La verifica delle assunzioni di base: metodi diagnostici

Andrea Onofri
Dipartimento di Scienze Agrarie ed Ambientali
Universitá degli Studi di Perugia

28 marzo 2014

Indice

1 Introduzione **1**

2 Procedure diagnostiche **2**
2.1 Analisi dei residui 2
2.1.1 Osservazioni aberranti 3
2.1.2 Disomogeneità delle varianze 3
2.1.3 Errori non-normali 5

3 Altri strumenti diagnostici **6**

4 Risultati contraddittori **6**

5 Azioni correttive: la procedura di BOX e COX **6**

6 La trasformazione dei dati **7**

7 Referenze bibliografiche per approfondimenti **8**

1 Introduzione

Nel momento in cui eseguiamo test d'ipotesi nell'ambito di un modello lineare, assumiamo implicitamente che i dati seguano una certa distribuzione di frequenza e rispondano ai seguenti requisiti:

1. il modello è corretto;
2. la risposta osservata è una funzione del modello più o meno l' errore sperimentale;

3. l' errore sperimentale è indipendente dal modello;
4. gli errori sono normalmente distribuiti, con media zero e varianze omogenee;
5. gli errori rilevati in esperimenti ripetuti sono tra loro indipendenti.
6. assenza di osservazioni aberranti;

Queste assunzioni di base non possono essere in genere verificate *a priori*, perchè per quasi nessuno dei fenomeni naturali sono note le vere relazioni causa-effetto. Per questo motivo, i procede in primo luogo alla stima dei parametri della funzione e, successivamente, si verifica il rispetto delle assunzioni di base.

Il problema è importante perché ogni deviazione rispetto agli anzidetti requisiti può inficiare la validità dei test d'ipotesi, modificando il livello di significatività e di protezione. A riguardo dei dati aberranti dobbiamo dire che, se è sbagliato correggerli arbitrariamente, senza aver prima accertato che siano effettivamente frutto di errore, è altrettanto sbagliato lasciarli nel dataset, in quanto essi possono influenzare in modo molto marcato il risultato dell'analisi. E'evidente comunque che la correzione non può che riguardare una larga minoranza dei dati sperimentali raccolti (uno o due dati), altrimenti si dovrà necessariamente pensare di ripetere l'esperimento.

2 Procedure diagnostiche

- PROCEDURA GENERALE: Ispezione grafica dei residui
- OUTLIERS: test di Anscombe e Tukey
- ETEROSCEDASTICITA': test di Bartlett e test di Levene
- VALUTAZIONE DI AZIONI CORRETTIVE: Procedura di Box e Cox

2.1 Analisi dei residui

La gran parte dei pre-requisiti fondamentali di un dataset riguardano la struttura dei residui e, di conseguenza, l'ispezione grafica di questi ultimi, eventualmente accompagnata da semplici strumenti algebrici, possono permetterci di evidenziare la gran parte delle 'patologie' di cui soffrono i dati sperimentali. Si può affermare che l'ispezione dei residui è uno strumento diagnostico fondamentale il cui impiego dovrebbe rientrare tra le metodiche di routine per ogni elaborazione statistica dei dati.

Si ricorda che i residui sono gli scostamenti tra i valori osservati e quelli attesi sulla base del modello in studio; il metodo grafico più utilizzato per il loro esame è quello di plottare i residui verso i valori attesi. Se non

vi sono problemi, i punti nel grafico dovrebbero essere distribuiti in modo assolutamente casuale, come in fig. 1.

