**	**0.52**
TORACECM	-0.01	-0.01	-0.28	0.10	**0.58**	**0.91**	**1.00**	**0.47**
CRANIOCM	0.11	0.11	-0.11	-0.00	**0.46**	**0.52**	**0.47**	**1.00**
BISACROM	0.16	0.17	-0.31	0.05	**0.75**	**0.75**	**0.70**	**0.49**
BITROCAN	0.04	0.05	-0.35	0.06	**0.73**	**0.84**	**0.77**	**0.47**
SPANCM	0.04	0.05	-0.39	0.07	**0.81**	**0.61**	**0.51**	**0.40**
GRASSO	-0.04	-0.05	0.01	0.16	-0.00	0.40	0.48	0.08
CARNAG	-0.16	-0.16	-0.01	-0.03	-0.10	0.03	0.06	-0.08
OCCHI	0.03	0.04	-0.02	-0.00	0.02	0.01	0.01	0.04
CAPELLI	0.03	0.03	-0.10	0.02	0.22	0.14	0.12	0.15
Fase_puberale	0.08	0.09	-0.41	0.05	0.70	0.53	0.43	0.35
PRT	-0.69	-0.67	-0.43	0.04	0.06	0.03	0.03	-0.07

Figure 15: Correlazioni

6.5 Problematiche statistiche (solo alcune!)

Non tutte saranno affrontate nel corso!

- Come interagiscono le variabili?
- Le relazioni fra le variabili antropometriche sono di tipo lineare?
- Che relazione c'è fra le variabili antropometriche e l'età dei soggetti?
- Un sottoinsieme di esse o loro combinazioni sarebbero sufficienti per dare buone informazioni sintetiche?
- le relazioni fra le variabili cambiano per le diverse fasi puberali?
- Alcune combinazioni di variabili potrebbero descrivere sufficientemente bene la fase dello sviluppo puberale di ciascun individuo?

7 Esempio delle prove dei Gran Premi (anno 2000)

Prendiamo ora un piccolo esempio tratto da avvenimenti sportivi:

esaminiamo i tempi di qualifica ottenuti dai vari piloti nei gran premi di formula uno (anno 1999).

Consideriamo i tempi di qualifica e non i risultati ottenuti in gara perché questi ultimi sono perturbati da diversi fattori, ed inoltre hanno una notevole percentuale di *dati censurati* o comunque indeterminati perché molti piloti non concludono la corsa; i dati relativi invece alle qualifiche sono in effetti più regolari anche perché ottenuti in condizioni più controllate.

```
"pilota""primaguida" "Scuderia" "Tempo ora" "gp"
"DE_LA_ROSA" 1 "(Arrows)" 01.32,323 "AUSTRALIA "
"DE_LA_ROSA" 1 "(Arrows)" 01.16,002 "BRASILE "
"DE_LA_ROSA" 1 "(Arrows)" 01.26,349 "SANMARINO"
"DE_LA_ROSA" 1 "(Arrows)" 01.28,135 "GRAN BRETAGNA"
"DE_LA_ROSA" 1 "(Arrows)" 01.22,185 "SPAGNA"
"DE_LA_ROSA" 1 "(Arrows)" 01.19,024 "EUROPA"
"DE_LA_ROSA" 1 "(Arrows)" 01.21,832 "MONACO (MONTECARLO)"
"DE_LA_ROSA" 1 "(Arrows)" 01.19,912 "CANADA "
"DE_LA_ROSA" 1 "(Arrows)" 01.17,279 "FRANCIA"
"DE_LA_ROSA" 1 "(Arrows)" 01.11,978 "AUSTRIA "
"DE_LA_ROSA" 1 "(Arrows)" 01.47,786 "GERMANIA"
```

```
"DE_LA_ROSA" 1 "(Arrows)" 01.19,897 "UNGHERIA "
"DE_LA_ROSA" 1 "(Arrows)" 01.53,237 "BELGIAN"
"DE_LA_ROSA" 1 "(Arrows)" 01.24,814 "ITALIAN"
"DE_LA_ROSA" 1 "(Arrows)" 01.16,143 "USA "
"DE_LA_ROSA" 1 "(Arrows)" 01.37,652 "GIAPPONE "
"DE_LA_ROSA" 1 "(Arrows)" 01.39,443 "MALESIA "
"VERSTAPPEN" "(Arrows)" 00.01.32 "AUSTRALIA "
"VERSTAPPEN" "(Arrows)" 01.15,704 "BRASILE "
"VERSTAPPEN" "(Arrows)" 01.26,845 "SANMARINO"
"VERSTAPPEN" "(Arrows)" 01.26,793 "GRAN BRETAGNA"
"VERSTAPPEN" "(Arrows)" 01.22,421 "SPAGNA"
"VERSTAPPEN" "(Arrows)" 01.19,19 "EUROPA"
"VERSTAPPEN" "(Arrows)" 01.21,738 "MONACO (MONTECARLO)"
"VERSTAPPEN" "(Arrows)" 01.20,107 "CANADA "
"VERSTAPPEN" "(Arrows)" 01.17,933 "FRANCIA"
```

I dati sono classificabili secondo più criteri:

- un criterio di classificazione è il particolare circuito; questo certamente andrà considerato come un effetto fisso la cui influenza va eliminata;
- un altro criterio di classificazione è la scuderia;
- un terzo criterio è costituito dai due piloti di ciascuna squadra (e questo potrebbe essere forse

un effetto casuale).

Il disegno è gerarchico perché i piloti di ciascuna scuderia sono sempre gli stessi (modello di analisi della varianza gerarchico con effetti misti, ossia effetti fissi e ed effetti casuali)

Potremmo chiederci:

- Qual è l'effetto medio di ciascun circuito?
- Quanto influente, o significativa, è la differenza fra le varie squadre?
- All'interno di ciascuna squadra, le differenze fra i piloti sono significative?
- Quale dei precedenti effetti è preponderante?
- È più ragionevole studiare i tempi o la velocità?
- L'ipotesi di normalità è ragionevole? Su questo aspetto una analisi dei residui empirici sarà di grande aiuto.
- Esiste interazione fra i circuiti e le squadre? È estremamente difficile che i dati relativi alle sessioni di prova possano fornirci tali informazioni, in quanto non sono stati rilevati tempi diversi per ciascun pilota.

8 Dati tratti da bilanci aziendali

I grafici che seguono sono tratti da archivi di dati reali, riguardanti 2835 aziende siciliane, operanti in Sicilia nel 1992.

I dati sono quelli relativi ai bilanci pubblicati per riguardano soltanto le società di capitale. Le variabili presenti sono:

- La ragione sociale;
- La provincia;
- Il fatturato annuo;
- Il numero di dipendenti;
- L'utile;
- I mezzi forniti dai terzi;
- Costo complessivo del lavoro;
- Codifica dell'attività svolta.

Come è ovvio, questo insieme di dati difficilmente può essere considerato un campione di aziende: in effetti, a meno di qualche errore materiale, si tratta di tutte le aziende siciliane costituite da società di capitale che hanno presentato un bilancio nel 92. Pertanto già per questo solo motivo è

impensabile trattare questi dati come un campione casuale semplice da una normale multivariata; inoltre le particolari variabili rilevate sono intrinsecamente non normali:

è noto infatti che la distribuzione delle aziende secondo la dimensione o secondo il numero di addetti è tipicamente asimmetrica come pure la distribuzione del fatturato; tuttavia è presumibile che alcune di queste variabili siano legate da correlazioni almeno approssimativamente lineari.