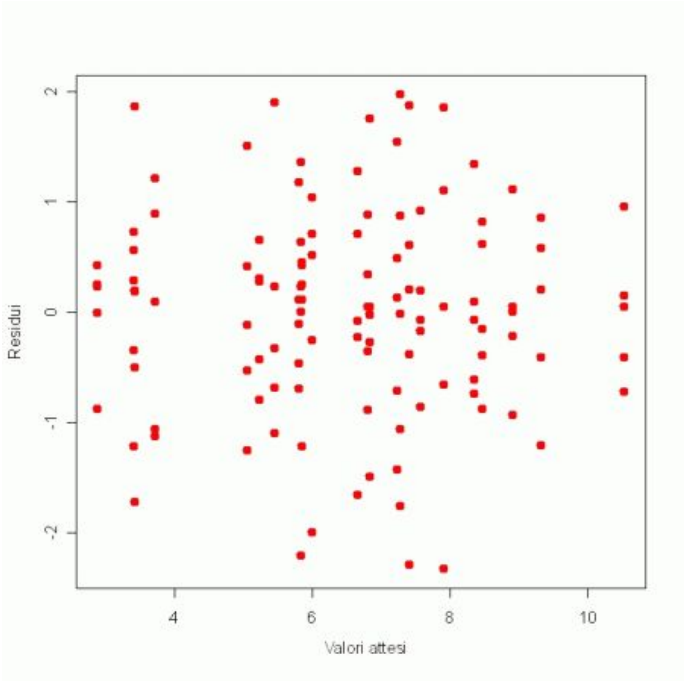


Figura 1: Residui distribuiti casualmente

2.1.1 Osservazioni aberranti

Le osservazioni aberranti (*outliers*) sono chiaramente indicate nel grafico dei residui come punti isolati rispetto agli altri (Figura 2).

Nel caso di outliers, l'ispezione grafica dei residui può essere supportata dalla procedura indicata da Anscombe e Tukey. Con questa, si va alla ricerca del residuo con valure assoluto più alto (che potrebbe quindi essere un outlier, cioè un dato aberrante) e lo si confronta con un valore massimo teorico stabilito sulla base della varianza dell'errore e del relativo numero di gradi di libertà (Si rimanda al lavoro citato per ulteriori informazioni).

2.1.2 Disomogeneità delle varianze

La disomogeneità delle varianze è chiaramente indicata da residui che si allargano o si stringono procedendo verso i margini del grafico (Fig. 3), facendo emergere una sorta di proporzionalitè tra media e varianza.

Questa evidenza può trovare conferma con il test di Bartlett, oppure con il test di Levene, più robusto del primo nei confronti di dati affetti da