8.1 esempio di dati

RKG	PROV	FATT92	DIPENDEN	UTILE	MEZZITER	COSTOLAV	ISTAT1
6	PA	337782		-67013	262558	91357	24
9	PA	224650	91	-88	5574	2693	51
20	AG	113788	396	-2330	23038	18439	45
23	PA	110926	211	-10707	33657	15925	61
41	RG	90297	36	617	22614	1306	51
64	PA	59578	39	56	15869	1609	15
69	RG	55479	17	565	7436	364	24
73	ME	52885	273	-3926	58978	16689	35
75	PA	52761	1	-90	16230	5451	51
95	PA	42722	28	893	38321	1393	15
102	RG	41364	34	735	2574	2574	50
109	AG	38937		309	7286	1088	51
127	PA	33333	255	827	3601	13935	74
130	PA	32823	35	65	1355	1265	63
136	CT	31756	53	160	2262	1914	51
147	PA	29987	5	9	1234	192	51
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
2699	RG	2124	6	56	347	136	52
2702	CT	2123	2	3	0	221	51
2703	AG	2122	6	-246	624	82	52
2708	PA	2116		-14	1150	250	45
2710	PA	2115	34	-564	684	1148	74
2715	CL	2108	1	-10	0	79	45
2719	CL	2101	11	-45	538	310	26
2735	PA	2086		3	89	45	52
2739	PA	2083	2	286	3179	178	74
2748	PA	2066	1	7	0	24	51
2766	CL	2054	6	37	132	220	52
2798	PA	2026	4	-11	54	198	63