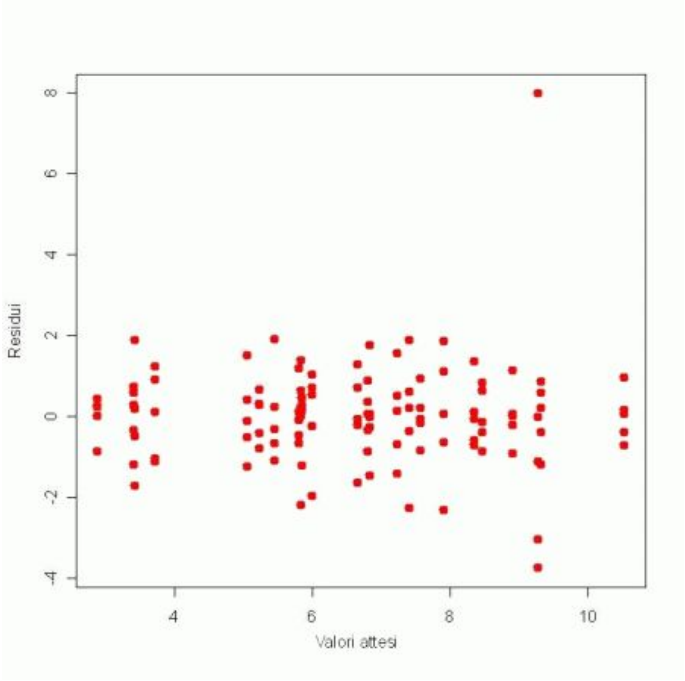


Figura 2: Presenza di un outlier

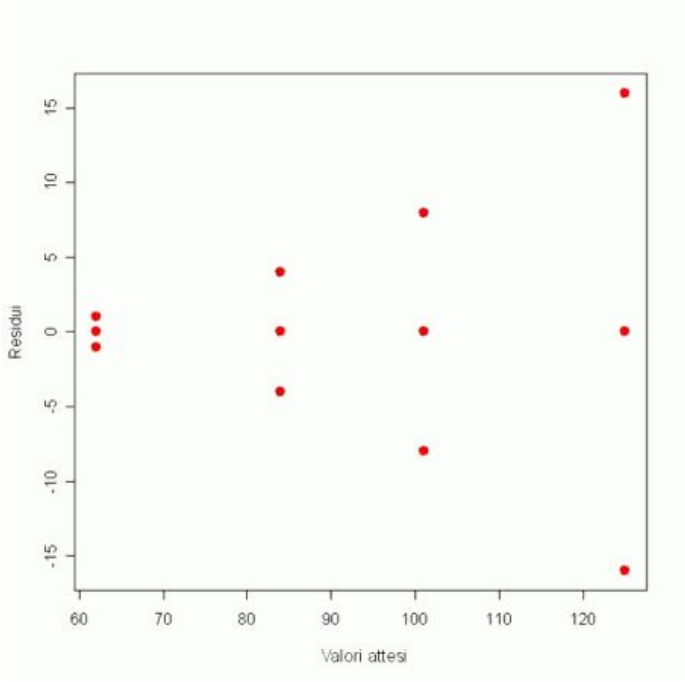


Figura 3: Eterogeneità delle varianze

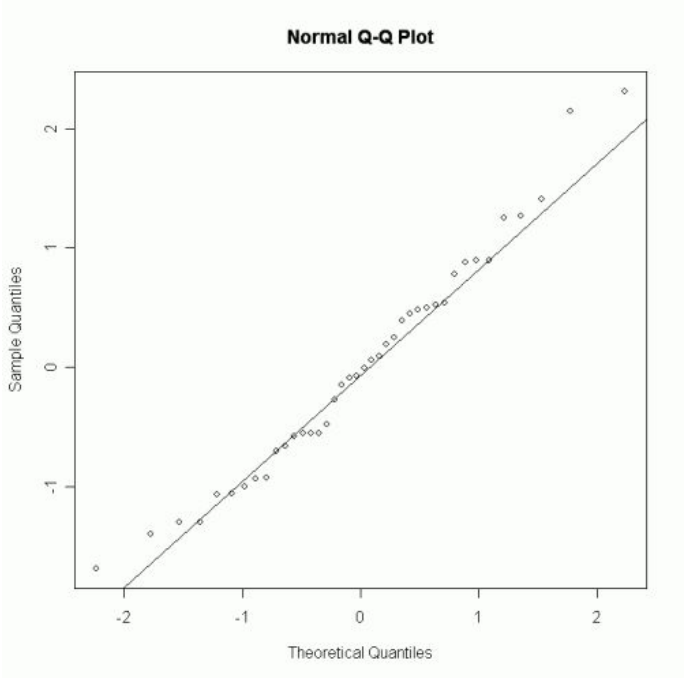


Figura 4: Esempio di QQPlot

problemi di non-normalità. in entrambi i casi, l'ipotesi nulla è l'assenza di problemi e questa può essere accettata se il valore di P è maggiore di 0.05. Viceversa, il fatto che uno o entrambi questi test diano P minore di 0.05 costituisce un campanello di allarme sulla omogeneità delle varianze.

2.1.3 Errori non-normali

Tra le possibili problematiche, la non-normalità delle distribuzioni non è particolarmente grave, dato che (come riconosciuto da molti autori) il test di F è sufficientemente robusto nei confronti delle deviazioni rispetto a questo assunto.

Tuttavia, nell'ambito di una valutazione preliminare del modello anche questo assunto dovrebbe essere verificato ricorrendo al QQplot, cioè plottando i residui standardizzati (divisi per la radice quadrata della varianza d'errore) e ordinati contro i valori di z corrispondenti ai relativi quantili della distribuzione normale standardizzata. Se i residui sono normali, essi dovrebbero giacere sulla bisettrice del primo quadrante (fig. 4). In questo corso trascureremo questa valutazione.

3 Altri strumenti diagnostici

La valutazioni precedentemente esposte sono di tipo grafico e sono considerate sufficientemente robuste per la maggior parte delle situazioni. Tuttavia, esistono anche test statistici che consentono di testare l'ipotesi nulla di 'assenza di deviazioni'. Questi test sono stati elencati in precedenza (test di Bartlett, test di Levene, test di Anscombe e Tukey) e, per la loro esecuzione, richiedono software specializzati. Per questo corso si può utilizzare l'add-in per Excel DSAASTAT, selezionando la voce: Strumenti/Diagnostic tools.