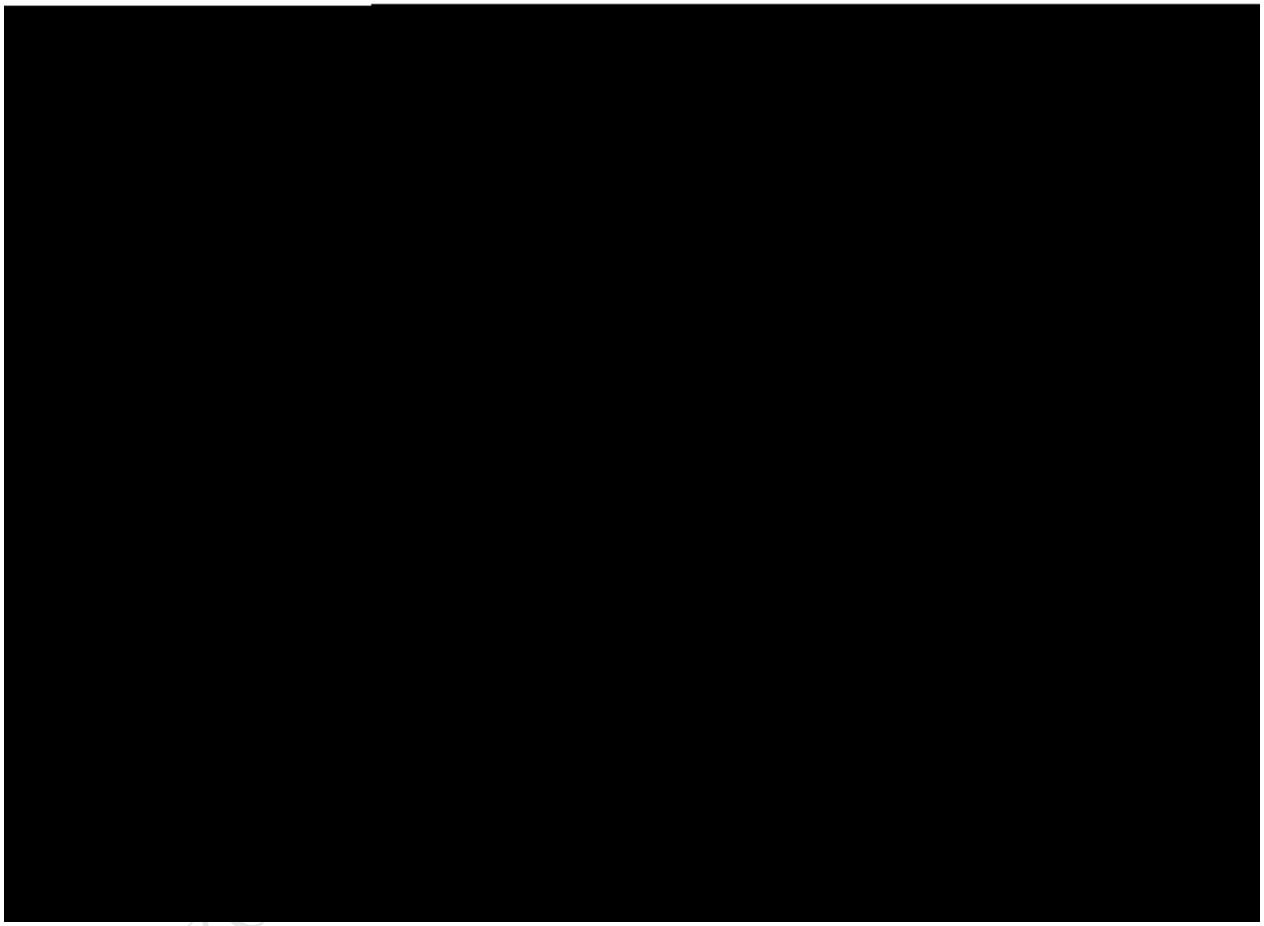

Figure 16: grafico a matrice delle coppie di variabili

Come si vede le distribuzioni sono molto asimmetriche e sono poco plausibili probabilmente le ipotesi di normalità delle distribuzioni (come peraltro si può immaginare data la natura delle variabili) e di linearità e omoscedasticità delle relazioni di regressione.

Solo a scopo esplorativo riporto qui anche gli stessi grafici in scala logaritmica: molte relazioni sembrano (ma è da verificare) più facilmente approssimabili da rette.

Figure 17: grafico a matrice delle coppie di variabili:scale logaritmiche per tutte le variabili

9 indici di 8 borse

Sono rappresentati nelle figure i grafici a matrice dei valori di chiusura giornaliera degli indici di 8 borse nell'arco di circa 20 anni, $x_{tj}, t = 1, 2, \ldots, 4959; j = 1, 2, \ldots, 8$