In tutti i casi, l'interpretazione è semplice, basta guardare il 'p value' del relativo test: se questo è inferiore a 0.05 l'ipotesi nulla (assenza di deviazioni) deve essere rifiutata e può essere necessario intraprendere azioni correttive.

4 Risultati contraddittori

La valutazione degli assunti di base è un passo fondamentale nell'analisi dei dati sperimentali è non può essere evitata in nessun modo. Il problema è importante perché ogni deviazione rispetto agli anzidetti requisiti può inficiare la validità dei test d'ipotesi, modificando il livello di significatività e di protezione.

Tuttavia, ricordiamo sempre che la 'verità vera' ci sfugge e, di conseguenza, le valutazioni sull'adozione di eventuali trasformazioni stabilizzanti debbono essere condotte con il massimo 'buon senso'!

In particolare nella pratica è molto facile incontrare situazioni dubbie, nelle quali l'analisi dei residui mostra deviazioni, mentre il test di Bartlett è significativo e quello di Levene no. Come comportarsi? Misurare sempre la forza dell'evidenza 'patologica': quanti campanelli di allarme abbiamo?

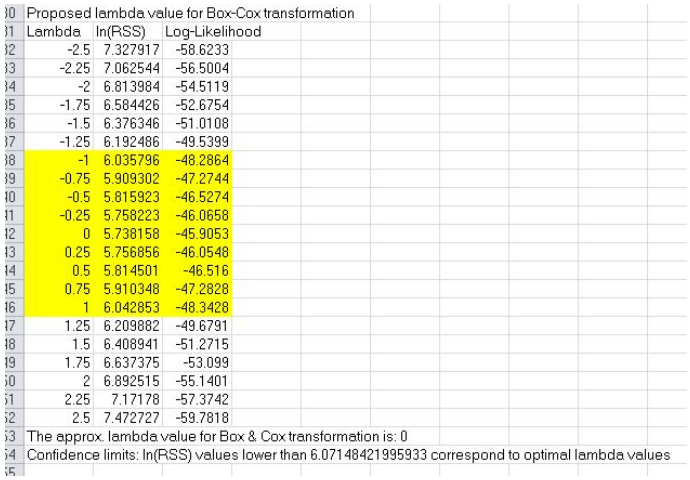
5 Azioni correttive: la procedura di BOX e COX

In genere, quando siano violate alcune delle assunzioni di base per il test d'ipotesi, piuttosto che ricorrere alla statistica non parametrica, si preferisce ricorrere alle trasformazioni correttive.

Invece che utilizzare dette trasformazioni in modo arbitrario (es. trasformazione logaritmica o in radice quadrata per le conte, e trasformazione angolare per le proporzioni), si può impiegare la procedura suggerita da Box e Cox (1964), che si basa su alcune famiglie di trasformazioni, tra cui la più diffusa è:

$$W = \begin{cases} Y^{\lambda} & per\ \lambda \neq 0 \\ \log(Y) & per\ \lambda = 0 \end{cases}$$

dove W è la variabile trasformata, Y è la variabile originale e λ è il parametro che definisce la trasformazione. In particolare, osserviamo che



30	Proposed lambda value for Box-Cox transformation		
31	Lambda	ln(RSS)	Log-Likelihood
32	-2.5	7.327917	-58.6233
33	-2.25	7.062544	-56.5004
34	-2	6.813984	-54.5119
35	-1.75	6.584426	-52.6754
36	-1.5	6.376346	-51.0108
37	-1.25	6.192486	-49.5399
38	-1	6.035796	-48.2864
39	-0.75	5.909302	-47.2744
40	-0.5	5.815923	-46.5274
41	-0.25	5.758223	-46.0658
42	0	5.738158	-45.9053
43	0.25	5.756856	-46.0548
44	0.5	5.814501	-46.516
45	0.75	5.910348	-47.2828
46	1	6.042853	-48.3428
47	1.25	6.209882	-49.6791
48	1.5	6.408941	-51.2715
49	1.75	6.637375	-53.099
50	2	6.892515	-55.1401
51	2.25	7.17178	-57.3742
52	2.5	7.472727	-59.7818
53	The approx. lambda value for Box & Cox transformation is: 0		
54	Confidence limits: ln(RSS) values lower than 6.07148421995933 correspond to optimal lambda values		
55			

Figura 5: Come leggere i risultati relativi alla scelta delle trasformazioni

se λ è pari ad 1 i dati non sono trasformati, se è pari a 0.5 abbiamo una trasformazione in radice, se è pari a 0 abbiamo la trasformazione logaritmica, se è pari a -1 abbiamo la trasformazione nel reciproco.