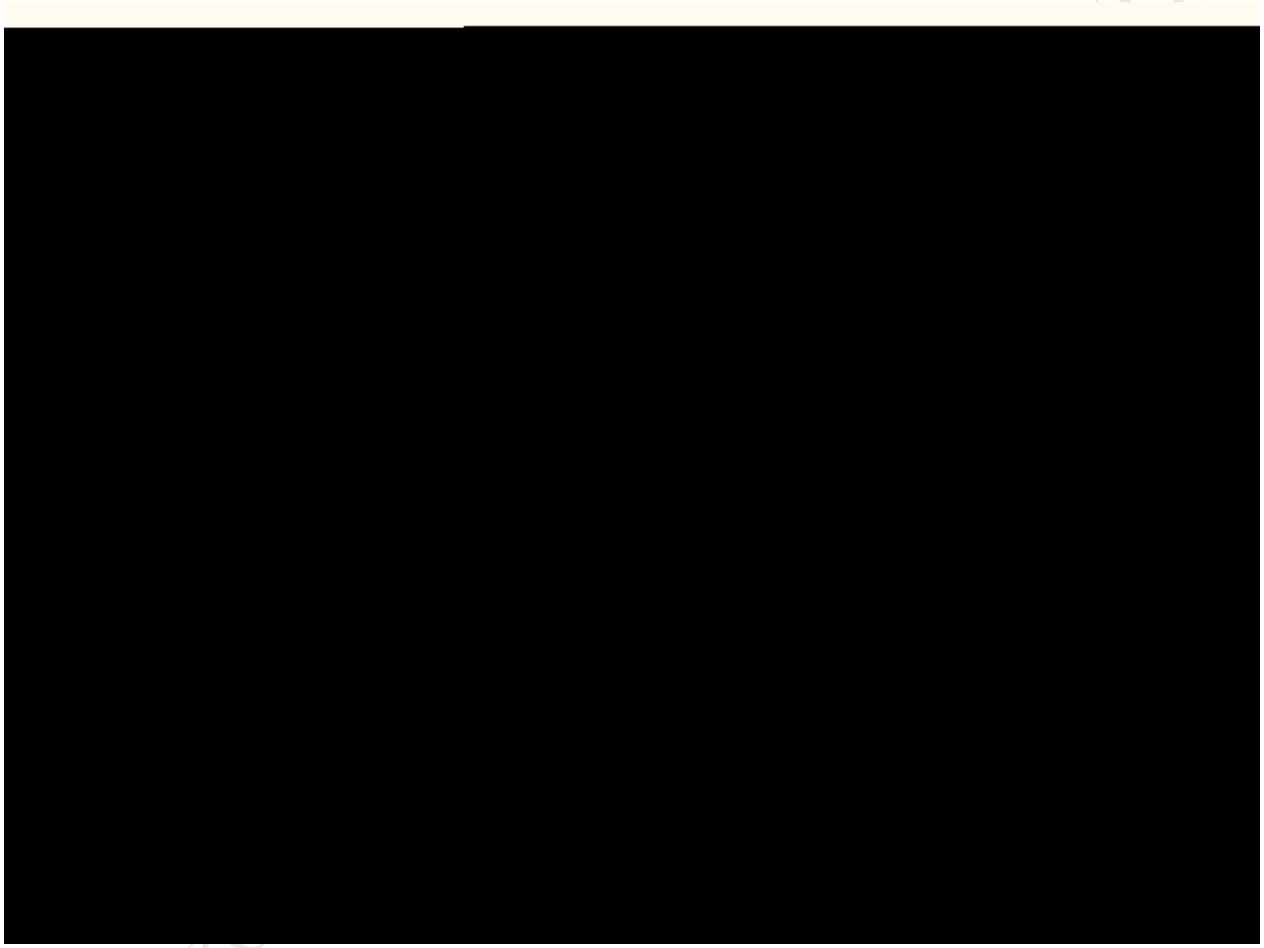

Figure 18: grafico a matrice delle coppie di variabili:valori giornalieri di 8 indici di 8 borse

Nel grafico successivo sono rappresentate le trasformate y_{tj} di questi stessi indici, ossia i rendimenti relativi giornalieri:

$$y_{tj} = \frac{x_{t+1,j} - x_{tj}}{x_{tj}} \qquad t = 1, 2, \ldots, 4958; j = 1, 2, \ldots, 8$$

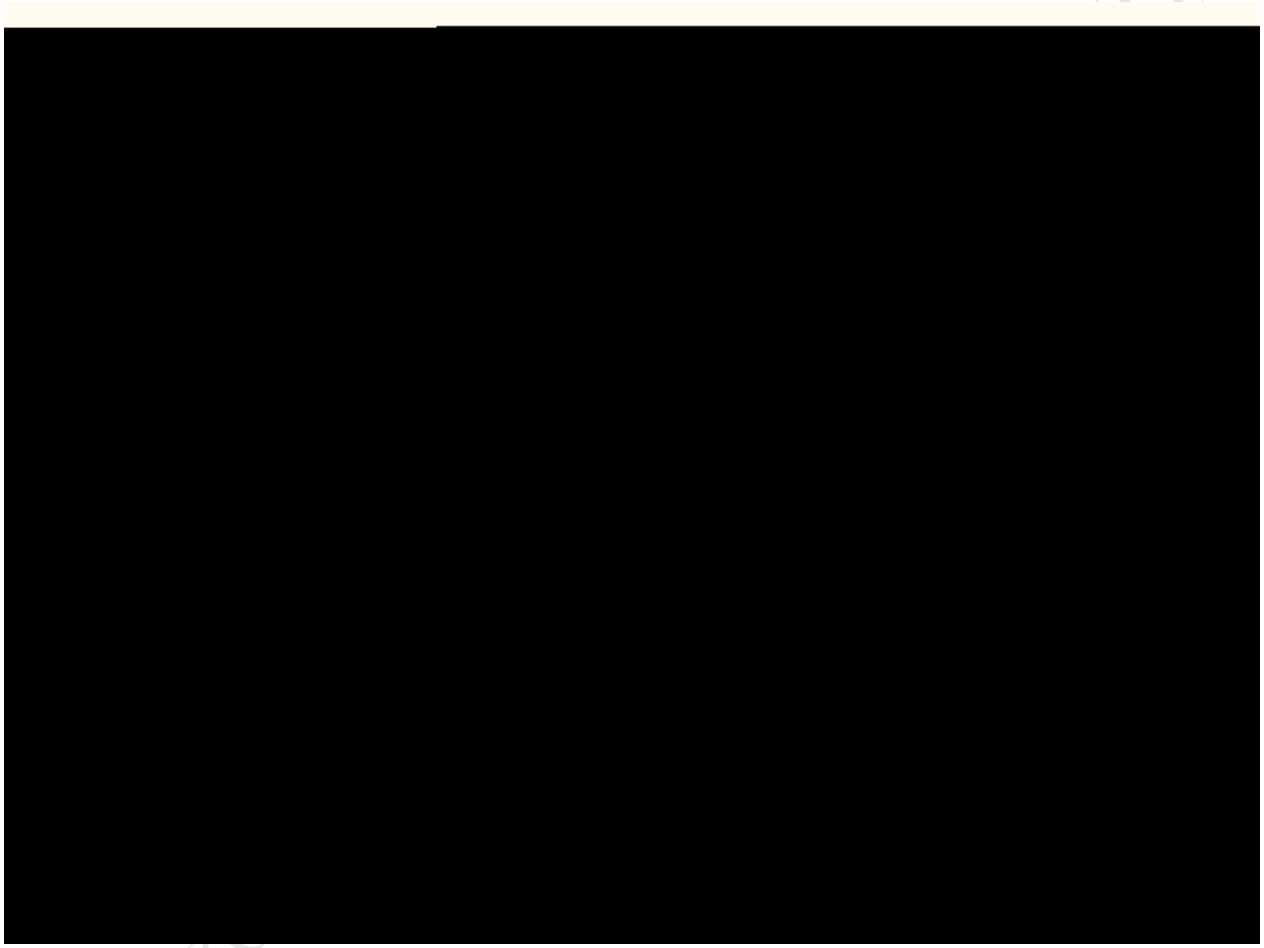

Figure 19: grafico a matrice delle coppie di variabili trasformate:valori giornalieri dei rendimenti relativi degli 8 indici di 8 borse

10 Discussione degli spunti statistici dei vari problemi

10.1 Elementi distintivi dei vari problemi

- la *variabile di risposta* (nei nostri esempi e nel nostro corso) è quantitativa (dove è identificabile chiaramente una risposta in funzione di variabili esplicative)
  - La produzione di frumento
  - Il peso alla nascita dei neonati
  - L'accrescimento di peso dei suini
  - Il miglior tempo sul giro di prova (o la velocità)
- Le *variabili esplicative* possono essere: quantitative, qualitative, miste.

  Esempi di variabili qualitative
  - Varietà di frumento
  - Tipo di dieta
  - Fertilizzante
  - Scuderia e circuito di Formula 1

  Esempi di variabili quantitative
  - Il numero di barbabietole (variabile non continua)
  - Il numero di settimane di gestazione (variabile continua discretizzata)

– L'altezza alla nascita

Sono miste le situazioni nelle quali sono presenti fra le variabili esplicative variabili sia quantitative che qualitative

- I dati possono provenire da *esperimenti pianificati*, in cui alcuni fattori sono tenuti sotto controllo, o da *studi osservazionali* in cui non è possibile tenere sotto controllo i fattori. E' di fondamentale importanza che lo statistico intervenga comunque nella fase di pianificazione dello studio, prima della rilevazione dei dati.
- Può interessare la verifica di una particolare ipotesi (o la costruzione di un intervallo o regione di confidenza) relativamente solo ad un gruppo di parametri, mentre altri parametri del modello giocheranno il ruolo di parametri di disturbo. Svolge spesso il ruolo di fattore di disturbo la particolare distribuzione di errori accidentali.
- La risposta che si vuole ottenere può essere soltanto di tipo comparativo (qual è il migliore fertilizzante fra A, B e C), oppure assoluto (qual è l'effetto medio del fertilizzante A?)
- Come attribuire i vari trattamenti alle singole unità?

10.2 Elementi comuni ai vari problemi

Problema generale

In generale si vuole studiare (sulla base di un campione di osservazioni) la dipendenza di un fenomeno (espresso spesso da una variabile quantitativa) da una molteplicità di fattori o variabili esplicative (quantitative e/o qualitative)
Sarà questo l'oggetto fondamentale del nostro corso.

La dipendenza sarà in generale studiata secondo il concetto *di dipendenza in media*

Ovviamente vorremo anche fare inferenza, ossia per esempio vedere se le diffrenze medie di produzione riscontrate sono statisticamente significative.