La scelta di λ viene eseguita in base al criterio della massima verosimiglianza. In pratica si tratta di cercare il valore di lambda che permette di ottenere la massima verosimiglianza del modello (log-likelihood) o il minimo della devianza del residuo (log-RSS). E'necessario comunque considerare che esiste un margine di incertezza intorno al valore di lambda, per il quale esiste un range di valori possibili. DSAASTAT esprime questo range nella forma: 'I valori di log-RSS inferiori a X corrispondono a valori di lambda ottimali', come nella figura (fig. 5).

In questo caso i valori di ln-RSS minori a 6.07 identificano valori di lambda compresi tra -1 ed 1 (indicati in giallo), che sono tutti ottimali per la trasformazione. Dato che il valore lambda = 1 corrisponde a non trasformare, questo risultato indicherebbe che la trasformazione non è, di fatto, necessaria e, probabilmente, le eventuali 'patologie' del dataset sono lievi a sufficienza da poter essere trascurate.

Se invece il range di valori ottimali di lambda non include 1, allora la trasformazione è necessaria e dovrà essere operata preferibilmente scegliendo nell'ambito dei valori ottimali qulli più facili da interpretare algebricamento (lambda = 0, lamda = 0.5, lambda = -1 e così via)

6 La trasformazione dei dati

Trasformare i dati è semplice: si prende la variabile da analizzare, la si trasforma come prescelto (ad esempio si fa il logaritmo di ogni valore) e si analizza la variabile trasformata.

ATTENZIONE! La trasformazione dei dati implica che i risultati debbono essere commentati nella loro scala trasformata e qundi l'interpretazione si complica!!!!!

A riguardo dei dati aberranti dobbiamo dire che, se è sbagliato correggerli arbitrariamente, senza aver prima accertato che siano effettivamente frutto di errore, è altrettanto sbagliato lasciarli nel dataset, in quanto essi possono influenzare in modo molto marcato il risultato dell'analisi. E'evidente comunque che la correzione non può che riguardare una larga minoranza dei dati sperimentali raccolti (uno o due dati), altrimenti si dovrà necessariamente pensare di ripetere l'esperimento.

7 Referenze bibliografiche per approfondimenti

Ahrens, W. H., D. J. Cox, and G. Budwar. 1990, Use of the arcsin and square root trasformation for subjectively determined percentage data. Weed science 452-458.

Anscombe, F. J. and J. W. Tukey. 1963, The examination and analysis of residuals. Technometrics 5: 141-160.

Babbini, M., B. Chiandotto, G. Chisci, R. d. Cristofaro, G. A. Maccararo, N. Montanaro, F. Nicolis, E. Ottaviano, F. Salvi, and M. Turri. 1978. Biometria: principi e metodi per studenti e ricercatori biologi. Padova: P. 552.

Box, G. E. P. and D. R. Cox. 1964, An analysis of transformations. Journal of the Royal Statistical Society, B-26, 211-243, discussion 244-252.

Camussi , A., F. Moller , E. Ottaviano , and M. Sarli Gorla . 1986, Metodi statistici per la sperimentazione biologica. Ed. Zanichelli.

Chatfield, C. 1985, The initial examination of data. J. R. Statist. Soc. A-148, 3, 214-253 A-148: 214-253.

D'Elia, A. 2001, Un metodo grafico per la trasformazione di Box-Cox: aspetti esplorativi ed inferenziali. STATISTICA LXI: 630-648.

Draper, N. R. and H. Smith. 1981, Applied regression. John Wiley & Sons, Inc. , New York, 2nd ed.

Saskia, R. M. 1992, The Box-Cox transformation technique: a review. Statistician 41: 169-178.

Snedecor, G. W. and W. G. Cochran. 1991. Statistical methods. AMES (Iowa): IOWA State University Press, VIII Edition. P. 503.